Copyrighted Material

CHARACTERIZATION OF MATERIALS

VOLUME 1

ELTON N. KAUFMANN

Copyrighted Material

CHARACTERIZATION OF MATERIALS

Characterization of Materials is available Online in full color at www.mrw.interscience.wiley.com/com.

WILEY-INTERSCIENCE
A John Wiley and Sons Publication

CHARACTERIZATION OF MATERIALS

EDITORIAL BOARD

Elton N. Kaufmann, (Editor-in-Chief)
Argonne National Laboratory
Argonne, IL

Reza Abbaschian
University of Florida at Gainesville
Gainesville, FL

Peter A. Barnes
Clemson University
Clemson, SC

Andrew B. Bocarsly
Princeton University
Princeton, NJ

Chia-Ling Chien
Johns Hopkins University
Baltimore, MD

David Dollimore
University of Toledo
Toledo, OH

Barney L. Doyle
Sandia National Laboratories
Albuquerque, NM

Brent Fultz
California Institute of Technology
Pasadena, CA

Alan I. Goldman
Iowa State University
Ames, IA

Ronald Gronsky
University of California at Berkeley
Berkeley, CA

Leonard Leibowitz
Argonne National Laboratory
Argonne, IL

Thomas Mason
Spallation Neutron Source Project
Oak Ridge, TN

Juan M. Sanchez
University of Texas at Austin
Austin, TX

Alan C. Samuels, Developmental Editor
Edgewood Chemical Biological Center
Aberdeen Proving Ground, MD

EDITORIAL STAFF

VP, STM Books: Janet Bailey
Executive Editor: Jacqueline I. Kroschwitz
Editor: Arza Seidel
Director, Book Production
and Manufacturing:
Camille P. Carter
Managing Editor: Shirley Thomas
Assistant Managing Editor: Kristen Parrish

CHARACTERIZATION OF MATERIALS

VOLUMES 1 AND 2

Characterization of Materials is available Online in full color at www.mrw.interscience.wiley.com/com.

WILEY-INTERSCIENCE
A John Wiley and Sons Publication

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging in Publication Data is available.

Characterization of Materials, 2 volume set
Elton N. Kaufmann, editor-in-chief

ISBN: 0-471-26882-8 (acid-free paper)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS, VOLUMES 1 AND 2

FOREWORD vii

PREFACE ix

CONTRIBUTORS xiii

COMMON CONCEPTS 1

Common Concepts in Materials Characterization, Introduction 1
General Vacuum Techniques 1
Mass and Density Measurements 24
Thermometry 30
Symmetry in Crystallography 39
Particle Scattering 51
Sample Preparation for Metallography 63

COMPUTATION AND THEORETICAL METHODS 71

Computation and Theoretical Methods, Introduction 71
Introduction to Computation 71
Summary of Electronic Structure Methods 74
Prediction of Phase Diagrams 90
Simulation of Microstructural Evolution Using the Field Method 112
Bonding in Metals 134
Binary and Multicomponent Diffusion 145
Molecular-Dynamics Simulation of Surface Phenomena 156
Simulation of Chemical Vapor Deposition Processes 166
Magnetism in Alloys 180
Kinematic Diffraction of X Rays 206
Dynamical Diffraction 224
Computation of Diffuse Intensities in Alloys 252

MECHANICAL TESTING 279

Mechanical Testing, Introduction 279
Tension Testing 279
High-Strain-Rate Testing of Materials 288
Fracture Toughness Testing Methods 302
Hardness Testing 316
Tribological and Wear Testing 324

THERMAL ANALYSIS 337

Thermal Analysis, Introduction 337
Thermal Analysis—Definitions, Codes of Practice, and Nomenclature 337
Thermogravimetric Analysis 344
Differential Thermal Analysis and Differential Scanning Calorimetry 362
Combustion Calorimetry 373
Thermal Diffusivity by the Laser Flash Technique 383
Simultaneous Techniques Including Analysis of Gaseous Products 392

ELECTRICAL AND ELECTRONIC MEASUREMENTS 401

Electrical and Electronic Measurement, Introduction 401
Conductivity Measurement 401
Hall Effect in Semiconductors 411
Deep-Level Transient Spectroscopy 418
Carrier Lifetime: Free Carrier Absorption, Photoconductivity, and Photoluminescence 427
Capacitance-Voltage (C-V) Characterization of Semiconductors 456
Characterization of *pn* Junctions 466
Electrical Measurements on Superconductors by Transport 472

MAGNETISM AND MAGNETIC MEASUREMENTS 491

Magnetism and Magnetic Measurement, Introduction 491
Generation and Measurement of Magnetic Fields 495
Magnetic Moment and Magnetization 511
Theory of Magnetic Phase Transitions 528
Magnetometry 531
Thermomagnetic Analysis 540
Techniques to Measure Magnetic Domain Structures 545
Magnetotransport in Metals and Alloys 559
Surface Magneto-Optic Kerr Effect 569

ELECTROCHEMICAL TECHNIQUES 579

Electrochemical Techniques, Introduction 579
Cyclic Voltammetry 580

Electrochemical Techniques for Corrosion Quantification 592
Semiconductor Photoelectrochemistry 605
Scanning Electrochemical Microscopy 636
The Quartz Crystal Microbalance in Electrochemistry 653

OPTICAL IMAGING AND SPECTROSCOPY 665

Optical Imaging and Spectroscopy, Introduction 665
Optical Microscopy 667
Reflected-Light Optical Microscopy 674
Photoluminescence Spectroscopy 681
Ultraviolet and Visible Absorption Spectroscopy 688
Raman Spectroscopy of Solids 698
Ultraviolet Photoelectron Spectroscopy 722
Ellipsometry 735
Impulsive Stimulated Thermal Scattering 744

RESONANCE METHODS 761

Resonance Methods, Introduction 761
Nuclear Magnetic Resonance Imaging 762
Nuclear Quadrupole Resonance 775
Electron Paramagnetic Resonance Spectroscopy 792
Cyclotron Resonance 805
Mössbauer Spectrometry 816

X-RAY TECHNIQUES 835

X-Ray Techniques, Introduction 835
X-Ray Powder Diffraction 835
Single-Crystal X-Ray Structure Determination 850
XAFS Spectroscopy 869
X-Ray and Neutron Diffuse Scattering Measurements 882
Resonant Scattering Techniques 905
Magnetic X-Ray Scattering 917
X-Ray Microprobe for Fluorescence and Diffraction Analysis 939
X-Ray Magnetic Circular Dichroism 953
X-Ray Photoelectron Spectroscopy 970
Surface X-Ray Diffraction 1007
X-Ray Diffraction Techniques for Liquid Surfaces and Monomolecular Layers 1027

ELECTRON TECHNIQUES 1049

Electron Techniques, Introduction 1049
Scanning Electron Microscopy 1050
Transmission Electron Microscopy 1063
Scanning Transmission Electron Microscopy: *Z*-Contrast Imaging 1090
Scanning Tunneling Microscopy 1111
Low-Energy Electron Diffraction 1120
Energy-Dispersive Spectrometry 1135
Auger Electron Spectroscopy 1157

ION-BEAM TECHNIQUES 1175

Ion-Beam Techniques, Introduction 1175
High-Energy Ion-Beam Analysis 1176
Elastic Ion Scattering for Composition Analysis 1179
Nuclear Reaction Analysis and Proton-Induced Gamma Ray Emission 1200
Particle-Induced X-Ray Emission 1210
Radiation Effects Microscopy 1223
Trace Element Accelerator Mass Spectrometry 1235
Introduction to Medium-Energy Ion Beam Analysis 1258
Medium-Energy Backscattering and Forward-Recoil Spectrometry 1259
Heavy-Ion Backscattering Spectrometry 1273

NEUTRON TECHNIQUES 1285

Neutron Techniques, Introduction 1285
Neutron Powder Diffraction 1285
Single-Crystal Neutron Diffraction 1307
Phonon Studies 1316
Magnetic Neutron Scattering 1328

INDEX 1341

FOREWORD

Whatever standards may have been used for materials research in antiquity, when fabrication was regarded more as an art than a science and tended to be shrouded in secrecy, an abrupt change occurred with the systematic discovery of the chemical elements two centuries ago by Cavendish, Priestly, Lavoisier, and their numerous successors. This revolution was enhanced by the parallel development of electrochemistry and eventually capped by the consolidating work of Mendeleyev which led to the periodic chart of the elements.

The age of materials science and technology had finally begun. This does not mean that empirical or trial and error work was abandoned as unnecessary. But rather that a new attitude had entered the field. The diligent fabricator of materials would welcome the development of new tools that could advance his or her work whether exploratory or applied. For example, electrochemistry became an intimate part of the armature of materials technology.

Fortunately, the physicist as well as the chemist were able to offer new tools. Initially these included such matters as a vast improvement of the optical microscope, the development of the analytic spectroscope, the discovery of x-ray diffraction and the invention of the electron microscope. Moreover, many other items such as isotopic tracers, laser spectroscopes and magnetic resonance equipment eventually emerged and were found useful in their turn as the science of physics and the demands for better materials evolved.

Quite apart from being used to re-evaluate the basis for the properties of materials that had long been useful, the new approaches provided much more important dividends. The ever-expanding knowledge of chemistry made it possible not only to improve upon those properties by varying composition, structure and other factors in controlled amounts, but revealed the existence of completely new materials that frequently turned out to be exceedingly useful. The mechanical properties of relatively inexpensive steels were improved by the additions of silicon, an element which had been produced first as a chemist's oddity. More complex ferrosilicon alloys revolutionized the performance of electric transformers. A hitherto all but unknown element, tungsten, provided a long-term solution in the search for a durable filament for the incandescent lamp. Eventually the chemists were to emerge with valuable families of organic polymers that replaced many natural materials.

The successes that accompanied the new approach to materials research and development stimulated an entirely new spirit of invention. What had once been dreams, such as the invention of the automobile and the airplane, were transformed into reality, in part through the modification of old materials and in part by creation of new ones. The growth in basic understanding of electromagnetic phenomena, coupled with the discovery that some materials possessed special electrical properties, encouraged the development of new equipment for power conversion and new methods of long-distance communication with the use of wired or wireless systems. In brief, the successes derived from the new approach to the development of materials had the effect of stimulating attempts to achieve practical goals which had previously seemed beyond reach. The technical base of society was being shaken to its foundations. And the end is not yet in sight.

The process of fabricating special materials for well defined practical missions, such as the development of new inventions or improving old ones, has, and continues to have, its counterpart in exploratory research that is carried out primarily to expand the range of knowledge and properties of materials of various types. Such investigations began in the field of mineralogy somewhat before the age of modern chemistry and were stimulated by the fact that many common minerals display regular cleavage planes and may exhibit unusual optical properties, such as different indices of refraction in different directions. Studies of this type became much broader and more systematic, however, once the variety of sophisticated exploratory tools provided by chemistry and physics became available. Although the groups of individuals involved in this work tended to live somewhat apart from the technologists, it was inevitable that some of their discoveries would eventually prove to be very useful. Many examples can be given. In the 1870s a young investigator who was studying the electrical properties of a group of poorly conducting metal sulfides, today classed among the family of semiconductors, noted that his specimens seemed to exhibit a different electrical conductivity when the voltage was applied in opposite directions. Careful measurements at a later date demonstrated that specially prepared specimens of silicon displayed this rectifying effect to an even more marked degree. Another investigator discovered a family of crystals that displayed surface

charges of opposite polarity when placed under unidirectional pressure, so called piezoelectricity. Natural radioactivity was discovered in a specimen of a uranium mineral whose physical properties were under study. Superconductivity was discovered incidentally in a systematic study of the electrical conductivity of simple metals close to the absolute zero of temperature. The possibility of creating a light-emitting crystal diode was suggested once wave mechanics was developed and began to be applied to advance our understanding of the properties of materials further. Actually, achievement of the device proved to be more difficult than its conception. The materials involved had to be prepared with great care.

Among the many avenues explored for the sake of obtaining new basic knowledge is that related to the influence of imperfections on the properties of materials. Some imperfections, such as those which give rise to temperature-dependent electrical conductivity in semiconductors, salts and metals could be ascribed to thermal fluctuations. Others were linked to foreign atoms which were added intentionally or occurred by accident. Still others were the result of deviations in the arrangement of atoms from that expected in ideal lattice structures. As might be expected, discoveries in this area not only clarified mysteries associated with ancient aspects of materials research, but provided tests that could have a bearing on the properties of materials being explored for novel purposes. The semiconductor industry has been an important beneficiary of this form of exploratory research since the operation of integrated circuits can be highly sensitive to imperfections.

In this connection, it should be added that the ever-increasing search for special materials that possess new or superior properties under conditions in which the sponsors of exploratory research and development and the prospective beneficiaries of the technological advance have parallel interests has made it possible for those engaged in the exploratory research to share in the funds directed toward applications. This has done much to enhance the degree of partnership between the scientist and engineer in advancing the field of materials research.

Finally, it should be emphasized again that whenever materials research has played a decisive role in advancing some aspect of technology, the advance has frequently been aided by the introduction of an increasingly sophisticated set of characterization tools that are drawn from a wide range of scientific disciplines. These tools usually remain a part of the array of test equipment.

FREDERICK SEITZ
President Emeritus, Rockefeller University
Past President, National Academy of Sciences, USA

PREFACE

Materials research is an extraordinarily broad and diverse field. It draws on the science, the technology, and the tools of a variety of scientific and engineering disciplines as it pursues research objectives spanning the very fundamental to the highly applied. Beyond the generic idea of a "material" *per se*, perhaps the single unifying element that qualifies this collection of pursuits as a field of research and study is the existence of a portfolio of characterization methods that is widely applicable irrespective of discipline or ultimate materials application. *Characterization of Materials* specifically addresses that portfolio with which researchers and educators must have working familiarity.

The immediate challenge to organizing the content for a methodological reference work is determining how best to parse the field. By far the largest number of materials researchers are focused on particular classes of materials and also perhaps on their uses. Thus a comfortable choice would have been to commission chapters accordingly. Alternatively, the objective and product of any measurement,—i.e., a materials property—could easily form a logical basis. Unfortunately, each of these approaches would have required mention of several of the measurement methods in just about every chapter. Therefore, if only to reduce redundancy, we have chosen a less intuitive taxonomy by arranging the content according to the type of measurement "probe" upon which a method relies. Thus you will find chapters focused on application of electrons, ions, x rays, heat, light, etc., to a sample as the generic thread tying several methods together. Our field is too complex for this not to be an oversimplification, and indeed some logical inconsistencies are inevitable.

We have tried to maintain the distinction between a property and a method. This is easy and clear for methods based on external independent probes such as electron beams, ion beams, neutrons, or x-rays. However many techniques rely on one and the same phenomenon for probe and property, as is the case for mechanical, electronic, and thermal methods. Many methods fall into both regimes. For example, light may be used to observe a microstructure, but may also be used to measure an optical property. From the most general viewpoint, we recognize that the properties of the measuring device and those of the specimen under study are inextricably linked. It is actually a joint property of the tool-plus-sample system that is observed. When both tool and sample each contribute their own *materials* properties—e.g., electrolyte and electrode, pin and disc, source and absorber, etc.—distinctions are blurred. Although these distinctions in principle ought not to be taken too seriously, keeping them in mind will aid in efficiently accessing content of interest in these volumes.

Frequently, the materials property sought is not what is directly measured. Rather it is deduced from direct observation of some other property or phenomenon that acts as a signature of what is of interest. These relationships take many forms. Thermal arrest, magnetic anomaly, diffraction spot intensity, relaxation rate and resistivity, to name only a few, might all serve as signatures of a phase transition and be used as "spectator" properties to determine a critical temperature. Similarly, inferred properties such as charge carrier mobility are deduced from basic electrical quantities and temperature-composition phase diagrams are deduced from observed microstructures. *Characterization of Materials*, being organized by technique, naturally places initial emphasis on the most directly measured properties, but authors have provided many application examples that illustrate the derivative properties a techniques may address.

First among our objectives is to help the researcher discriminate among alternative measurement modalities that may apply to the property under study. The field of possibilities is often very wide, and although excellent texts treating each possible method in great detail exist, identifying the most appropriate method before delving deeply into any one seems the most efficient approach. *Characterization of Materials* serves to sort the options at the outset, with individual articles affording the researcher a description of the method sufficient to understand its applicability, limitations, and relationship to competing techniques, while directing the reader to more extensive resources that fit specific measurement needs.

Whether one plans to perform such measurements oneself or whether one simply needs to gain sufficient familiarity to effectively collaborate with experts in the method, *Characterization of Materials* will be a useful reference. Although our expert authors were given great latitude to adjust their presentations to the "personalities" of their specific methods, some uniformity and circumscription of content was sought. Thus, you will find most

units organized in a similar fashion. First, an introduction serves to succinctly describe for what properties the method is useful and what alternatives may exist. Underlying physical principles of the method and practical aspects of its implementation follow. Most units will offer examples of data and their analyses as well as warnings about common problems of which one should be aware. Preparation of samples and automation of the methods are also treated as appropriate.

As implied above, the level of presentation of these volumes is intended to be intermediate between cursory overview and detailed instruction. Readers will find that, in practice, the level of coverage is also very much dictated by the character of the technique described. Many are based on quite complex concepts and devices. Others are less so, but still, of course, demand a precision of understanding and execution. What is or is not included in a presentation also depends on the technical background assumed of the reader. This obviates the need to delve into concepts that are part of rather standard technical curricula, while requiring inclusion of less common, more specialized topics.

As much as possible, we have avoided extended discussion of the science and application of the materials properties themselves, which, although very interesting and clearly the motivation for research in first place, do not generally speak to efficacy of a method or its accomplishment.

This is a *materials-oriented* volume, and as such, must overlap fields such as physics, chemistry, and engineering. There is no sharp delineation possible between a "physics" property (e.g., the band structure of a solid) and the materials consequences (e.g., conductivity, mobility, etc.) At the other extreme, it is not at all clear where a materials property such as toughness ends and an engineering property associated with performance and life-cycle begins. The very attempt to assign such concepts to only one disciplinary category serves no useful purpose. Suffice it to say, therefore, that *Characterization of Materials* has focused its coverage on a core of materials topics while trying to remain inclusive at the boundaries of the field.

Processing and fabrication are also important aspect of materials research. *Characterization of Materials* does not deal with these methods *per se* because they are not strictly measurement methods. However, here again no clear line is found and in such methods as electrochemistry, tribology, mechanical testing, and even ion-beam irradiation, where the processing can *be* the measurement, these aspects are perforce included.

The second chapter is unique in that it collects methods that are not, literally speaking, measurement methods; these articles do not follow the format found in subsequent chapters. As theory or simulation or modeling methods, they certainly serve to augment experiment. They may be a necessary corollary to an experiment to understand the result after the fact or to predict the result and thus help direct an experimental search in advance. More than this, as equipment needs of many experimental studies increase in complexity and cost, as the materials themselves become more complex and multicomponent in nature, and as computational power continues to expand, simulation of properties *will* in fact become the measurement method of choice in many cases.

Another unique chapter is the first, covering "common concepts." It collects some of the ubiquitous aspects of measurement methods that would have had to be described repeatedly and in more detail in later units. Readers may refer back to this chapter as related topics arise around specific methods, or they may use this chapter as a general tutorial. The Common Concepts chapter, however, does not and should not eliminate *all* redundancies in the remaining chapters. Expositions within individual articles attempt to be somewhat self-contained and the details as to how a common concept actually relates to a given method are bound to differ from one to the next. Although *Characterization of Materials* is directed more toward the research lab than the classroom, the focused units in conjunction with chapters one and two can serve as a useful educational tool.

The content of *Characterization of Materials* had previously appeared as *Methods in Materials Research*, a loose-leaf compilation amenable to updating. To retain the ability to keep content as up to date as possible, *Characterization of Materials* is also being published on-line where several new and expanded topics will be added over time.

ACKNOWLEDGMENTS

First we express our appreciation to the many expert authors who have contributed to *Characterization of Materials*. On the production side of the predecessor publication, *Methods in Materials Research*, we are pleased to acknowledge the work of a great many staff of the Current Protocols division of John Wiley & Sons, Inc. We also thank the previous series editors, Dr. Virginia Chanda and Dr. Alan Samuels. Republication in the present on-line and hard-bound forms owes its continuing quality to staff of the Major Reference Works group of John Wiley & Sons, Inc., most notably Dr. Jacqueline Kroschwitz and Dr. Arza Seidel.

For the editors,

ELTON N. KAUFMANN
Editor-in-Chief

CONTRIBUTORS

Reza Abbaschian
University of Florida at Gainesville
Gainesville, FL
Mechanical Testing, Introduction

John Ågren
Royal Institute of Technology, KTH
Stockholm, SWEDEN
Binary and Multicomponent Diffusion

Stephen D. Antolovich
Washington State University
Pullman, WA
Tension Testing

Samir J. Anz
California Institute of Technology
Pasadena, CA
Semiconductor Photoelectrochemistry

Georgia A. Arbuckle-Keil
Rutgers University
Camden, NJ
The Quartz Crystal Microbalance In Electrochemistry

Ljubomir Arsov
University of Kiril and Metodij
Skopje, MACEDONIA
Ellipsometry

Albert G. Baca
Sandia National Laboratories
Albuquerque, NM
Characterization of pn Junctions

Sam Bader
Argonne National Laboratory
Argonne, IL
Surface Magneto-Optic Kerr Effect

James C. Banks
Sandia National Laboratories
Albuquerque, NM
Heavy-Ion Backscattering Spectrometry

Charles J. Barbour
Sandia National Laboratory
Albuquerque, NM
Elastic Ion Scattering for Composition Analysis

Peter A. Barnes
Clemson University
Clemson, SC
Electrical and Electronic Measurements, Introduction
Capacitance-Voltage (C-V) Characterization of Semiconductors

Jack Bass
Michigan State University
East Lansing, MI
Magnetotransport in Metals and Alloys

Bob Bastasz
Sandia National Laboratories
Livermore, CA
Particle Scattering

Raymond G. Bayer
Consultant
Vespal, NY
Tribological and Wear Testing

Goetz M. Bendele
SUNY Stony Brook
Stony Brook, NY
X-Ray Powder Diffraction

Andrew B. Bocarsly
Princeton University
Princeton, NJ
Cyclic Voltammetry
Electrochemical Techniques, Introduction

Mark B.H. Breese
University of Surrey, Guildford
Surrey, UNITED KINGDOM
Radiation Effects Microscopy

Iain L. Campbell
University of Guelph
Guelph, Ontario CANADA
Particle-Induced X-Ray Emission

Gerbrand Ceder
Massachusetts Institute of Technology
Cambridge, MA
Introduction to Computation

Robert Celotta
National Institute of Standards and
Technology Gaithersburg, MD
Techniques to Measure Magnetic Domain Structures

Gary W. Chandler
University of Arizona
Tucson, AZ
Scanning Electron Microscopy

Haydn H. Chen
University of Illinois
Urbana, IL
Kinematic Diffraction of X Rays

Long-Qing Chen
Pennsylvania State University
University Park, PA
Simulation of Microstructural Evolution Using the
Field Method

Chia-Ling Chien
Johns Hopkins University
Baltimore, MD
Magnetism and Magnetic Measurements, Introduction

J.M.D. Coey
University of Dublin, Trinity College
Dublin, IRELAND
Generation and Measurement of Magnetic Fields

Richard G. Connell
University of Florida
Gainesville, FL
Optical Microscopy Reflected-Light
Optical Microscopy

Didier de Fontaine
University of California
Berkeley, CA
Prediction of Phase Diagrams

T.M. Devine
University of California
Berkeley, CA
Raman Spectroscopy of Solids

David Dollimore
University of Toledo
Toledo, OH
Mass and Density Measurements Thermal Analysis-
Definitions, Codes of Practice, and Nomenclature
Thermometry
Thermal Analysis, Introduction

Barney L. Doyle
Sandia National Laboratory
Albuquerque, NM
High-Energy Ion Beam Analysis
Ion-Beam Techniques, Introduction

Jeff G. Dunn
University of Toledo
Toledo, OH
Thermogravimetric Analysis

Gareth R. Eaton
University of Denver
Denver, CO
Electron Paramagnetic Resonance
Spectroscopy

Sandra S. Eaton
University of Denver
Denver, CO
Electron Paramagnetic Resonance
Spectroscopy

Fereshteh Ebrahimi
University of Florida
Gainesville, FL
Fracture Toughness Testing Methods

Wolfgang Eckstein
Max-Planck-Institut fur Plasmaphysik
Garching, GERMANY
Particle Scattering

Arnel M. Fajardo
California Institute of Technology
Pasadena, CA
Semiconductor Photoelectrochemistry

Kenneth D. Finkelstein
Cornell University
Ithaca, NY
Resonant Scattering Technique

Simon Foner
Massachusetts Institute of Technology
Cambridge, MA
Magnetometry

Brent Fultz
California Institute of Technology
Pasadena, CA
Electron Techniques, Introduction
Mössbauer Spectrometry
Resonance Methods, Introduction
Transmission Electron Microscopy

Jozef Gembarovic
Thermophysical Properties Research Laboratory
West Lafayette, IN
Thermal Diffusivity by the Laser
Flash Technique

Craig A. Gerken
University of Illinois
Urbana, IL
Low-Energy, Electron Diffraction

Atul B. Gokhale
MetConsult, Inc.
Roosevelt Island, NY
Sample Preparation for Metallography

Alan I. Goldman
Iowa State University
Ames, IA
X-Ray Techniques, Introduction
Neutron Techniques, Introduction

John T. Grant
University of Dayton
Dayton, OH
Auger Electron Spectroscopy

George T. Gray
Los Alamos National Laboratory
Los Alamos, NM
High-Strain-Rate Testing of Materials

Vytautas Grivickas
Vilnius University
Vilnius, LITHUANIA
Carrier Lifetime: Free Carrier Absorption, Photoconductivity, and Photoluminescence

Robert P. Guertin
Tufts University
Medford, MA
Magnetometry

Gerard S. Harbison
University of Nebraska
Lincoln, NE
Nuclear Quadrupole Resonance

Steve Heald
Argonne National Laboratory
Argonne, IL
XAFS Spectroscopy

Bruno Herreros
University of Southern California
Los Angeles, CA
Nuclear Quadrupole Resonance

John P. Hill
Brookhaven National Laboratory
Upton, NY
Magnetic X-Ray Scattering
Ultraviolet Photoelectron Spectroscopy

Kevin M. Horn
Sandia National Laboratories
Albuquerque, NM
Ion Beam Techniques, Introduction

Joseph P. Hornak
Rochester Institute of Technology
Rochester, NY
Nuclear Magnetic Resonance Imaging

James M. Howe
University of Virginia
Charlottesville, VA
Transmission Electron Microscopy

Gene E. Ice
Oak Ridge National Laboratory
Oak Ridge, TN
X-Ray Microprobe for Fluorescence and Diffraction
Analysis X-Ray and Neutron Diffuse Scattering Measurements

Robert A. Jacobson
Iowa State University
Ames, IA
Single-Crystal X-Ray Structure Determination

Duane D. Johnson
University of Illinois
Urbana, IL
Computation of Diffuse Intensities in Alloys
Magnetism in Alloys

Michael H. Kelly
National Institute of Standards and Technology
Gaithersburg, MD
Techniques to Measure Magnetic Domain Structures

Elton N. Kaufmann
Argonne National Laboratory
Argonne, IL
Common Concepts in Materials Characterization, Introduction

Janice Klansky
Beuhler Ltd.
Lake Bluff, IL
Hardness Testing

Chris R. Kleijn
Delft University of Technology
Delft, THE NETHERLANDS
Simulation of Chemical Vapor Deposition Processes

James A. Knapp
Sandia National Laboratories
Albuquerque, NM
Heavy-Ion Backscattering Spectrometry

Thomas Koetzle
Brookhaven National Laboratory
Upton, NY
Single-Crystal Neutron Diffraction

Junichiro Kono
Rice University
Houston, TX
Cyclotron Resonance

Phil Kuhns
Florida State University
Tallahassee, FL
Generation and Measurement of Magnetic Fields

Jonathan C. Lang
Argonne National Laboratory
Argonne, IL
X-Ray Magnetic Circular Dichroism

David E. Laughlin
Carnegie Mellon University
Pittsburgh, PA
Theory of Magnetic Phase Transitions

Leonard Leibowitz
Argonne National Laboratory
Argonne, IL
Differential Thermal Analysis and Differential Scanning Calorimetry

Supaporn Lerdkanchanaporn
University of Toledo
Toledo, OH
Simultaneouse Techniques Including Analysis of Gaseous Products

Nathan S. Lewis
California Institute of Technology
Pasadena, CA
Semiconductor Photoelectrochemistry

Dusan Lexa
Argonne National Laboratory
Argonne, IL
Differential Thermal Analysis and Differential Scanning Calorimetry

Jan Linnros
Royal Institute of Technology
Kista-Stockholm, SWEDEN
Carrier Liftime: Free Carrier Absorption, Photoconductivity, and Photoluminescene

David C. Look
Wright State University
Dayton, OH
Hall Effect in Semiconductors

Jeffery W. Lynn
University of Maryland
College Park, MD
Magentic Neutron Scattering

Kosta Maglic
Institute of Nuclear Sciences "Vinca"
Belgrade, YUGOSLAVIA
Thermal Diffusivity by the Laser Flash Technique

Floyd McDaniel
University of North Texas
Denton, TX
Trace Element Accelerator Mass Spectrometry

Michael E. McHenry
Carnegie Mellon University
Pittsburgh, PA
Magnetic Moment and Magnetization
Thermomagnetic Analysis
Theory of Magnetic Phase Transitions

Keith A. Nelson
Massachusetts Institute of Technology
Cambridge, MA
Impulsive Stimulated Thermal Scattering

Dale E. Newbury
National Institute of Standards and Technology
Gaithersburg, MD
Energy-Dispersive Spectrometry

P.A.G. O'Hare
Darien, IL
Combustion Calorimetry

Stephen J. Pennycook
Oak Ridge National Laboratory
Oak Ridge, TN
Scanning Transmission Electron Microscopy: Z-Contrast Imaging

Daniel T. Pierce
National Institute of Standards and Technology
Gaithersburg, MD
Techniques to Measure Magnetic Domain Structures

Frank J. Pinski
University of Cincinnati
Cincinnati, OH
Magnetism in Alloys
Computation of Diffuse Intensities in Alloys

Branko N. Popov
University of South Carolina
Columbia, SC
Ellipsometry

Ziqiang Qiu
University of California at Berkeley
Berkeley, CA
Surface Magneto-Optic Kerr Effect

Talat S. Rahman
Kansas State University
Manhattan, Kansas
Molecular-Dynamics Simulation of Surface Phenomena

T.A. Ramanarayanan
Exxon Research and Engineering Corp.
Annandale, NJ
Electrochemical Techniques for Corrosion Quantification

M. Ramasubramanian
University of South Carolina
Columbia, SC
Ellipsometry

S.S.A. Razee
University of Warwick
Coventry, UNITED KINGDOM
Magnetism in Alloys

James L. Robertson
Oak Ridge National Laboratory
Oak Ridge, TN
X-Ray and Neutron Diffuse Scattering Measurements

Ian K. Robinson
University of Illinois
Urbana, IL
Surface X-Ray Diffraction

John A. Rogers
Bell Laboratories, Lucent Technologies
Murray Hill, NJ
Impulsive Stimulated Thermal Scattering

William J. Royea
California Institute of Technology
Pasadena, CA
Semiconductor Photoelectrochemistry

Larry Rubin
Massachusetts Institute of Technology
Cambridge, MA
Generation and Measurement of Magnetic Fields

Miquel Salmeron
Lawrence Berkeley National Laboratory
Berkeley, CA
Scanning Tunneling Microscopy

Alan C. Samuels
Edgewood Chemical Biological Center
Aberdeen Proving Ground, MD
Mass and Density Measurements
Optical Imaging and Spectroscopy, Introduction
Thermometry

Juan M. Sanchez
University of Texas at Austin
Austin, TX
Computational and Theoretical Methods, Introduction

Hans J. Schneider-Muntau
Florida State University
Tallahassee, FL
Generation and Measurement of Magnetic Fields

Christian Schott
Swiss Federal Institute of Technology
Lausanne, SWITZERLAND
Generation and Measurement of Magnetic Fields

Justin Schwartz
Florida State University
Tallahassee, FL
Electrical Measurements on Superconductors by Transport

Supapan Seraphin
University of Arizona
Tucson, AZ
Scanning Electron Microscopy

Qun Shen
Cornell University
Ithaca, NY
Dynamical Diffraction Y

Jack Singleton
Consultant
Monroeville, PA
General Vacuum Techniques

Gabor A. Somorjai
University of California & Lawrence Berkeley National Laboratory
Berkeley, CA
Low-Energy Electron Diffraction

Cullie J. Sparks
Oak Ridge National Laboratory
Oak Ridge, TN
X-Ray and Neutron Diffuse Scattering Measurements

Costas Stassis
Iowa State University
Ames, IA
Phonon Studies

Julie B. Staunton
University of Warwick
Coventry, UNITED KINGDOM
Computation of Diffuse Intensities in Alloys
Magnetism in Alloys

Hugo Steinfink
University of Texas
Austin, TX
Symmetry in Crystallography

Peter W. Stephens
SUNY Stony Brook
Stony Brook, NY
X-Ray Powder Diffraction

Ray E. Taylor
Thermophysical Properties Research Laboratory
West Lafayette, IN 47906
Thermal Diffusivity by the Laser Flash Technique

Chin-Che Tin
Auburn University
Auburn, AL
Deep-Level Transient Spectroscopy

Brian M. Tissue
Virginia Polytechnic Institute & State University
Blacksburg, VA
Ultraviolet and Visible Absorption Spectroscopy

James E. Toney
Applied Electro-Optics Corporation
Bridgeville, PA
Photoluminescene Spectroscopy

John Unguris
National Institute of Standards and Technology
Gaithersburg, MD
Techniques to Measure Magnetic Domain Structures

David Vaknin
Iowa State University
Ames, IA
X-Ray Diffraction Techniques for Liquid Surfaces and Monomolecular Layers

Mark van Schilfgaarde
SRI International
Menlo Park, California
Summary of Electronic Structure Methods

György Vizkelethy
Sandia National Laboratories
Albuquerque, NM
Nuclear Reaction Analysis and Proton-Induced Gamma Ray Emission

Thomas Vogt
Brookhaven National Laboratory
Upton, NY
Neutron Powder Diffraction

Yunzhi Wang
Ohio State University
Columbus, OH
Simulation of Microstructural Evolution Using the Field Method

Richard E. Watson
Brookhaven National Laboratory
Upton, NY
Bonding in Metals

Huub Weijers
Florida State University
Tallahassee, FL
Electrical Measurements on Superconductors by Transport

Jefferey Weimer
University of Alabama
Huntsville, AL
X-Ray Photoelectron Spectroscopy

Michael Weinert
Brookhaven National Laboratory
Upton, NY
Bonding in Metals

Robert A. Weller
Vanderbilt University
Nashville, TN
Introduction To Medium-Energy Ion Beam Analysis
Medium-Energy Backscattering and Forward-Recoil Spectrometry

Stuart Wentworth
Auburn University
Auburn University, AL
Conductivity Measurement

David Wipf
Mississippi State University
Mississippi State, MS
Scanning Electrochemical Microscopy

Gang Xiao
Brown University
Providence, RI
Magnetism and Magnetic Measurements, Introduction

CHARACTERIZATION OF MATERIALS

This page intentionally left blank

COMMON CONCEPTS

COMMON CONCEPTS IN MATERIALS CHARACTERIZATION, INTRODUCTION

From a tutorial standpoint, one may view this chapter as a good preparatory entrance to subsequent chapters of *Characterization of Materials*. In an educational setting, the generally applicable topics of the units in this chapter can play such a role, notwithstanding that they are each quite independent without having been sequenced with any pedagogical thread in mind.

In practice, we expect that each unit of this chapter will be separately valuable to users of *Characterization of Materials* as they choose to refer to it for concepts underlying many of those exposed in units covering specific measurement methods.

Of course, not every topic covered by a unit in this chapter will be relevant to *every* measurement method covered in subsequent chapters. However, the concepts in this chapter are sufficiently common to appear repeatedly in the pursuit of materials research. It can be argued that the units treating vacuum techniques, thermometry, and sample preparation do not deal directly with the materials properties to be measured at all. Rather, they are crucial to preparation and implementation of such a measurement. It is interesting to note that the properties of materials nevertheless play absolutely crucial roles for each of these topics as they rely on materials performance to accomplish their ends.

Mass/density measurement does of course relate to a most basic materials property, but is itself more likely to be an ancillary necessity of a measurement protocol than to be the end goal of a measurement (with the important exceptions of properties related to porosity, defect density, etc.). In temperature and mass measurement, appreciating the role of standards and definitions is central to proper use of these parameters.

It is hard to think of a materials property that does not depend on the crystal structure of the materials in question. Whether the structure is a known part of the explanation of the value of another property or its determination is itself the object of the measurement, a good grounding in essentials of crystallographic groups and syntax is a common need in most measurement circumstances. A unit provided in this chapter serves that purpose well.

Several chapters in *Characterization of Materials* deal with impingement of projectiles of one kind or another on a sample, the reaction to which reflects properties of interest in the target. Describing the scattering of the projectiles is necessary in all these cases. Many concepts in such a description are similar regardless of projectile type, while the details differ greatly among ions, electrons, neutrons, and photons. Although the particle scattering unit in this chapter emphasizes the charged particle and ions in particular, the concepts are somewhat portable. A good deal of generic scattering background is provided in the chapters covering neutrons, x rays, and electrons as projectiles as well.

As *Characterization of Materials* evolves, additional common concepts will be added. However, when it seems more appropriate, such content will appear more closely tied to its primary topical chapter.

ELTON N. KAUFMANN

GENERAL VACUUM TECHNIQUES

INTRODUCTION

In this unit we discuss the procedures and equipment used to maintain a vacuum system at pressures in the range from 10^{-3} to 10^{-11} torr. Total and partial pressure gauges used in this range are also described.

Because there is a wide variety of equipment, we describe each of the various components, including details of their principles and technique of operation, as well as their recommended uses.

SI units are not used in this unit. The American Vacuum Society attempted their introduction many years ago, but the more traditional units continue to dominate in this field in North America. Our usage will be consistent with that generally found in the current literature. The following units will be used.

Pressure is given in torr. 1 torr is equivalent to 133.32 pascal (Pa).

Volume is given in liters (L), and time in seconds (s).

The flow of gas through a system, i.e., the "throughput" (Q), is given in torr-L/s.

Pumping speed (S) and conductance (C) are given in L/s.

PRINCIPLES OF VACUUM TECHNOLOGY

The most difficult step in designing and building a vacuum system is defining precisely the conditions required to fulfill the purpose at hand. Important factors to consider include:

1. The required system operating pressure and the gaseous impurities that must be avoided;
2. The frequency with which the system must be vented to the atmosphere, and the required recycling time;
3. The kind of access to the vacuum system needed for the insertion or removal of samples.

For systems operating at pressures of 10^{-6} to 10^{-7} torr, venting the system is the simplest way to gain access, but for ultrahigh vacuum (UHV), e.g., below 10^{-8} torr, the pumpdown time can be very long, and system bakeout would usually be required. A vacuum load-lock antechamber for the introduction and removal of samples may be essential in such applications.

Because it is difficult to address all of the above questions, a viable specification of system performance is often neglected, and it is all too easy to assemble a more sophisticated and expensive system than necessary, or, if budgets are low, to compromise on an inadequate system that cannot easily be upgraded.

Before any discussion of the specific components of a vacuum system, it is instructive to consider the factors that govern the ultimate, or base, pressure. The pressure can be calculated from

$$P = \frac{Q}{S} \tag{1}$$

where P is the pressure in torr, Q is the total flow, or throughput of gas, in torr-L/s, and S is the pumping speed in L/s.

The influx of gas, Q, can be a combination of a deliberate influx of process gas from an exterior source and gas originating in the system itself. With no external source, the base pressure achieved is frequently used as the principle indicator of system performance. The most important internal sources of gas are outgassing from the walls and permeation from the atmosphere, most frequently through elastomer O-rings. There may also be leaks, but these can readily be reduced to negligible levels by proper system design and construction. Vacuum pumps also contribute to background pressure, and here again careful selection and operation will minimize such problems.

The Problem of Outgassing

Of the sources of gas described above, outgassing is often the most important. With a new system, the origin of outgassing may be in the manufacture of the materials used in construction, in handling during construction, and in exposure of the system to the atmosphere. In general these sources scale with the area of the system walls, so that it is wise to minimize the surface area and to avoid porous materials in construction. For example, aluminum is an excellent choice for use in vacuum systems, but anodized aluminum has a porous oxide layer that provides an internal surface for gas adsorption many times greater than the apparent surface, making it much less suitable for use in vacuum.

The rate of outgassing in a new, unbaked system, fabricated from materials such as aluminum and stainless steel, is initially very high, on the order of 10^{-6} to 10^{-7} torr-L/s $\cdot$ cm^2 of surface area after one hour of exposure to vacuum (O'Hanlon, 1989). With continued pumping, the rate falls by one or two orders of magnitude during the first 24 hr, but thereafter drops very slowly over many months. Typically the main residual gas is water vapor. In a clean vacuum system, operating at ambient temperature and containing only a moderate number of O-rings, the lowest achievable pressure is usually 10^{-7} to mid-10^{-8} torr. The limiting factor is generally residual outgassing, not the capability of the high-vacuum pump.

The outgassing load is highest when a new system is put into service, but with steady use the sins of construction are slowly erased, and on each subsequent evacuation, the system will reach its typical base pressure more rapidly. However, water will persist as the major outgassing load. Every time a system is vented to air, the walls are exposed to moisture and one or more layers of water will adsorb virtually instantaneously. The amount adsorbed will be greatest when the relative humidity is high, increasing the time needed to reach base pressure. Water is bound by physical adsorption, a reversible process, but the binding energy of adsorption is so great that the rate of desorption is slow at ambient temperature. Physical adsorption involves van der Waal's forces, which are relatively weak. Physical adsorption should be distinguished from chemisorption, which typically involves the formation of chemical-type bonding of a gas to an atomically clean surface—for example, oxygen on a stainless steel surface. Chemisorption of gas is irreversible under all conditions normally encountered in a vacuum system. After the first few minutes of pumping, pressures are almost always in the free molecular flow regime, and when a water molecule is desorbed, it experiences only collisions with the walls, rather than with other molecules. Consequently, as it leaves the system, it is readsorbed many times, and on each occasion desorption is a slow process.

One way of accelerating the removal of adsorbed water is by purging at a pressure in the viscous flow region, using a dry gas such as nitrogen or argon. Under viscous flow conditions, the desorbed water molecules rarely reach the system walls, and readsorption is greatly reduced. A second method is to heat the system above its normal operating temperature.

Any process that reduces the adsorption of water in a vacuum system will improve the rate of pumpdown. The simplest procedure is to vent a vacuum system with a dry gas rather than with atmospheric air, and to minimize the time the system remains open following such a procedure. Dry air will work well, but it is usually more convenient to substitute nitrogen or argon.

From Equation 1, it is evident that there are two approaches to achieving a lower ultimate pressure, and hence a low impurity level, in a system. The first is to increase the effective pumping speed, and the second is to reduce the outgassing rate. There are severe limitations to the first approach. In a typical system, most of one wall of the chamber will be occupied by the connection to the high-vacuum pump; this limits the size of pump that can be used, imposing an upper limit on the achievable pumping speed. As already noted, the ultimate pressure achieved in an unbaked system having this configuration will rarely reach the mid-10^{-8} torr range. Even if one could mount a similar-sized pump on every side, the best to be expected would be a 6-fold improvement, achieving a base pressure barely into the 10^{-9} torr range, even after very long exhaust times.

It is evident that, to routinely reach pressures in the 10^{-10} torr range in a realistic period of time, a reduction in the rate of outgassing is necessary—e.g., by heating the vacuum system. Baking an entire system to 400°C for 16 hr can produce outgassing rates of 10^{-15} torr-L/s $\cdot$ cm^2 (Alpert, 1959), a reduction of 10^8 from those found after 1 hr of pumping at ambient temperature. The magnitude of this reduction shows that as large a portion as

possible of a system should be heated to obtain maximum advantage.

PRACTICAL ASPECTS OF VACUUM TECHNOLOGY

Vacuum Pumps

The operation of most vacuum systems can be divided into two regimes. The first involves pumping the system from atmosphere to a pressure at which a high-vacuum pump can be brought into operation. This is traditionally known as the rough vacuum regime and the pumps used are commonly referred to as roughing pumps. Clearly, a system that operates at an ultimate pressure within the capability of the roughing pump will require no additional pumps.

Once the system has been roughed down, a high-vacuum pump must be used to achieve lower pressures. If the high-vacuum pump is the type known as a transfer pump, such as a diffusion or turbomolecular pump, it will require the continuous support of the roughing pump in order to maintain the pressure at the exit of the high-vacuum pump at a tolerable level (in this phase of the pumping operation the function of the roughing pump has changed, and it is frequently referred to as a backing or forepump). Transfer pumps have the advantage that their capacity for continuous pumping of gas, within their operating pressure range, is limited only by their reliability. They do not accumulate gas, an important consideration where hazardous gases are involved. Note that the reliability of transfer pumping systems depends upon the satisfactory performance of two separate pumps. A second class of pumps, known collectively as capture pumps, require no further support from a roughing pump once they have started to pump. Examples of this class are cryogenic pumps and sputter-ion pumps. These types of pump have the advantage that the vacuum system is isolated from the atmosphere, so that system operation depends upon the reliability of only one pump. Their disadvantage is that they can provide only limited storage of pumped gas, and as that limit is reached, pumping will deteriorate. The effect of such a limitation is quite different for the two examples cited. A cryogenic pump can be totally regenerated by a brief purging at ambient temperature, but a sputter-ion pump requires replacement of its internal components. One aspect of the cryopump that should not be overlooked is that hazardous gases are stored, unchanged, within the pump, so that an unexpected failure of the pump can release these accumulated gases, requiring provision for their automatic safe dispersal in such an emergency.

Roughing Pumps

Two classes of roughing pumps are in use. The first type, the oil-sealed mechanical pump, is by far the most common, but because of the enormous concern in the semiconductor industry about oil contamination, a second type, the so-called "dry" pump, is now frequently used. In this context, "dry" implies the absence of volatile organics in the part of the pump that communicates with the vacuum system.

Oil-Sealed Pumps

The earliest roughing pumps used either a piston or liquid to displace the gas. The first production methods for incandescent lamps used such pumps, and the development of the oil-sealed mechanical pump by Gaede, around 1907, was driven by the need to accelerate the pumping process.

Applications. The modern versions of this pump are the most economic and convenient for achieving pressures as low as the 10^{-4} torr range. The pumps are widely used as a backing pump for both diffusion and turbomolecular pumps; in this application the backstreaming of mechanical pump oil is intercepted by the high vacuum pump, and a foreline trap is not required.

Operating Principles. The oil-sealed pump is a positive-displacement pump, of either the vane or piston type, with a compression ratio of the order of 10^5:1 (Dobrowolski, 1979). It is available as a single or two-stage pump, capable of reaching base pressures in the 10^{-2} and 10^{-4} torr range, respectively. The pump uses oil to maintain sealing, and to provide lubrication and heat transfer, particularly at the contact between the sliding vanes and the pump wall. Oil also serves to fill the significant dead space leading to the exhaust valve, essentially functioning as a hydraulic valve lifter and permitting the very high compression ratio.

The speed of such pumps is often quoted as the "free-air displacement," which is simply the volume swept by the pump rotor. In a typical two-stage pump this speed is sustained down to $\sim 1 \times 10^{-1}$ torr; below this pressure the speed decreases, reaching zero in the 10^{-5} torr range. If a pump is to sustain pressures near the bottom of its range, the required pump size must be determined from published pumping-speed performance data. It should be noted that mechanical pumps have relatively small pumping speed, at least when compared with typical high-vacuum pumps. A typical laboratory-sized pump, powered by a 1/3 hp motor, may have a speed of $\sim$3.5 cubic feet per minute (cfm), or rather less than 2 L/s, as compared to the smallest turbomolecular pump, which has a rated speed of 50 L/s.

Avoiding Oil Contamination from an Oil-Sealed Mechanical Pump. The versatility and reliability of the oil-sealed mechanical pump carries with it a serious penalty. When used improperly, contamination of the vacuum system is inevitable. These pumps are probably the most prevalent source of oil contamination in vacuum systems. The problem arises when thay are untrapped and pump a system down to its ultimate pressure, often in the free molecular flow regime. In this regime, oil molecules flow freely into the vacuum chamber. The problem can readily be avoided by careful control of the pumping procedures, but possible system or operator malfunction, leading to contamination, must be considered. For many years, it was common practice to leave a system in the standby condition evacuated only by an untrapped mechanical pump, making contamination inevitable.

Mechanical pump oil has a vapor pressure, at room temperature, in the low 10^{-5} torr range when first installed, but this rapidly deteriorates up to two orders of magnitude as the pump is operated (Holland, 1971). A pump operates at temperatures of 60°C, or higher, so the oil vapor pressure far exceeds 10^{-3} torr, and evaporation results in a substantial flux of oil into the roughing line. When a system at atmospheric pressure is connected to the mechanical pump, the initial gas flow from the vacuum chamber is in the viscous flow regime, and oil molecules are driven back to the pump by collisions with the gas being exhausted (Holland, 1971; Lewin, 1985). Provided the roughing process is terminated while the gas flow is still in the viscous flow regime, no significant contamination of the vacuum chamber will occur. The condition for viscous flow is given by the equation

$$PD \geq 0.5 \tag{2}$$

where P is the pressure in torr and D is the internal diameter of the roughing line in centimeters.

Termination of the roughing process in the viscous flow region is entirely practical when the high-vacuum pump is either a turbomolecular or modern diffusion pump (see precautions discussed under Diffusion Pumps and Turbomolecular Pumps, below). Once these pumps are in operation, they function as an effective barrier against oil migration into the system from the forepump. Hoffman (1979) has described the use of a continuous gas purge on the foreline of a diffusion-pumped system as a means of avoiding backstreaming from the forepump.

Foreline Traps. A foreline trap is a second approach to preventing oil backstreaming. If a liquid nitrogen–cooled trap is always in place between a forepump and the vacuum chamber, cleanliness is assured. But the operative word is "always." If the trap warms to ambient temperature, oil from the trap will migrate upstream, and this is much more serious if it occurs while the line is evacuated. A different class of trap uses an adsorbent for oil. Typical adsorbents are activated alumina, molecular sieve (a synthetic zeolite), a proprietary ceramic (Micromaze foreline traps; Kurt J. Lesker Co.), and metal wool. The metal wool traps have much less capacity than the other types, and unless there is evidence of their efficacy, they are best avoided. Published data show that activated alumina can trap 99% of the backstreaming oil molecules (Fulker, 1968). However, one must know when such traps should be reactivated. Unequivocal determination requires insertion of an oil-detection device, such as a mass spectrometer, on the foreline. The saturation time of a trap depends upon the rate of oil influx, which in turn depends upon the vapor pressure of oil in the pump and the conductance of the line between pump and trap. The only safe procedure is frequent reactivation of traps on a conservative schedule. Reactivation may be done by venting the system, replacing the adsorbent with a new charge, or by baking the adsorbent in a stream of dry air or inert gas to a temperature of ~300°C for several hours. Some traps can be regenerated by heating *in situ*, but only using a stream of inert gas, at a pressure in the viscous flow region, flowing from the system side of the trap to the pump (D.J. Santeler, pers. comm.). The foreline is isolated from the rest of the system and the gas flow is continued throughout the heating cycle, until the trap has cooled back to ambient temperature. An adsorbent foreline trap must be optically dense, so the oil molecules have no path past the adsorbent; commercial traps do not always fulfill this basic requirement. Where regeneration of the foreline trap has been totally neglected, acceptable performance may still be achieved simply because a diffusion pump or turbomolecular pump serves as the true "trap," intercepting the oil from the forepump.

Oil contamination can also result from improperly turning a pump off. If it is stopped and left under vacuum, oil frequently leaks slowly across the exhaust valve into the pump. When it is partially filled with oil, a hydraulic lock may prevent the pump from starting. Continued leakage will drive oil into the vacuum system itself; an interesting procedure for recovery from such a catastrophe has been described (Hoffman, 1979).

Whenever the pump is stopped, either deliberately or by power failure or other failure, automatic controls that first isolate it from the vacuum system, and then vent it to atmospheric pressure, should be used.

Most gases exhausted from a system, including oxygen and nitrogen, are readily removed from the pump oil, but some can liquify under maximum compression just before the exhaust valve opens. Such liquids mix with the oil and are more difficult to remove. They include water and solvents frequently used to clean system components. When pumping large volumes of air from a vacuum chamber, particularly during periods of high humidity (or whenever solvent residues are present), it is advantageous to use a gas-ballast feature commonly fitted to two-stage and also to some single-stage pumps. This feature admits air during the final stage of compression, raising the pressure and forcing the exhaust valve to open before the partial pressure of water has reached saturation. The ballast feature minimizes pump contamination and reduces pump-down time for a chamber exposed to humid air, although at the cost of about ten-times-poorer base pressure.

Oil-Free ("Dry") Pumps

Many different types of oil-free pumps are available. We will emphasize those that are most useful in analytical and diagnostic applications.

Diaphragm Pumps

Applications: Diaphragm pumps are increasingly used where the absence of oil is an imperative, for example, as the forepump for compound turbomolecular pumps that incorporate a molecular drag stage. The combination renders oil contamination very unlikely. Most diaphragm pumps have relatively small pumping speeds. They are adequate once the system pressure reaches the operating range of a turbomolecular pump, usually well below 10^{-2} torr, but not for rapidly roughing down a large volume. Pumps are available with speeds up to several liters per second, and base pressures from a few torr to as low as 10^{-3} torr, lower ultimate pressures being associated with the lower-speed pumps.

Operating Principles: Four diaphragm modules are often arranged in three separate pumping stages, with the lowest-pressure stage served by two modules in tandem to boost the capacity. Single modules are adequate for subsequent stages, since the gas has already been compressed to a smaller volume. Each module uses a flexible diaphragm of Viton or other elastomer, as well as inlet and outlet valves. In some pumps the modules can be arranged to provide four stages of pumping, providing a lower base pressure, but at lower pumping speed because only a single module is employed for the first stage. The major required maintenance in such pumps is replacement of the diaphragm after 10,000 to 15,000 hr of operation.

Scroll Pumps

Applications: Scroll pumps (Coffin, 1982; Hablanian, 1997) are used in some refrigeration systems, where the limited number of moving parts is reputed to provide high reliability. The most recent versions introduced for general vacuum applications have the advantages of diaphragm pumps, but with higher pumping speed. Published speeds on the order of 10 L/s and base pressures below 10^{-2} torr make this an appealing combination. Speeds decline rapidly at pressures below $\sim 2 \times 10^{-2}$ torr.

Operating Principles: Scroll pumps use two enmeshed spiral components, one fixed and the other orbiting. Successive crescent-shaped segments of gas are trapped between the two scrolls and compressed from the inlet (vacuum side) toward the exit, where they are vented to the atmosphere. A sophisticated and expensive version of this pump has long been used for processes where leaktight operation and noncontamination are essential, for example, in the nuclear industry for pumping radioactive gases. An excellent description of the characteristics of this design has been given by Coffin (1982). In this version, extremely close tolerances (10 μm) between the two scrolls minimize leakage between the high- and low-pressure ends of the scrolls. The more recent pump designs, which substitute Teflon-like seals for the close tolerances, have made the pump an affordable option for general oil-free applications. The life of the seals is reported to be in the same range as that of the diaphragm in a diaphragm pump.

Screw Compressor. Although not yet widely used, pumps based on the principle of the screw compressor, such as that used in supercharging some high-performance cars, appear to offer some interesting advantages: i.e., pumping speeds in excess of 10 L/s, direct discharge to the atmosphere, and ultimate pressures in the 10^{-3} torr range. If such pumps demonstrate high reliability in diverse applications, they constitute the closest alternative, in a single-unit "dry" pump, to the oil-sealed mechanical pump.

Molecular Drag Pump

Applications: The molecular drag pump is useful for applications requiring pressures in the 1 to 10^{-7} torr range and freedom from organic contamination. Over this range the pump permits a far higher throughput of gas, compared to a standard turbomolecular pump. It has also been used in the compound turbomolecular pump as an integral backing stage. This will be discussed in detail under Turbomolecular Pumps.

Operating Principles: The pump uses one or more drums rotating at speeds as high as 90,000 rpm inside stationary, coaxial housings. The clearance between drum and housing is ~0.3 mm. Gas is dragged in the direction of rotation by momentum transfer to the pump exit along helical grooves machined in the housing. The bearings of these devices are similar to those in turbomolecular pumps (see discussion of Turbomolecular Pumps, below). An internal motor avoids difficulties inherent in a high-speed vacuum seal. A typical pump uses two or more separate stages, arranged in series, providing a compression ratio as high as $1{:}10^7$ for air, but typically less than $1{:}10^3$ for hydrogen. It must be supported by a backing pump, often of the diaphragm type, that can maintain the forepressure below a critical value, typically 10 to 30 torr, depending upon the particular design. The much lower compression ratio for hydrogen, a characteristic shared by all turbomolecular pumps, will increase its percentage in a vacuum chamber, a factor to consider in rare cases where the presence of hydrogen affects the application.

Sorption Pumps

Applications: Sorption pumps were introduced for roughing down ultrahigh vacuum systems prior to turning on a sputter-ion pump (Welch, 1991). The pumping speed of a typical sorption pump is similar to that of a small oil-sealed mechanical pump, but they are rather awkward in application. This is of little concern in a vacuum system likely to run many months before venting to the atmosphere. Occasional inconvenience is a small price for the ultimate in contamination-free operation.

Operating Principles: A typical sorption pump is a cannister containing ~3 lb of a molecular sieve material that is cooled to liquid nitrogen temperature. Under these conditions the molecular sieve can adsorb $\sim 7.6 \times 10^4$ torr-liter of most atmospheric gases; exceptions are helium and hydrogen, which are not significantly adsorbed, and neon, which is adsorbed to a limited extent. Together, these gases, if not pumped, would leave a residual pressure in the 10^{-2} torr range. This is too high to guarantee the trouble-free start of a sputter-ion pump, but the problem is readily avoided. For example, a sorption pump connected to a vacuum chamber of ~100 L volume exhausts air to a pressure in the viscous flow region, say 5 torr, and then is valved off. The nonadsorbing gases are swept into the pump along with the adsorbed gases; the pump now contains a fraction (760–5)/760 or 99.3% of the nonadsorbable gases originally present, leaving hydrogen, helium, and neon in the low 10^{-4} torr range in the vacuum chamber. A second sorption pump on the vacuum chamber will then readily achieve a base pressure below 5×10^{-4} torr, quite adequate to start even a recalcitrant ion pump.

High-Vacuum Pumps

Four types of high-vacuum pumps are in general use: diffusion, turbomolecular, cryosorption, and sputter-ion.

Each of these classes has advantages, and also some problems, and it is vital to consider both sides for a particular application. Any of these pumps can be used for ultimate pressures in the ultrahigh vacuum region and to maintain a working chamber that is substantially free from organic contamination. The choice of system rests primarily on the ease and reliability of operation in a particular environment, and inevitably on the capital and running costs.

Diffusion Pumps

Applications: The practical diffusion pump was invented by Langmuir in 1916, and this is the most common high-vacuum pump when all vacuum applications are considered. It is far less dominant where avoidance of organic contamination is essential. Diffusion pumps are available in a wide range of sizes, with speeds of up to 50,000 L/s; for such high-speed pumping only the cryopump seriously competes.

A diffusion pump can give satisfactory service in a number of situations. One such case is in a large system in which cleanliness is not critical. Contamination problems of diffusion-pumped systems have actually been somewhat overstated. Commercial processes using highly reactive metals are routinely performed using diffusion pumps. When funds are scarce, a diffusion pump, which incurs the lowest capital cost of any of the high-vacuum alternatives, is often selected. The continuing costs of operation, however, are higher than for the other pumps, a factor not often considered.

An excellent detailed discussion of diffusion pumps is available (Hablanian, 1995).

Operating Principles: A diffusion pump normally contains three or more oil jets operating in series. It can be operated at a maximum inlet pressure of $\sim 1 \times 10^{-3}$ torr and maintains a stable pumping speed down to 10^{-10} torr or lower. As a transfer pump, the total amount of gas it can pump is limited only by its reliability, and accumulation of any hazardous gas is not a problem. However, there are a number of key requirements in maintaining its operation. First, the outlet of the pump must be kept below some maximum pressure, which can, however, be as high as the mid-10^{-1} torr range. If the pressure exceeds this limit, all oil jets in the pump collapse and the pumping stops. Consequently the forepump (often called the backing pump) must operate continuously. Other services that must be maintained without interruption include water or air cooling, electrical power to the heater, and refrigeration, if a trap is used, to prevent oil backstreaming. A major drawback of this type of pump is the number of such criteria.

The pump oil undergoes continuous thermal degradation. However, the extent of such degradation is small, and an oil charge can last for many years. Oil decomposition products have considerably higher vapor pressure than their parent molecules. Therefore modern pumps are designed to continuously purify the working fluid, ejecting decomposition products toward the forepump. In addition, any forepump oil reaching the diffusion pump has a much higher vapor pressure than the working fluid, and it too must be ejected. The purification mechanism primarily involves the oil from the pump jet, which is cooled at the pump wall and returns, by gravity, to the boiler. The cooling area extends only past the lowest pumping jet, below which returning oil is heated by conduction from the boiler, boiling off any volatile fraction, so that it flows toward the forepump. This process is greatly enhanced if the pump is fitted with an ejector jet, directed toward the foreline; the jet exhausts the volume directly over the boiler, where the decomposition fragments are vaporized. A second step to minimize the effect of oil decomposition is to design the heater and supply tubes to the jets so that the uppermost jet, i.e., that closest to the vacuum chamber, is supplied with the highest-boiling-point oil fraction. This oil, when condensed on the upper end of the pump wall, has the lowest possible vapor pressure. It is this film of oil that is a major source of backstreaming into the vacuum chamber.

The selection of the oil used is important (O'Hanlon, 1989). If minimum backstreaming is essential, one can select an oil that has a very low vapor pressure at room temperature. A polyphenyl ether, such as Santovac 5, or a silicone oil, such as DC705, would be appropriate. However, for the most oil-sensitive applications, it is wise to use a liquid nitrogen (LN2) temperature trap between pump and vacuum chamber. Any cold trap will reduce the system base pressure, primarily by pumping water vapor, but to remove oil to a partial pressure well below 10^{-11} torr it is essential that molecules make at least two collisions with surfaces at LN2 temperature. Such traps are thermally isolated from ambient temperature and only need cryogen refills every 8 hr or more. With such a trap, the vapor pressure of the pump oil is secondary, and a less expensive oil may be used.

If a pump is exposed to substantial flows of reactive gases or to oxygen, either because of a process gas flow or because the chamber must be frequently pumped down after venting to air, the chemical stability of the oil is important. Silicone oils are very resistant to oxidation, while perfluorinated oils are stable against both oxygen and many reactive gases.

When a vacuum chamber includes devices such as mass spectrometers, which depend upon maintaining uniform electrical potential on electrodes, silicone oils can be a problem, because on decomposition they may deposit insulating films on electrodes.

Operating Procedures: A vacuum chamber free from organic contamination pumped by a diffusion pump requires stringent operating procedures. While the pump is warming, high backstreaming occurs until all jets are in full operation, so the chamber must be protected during this phase, either by a LN2 trap, before the pressure falls below the viscous flow regime, or by an isolation valve.

The chamber must be roughed down to some predetermined pressure before opening to the diffusion pump. This cross-over pressure requires careful consideration. Procedures to minimize the backstreaming for the frequently used oil-sealed mechanical pump have already been discussed (see Oil-Sealed Pumps). If a trap is used, one can safely rough down the chamber to the ultimate pressure of the pump. Alternatively, backstreaming can be minimized

by limiting the exhaust to the viscous flow regime. This procedure presents a potential problem. The vacuum chamber will be left at a pressure in the 10^{-1} torr range, but sustained operation of the diffusion pump must be avoided when its inlet pressure exceeds 10^{-3} torr. Clearly, the moment the isolation valve between diffusion pump and the roughed-down vacuum chamber is opened, the pump will suffer an overload of at least two decades pressure. In this condition, the upper jet of the pump will be overwhelmed and backstreaming will rise. If the diffusion pump is operated with a LN2 trap, this backstreaming will be intercepted. But, even with an untrapped diffusion pump, the overload condition rarely lasts more than 10 to 20 s, because the pumping speed of a diffusion pump is very high, even with one inoperative jet. Consequently, the backstreaming from roughing and high-vacuum pumps remains acceptable for many applications.

Where large numbers of different operators use a system, fully automatic sequencing and safety interlocks are recommended to reduce the possibility of operator error. Diffusion pumps are best avoided if simplicity of operation is essential and freedom from organic contamination is paramount.

Turbomolecular Pumps

Applications: Turbomolecular pumps were introduced in 1958 (Becker, 1959) and were immediately hailed as the solution to all of the problems of the diffusion pump. Provided that recommended procedures are used, these pumps live up to the original high expectations.

These are reliable, general-purpose pumps requiring simple operating procedures and capable of maintaining clean vacuum down to the 10^{-10} torr range. Pumping speeds up to 10,000 L/s are available.

Operating Principles: The pump is a multistage axial compressor, operating at rotational speeds from around 20,000 to 90,000 rpm. The drive motor is mounted inside the pump housing, avoiding the shaft seal needed with an external drive. Modern power supplies sense excessive loading of the motor, as when operating at too high an inlet pressure, and reduce the motor speed to avoid overheating and possible failure. Occasional failure of the frequency control in the supply has resulted in excessive speeds and catastrophic failure of the rotor.

At high speeds, the dominant problem is maintenance of the rotational bearings. Careful balancing of the rotor is essential; in some models bearings can be replaced in the field, if rigorous cleanliness is assured, preferably in a clean environment such as a laminar-flow hood. In other designs, the pump must be returned to the manufacturer for bearing replacement and rotor rebalancing. This service factor should be considered in selecting a turbomolecular pump, since few facilities can keep a replacement pump on hand.

Several different types of bearings are common in turbomolecular pumps:

1. *Oil Lubrication*. All first-generation pumps used oil-lubricated bearings that often lasted several years in continuous operation. These pumps were mounted horizontally with the gas inlet between two sets of blades. The bearings were at the ends of the rotor shaft, on the forevacuum side. This type of pump, and the magnetically levitated designs discussed below, offer minimum vibration.

   Second-generation pumps are vertically mounted and single-ended. This is more compact, facilitating easy replacement of a diffusion pump. Many of these pumps rely on gravity return of lubrication oil to the reservoir and thus require vertical orientation. Using a wick as the oil reservoir both localizes the liquid and allows more flexible pump orientation.
2. *Grease Lubrication*: A low-vapor-pressure grease lubricant was introduced to reduce transport of oil into the vacuum chamber (Osterstrom, 1979) and to permit orientation of the pump in any direction. Grease has lower frictional loss and allows a lower-power drive motor, with consequent drop in operating temperature.
3. *Ceramic Ball Bearings*: Most bearings now use a ceramic-balls/steel-race combination; the lighter balls reduce centrifugal forces and the ceramic-to-steel interface minimizes galling. There appears to be a significant improvement in bearing life for both oil and grease lubrication systems.
4. *Magnetic Bearings*: Magnetic suspension systems have two advantages: a non-contact bearing with a potentially unlimited life, and very low vibration. First-generation pumps used electromagnetic suspension with a battery backup. When nickel-cadmium batteries were used, this backup was not continuously available; incomplete discharge before recharging cycles often reduces discharge capacity. A second generation using permanent magnets was more reliable and of lower cost. Some pumps now offer an improved electromagnetic suspension with better active balancing of the rotor on all axes. In some designs, the motor is used as a generator when power is interrupted, to assure safe shutdown of the magnetic suspension system. Magnetic bearing pumps use a second set of "touch-down" bearings for support when the pump is stationary. The bearings use a solid, low-vapor-pressure lubricant (O'Hanlon, 1989) and further protect the pump in an emergency. The life of the touch-down bearings is limited, and their replacement may be a nuisance; it is, however, preferable to replacing a shattered pump rotor and stator assembly.
5. *Combination Bearings Systems*: Some designs use combinations of different types of bearings. One example uses a permanent-magnet bearing at the high-vacuum end and an oil-lubricated bearing at the forevacuum end. A magnetic bearing does not contaminate the system and is not vulnerable to damage by aggressive gases as is a lubricated bearing. Therefore it can be located at the very end of the rotor shaft, while the oil-fed bearing is at the opposite forevacuum end. This geometry has the advantage of minimizing vibration.

Problems with Pumping Reactive Gases: Very reactive gases, common in the semiconductor industry, can result in rapid bearing failure. A purge with nonreactive gas, in the viscous flow regime, can prevent the pumped gases from contacting the bearings. To permit access to the bearing for a purge, pump designs move the upper bearing below the turbine blades, which often cantilevers the center of mass of the rotor beyond the bearings. This may have been a contributing factor to premature failure seen in some pump designs.

The turbomolecular pump shares many of the performance characteristics of the diffusion pump. In the standard construction, it cannot exhaust to atmospheric pressure, and must be backed at all times by a forepump. The critical backing pressure is generally in the 10^{-1} torr, or lower, region, and an oil-sealed mechanical pump is the most common choice. Failure to recognize the problem of oil contamination from this pump was a major factor in the problems with early applications of the turbomolecular pump. But, as with the diffusion pump, an operating turbomolecular pump prevents significant backstreaming from the forepump and its own bearings. A typical turbomolecular pump compression ratio for heavy oil molecules, $\sim 10^{12}$:1, ensures this. The key to avoiding oil contamination during evacuation is the pump reaching its operating speed as soon as is possible.

In general, turbomolecular pumps can operate continuously at pressures as high as 10^{-2} torr and maintain constant pumping speed to at least 10^{-10} torr. As the turbomolecular pump is a transfer pump, there is no accumulation of hazardous gas, and less concern with an emergency shutdown situation. The compression ratio is $\sim 10^{8}$:1 for nitrogen, but frequently below 1000:1 for hydrogen. Some first-generation pumps managed only 50:1 for hydrogen. Fortunately, the newer compound pumps, which add an integral molecular drag backing pump, often have compression ratios for hydrogen in excess of 10^{5}:1. The large difference between hydrogen (and to a lesser extent helium) and gases such as nitrogen and oxygen leaves the residual gas in the chamber enriched in the lighter species. If a low residual hydrogen pressure is an important consideration, it may be necessary to provide supplementary pumping for this gas, such as a sublimation pump or nonevaporable getter (NEG), or to use a different class of pump.

The demand for negligible organic compound contamination has led to the compound pump, comprising a standard turbomolecular stage backed by a molecular drag stage, mounted on a common shaft. Typically, a backing pressure of only 10 torr or higher, conveniently provided by an oil-free ("dry") diaphragm pump, is needed (see discussion of Oil-Free Pumps). In some versions, greased or oil-lubricated bearings are used (on the high-pressure side of the rotor); magnetic bearings are also available. Compound pumps provide an extremely low risk of oil contamination and significantly higher compression ratios for light gases.

Operation of a Turbomolecular Pump System: Freedom from organic contamination demands care during both the evacuation and venting processes. However, if a pump is contaminated with oil, the cleanup requires disassembly and the use of solvents.

The following is a recommended procedure for a system in which an untrapped oil-sealed mechanical roughing/backing pump is combined with an oil-lubricated turbomolecular pump, and an isolation valve is provided between the vacuum chamber and the turbomolecular pump.

1. *Startup*: Begin roughing down and turn on the pump as soon as is possible without overloading the drive motor. Using a modern electronically controlled supply, no delay is necessary, because the supply will adjust power to prevent overload while the pressure is high. With older power supplies, the turbomolecular pump should be started as soon as the pressure reaches a tolerable level, as given by the manufacturer, probably in the 10 torr region. A rapid startup ensures that the turbomolecular pump reaches at least 50% of the operating speed while the pressure in the foreline is still in the viscous flow regime, so that no oil backstreaming can enter the system through the turbomolecular pump.

   Before opening to the turbomolecular pump, the vacuum chamber should be roughed down using a procedure to avoid oil contamination, as was described for diffusion pump startup (see discussion above).
2. *Venting*: When the entire system is to be vented to atmospheric pressure, it is essential that the venting gas enter the turbomolecular pump at a point on the system side of any lubricated bearings in the pump. This ensures that oil liquid or vapor is swept away from the system towards the backing system. Some pumps have a vent midway along the turbine blades, while others have vents just above the upper, system-side, bearings. If neither of these vent points are available, a valve must be provided on the vacuum chamber itself. Never vent the system from a point on the foreline of the turbomolecular pump; that can flush both mechanical pump oil and turbomolecular pump oil into the turbine rotor and stator blades and the vacuum chamber. Venting is best started immediately after turning off the power to the turbomolecular pump and adjusting so the chamber pressure rises into the viscous flow region within a minute or two. Too-rapid venting exposes the turbine blades to excessive pressure in the viscous flow regime, with unnecessarily high upward force on the bearing assembly (often called the "helicopter" effect). When venting frequently, the turbomolecular pump is usually left running, isolated from the chamber, but connected to the forepump.

The major maintenance is checking the oil or grease lubrication, as recommended by the pump manufacturer, and replacing the bearings as required. The stated life of bearings is often $\sim$2 years continuous operation, though an actual life of $\geq$5 years is not uncommon. In some facilities, where multiple pumps are used in production, bearings are checked by monitoring the amplitude of the

vibration frequency associated with the bearings. A marked increase in amplitude indicates the approaching end of bearing life, and the pump is removed for maintenance.

Cryopumps

Applications: Cryopumping was first extensively used in the space program, where test chambers modeled the conditions encountered in outer space, notably that by which any gas molecule leaving the vehicle rarely returns. This required all inside surfaces of the chamber to function as a pump, and led to liquid-helium-cooled shrouds in the chambers on which gases condensed. This is very effective, but is not easily applicable to individual systems, given the expense and difficulty of handling liquid helium. However, the advent of reliable closed-cycle mechanical refrigeration systems, achieving temperatures in the 10 to 20 K range, allow reliable, contamination-free pumps, with a wide range of pumping speeds, and which are capable of maintaining pressures as low as the 10^{-10} torr range (Welch, 1991).

Cryopumps are general purpose and available with very high pumping speeds (using internally mounted cryopanels), so they work for all chamber sizes. These are capture pumps, and, once operating, are totally isolated from the atmosphere. All pumped gas is stored in the body of the pump. They must be regenerated on a regular basis, but the quantity of gas pumped before regeneration is very large for all gases that are captured by condensation. Only helium, hydrogen, and neon are not effectively condensed. They must be captured by adsorption, for which the capacity is far smaller. Indeed, if pumping any significant quantity of helium, regeneration would have to be so frequent that another type of pump should be selected. If the refrigeration fails due to a power interruption or a mechanical failure, the pumped gas will be released within minutes. All pumps are fitted with a pressure relief valve to avoid explosion, but provision must be made for the safe disposal of any hazardous gases released.

Operating Principles: A cryopump uses a closed-cycle refrigeration system with helium as the working gas. An external compressor, incorporating a heat exchanger that is usually water-cooled, supplies helium at ~300 psi to the cold head, which is mounted on the vacuum system. The helium is cooled by passing through a pair of regenerative heat exchangers in the cold head, and then allowed to expand, a process which cools the incoming gas, and in turn, cools the heat exchangers as the low-pressure gas returns to the compressor. Over a period of several hours, the system develops two cold zones, nominally 80 and 15 K. The ~80 K zone is used to cool a shroud through which gas molecules pass into its interior; water is pumped by this shroud, and it also minimizes the heat load on the second-stage array from ambient temperature radiation. Inside the shroud is an array at ~15 K, on which most other gases are condensed. The energy available to maintain the 15 K temperature is just a few watts.

The second stage should typically remain in the range 10 to 20 K, low enough to pump most common gases to well below 10^{-10} torr. In order to remove helium, hydrogen, and neon the modern cryopump incorporates a bed of charcoal, having a very large surface area, cooled by the second-stage array. This bed is so positioned that most gases are first removed by condensation, leaving only these three to be physically adsorbed.

As already noted, the total pumping capacity of a cryopump is very different for the gases that are condensed, as compared to those that are adsorbed. The capacity of a pump is frequently quoted for argon, commonly used in sputtering systems. For example, a pump with a speed of ~1000 L/s will have the capability of pumping $\sim 3 \times 10^5$ torr-liter of argon before requiring regeneration. This implies that a 200-L volume could be pumped down from a typical roughing pressure of 2.5×10^{-1} torr ~6000 times. The pumping speed of a cryopump remains constant for all gases that are condensable at 20 K, down to the 10^{-10} torr range, so long as the temperature of the second-stage array does not exceed 20 K. At this temperature the vapor pressure of nitrogen is $\sim 1 \times 10^{-11}$ torr, and that of all other condensable gases lies well below this figure.

The capacity for adsorption-pumped gases is not nearly so well defined. The capacity increases both with decreasing temperature and with the pressure of the adsorbing gas. The temperature of the second-stage array is controlled by the balance between the refrigeration capacity and generation of heat by both condensation and adsorption of gases. Of necessity, the heat input must be limited so that the second-stage array never exceeds 20 K, and this translates into a maximum permissible gas flow into the pump. The lowest temperature of operation is set by the pump design, nominally ~10 K. Consequently the capacity for adsorption of a gas such as hydrogen can vary by a factor of four or more when between these two temperature extremes. For a given flow of hydrogen, if this is the only gas being pumped, the heat input will be low, permitting a higher pumping capacity, but if a mixture of gases is involved, then the capacity for hydrogen will be reduced, simply because the equilibrium operating temperature will be higher. A second factor is the pressure of hydrogen that must be maintained in a particular process. Because the adsorption capacity is determined by this pressure, a low hydrogen pressure translates into a reduced adsorptive capacity, and therefore a shorter operating time before the pump must be regenerated. The effect of these factors is very significant for helium pumping, because the adsorption capacity for this gas is so limited. A cryopump may be quite impractical for any system in which there is a deliberate and significant inlet of helium as a process gas.

Operating Procedure: Before startup, a cryopump must first be roughed down to some recommended pressure, often $\sim 1 \times 10^{-1}$ torr. This serves two functions. First, the vacuum vessel surrounding the cold head functions as a Dewar, thermally isolating the cold zone. Second, any gas remaining must be pumped by the cold head as it cools down; because adsorption is always effective at a much higher temperature than condensation, the gas is adsorbed in the charcoal bed of the 20 K array, partially saturating it, and limiting the capacity for subsequently adsorbing helium, hydrogen, and neon. It is essential to

avoid oil contamination when roughing down, because oil vapors adsorbed on the charcoal of the second-stage array cannot be removed by regeneration and irreversibly reduce the adsorptive capacity. Once the required pressure is reached, the cryopump is isolated from the roughing line and the refrigeration system is turned on. When the temperature of the second-stage array reaches 20 K, the pump is ready for operation, and can be opened to the vacuum chamber, which has previously been roughed down to a selected cross-over pressure. This cross-over pressure can readily be calculated from the figure for the impulse gas load, specified by the manufacturer, and the volume of the chamber. The impulse load is simply the quantity of gas to which the pump can be exposed without increasing the temperature of the second-stage array above 20 K.

When the quantity of gas that has been pumped is close to the limiting capacity, the pump must be regenerated. This procedure involves isolation from the system, turning off the refrigeration unit, and warming the first- and second-stage arrays until all condensed and adsorbed gas has been removed. The most common method is to purge these gases using a warm (~60°C) dry gas, such as nitrogen, at atmospheric pressure. Internal heaters were deliberately avoided for many years, to avoid an ignition source in the event that explosive gas mixtures, such as hydrogen and oxygen, were released during regeneration. To the same end, the use of any pressure sensor having a hot surface was, and still is, avoided in the regeneration procedure. Current practice has changed, and many pumps now incorporate a means of independently heating each of the refrigerated surfaces. This provides the flexibility to heat the cold surfaces only to the extent that adsorbed or condensed gases are rapidly removed, greatly reducing the time needed to cool back to the operating temperature. Consider, for example, the case where argon is the predominant gas load. At the maximum operating temperature of 20 K, its vapor pressure is well below 10^{-11} torr, but warming to 90 K raises the vapor pressure to 760 torr, facilitating rapid removal.

In certain cases, the pumping of argon can cause a problem commonly referred to as argon hangup. This occurs after a high pressure of argon, e.g., $>1 \times 10^{-3}$ torr, has been pumped for some time. When the argon influx stops, the argon pressure remains comparatively high instead of falling to the background level. This happens when the temperature of the pump shroud is too low. At 40 K, in contrast to 80 K, argon condenses on the outer shroud instead of being pumped by the second-stage array. Evaporation from the shroud at the argon vapor pressure of 1×10^{-3} torr keeps the partial pressure high until all of the gas has desorbed. The problem arises when the refrigeration capacity is too large, for example, when several pumps are served by a single compressor and the helium supply is improperly proportioned. An internal heater to increase the shroud temperature is an easy solution.

A cryopump is an excellent general-purpose device. It can provide an extremely clean environment at base pressures in the low 10^{-10} torr range. Care must be taken to ensure that the pressure-relief valve is always operable, and to ensure that any hazardous gases are safely handled in the event of an unscheduled regeneration. There is some possibility of energetic chemical reactions during regeneration. For example, ozone, which is generated in some processes, may react with combustible materials. The use of a nonreactive purge gas will minimize hazardous conditions if the flow is sufficient to dilute the gases released during regeneration. The pump has a high capital cost and fairly high running costs for power and cooling.

Maintenance of a cryopump is normally minimal. Seals in the displacer piston in the cold head must be replaced as required (at intervals of one year or more, depending on the design); an oil-adsorber cartridge in the compressor housing requires a similar replacement schedule.

Sputter-Ion Pumps

Applications: These pumps were originally developed for ultrahigh vacuum (UHV) systems and are admirably suited to this application, especially if the system is rarely vented to atmospheric pressure. Their main advantages are as follows.

1. High reliability, because of no moving parts.
2. The ability to bake the pump up to 400°C, facilitating outgassing and rapid attainment of UHV conditions.
3. Fail-safe operation if on a leak-tight UHV system. If the power is interrupted, a moderate pressure rise will occur; the pump retains some pumping capacity by gettering. When power is restored, the base pressure is normally reestablished rapidly.
4. The pump ion current indicates the pressure in the pump itself, which is useful as a monitor of performance.

Sputter-ion pumps are not suitable for the following uses.

1. On systems with a high, sustained gas load or frequent venting to atmosphere.
2. Where a well-defined pumping speed for all gases is required. This limitation can be circumvented with a severely conductance-limited pump, so the speed is defined by conductance rather than by the characteristics of the pump itself.

Operating Principles: The operating mechanisms of sputter-ion pumps are very complex indeed (Welch, 1991). Crossed electrostatic and magnetic fields produce a confined discharge using a geometry originally devised by Penning (1937) to measure pressure in a vacuum system. A trapped cloud of electrons is produced, the density of which is highest in the 10^{-4} torr region, and falls off as the pressure decreases. High-energy ions, produced by electron collision, impact on the pump cathodes, sputtering reactive cathode material (titanium, and to a lesser extent, tantalum), which is deposited on all surfaces within line-of sight of the impact area. The pumping mechanisms include the following.

1. Chemisorption on the sputtered cathode material, which is the predominant pumping mechanism for reactive gases.

2. Burial in the cathodes, which is mainly a transient contributor to pumping. With the exception of hydrogen, the atoms remain close to the surface and are released as pumping/sputtering continues. This is the source of the "memory" effect in diode ion pumps; previously pumped species show up as minor impurities when a different gas is pumped.
3. Burial of ions back-scattered as neutrals, in all surfaces within line-of sight of the impact area. This is a crucial mechanism in the pumping of argon and other noble gases (Jepsen, 1968).
4. Dissociation of molecules by electron impact. This is the mechanism for pumping methane and other organic molecules.

The pumping speed of these pumps is variable. Typical performance curves show the pumping of a single gas under steady-state conditions. Figure 1 shows the general characteristic as a function of pressure. Note the pronounced drop with falling pressure. The original commercial pumps used anode cells the order of 1.2 cm in diameter and had very low pumping speeds even in the 10^{-9} torr range. However, newer pumps incorporate at least some larger anode cells, up to ~2.5 cm diameter, and the useful pumping speed is extended into the 10^{-11} torr range (Rutherford, 1963).

The pumping speed of hydrogen can change very significantly with conditions, falling off drastically at low pressures and increasing significantly at high pressures (Singleton, 1969, 1971; Welch, 1994). The pumped hydrogen can be released under some conditions, primarily during the startup phase of a pump. When the pressure is 10^{-3} torr or higher, the internal temperatures can readily reach 500°C (Snouse, 1971). Hydrogen is released, increasing the pressure and frequently stalling the pumpdown.

Rare gases are not chemisorbed, but are pumped by burial (Jepsen, 1968). Argon is of special importance, because it can cause problems even when pumping air. The release of argon, buried as atoms in the cathodes, sometimes causes a sudden increase in pressure of as much as three decades, followed by renewed pumping, and a concomitant drop in pressure. The unstable behavior is repeated at regular intervals, once initiated (Brubaker, 1959). This problem can be avoided in two ways.

1. By use of the "differential ion" or DI pump (Tom and James, 1969), which is a standard diode pump in which a tantalum cathode replaces one titanium cathode.
2. By use of the triode sputter-ion pump, in which a third electrode is interposed between the ends of the cylindrical anode and the pump walls. The additional electrode is maintained at a high negative potential, serving as a sputter cathode, while the anode and walls are maintained at ground potential. This pump has the additional advantage that the "memory" effect of the diode pump is almost completely suppressed.

The operating life of a sputter-ion pump is inversely proportional to the operating pressure. It terminates when the cathodes are completely sputtered through at a small area on the axis of each anode cell where the ions impact. The life therefore depends upon the thickness of the cathodes at the point of ion impact. For example, a conventional triode pump has relatively thin cathodes as compared to a diode pump, and this is reflected in the expected life at an operating pressure of 1×10^{-6} torr, i.e., 35,000 as compared to 50,000 hr.

The fringing magnetic field in older pumps can be very significant. Some newer pumps greatly reduce this problem. A vacuum chamber can be exposed to ultraviolet and x radiation, as well as ions and electrons produced by an ion pump, so appropriate electrical and optical shielding may be required.

Operating Procedures: A sputter-ion pump must be roughed down before it can be started. Sorption pumps or any other clean technique can be used. For a diode pump, a pressure in the 10^{-4} torr range is recommended, so that the Penning discharge (and associated pumping mechanisms) will be immediately established. A triode pump can safely be started at pressures about a decade higher than the diode, because the electrostatic fields are such that the walls are not subjected to ion bombardment

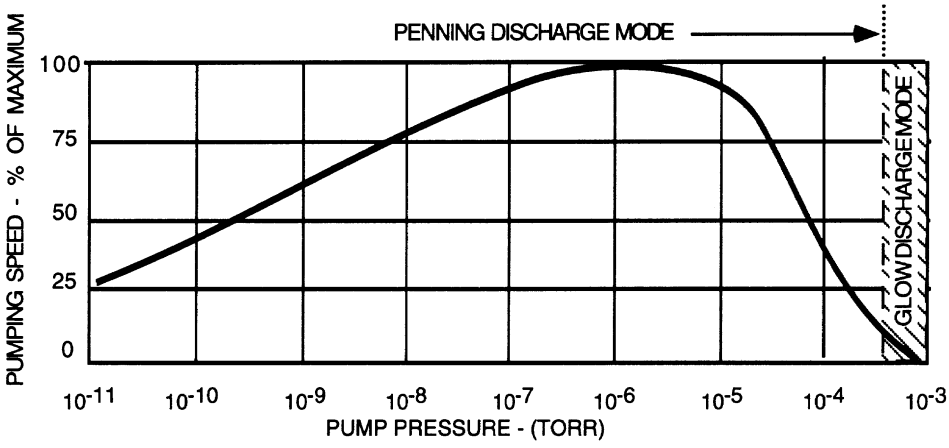


Figure 1. Schematic representation of the pumping speed of a diode sputter-ion pump as a function of pressure.

(Snouse, 1971). An additional problem develops in pumps that have operated in hydrogen or water vapor. Hydrogen accumulates in the cathodes and this gas is released when the cathode temperatures increase during startup. The higher the pressure, the greater the temperature; temperatures as high as 900°C have been measured at the center of cathodes under high gas loads (Jepsen, 1967).

An isolation valve should be used to avoid venting the pump to atmospheric pressure. The sputtered deposits on the walls of a pump adsorb gas with each venting, and the bonding of subsequently sputtered material will be reduced, eventually causing flaking of the deposits. The flakes can serve as electron emitters, sustaining localized (non-pumping) discharges and can also short out the electrodes.

Getter Pumps. Getter pumps depend upon the reaction of gases with reactive metals as a pumping mechanism; such metals were widely used in electronic vacuum tubes, being described as getters (Reimann, 1952). Production techniques for the tubes did not allow proper outgassing of tube components, and the getter completed the initial pumping on the new tube. It also provided continuous pumping for the life of the device.

Some practical getters used a "flash getter," a stable compound of barium and aluminum that could be heated, using an RF coil, once the tube had been sealed, to evaporate a mirror-like barium deposit on the tube wall. This provided a gettering surface that operated close to ambient temperature. Such films initially offer rapid pumping, but once the surface is covered, a much slower rate of pumping is sustained by diffusion into the bulk of the film. These getters are the forerunners of the modern sublimation pump.

A second type of getter used a reactive metal, such as titanium or zirconium wire, operated at elevated temperature; gases react at the metal surface to produce stable, low-vapor-pressure compounds that then diffuse into the interior, allowing a sustained reaction at the surface. These getters are the forerunners or the modern nonevaporable getter (NEG).

Sublimation pumps

Applications: Sublimation pumps are frequently used in combination with a sputter-ion pump, to provide high-speed pumping for reactive gases with a minimum investment (Welch, 1991). They are more suitable for ultrahigh vacuum applications than for handling large pumping loads. These pumps have been used in combination with turbomolecular pumps to compensate for the limited hydrogen-pumping performance of older designs. The newer, compound turbomolecular pumps avoid this need.

Operating Principles: Most sublimation pumps use a heated titanium surface to sublime a layer of atomically clean metal onto a surface, commonly the wall of a vacuum chamber. In the simplest version, a wire, commonly 85% Ti/15% Mo (McCracken and Pashley, 1966; Lawson and Woodward, 1967) is heated electrically; typical filaments deposit ~1 g before failure. It is normal to mount two or three filaments on a common flange for longer use before replacement. Alternatively, a hollow sphere of titanium is radiantly heated by an internal incandescent lamp filament, providing as much as 30 g of titanium. In either case, a temperature of ~1500°C is required to establish a useable sublimation rate. Because each square centimeter of a titanium film provides a pumping speed of several liters per second at room temperature (Harra, 1976), one can obtain large pumping speeds for reactive gases such as oxygen and nitrogen. The speed falls dramatically as the surface is covered by even one monolayer. Although the sublimation process must be repeated periodically to compensate for saturation, in an ultrahigh vacuum system the time between sublimation cycles can be many hours. With higher gas loads the sublimation cycles become more frequent, and continuous sublimation is required to achieve maximum pumping speed. A sublimator can only pump reactive gases and must always be used in combination with a pump for remaining gases, such as the rare gases and methane. Do not heat a sublimator when the pressure is too high, e.g., 10^{-3} torr; pumping will start on the heated surface, and can suppress the rate of sublimation completely. In this situation the sublimator surface becomes the only effective pump, functioning as a nonevaporable getter, and the effective speed will be very small (Kuznetsov et al., 1969).

Nonevaporable Getter Pumps (NEGs)

Applications: In vacuum systems, NEGs can provide supplementary pumping of reactive gases, being particularly effective for hydrogen, even at ambient temperature. They are most suitable for maintaining low pressures. A niche application is the removal of reactive impurities from rare gases such as argon. NEGs find wide application in maintaining low pressures in sealed-off devices, in some cases at ambient temperature (Giorgi et al., 1985; Welch, 1991).

Operating Principles: In one form of NEG, the reactive metal is carried as a thin surface layer on a supporting substrate. An example is an alloy of Zr/16%Al supported on either a soft iron or nichrome substrate. The getter is maintained at a temperature of around ~400°C, either by indirect or ohmic heating. Gases are chemisorbed at the surface and diffuse into the interior. When a getter has been exposed to the atmosphere, for example, when initially installed in a system, it must be activated by heating under vacuum to a high temperature, 600° to 800°C. This permits adsorbed gases such as nitrogen and oxygen to diffuse into the bulk. With use, the speed falls off as the near-surface getter becomes saturated, but the getter can be reactivated several times by heating. Hydrogen is evolved during reactivation; consequently reactivation is most effective when hydrogen can be pumped away. In a sealed device, however, the hydrogen is readsorbed on cooling.

A second type of getter, which has a porous structure with far higher accessible surface area, effectively pumps reactive gases at temperatures as low as ambient. In many cases, an integral heater is embedded in the getter.

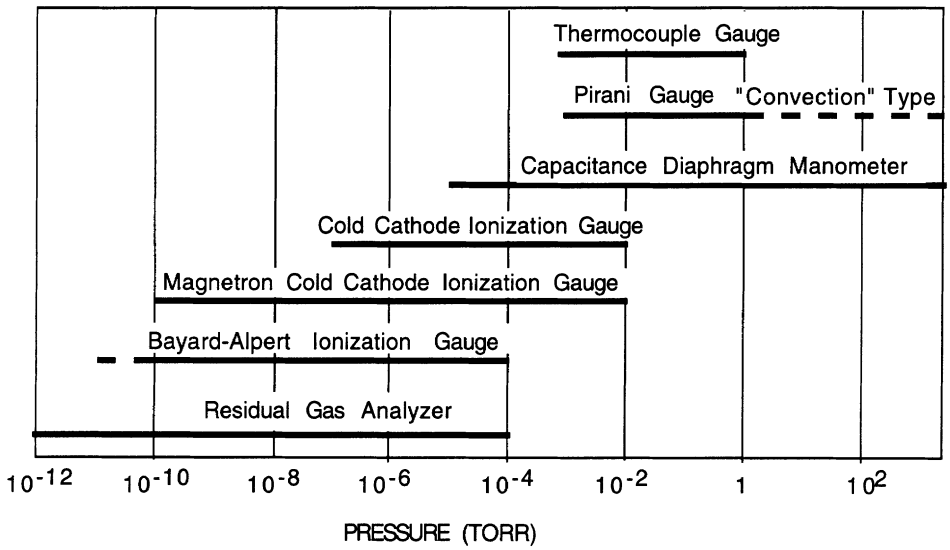


Figure 2. Approximate pressure ranges of total and partial pressure gauges. Note that only the capacitance manometer is an absolute gauge. Based, with permission, on Short Course Notes of the American Vacuum Society.

Total and Partial Pressure Measurement

Figure 2 provides a summary of the approximate range of pressure measurement for modern gauges. Note that only the capacitance diaphragm manometers are absolute gauges, having the same calibration for all gases. In all other gauges, the response depends on the specific gas or mixture of gases present, making it impossible to determine the absolute pressure without knowing gas composition.

Capacitance Diaphragm Manometers. A very wide range of gauges are available. The simplest are signal or switching devices with limited accuracy and reproducibility. The most sophisticated have the ability to measure over a range of $1{:}10^4$, with an accuracy exceeding 0.2% of reading, and a long-term stability that makes them valuable for calibration of other pressure gauges (Hyland and Shaffer, 1991). For vacuum applications, they are probably the most reliable gauge for absolute pressure measurement. The most sensitive can measure pressures from 1 torr down to the 10^{-4} torr range and can sense changes in the 10^{-5} torr range. Another advantage is that some models use stainless-steel and inconel parts, which resist corrosion and cause negligible contamination.

Operating Principles: These gauges use a thin metal, or in some cases, ceramic diaphragm, which separates two chambers, one connected to the vacuum system and the other providing the reference pressure. The reference chamber is commonly evacuated to well below the lowest pressure range of the gauge, and has a getter to maintain that pressure. The deflection of the diaphragm is measured using a very sensitive electrical capacitance bridge circuit that can detect changes of $\sim 2 \times 10^{-10}$ m. In the most sensitive gauges the device is thermostatted to avoid drifts due to temperature change; in less sensitive instruments there is no temperature control.

Operation: The bridge must be periodically zeroed by evacuating the measuring side of the diaphragm to a pressure below the lowest pressure to be measured.

Any gauge that is not thermostatically controlled should be placed in such a way as to avoid drastic temperature changes, such as periodic exposure to direct sunlight.

The simplest form of the capacitance manometer uses a capacitance electrode on both the reference and measurement sides of the diaphragm. In applications involving sources of contamination, or a radioactive gas such as tritium, this can lead to inaccuracies, and a manometer with capacitance probes only on the reference side should be used.

When a gauge is used for precision measurements, it must be corrected for the pressure differential that results when the thermostatted gauge head is operating at a different temperature than the vacuum system (Hyland and Shaffer, 1991).

Gauges Using Thermal Conductivity for the Measurement of Pressure

Applications: Thermal conductivity gauges are relatively inexpensive. Many operate in a range of $\sim 1 \times 10^{-3}$ to 20 torr. This range has been extended to atmospheric pressure in some modifications of the "traditional" gauge geometry. They are valuable for monitoring and control, for example, during the processes of roughing down from atmospheric pressure and for the cross-over from roughing pump to high-vacuum pump. Some are subject to drift over time, for example, as a result of contamination from mechanical pump oil, but others remain surprising stable under common system conditions.

Operating Principles: In most gauges, a ribbon or filament serves as the heated element. Heat loss from this element to the wall is measured either by the change in element temperature, in the thermocouple gauge, or as a change in electrical resistance, in the Pirani gauge.

Heat is lost from a heated surface in a vacuum system by energy transfer to individual gas molecules at low pressures (Peacock, 1998). This process has been used in the "traditional" types of gauges. At pressures well above 20 torr, convection currents develop. Heat loss in this mode has recently been used to extend the pressure measurement range up to atmospheric. Thermal radiation heat loss from the heated element is independent of the presence of gas, setting a lower limit to the measurement of pressure. For most practical gauges this limit is in the mid- to upper-10^{-4} torr range.

Two common sources of drift in the pressure indication are changes in ambient temperature and contamination of the heated element. The first is minimized by operating the heated element at 300°C or higher. However, this increases chemical interactions at the element, such as the decomposition of organic vapors into deposits of tars or carbon; such deposits change the thermal accommodation coefficient of gases on the element, and hence the gauge sensitivity. More satisfactory solutions to drift in the ambient temperature include a thermostatically controlled envelope temperature or a temperature-sensing element that compensates for ambient temperature changes. The problem of changes in the accommodation coefficient is reduced by using chemically stable heating elements, such as the noble metals or gold-plated tungsten.

Thermal conductivity gauges are commonly calibrated for air, and it is important to note that this changes significantly with the gas. The gauge sensitivity is higher for hydrogen and lower for argon. Thus, if the gas composition is unknown, the gauge reading may be in error by a factor of two or more.

Thermocouple Gauge. In this gauge, the element is heated at constant power, and its change in temperature, as the pressure changes, is directly measured using a thermocouple. In many geometries the thermocouple is spot welded directly at the center of the element; the additional thermal mass of the couple reduces the response time to pressure changes. In an ingenious modification, the thermocouple itself (Benson, 1957) becomes the heated element, and the response time is improved.

Pirani Gauge. In this gauge, the element is heated electrically, but the temperature is sensed by measuring its resistance. The absence of a thermocouple permits a faster time constant. A further improvement in response results if the element is maintained at constant temperature, and the power required becomes the measure of pressure.

Gauges capable of measurement over a range extending to atmospheric pressure use the Pirani principle. Those relying on convection are sensitive to gauge orientation, and the recommendation of the manufacturer must be observed if calibration is to be maintained. A second point, of great importance for safe operation, arises from the difference in gauge calibration with different gases. Such gauges have been used to control the flow of argon into a sputtering system measuring the pressure on the high-pressure side of a flow restriction. If pressure is set close to atmospheric, it is crucial to use a gauge calibrated for argon, or to apply the appropriate correction; using a gauge reading calibrated for air to adjust the argon to atmospheric results in an actual argon pressure well above one atmosphere, and the danger of explosion becomes significant.

A second technique that extends the measurement range to atmospheric pressure is drastic reduction of gauge dimensions so that the spacing between the heated element and the room temperature gauge wall is only 5 μm (Alvesteffer et al., 1995).

Ionization Gauges: Hot Cathode Type. The Bayard-Alpert gauge (Redhead et al., 1968) is the principal gauge used for accurate indication of pressure from 10^{-4} to 10^{-10} torr. Over this range, a linear relationship exists between the measured ion current and pressure. The gauge has a number of problems, but they are fairly well understood and to some extent can be avoided. Modifications of the gauge structure, such as the Redhead Extractor Gauge (Redhead et al., 1968) permit measurement into the high 10^{-13} torr region, and minimize errors due to electron-stimulated desorption (see below).

Operating Principles: In a typical Bayard-Alpert gauge configuration, shown in Figure 3A, a current of electrons, between 1 and 10 mA, from a heated cathode, is accelerated towards an anode grid by a potential of 150 V. Ions produced by electron collision are collected on an axial, fine-wire ion collector, which is maintained 30 V negative with respect to the cathode. The electron energy of 150 V is selected for the maximum ionization probability with most common gases.

The equation describing the gauge operation is

$$P = \frac{i^+}{(i^-)(K)} \tag{3}$$

where P is pressure, in torr, i^+ is the ion current, i^- is the electron current, and K, in torr^{-1}, is the gauge constant for the specific gas.

The original design of the ionization gauge, the triode gauge, shown in Figure 3B, cannot read below $\sim 1 \times 10^{-8}$ torr because of a spurious current, known as the

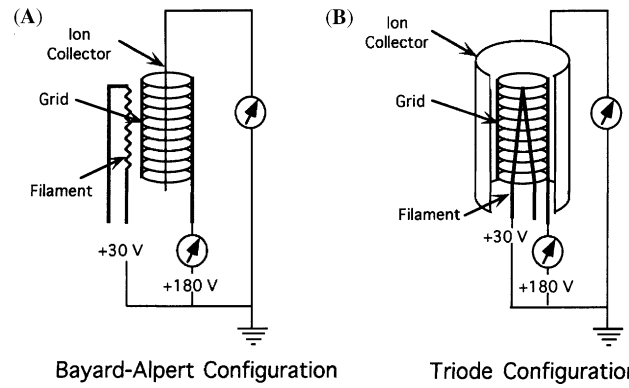


Figure 3. Comparison of the (**A**) Bayard-Alpert and (**B**) triode ion gauge geometries. Based, with permission, on Short Course Notes of the American Vacuum Society.

x-ray effect. The electron impact on the grid produces soft x rays, many of which strike the coaxial ion collector cylinder, generating a flux of photoelectrons; an electron ejected from the ion collector cannot be distinguished from an arriving ion by the current-measuring circuit. The existence of the x ray effect was first proposed by Nottingham (1947), and studies stimulated by his proposal led directly to the development of the Bayard-Alpert gauge, which simply inverted the geometry of the triode gauge. The sensitivity of the gauge is little changed from that of the triode, but the area of the ion collector, and presumably the x ray-induced spurious current, is reduced by a factor of 300, extending the usable range of the gauge to the order of 1×10^{-10} torr.

The gauge and associated electronics are normally calibrated for nitrogen gas, but, as with the thermal conductivity gauge, the sensitivity varies with gas, so the gas composition must be known for an absolute pressure reading. Gauge constants for various gases can be found in many texts (Redhead et al., 1968).

A gauge can affect the pressure in a system in three important ways.

1. An operating gauge functions as a small pump; at an electron emission of 10 mA the pumping speed is the order of 0.1 L/s. In a small system this can be a significant part of the pumping. In systems that are pumped at relatively large speeds, the gauge has negligible effect, but if the gauge is connected to the system by a long tube of small diameter, the limited conductance of the connection will result in a pressure drop, and the gauge will record a pressure lower than that in the system. For example, a gauge pumping at 0.1 L/s, connected to a chamber by a 100-cm-long, 1-cm-diameter tube, with a conductance of $\sim$0.2 L/s for air, will give a reading 33% lower than the actual chamber pressure. The solution is to connect all gauges using short and fat (i.e., high-conductance) tubes, and/or to run the gauge at a lower emission current.
2. A new gauge is a source of significant outgassing, which increases further when turned on as its temperature increases. Whenever a well-outgassed gauge is exposed to the atmosphere, gas adsorption occurs, and once again significant outgassing will result after system evacuation. This affects measurements in any part of the pressure range, but is more significant at very low pressures. Provision is made for outgassing all ionization gauges. For gauges especially suitable for pressures higher than the low 10^{-7} torr range, the grid of the gauge is a heavy non-sag tungsten or molybdenum wire that can be heated using a high-current, low-voltage supply. Temperatures of $\sim$1300°C can be achieved, but higher temperatures, desirable for UHV applications, can cause grid sagging; the radiation from the grid accelerates the outgassing of the entire gauge structure, including the envelope. The gauge remains in operation throughout the outgassing, and when the system pressure falls well below that existing before starting the outgas, the process can be terminated. For a system operating in the 10^{-7} torr range, 30 to 60 min should be adequate. The higher the operating pressure, the lower is the importance of outgassing. For pressures in the ultra-high vacuum region ($<1 \times 10^{-8}$ torr), the outgassing requirements become more rigorous, even when the system has been baked to 250°C or higher. In general, it is best to use a gauge designed for electron bombardment outgassing; the grid structure follows that of the original Bayard-Alpert design in having a thin wire grid reinforced with heavier support rods. The grid and ion collector are connected to a 500 to 1000 V supply and the electron emission is slowly increased to as high as 200 mA, achieving temperatures above 1500°C. A minimum of 60 min outgassing, using each of the two electron sources (for tungsten filament gauges), in turn, is adequate. During this procedure pressure measurements are not possible.

   After outgassing, the clean gauge will adsorb gas as it cools, and consequently the pressure will often fall significantly below its ultimate equilibrium value; it is the equilibrium value that provides a true measure of the system operation, and adequate time should be allowed for the gauge to readsorb gas and reestablish steady-state coverage. In the ultra-high vacuum region this may take an hour or more, but at higher pressure, say 10^{-6} torr, it is a transient effect.
3. The high-temperature electron emitter can interact with gases in the system in a number of ways. A tungsten filament, operating at 10 mA emission current, has a temperature on the order of 1825°C. Effects observed include the pumping of hydrogen and oxygen at speeds of 1 L/s or more; decomposition of organics, such as acetone, produce multiple fragments and can result in a substantial increase in the apparent pressure. Oxygen also reacts with trace amounts of carbon invariably present in tungsten filaments, producing carbon monoxide and carbon dioxide. To minimize such reactions, low-work-function electron emitters such as thoria-coated or thoriated tungsten and rhenium can be used, typically operating at temperatures on the order of 1450°C. An additional advantage is that the power used by a thoria-coated emitter is quite low; for example, at 10 mA emission it is $\sim$3 W for thoria-coated tungsten, as compared to $\sim$27 W for pure tungsten, so that the gauge operates at a lower overall temperature and outgassing is reduced (P.A. Redhead, pers. comm.). One disadvantage of such filaments is that the calibration stability is possibly less than that for tungsten filament gauges (Filippelli and Abbott, 1995). Additionally, thoria is radioactive and it appears possible that it will eventually be replaced by other emitters, such as yttria.

Tungsten filaments will instantaneously fail on exposure to a high pressure of air or oxygen. Where such exposure is a possibility, failure can be avoided with a rhenium filament.

A further, if minor, perturbation of an operating gauge results from greatly decreased electron emission in the presence of certain gases such as oxygen. For this reason, all gauge supplies control the electron emission by changes in temperature of the filament, typically increasing it in the presence of such gases, and consequently affecting the rate of reaction at the filament.

Electron Stimulated Desorption (ESD): The normal mechanism of gauge operation is the production of ions by electron impact on gas molecules or atoms. However, electron impact on molecules and atoms adsorbed on the grid structure results in their desorption, and a small fraction of the desorption occurs as ions, some of which will reach the ion collector. For some gases, such as oxygen on a molybdenum grid operating at low pressure, the ion desorption can actually exceed the gas-phase ion production, resulting in a large error in pressure measurement (Redhead et al., 1968). The spurious current increases with the gas coverage of the grid surface, and can be minimized by operating at a high electron-emission level, such that the rate of electron-stimulated desorption exceeds the rate of gas adsorption. Thus, one technique for detecting serious errors due to ESD is to change the electron emission. For example, an increase in current will result in a drop in the indicated pressure as the adsorption coverage is reduced. A second and more valuable technique is to use a gauge fitted with a modulator electrode, as first described by Redhead (1960). This simple procedure can clearly identify the presence of an ESD-based error and provide a more accurate reading of pressure; unfortunately, modulated gauges are not commonly available. Ultimately, low-pressure measurements are best determined using a calibrated mass spectrometer, or by a combination of a mass spectrometer and calibrated ion gauge.

Stability of Ion Gauge Calibration: The calibration of a gauge depends on the geometry of the gauge elements, including the relative positioning of the filaments and grid. For this reason, it is always desirable before calibrating a gauge to stabilize its structure by outgassing rigorously. This allows the grid structure to anneal, and, with tungsten filaments, accelerates crystal growth, which ultimately stabilizes the filament shape. In work demanding high-precision pressure measurement, comparison against a reference gauge that has a calibration traceable to the National Institutes of Standards and Technology (NIST) is essential. Clearly such a reference standard should only be used for calibration and not under conditions that are subject to the possible problems associated with use in a working environment. Recent work describes the development of a highly stable gauge structure, reportedly providing high reliability without recourse to such time-consuming calibration procedures (Arnold et al., 1994; Tilford et al., 1995).

Ionization Gauges: Cold Cathode Type. The cold cathode gauge uses a confined discharge to sustain a circulating electron current for the ionization of gases. The absence of a hot cathode provides a far more rugged gauge, the discharge requires less power, outgassing is much reduced, and the sensitivity of the gauge is high, so that simpler electronics can be used.

Based on these facts, the gauges would appear ideal substitutes for the hot-cathode type of gauge. The fact that this is not yet so where precise measurement of pressure is required is related to the history of cold cathode gauges. Some older designs of such gauges are known to exhibit serious instabilities (Lange et al., 1966). An excellent review of these gauges has been published (Peacock et al., 1991), and a recent paper (Kendall and Drubetsky, 1997) provides reassurance that instabilities should not be a major concern with modern gauges.

Operating Principles: The first widely used commercial cold cathode gauge was developed by Penning (Penning, 1937; Penning and Nienhuis, 1949). It uses an anode ring or cylinder at a potential of 2000 V placed between cathode plates at ground potential. A magnetic field of 0.15 tesla is directed along the axis of the cathode. The confined Penning discharge traps a circulating cloud of electrons, substantially at cathode potential, along the axis of the anode. The electron path lengths are very long, as compared to the hot cathode gauge, so that the pressure measuring sensitivity is very high, permitting a simple microammeter to be used for readout. The gauge is known variously as the Penning (or PIG), Philips, or simply cold cathode gauge and is widely used where a rugged gauge is required. The operating range is from 10^{-2} to 10^{-7} torr.

Note that any discharge current in the Penning and other cold cathode discharge gauges is extinguished at pressures of a few torr. When a gauge does not give a pressure indication, this means that either the pressure is below 10^{-7} torr or at many torr, a rather significant difference. Thus, it is necessary to use an additional gauge that is responsive in the blind spot of the Penning—i.e., between atmospheric pressure and 10^{-2} torr.

The Penning gauge has a number of limitations which preclude its use for the precise measurement of pressure. It is subject to some instability at the lower end of its range, because the discharge tends to extinguish. Discontinuities in the pressure indication may also occur throughout the range. Both of these characteristics were also evident in a discharge gauge that was designed to measure pressures into the 10^{-10} torr range, where discontinuities were detected throughout the entire operating range (Lange et al., 1966). These appear to result from changes between two or more modes of discharge. Note, however, that instabilities may simply indicate that the gauge is dirty.

The Penning has a higher pumping speed (up to 0.5 L/s) than a Bayard-Alpert gauge, so it is even more essential to provide a high-conductance connection between gauge and the vacuum chamber.

A number of refinements of the cold cathode gauge have been introduced by Redhead (Redhead et al., 1968), and commercial versions of these and other gauges are available for use to at least 10^{-10} torr. The discontinuities in such gauges are far less than those discussed above so that they are becoming widely used.

Mass Spectrometers. A mass spectrometer for use on vacuum systems is variously referred to as a partial

pressure analyzer (PPA), residual gas analyzer (RGA), or quad. These are indispensable for monitoring the composition of the gas in a vacuum system, for troubleshooting, and for detecting leaks. It is however a relatively difficult procedure for obtaining quantitative readings.

Magnetic Sector Mass Spectrometers: Commercial instruments were originally developed for analytical purposes but were generally too large for general application to vacuum systems. Smaller versions provided excellent performance, but the presence of a large permanent magnet or electromagnet remained a serious limitation. Such instruments are now mainly used in helium leak detectors, where a compact instrument using a small magnet is perfectly satisfactory for resolving the helium-4 peak from the only common adjacent peak, that due to hydrogen-2.

Quadrupole Mass Filter: Compact instruments are available in many versions, combined with control units that are extremely versatile (Dawson, 1995). Mass selection is accomplished through the application of a combined DC and RF electrical voltage to two pairs of quadrupole rods. The mass peaks are selected, in turn, by changing the voltage. High sensitivity is achieved using a variety of ion sources, all employing electron bombardment ionization, and by an electron-multiplier-based ion detector with a gain as high as 10^5. By adjusting the ratio of the DC and RF electrical potentials the quadrupole resolution can be changed to provide either high sensitivity to detect very low partial pressures, at the expense of the ability to completely resolve adjacent mass peaks, or low sensitivity, with more complete mass resolution.

A major problem with the quadrupole spectrometers is that not all compact versions possess the stability for quantitative applications. Very wide variations in detection sensitivity require frequent recalibration, and spurious peaks are sometimes present (Lieszkovszky et al., 1990; Tilford, 1994).

Despite such concerns, rapidly scanning the ion peaks from the gas in a system can often determine the gas composition. Interpretation is simplified by the limited number of residual gases commonly found in a system (Drinkwine and Lichtman, 1980).

VACUUM SYSTEM CONSTRUCTION AND DESIGN

Materials for Construction

Stainless steel, aluminum, copper, brass, alumina, and borosilicate glass are among the materials that have been used for the construction of vacuum envelopes. Excellent sources of information on these and other materials, including refractory metals suitable for internal electrodes and heaters, are available (Kohl, 1967; Rosebury, 1965). The present discussion will be limited to those most commonly used.

Perhaps the most common material is stainless steel, most frequently 304L. The material is available in a wide range of shapes and sizes, and is easily fabricated by heliarc welding or brazing. Brazing should be done in hydrogen or vacuum. Torch brazing with the use of a flux should be avoided, or reserved for rough vacuum applications. The flux is extremely difficult to remove, and the washing procedure leaves any holes plugged with water, so that helium leak detection is often unsuccessful. However, note that any imperfections in this and other metals tend to extend along the rolling or extrusion direction. For example, the defects in bar stock extend along the length. If a thin-walled flat disc or conical shape is machined from such stock, it will often develop holes through the wall. Construction from rolled plate or sheet stock is preferable, with cones or cylinders made by full-penetration seam welding.

Aluminum is an excellent choice for the vacuum envelope, particularly for high-atomic-radiation environments. But joining by either arc welding or brazing is exceedingly demanding.

Kovar is a material having an expansion coefficient closely matched to borosilicate glass, which is widely used in sealing windows and alumina-insulating sections for electrical feedthroughs.

Borosilicate glass remains a valid choice for construction. Although outgassing is a serious problem in an unbaked system, a bakeout to as high as 400°C reduces the outgassing so UHV conditions can readily be achieved.

High-alumina ceramic provides a high-temperature vacuum wall. It can be coated with a thin metallic film, such as titanium, and then brazed to a Kovar header to make electrical feedthroughs. Alumina stock tubing does not have close dimension tolerances. Shaping for vacuum use requires time-consuming and expensive procedures such as grinding and cavitron milling. It is useful for internal electrically insulating components, and can be used at temperatures up to at least 1800°C. In many cases where lower-temperature performance is acceptable, such components are more conveniently fabricated from either boron nitride (1650°C) or Macor (1000°C) using conventional metal-cutting techniques.

Elastomers are widely used in O-ring seals. Buna N and Viton are most commonly used. The main advantage of Viton is that it can be used up to 200°C, as compared to 80°C for Buna. The higher temperature capability is valuable for rapid degassing of the O-ring prior to installation, and also for use at an elevated temperature. Polyimide can be used as high as 300°C and is used as the sealing surface on the nose of vacuum valves intended for UHV applications, permitting a higher-temperature bakeout.

Hardware Components

Use of demountable seals allows greater flexibility in system assembly and change. Replacement of a flange-mounted component is a relatively trivial problem as compared to that where the component is brazed or welded in place. A wide variety of seals are available, but only two examples will be discussed.

O-ring Flange Seals. The O-ring seal is by far the most convenient technique for high-vacuum use. It should be generally limited to systems at pressures in the 10^{-7} torr range and higher. Figure 4A shows the geometry of a simple seal. Note that the sealing surfaces are those in contact with the two flanges. The inner surface serves only to

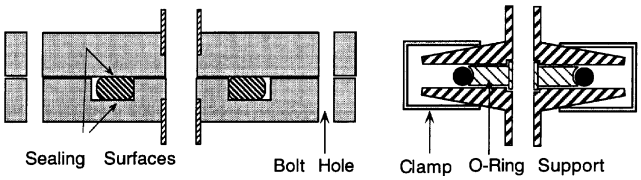


Figure 4. Simple O-ring flange (**A**) and International Standards Organization Type KF O-ring flange (**B**). Based, with permission, on Short Course Notes of the American Vacuum Society.

support the O-ring. The compressed O-ring does not completely fill the groove. The groove for vacuum use is machined so that the inside diameter fits the inside diameter of the O-ring. The pressure differential across the ring pushes the ring against the inner surface. For hydraulic applications the higher pressure is on the inside, so the groove is machined to contact the outside surface of the O-ring. Use of O-rings for moving seals is commonplace; this is acceptable for rotational motion, but is undesirable for translational motion. In the latter case, scuffing of the O-ring leads to leakage, and each inward motion of the rod transfers moisture and other atmospheric gases into the system. Many inexpensive valves use such a seal on the valve shaft, and it is the main point of failure. In some cases, a double O-ring with a grease fill between the rings is used; this is unacceptable if a clean, oil-free system is required.

Modern O-ring seal assemblies have largely standardized on the ISO KF type geometry (Fig. 4B), where the two flanges are identical. This permits flexibility in assembly, and has the great advantage that components from different manufacturers may readily be interchanged.

Acetone should never be used for cleaning O-rings for vacuum use; it permeates into the rubber, creating a long-lasting source of the solvent in the vacuum system. It is common practice to clean new O-rings by wiping with a lint-free cloth. O-rings may be lubricated with a low-vapor-pressure vacuum grease such as Apiezon M; the grease should provide a shiny surface, but no accumulation of grease. Such a film is not required to produce a leak tight seal, and is widely thought to function only as a lubricant.

All elastomers are permeable to the gases in air, and an O-ring seal is always a source of gas permeating into a vacuum system. The approximate permeation rate for air, at ambient temperature, through a Viton O-ring is 8×10^{-9} torr-L/s for each centimeter length of O-ring. This permeation load is rarely of concern in systems at pressures in the 10^{-7} torr range or higher, if the system does not include an inordinate number of such seals. But in a UHV system they become the dominant source of gas and must generally be excluded. A double O-ring geometry, with the space between the rings continuously evacuated, constitutes a useful improvement. The guard ring vacuum minimizes air permeation and minimizes the effect of leaks during motion. Such a seal extends the use of O-rings to the UHV range, and can be helpful if frequent access to the system is essential.

All-metal Flange Seals. All-metal seals were developed for UHV applications, but the simplicity and reliability of these seals finds application where the absence of leakage over a wide range of operating conditions is essential. The Conflat seal, developed by Wheeler (1963), uses a partially annealed copper gasket that is elastically deformed by capture between the knife edges on the two flanges and along the outside diameter of the gasket. Such a seal remains leak-tight at temperatures from -195° to 450°C. A new gasket should be used every time the seal is made, and the flanges pulled down metal-to-metal. For flanges with diameters above $\sim$10 in., the weight of the flange becomes excessive, and a copper-wire seal using a similar capture geometry (Wheeler, 1963) is easier to lift, if more difficult to assemble.

Vacuum Valves. The most important factor in the selection of vacuum valves is avoiding an O-ring on a sliding shaft, choosing instead a bellows seal. The bonnet of the valve is usually O-ring–sealed, except where bakeability is necessary, in which case a metal seal such as the Conflat is substituted. The nose of the valve is usually an O-ring or molded gasket of Viton, polyimide for higher temperature, or metal for the most stringent UHV applications. Sealing techniques for the a UHV valve include a knife edge nose/copper seat and a capture geometry such as silver nose/stainless steel seat (Bills and Allen, 1955).

When a valve is used between a high-vacuum pump and a vacuum chamber, high conductance is essential. This normally mandates some form of gate valve. The shaft seal is again crucial for reliable, leak-tight operation. Because of the long travel of the actuating rod, a grease-filled double O-ring has frequently been used, but evacuation would be preferable. A lever-operated drive requiring only simple rotation using an O-ring seal, or a bellows-sealed drive, provides superior performance.

Electrical Feedthroughs. Electrical feedthroughs of high reliability are readily available, from single to multiple pin, with a wide range of current and voltage ratings. Many can be brazed or welded in place, or can be obtained mounted in a flange with elastomer or metal seals. Many are compatible with high-temperature bakeout. One often overlooked point is the use of feedthroughs for thermocouples. It is common practice to terminate the thermocouple leads on the internal pins of any available feedthrough, continuing the connection on the outside of these same pins using the appropriate thermocouple compensating leads. This is acceptable if the temperature at the two ends of each pin is the same, so the EMF generated at each bimetallic connection is identical, but in opposition. Such uniformity in temperature is not the norm, and an error in the measurement of temperature is thus inevitable. Thermocouple feedthroughs with pins of the appropriate compensating metal are readily available. Alternatively, the thermocouple wires can pass through hollow pins on a feedthrough, using a brazed seal on the vacuum side of the pins.

Rotational and Translational Motion Feedthroughs. For systems operating at pressures of 10^{-7} torr and higher, O-ring seals are frequently used. They are reasonably satisfactory for rotational motion but prone to leakage in translational motion. In either case, the use of a double

O-ring, with an evacuated space between the O-rings, reduces both permeation and leakage from the atmosphere, as described above. Magnetic coupling across the vacuum wall avoids any problem of leakage and is useful where the torque is low and where precise motion is not required. If high-speed precision motion or very high torque is required, feedthroughs are available where a drive shaft passes through the wall, using a magnetic liquid seal; such seals use low-vapor-pressure liquids (polyphenylether diffusion pump fluids: see discussion of Diffusion Pumps) and are helium-leak-tight, but cannot sustain a significant bakeout. For UHV applications, bellows-sealed feedthroughs are available for precise control of both rotational and translational motion. O'Hanlon (1989) gives a detailed discussion of this.

Assembly, Processing, and Operation of Vacuum Systems

As a general rule, all parts of a vacuum system should be kept as clean as possible through fabrication and until final system assembly. This applies with equal importance if the system is to be baked. Avoid the use of sulfur-containing cutting oil in machining components; sulfur gives excellent adhesion during machining, but is difficult to remove. Rosebury (1965) provides guidance on this point.

Traditional cleaning procedures have been reviewed by Sasaki (1991). Many of these use a vapor degreaser to remove organic contamination and detergent scrubbing to remove inorganic contamination. Restrictions on the use of chlorinated solvents such as trichloroethane, and limits on the emission of volatile organics, have resulted in a switch to more environmentally acceptable procedures. Studies at the Argonne National Laboratory led to two cleaning agents which were used in the assembly of the Advanced Photon Source (Li et al., 1995; Rosenburg et al., 1994). These were a mild alkaline detergent (Almeco 18) and a citric acid-based material (Citronox). The cleaning baths were heated to 50° to 65°C and ultrasonic agitation was used. The facility operates in the low 10^{-10} torr range. Note that cleanliness was observed during all stages, so that the cleaning procedures were not required to reverse careless handling.

The initial pumpdown of a new vacuum system is always slower than expected. Outgassing will fall off with time by perhaps a factor of 10^3, and can be accelerated by increasing the temperature of the system. A bakeout to 400°C will accomplish a reduction in the outgassing rate by a factor of 10^7 in 15 hr. But to be effective, the entire system must be heated.

Once the system has been outgassed any pumpdown will be faster than the initial one as long as the system is not left open to the atmosphere for an extended period. Always vent a system to dry gas, and minimize exposure to atmosphere, even by continuing the purge of dry gas when the system is open. To the extent that the ingress of moist air is prevented, so will the subsequent pumpdown be accelerated.

Current vacuum practice for achieving ultrahigh vacuum conditions frequently involves baking to a more moderate temperature, on the order of 200°C, for a period of several days. Thereafter, samples are introduced or removed from the system by use of an intermediate vacuum entrance, the so-called load-lock system. With this technique the pressure in the main vacuum system need never rise by more than a decade or so from the operating level. If the need for ultrahigh vacuum conditions in a project is anticipated at the outset, the construction and operation of the system is not particularly difficult, but does involve a commitment to use materials and components that can be subjected to bakeout temperature without deterioration.

Design of Vacuum Systems

In any vacuum system, the pressure of interest is that in the chamber where processes occur, or where measurements are carried out, and not at the mouth of the pump. Thus, the pumping speed used to estimate the system operating pressure should be that at the exit of the vacuum chamber, not simply the speed of the pump itself. Calculation of this speed requires a knowledge of the substantial pressure drop often present in the lines connecting the pump to the chamber. In many system designs, the effective pumping speed at the chamber opening falls below 40% of the pump speed; the further the pump is located from the vacuum chamber the greater is the loss.

One can estimate system performance to avoid any serious errors in component selection. It is essential to know the conductance, C, of all components in the pumping system.

Conductance is defined by the expression

$$C = \frac{Q}{P_1 - P_2} \tag{4}$$

where C is the conductance in L/s, $(P_1 - P_2)$ is the pressure drop across the component in torr, and Q is the total throughput of gas, in torr-L/s.

The conductance of tubes with a circular cross-section can be calculated from the dimensions, although the conductance does depend upon the type of gas flow. Most high-vacuum systems operate in the molecular flow regime, which often begins to dominate once the pressure falls below 10^{-2} torr. An approximate expression for the pressure at which molecular flow becomes dominant is

$$P \geq \frac{0.01}{D} \tag{5}$$

where D is the internal diameter of the conductance element in cm and P is the pressure in torr.

In molecular flow, the conductance for air, of a long tube (where $L/D > 50$), is given by

$$C = \frac{12.1(D^3)}{L} \tag{6}$$

where L is the length in centimeters and C is in L/s.

Molecular flow occurs in virtually all high-vacuum systems. Note that the conductance in this regime is independent of pressure. The performance of pumping systems is frequently limited by practical conductance limits. For any component, conductance in the low-pressure regime is lower than in any other pressure regime, so careful design consideration is necessary.

At higher pressures ($PD \geq 0.5$) the flow becomes viscous. For long tubes, where laminar flow is fully developed ($L/D \gg 100$), the conductance is given by

$$C = \frac{(182)(P)(D^4)}{L} \tag{7}$$

As can be seen from this equation, in viscous flow, the conductance is dependent on the fourth power of the diameter, and is also dependent upon the average pressure in the tube. Because the vacuum pumps used in the higher-pressure range normally have significantly smaller pumping speeds than do those for high vacuum, the problems associated with the vacuum plumbing are much simpler. The only time that one must pay careful attention to the higher-pressure performance is when system cycling time is important, or when the entire process operates in the viscous flow regime.

When a group of components are connected in series, the net conductance of the group can be approximated by the expression

$$\frac{1}{C_{\text{total}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots \tag{8}$$

From this expression, it is clear that the limiting factor in the conductance of any string of components is the smallest conductance of the set. It is not possible to compensate low conductance, e.g., in a small valve, by increasing the conductance of the remaining components. This simple fact has escaped very many casual assemblers of vacuum systems.

The vacuum system shown in Figure 5 is assumed to be operating with a fixed input of gas from an external source, which dominates all other sources of gas such as outgassing or leakage. Once flow equilibrium is established, the throughput of gas, Q, will be identical at any plane drawn through the system, since the only source of gas is the external source, and the only sink for gas is the pump.

The pressure at the mouth of the pump is given by

$$P_2 = \frac{Q}{S_{\text{pump}}} \tag{9}$$

and the pressure in the chamber will be given by

$$P_1 = \frac{Q}{S_{\text{chamber}}} \tag{10}$$

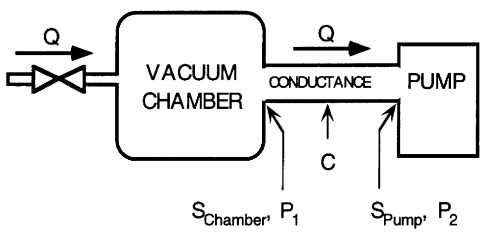


Figure 5. Pressures and pumping speeds developed by a steady throughput of gas (Q) through a vacuum chamber, conductance (C) and pump.

Combining this with Equation 4, to eliminate pressure, we have

$$\frac{1}{S_{\text{chamber}}} = \frac{1}{S_{\text{pump}}} + \frac{1}{C} \tag{11}$$

For the case where there are a series of separate components in the pumping line, the expression becomes

$$\frac{1}{S_{\text{chamber}}} = \frac{1}{S_{\text{pump}}} + \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots \tag{12}$$

The above discussion is intended only to provide an understanding of the basic principles involved and the type of calculations necessary to specify system components. It does not address the significant deviations from this simple framework that must be corrected for, in a precise calculation (O'Hanlon, 1989).

The estimation of the base pressure requires a determination of gas influx from all sources and the speed of the high-vacuum pump at the base pressure. The outgassing contributed by samples introduced into a vacuum system should not be neglected. The critical sources are outgassing and permeation. Leaks can be reduced to negligible levels using good assembly techniques. Published outgassing and permeation rates for various materials can vary by as much as a factor of two (O'Hanlon, 1989; Redhead et al., 1968; Santeler et al., 1966). Several computer programs, such as that described by Santeler (1987), are available for more precise calculation.

LEAK DETECTION IN VACUUM SYSTEMS

Before assuming that a vacuum system leaks, it is useful to consider if any other problem is present. The most important tool in such a consideration is a properly maintained log book of the operation of the system. This is particularly the case if several people or groups use a single system. If key check points in system operation are recorded weekly, or even monthly, then the task of detecting a slow change in performance is far easier.

Leaks develop in cracked braze joints, or in torch-brazed joints once the flux has finally been removed. Demountable joints leak if the sealing surfaces are badly scratched, or if a gasket has been scuffed, by allowing the flange to rotate relative to the gasket as it is compressed. Cold flow of Teflon or other gaskets slowly reduces the compression and leaks develop. These are the easy leaks to detect, since the leak path is from the atmosphere into the vacuum chamber, and a trace gas can be used for detection.

A second class of leaks arise from faulty construction techniques; they are known as virtual leaks. In all of these, a volume or void on the inside of a vacuum system communicates to that system only through a small leak path. Every time the system is vented to the atmosphere, the void fills with venting gas, then in the pumpdown this gas flows back into the chamber with a slowly decreasing throughput, as the pressure in the void falls. This extends

the system pumpdown. A simple example of such a void is a screw placed in a blind tapped hole. A space always remains at the bottom of the hole and the void is filled by gas flowing along the threads of the screw. The simplest solution is a screw with a vent hole through the body, providing rapid pumpout. Other examples include a double O-ring in which the inside O-ring is defective, and a double weld on the system wall with a defective inner weld. A mass spectrometer is required to confirm that a virtual leak is present. The pressure is recorded during a routine exhaust, and the residual gas composition is determined as the pressure is approaching equilibrium. The system is again vented using the same procedure as in the preceding vent, but the vent uses a gas that is not significant in the residual gas composition; the gas used should preferably be nonadsorbing, such as a rare gas. After a typical time at atmospheric pressure, the system is again pumped down. If gas analysis now shows significant vent gas in the residual gas composition, then a virtual leak is probably present, and one can only look for the culprit in faulty construction.

Leaks most often fall in the range of 10^{-4} to 10^{-6} torr-L/s. The traditional leak rate is expressed in atmospheric cubic centimeters per second, which is 1.3 torr-L/s. A variety of leak detectors are available with practical sensitivities varying from around 1×10^{-3} to 2×10^{-11} torr-L/s.

The simplest leak detection procedure is to slightly pressurize the system and apply a detergent solution, similar to that used by children to make soap bubbles, to the outside of the system. With a leak of $\sim 1 \times 10^{-3}$ torr-L/s, bubbles should be detectable in a few seconds. Although the lower limit of detection is at least one decade lower than this figure, successful use at this level demands considerable patience.

A similar inside-out method of detection is to use the kind of halogen leak detector commonly available for refrigeration work. The vacuum system is partially backfilled with a freon and the outside is examined using a sniffer hose connected to the detector. Leaks the order of 1×10^{-5} torr-L/s can be detected. It is important to avoid any significant drafts during the test, and the response time can be many seconds, so the sniffer must be moved quite slowly over the suspect area of the system. A far more sensitive instrument for this procedure is a dedicated helium leak detector (see below) with a sniffer hose testing a system partially back-filled with helium.

A pressure gauge on the vacuum system can be used in the search for leaks. The most productive approach applies if the system can be segmented by isolation valves. By appropriate manipulation, the section of the system containing the leak can be identified. A second technique is not so straightforward, especially in a nonbaked system. It relies on the response of ion or thermal conductivity gauges differing from gas to gas. For example, if the flow of gas through a leak is changed from air to helium by covering the suspected area with helium, then the reading of an ionization gauge will change, since the helium sensitivity is only ~16% of that for air. Unfortunately, the flow of helium through the leak is likely to be ~2.7 times that for air, assuming a molecular flow leak, which partially offsets the change in gauge sensitivity. A much greater problem is that the search for a leak is often started just after exposure to the atmosphere and pumpdown. Consequently outgassing is an ever-changing factor, decreasing with time. Thus, one must detect a relatively small decrease in a gauge reading, due to the leak, against a decreasing background pressure. This is not a simple process; the odds are greatly improved if the system has been baked out, so that outgassing is a much smaller contributor to the system pressure.

A far more productive approach is possible if a mass spectrometer is available on the system. The spectrometer is tuned to the helium-4 peak, and a small helium probe is moved around the system, taking the precautions described later in this section. The maximum sensitivity is obtained if the pumping speed of the system can be reduced by partially closing the main pumping valve to increase the pressure, but no higher than the mid-10^{-5} torr range, so that the full mass spectrometer resolution is maintained. Leaks in the 1×10^{-8} torr-L/s range should be readily detected.

The preferred method of leak detection uses a standalone helium mass spectrometer leak detector (HMSLD). Such instruments are readily available with detection limits of 2×10^{-10} torr-L/s or better. They can be routinely calibrated so the absolute size of a leak can be determined. In many machines this calibration is automatically performed at regular intervals. Given this, and the effective pumping speed, one can find, using Equation 1, whether the leak detected is the source of the observed deterioration in the system base pressure.

In an HMSLD, a small mass spectrometer tuned to detect helium is connected to a dedicated pumping system, usually a diffusion or turbomolecular pump. The system or device to be checked is connected to a separately pumped inlet system, and once a satisfactory pressure is achieved, the inlet system is connected directly to the detector and the inlet pump is valved off. In this mode, all of the gas from the test object passes directly to the helium leak detector. The test object is then probed with helium, and if a leak is detected, and is covered entirely with a helium blanket, the reading of the detector will provide an absolute indication of the leak size. In this detection mode, the pressure in the leak detector module cannot exceed $\sim 10^{-4}$ torr, which places a limit on the gas influx from the test object. If that influx exceeds some critical value, the flow of gas to the helium mass spectrometer must be restricted, and the sensitivity for detection will be reduced. This mode, of leak detection is not suitable for dirty systems, since the gas flows from the test object directly to the detector, although some protection is usually provided by interposing a liquid nitrogen cold trap.

An alternative technique using the HMSLD is the so-called counterflow mode. In this, the mass spectrometer tube is pumped by a diffusion or turbomolecular pump which is designed to be an ineffective pump for helium (and for hydrogen), while still operating at normal efficiency for all higher-molecular-weight gases. The gas from the object under test is fed to the roughing line of the mass spectrometer high-vacuum pump, where a higher pressure can be tolerated (on the order of 0.5 torr). Contaminant gases, such as hydrocarbons, as well as air, cannot reach the spectrometer tube. The sensitivity of an

HMSLD in this mode is reduced about an order of magnitude from the conventional mode, but it provides an ideal method of examining quite dirty items, such as metal drums or devices with a high outgassing load.

The procedures for helium leak detection are relatively simple. The HMSLD is connected to the test object for maximum possible pumping speed. The time constant for the buildup of a leak signal is proportional to V/S, where V is the volume of the test system and S the effective pumping speed. A small time constant allows the helium probe to be moved more rapidly over the system.

For very large systems, pumped by either a turbomolecular or diffusion pump, the response time can be improved by connecting the HMSLD to the foreline of the system, so the response is governed by the system pump rather than the relatively small pump of the HMSLD. With pumping systems that use a capture-type pump, this procedure cannot be used, so a long time constant is inevitable. In such cases, use of an HMSLD and helium sniffer to probe the outside of the system, after partially venting to helium, may be a better approach. Further, a normal helium leak check is not possible with an operating cryopump; the limited capacity for pumping helium can result in the pump serving as a low-level source of helium, confounding the test.

Rubber tubing must be avoided in the connection between system and HMSLD, since helium from a large leak will quickly permeate into the rubber and thereafter emit a steadily declining flow of helium, thus preventing use of the most sensitive detection scale. Modern leak detectors can offset such background signals, if they are relatively constant with time.

With the HMSLD operating at maximum sensitivity, a probe, such as a hypodermic needle with a very slow flow of helium, is passed along any suspected leak locations, starting at the top of the system, and avoiding drafts. Whenever a leak signal is first heard, and the presence of a leak is quite apparent, the probe is removed, allowing the signal to decay; checking is resumed, using the probe with no significant helium flow, to pinpoint the exact location of the leak. Ideally, the leak should be fixed before the probe is continued, but in practice the leak is often plugged with a piece of vacuum wax (sometimes making the subsequent repair more difficult), and the probe is completed before any repair is attempted. One option, already noted, is to blanket the leak site with helium to obtain a quantitative measure of its size, and then calculate whether this is the entire problem. This is not always the preferred procedure, because a large slug of helium can lead to a lingering background in the detector, precluding a check for further leaks at maximum detector sensitivity.

A number of points need to be made with regard to the detection of leaks:

1. Bellows should be flexed while covered with helium.
2. Leaks in water lines are often difficult to locate. If the water is drained, evaporative cooling may cause ice to plug a leak, and helium will permeate through the plug only slowly. Furthermore, the evaporating water may leave mineral deposits that plug the hole. A flow of warm gas through the line, overnight, will often open up the leak and allow helium leak detection. Where the water lines are internal to the system, the chamber must be opened so that the entire line is accessible for a normal leak check. However, once the lines can be viewed, the location of the leak is often signaled by the presence of discoloration.
3. Do not leave a helium probe near an O-ring for more than a few seconds; if too much helium goes into solution in the elastomer, the delayed permeation that develops will cause a slow flow of helium into the system, giving a background signal which will make further leak detection more difficult.
4. A system with a high background of hydrogen may produce a false signal in the HMSLD because of inadequate resolution of the helium and hydrogen peaks. A system that is used for the hydrogen isotopes deuterium or tritium will also give a false signal because of the presence of D_2 or HT, both of which have their major peaks at mass 4. In such systems an alternate probe gas such as argon must be used, together with a mass spectrometer which can be tuned to the mass 40 peak.

Finally, if a leak is found in a system, it is wise to fix it properly the first time lest it come back to haunt you!

LITERATURE CITED

Alpert, D. 1959. Advances in ultrahigh vacuum technology. *In* Advances in Vacuum Science and Technology, vol. 1: Proceedings of the 1st International Conference on Vacuum Technology (E. Thomas, ed.) pp. 31–38. Pergamon Press, London.

Alvesteffer, W. J., Jacobs, D. C., and Baker, D. H., 1995. Miniaturized thin film thermal vacuum sensor. *J. Vac. Sci. Technol.* A13:2980–298.

Arnold, P. C., Bills, D. G., Borenstein, M. D., and Borichevsky, S. C. 1994. Stable and reproducible Bayard-Alpert ionization gauge. *J. Vac. Sci. Technol.* A12:580–586.

Becker, W. 1959. Über eine neue Molekularpumpe. *In* Advances in Ultrahigh Vacuum Technology. Proc. 1st. Int. Cong. on Vac. Tech. (E. Thomas, ed.) pp. 173–176. Pergamon Press, London.

Benson, J. M., 1957. Thermopile vacuum gauges having transient temperature compensation and direct reading over extended ranges. *In* National Symp. on Vac. Technol. Trans. (E. S. Perry and J. H. Durrant, eds.) pp. 87–90. Pergamon Press, London.

Bills, D. G. and Allen, F. G., 1955. Ultra-high vacuum valve. *Rev. Sci. Instrum.* 26:654–656.

Brubaker, W. M. 1959. A method of greatly enhancing the pumping action of a Penning discharge. *In* Proc. 6th. Nat. AVS Symp. pp. 302–306. Pergamon Press, London.

Coffin, D. O. 1982. A tritium-compatible high-vacuum pumping system. *J. Vac. Sci. Technol.* 20:1126–1131.

Dawson, P. T. 1995. Quadrupole Mass Spectrometry and its Applications. AVS Classic Series in Vacuum Science and Technology. Springer-Verlag, New York.

Drinkwine, M. J. and Lichtman, D. 1980. Partial pressure analyzers and analysis. American Vacuum Society Monograph Series, American Vacuum Society, New York.

Dobrowolski, Z. C. 1979. Fore-Vacuum Pumps. *In* Methods of Experimental Physics, Vol. 14 (G. L. Weissler and R. W. Carlson, eds.) pp. 111–140. Academic Press, New York.

Filippelli, A. R. and Abbott, P. J. 1995. Long-term stability of Bayard-Alpert gauge performance: Results obtained from repeated calibrations against the National Institute of Standards and Technology primary vacuum standard. *J. Vac. Sci. Technol.* A13:2582–2586.

Fulker, M. J. 1968. Backstreaming from rotary pumps. *Vacuum* 18:445–449.

Giorgi, T. A., Ferrario, B., and Storey, B., 1985. An updated review of getters and gettering. *J. Vac. Sci. Technol.* A3:417–423.

Hablanian, M. H. 1997. High-Vacuum Technology, 2nd ed., Marcel Dekker, New York.

Hablanian, M. H. 1995. Diffusion pumps: Performance and operation. American Vacuum Society Monograph, American Vacuum Society, New York.

Harra, D. J. 1976. Review of sticking coefficients and sorption capacities of gases on titanium films. *J. Vac. Sci. Technol.* 13: 471–474.

Hoffman, D. M. 1979. Operation and maintenance of a diffusion-pumped vacuum system. *J. Vac. Sci. Technol.* 16:71–74.

Holland, L. 1971. Vacua: How they may be improved or impaired by vacuum pumps and traps. *Vacuum* 21:45–53.

Hyland, R. W. and Shaffer, R. S. 1991. Recommended practices for the calibration and use of capacitance diaphragm gages as transfer standards. *J. Vac. Sci. Technol.* A9:2843–2863.

Jepsen, R. L. 1967. Cooling apparatus for cathode getter pumps. U. S. patent 3,331,975, July 16, 1967.

Jepsen, R. L., 1968. The physics of sputter-ion pumps. Proc. 4th. Int. Vac. Congr. : Inst. Phys. Conf. Ser. No. 5. pp. 317–324. The Institute of Physics and the Physical Society, London.

Kendall, B. R. F. and Drubetsky, E. 1997. Cold cathode gauges for ultrahigh vacuum measurements. *J. Vac. Sci. Technol.* A15: 740–746.

Kohl, W. H., 1967. Handbook of Materials and Techniques for Vacuum Devices. AVS Classic Series in Vacuum Science and Technology. Springer-Verlag, New York.

Kuznetsov, M. V., Nazarov, A. S., and Ivanovsky, G. F. 1969. New developments in getter-ion pumps in the U. S. S. R. *J. Vac. Sci. Technol.* 6:34–39.

Lange, W. J., Singleton, J. H., and Eriksen, D. P., 1966. Calibration of a low pressure Penning discharge type gauges. *J. Vac. Sci. Technol.* 3:338–344.

Lawson, R. W. and Woodward, J. W. 1967. Properties of titanium-molybdenum alloy wire as a source of titanium for sublimation pumps. *Vacuum* 17:205–209.

Lewin, G. 1985. A quantitative appraisal of the backstreaming of forepump oil vapor. *J. Vac. Sci. Technol.* A3:2212–2213.

Li, Y., Ryding, D., Kuzay, T. M., McDowell, M. W., and Rosenburg, R. A., 1995. X-ray photoelectron spectroscopy analysis of cleaning procedures for synchrotron radiation beamline materials at the Advanced Proton Source. *J. Vac. Sci. Technol.* A13:576–580.

Lieszkovszky, L., Filippelli, A. R., and Tilford, C. R. 1990. Metrological characteristics of a group of quadrupole partial pressure analyzers. *J. Vac. Sci. Technol.* A8:3838–3854.

McCracken,. G. M. and Pashley, N. A., 1966. Titanium filaments for sublimation pumps. *J. Vac. Sci. Technol.* 3:96–98.

Nottingham, W. B. 1947. 7th. Annual Conf. on Physical Electronics, M.I.T.

O'Hanlon, J. F. 1989. A User's Guide to Vacuum Technology. John Wiley & Sons, New York.

Osterstrom, G. 1979. Turbomolecular vacuum pumps. *In* Methods of Experimental Physics, Vol. 14 (G. L. Weissler and R. W. Carlson, eds.) pp. 111–140. Academic Press, New York.

Peacock, R. N. 1998. Vacuum gauges. *In* Foundations of Vacuum Science and Technology (J. M. Lafferty, ed.) pp. 403–406. John Wiley & Sons, New York.

Peacock, R. N., Peacock, N. T., and Hauschulz, D. S., 1991. Comparison of hot cathode and cold cathode ionization gauges. *J. Vac. Sci. Technol.* A9: 1977–1985.

Penning, F. M. 1937. High vacuum gauges. *Philips Tech. Rev.* 2:201–208.

Penning, F. M. and Nienhuis, K. 1949. Construction and applications of a new design of the Philips vacuum gauge. *Philips Tech. Rev.* 11:116–122.

Redhead, P. A. 1960. Modulated Bayard-Alpert Gauge Rev. *Sci. Instr.* 31:343–344.

Redhead, P. A., Hobson, J. P., and Kornelsen, E. V. 1968. The Physical Basis of Ultrahigh Vacuum AVS Classic Series in Vacuum Science and Technology. Springer-Verlag, New York.

Reimann, A. L. 1952. Vacuum Technique. Chapman & Hall, London.

Rosebury, F., 1965. Handbook of Electron Tube and Vacuum Technique. AVS Classic Series in Vacuum Science and Technology. Springer-Verlag, New York.

Rosenburg, R. A., McDowell, M. W., and Noonan, J. R., 1994. X-ray photoelectron spectroscopy analysis of aluminum and copper cleaning procedures for the Advanced Proton Source. *J. Vac. Sci. Technol.* A12:1755–1759.

Rutherford, 1963. Sputter-ion pumps for low pressure operation. *In* Proc. 10th. Nat. AVS Symp. pp. 185–190. The Macmillan Company, New York.

Santeler, D. J. 1987. Computer design and analysis of vacuum systems. *J. Vac. Sci. Technol.* A5:2472–2478.

Santeler, D. J., Jones, W. J., Holkeboer, D. H., and Pagano, F. 1966. AVS Classic Series in Vacuum Science and Technology. Springer-Verlag, New York.

Sasaki, Y. T. 1991. A survey of vacuum material cleaning procedures: A subcommittee report of the American Vacuum Society Recommended Practices Committee. *J. Vac. Sci. Technol.* A9:2025–2035.

Singleton, J. H. 1969. Hydrogen pumping speed of sputter-ion pumps. *J. Vac. Sci. Technol.* 6:316–321.

Singleton, J. H. 1971. Hydrogen pumping speed of sputter-ion pumps and getter pumps. *J. Vac. Sci. Technol.* 8:275–282.

Snouse, T. 1971. Starting mode differences in diode and triode sputter-ion pumps *J. Vac. Sci. Technol.* 8:283–285.

Tilford, C. R. 1994. Process monitoring with residual gas analyzers (RGAs): Limiting factors. *Surface and Coatings Technol.* 68/69: 708–712.

Tilford, C. R., Filippelli, A. R., and Abbott, P. J. 1995. Comments on the stability of Bayard-Alpert ionization gages. *J. Vac. Sci. Technol.* A13:485–487.

Tom, T. and James, B. D. 1969. Inert gas ion pumping using differential sputter-yield cathodes. *J. Vac. Sci. Technol.* 6:304–307.

Welch, K. M. 1991. Capture pumping technology. Pergamon Press, Oxford, U. K.

Welch, K. M. 1994. Pumping of helium and hydrogen by sputter-ion pumps. II. Hydrogen pumping. *J. Vac. Sci. Technol.* A12:861–866.

Wheeler, W. R. 1963. Theory And Application Of Metal Gasket Seals. Trans. 10th. Nat. Vac. Symp. pp. 159–165. Macmillan, New York.

KEY REFERENCES

Dushman, 1962. See above.

Provides the scientific basis for all aspects of vacuum technology.

Hablanian, 1997. See above.

Excellent general practical guide to vacuum technology.

Kohl, 1967. See above.

A wealth of information on materials for vacuum use, and on electron sources.

Lafferty, J. M. (ed.). 1998. Foundations of Vacuum Science and Technology. John Wiley & Sons, New York.

Provides the scientific basis for all aspects of vacuum technology.

O'Hanlon, 1989. See above.

Probably the best general text for vacuum technology; SI units are used throughout.

Redhead et al., 1968. See above.

The classic text on UHV; a wealth of information.

Rosebury, 1965. See above.

An exceptional practical book covering all aspects of vacuum technology and the materials used in system construction.

Santeler et al., 1966. See above.

A very practical approach, including a unique treatment of outgassing problems; suffers from lack of an index.

JACK H. SINGLETON
Consultant Monroeville
Pennsylvania

MASS AND DENSITY MEASUREMENTS

INTRODUCTION

The precise measurement of mass is one of the more challenging measurement requirements that materials scientists must deal with. The use of electronic balances has become so widespread and routine that the accurate measurement of mass is often taken for granted. While government institutions such as the National Institutes of Standards and Technology (NIST) and state metrology offices enforce controls in the industrial and legal sectors, no such rigors generally affect the research laboratory. The process of peer review seldom makes assessments of the accuracy of an underlying measurement involved unless an egregious problem is brought to the surface by the reported results. In order to ensure reproducibility, any measurement process in a laboratory should be subjected to a rigorous and frequent calibration routine. This unit will describe the options available to the investigator for establishing and executing such a routine; it will define the underlying terms, conditions, and standards, and will suggest appropriate reporting and documenting practices. The measurement of mass, which is a fundamental measurement of the amount of material present, will constitute the bulk of the discussion. However, the measurement of derived properties, particularly density, will also be discussed, as well as some indirect techniques used particularly by materials scientists in the determination of mass and density, such as the quartz crystal microbalance for mass measurement and the analysis of diffraction data for density determination.

INDIRECT MASS MEASUREMENT TECHNIQUES

A number of differential and equivalence methods are frequently used to measure mass, or obtain an estimate of the change in mass during the course of a process or analysis. Given knowledge of the system under study, it is often possible to ascertain with reasonable accuracy the quantity of material using chemical or physical equivalence, such as the evolution of a measurable quantity of liquid or vapor by a solid upon phase transition, or the titrimetric oxidation of the material. Electroanalytical techniques can provide quantitative numbers from coulometry during an electrodeposition or electrodissolution of a solid material. Magnetometry can provide quantitative information on the amount of material when the magnetic susceptibility of the material is known.

A particularly important indirect mass measurement tool is the quartz crystal microbalance (QCM). The QCM is a piezoelectric quartz crystal routinely incorporated in vacuum deposition equipment to monitor the buildup of films. The QCM is operated at a resonance frequency that changes (shifts) as the mass of the crystal changes, providing the valuable information needed to estimate mass changes on the order of 10^{-9} to 10^{-10} g/cm^2, giving these devices a special niche in the differential mass measurement arena (Baltes et al., 1998). QCMs may also be coupled with analytical techniques such as electrochemistry or differential thermal analysis to monitor the simultaneous buildup or removal of a material under study.

DEFINITION OF MASS, WEIGHT, AND DENSITY

Mass has already been defined as a measure of the amount of material present. Clearly, there is no direct way to answer the fundamental question "what is the mass of this material?" Instead, the question must be answered by employing a tool (a balance) to compare the mass of the material to be measured to a known mass. While the SI unit of mass is the kilogram, the convention in the scientific community is to report mass or weight measurements in the metric unit that more closely yields a whole number for the amount of material being measured (e.g., grams, milligrams, or micrograms). Many laboratory balances contain "internal standards," such as metal rings of calibrated mass or an internally programmed electronic reference in the case of magnetic force compensation balances. To complicate things further, most modern electronic balances apply a set of empirically derived correction factors to the differential measurement (of the sample versus the internal standard) to display a result on the readout of the balance. This readout, of course, is what the investigator is to take on faith, and record the amount of material present

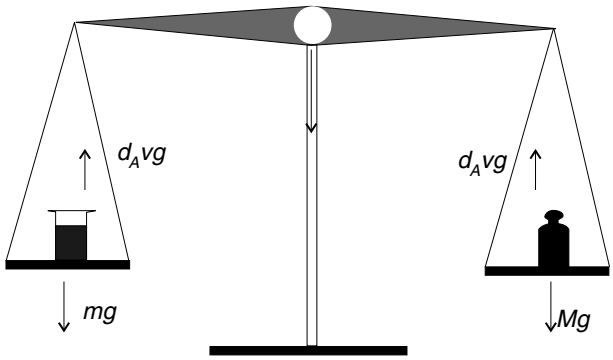


Figure 1. Schematic diagram of an equal-arm two-pan balance.

to as many decimal places as appeared on the display. In truth one must consider several concerns: what is the actual accuracy of the balance? How many of the figures in the display are significant? What are the tolerances of the internal standards? These and other relevant issues will be discussed in the sections to follow.

One type of balance does not cloak its modus operandi in internal standards and digital circuitry: the equal arm balance. A schematic diagram of an equal arm balance is shown in Figure 1. This instrument is at the origin of the term "balance," which is derived from a Latin word meaning having two pans. This elegantly simple device clearly compares the mass of the unknown to a known mass standard (see discussion of Weight Standards, below) by accurately indicating the deflection of the lever from the equilibrium state (the "balance point"). We quickly draw two observations from this arrangement. First, the lever is affected by a force, not a mass, so the balance can only operate in the presence of a gravitational field. Second, if the sample and reference mass are in a gaseous atmosphere, then each will have buoyancy characterized by the mass of the air displaced by each object. The amount of displaced air will depend on such factors as sample porosity, but for simplicity we assume here (for definition purposes) that neither the sample nor the reference mass are porous and the volume of displaced air equals the volume of the object.

We are now in a position to define the weight of an object. The weight (W) is effectively the force exerted by a mass (M) under the influence of a gravitational field, i.e., $W = Mg$, where g is the acceleration due to gravity (9.80665 m/s^2). Thus, a mass of exactly 1 g has a weight in centimeter–gram–second (cgs) units of 1 g $\times$ 980.665 cm/s^2 = 980.665 dyn, neglecting buoyancy due to atmospheric displacement. It is common to state that the object "weighs" 1 g (colloquially equating the gram to the force exerted by gravity on one gram), and to do so neglects any effect due to atmospheric buoyancy. The American Society for Testing and Materials (ASTM, 1999) further defines the force (F) exerted by a weight measured in the air as

$$F = \frac{Mg}{9.80665}\left(1 - \frac{d_A}{D}\right) \quad (1)$$

where d_A is the density of air, and D is the density of the weight (standard E4). The ASTM goes on to define a set of units to use in reporting force measurements as mass-force quantities, and presents a table of correction factors that take into account the variation of the Earth's gravitational field as a function of altitude above (or below) sea level and geographic latitude. Under this custom, forces are reported by relation to the newton, and by definition, one kilogram-force (kgf) unit is equal to 9.80665 N. The kgf unit is commonly encountered in the mechanical testing literature (see HARDNESS TESTING). It should be noted that the ASTM table considers only the changes in gravitational force and the density of dry air; i.e., the influence of humidity and temperature, for example, on the density of air is not provided. The Chemical Rubber Company's Handbook of Chemistry and Physics (Lide, 1999) tabulates the density of air as a function of these parameters. The International Committee for Weights and Measures (CIPM) provides a formula for air density for use in mass calibration. The CIPM formula accounts for temperature, pressure, humidity, and carbon dioxide concentration. The formula and description can be found in the International Organization for Legal Metrology (OIML) recommendation R 111 (OIML, 1994).

The "balance condition" in Figure 1 is met when the forces on both pans are equivalent. Taking M to be the mass of the standard, V to be the volume of the standard, m to be the mass of the sample, and v to be the volume of the sample, then the balance condition is met when $mg - d_A vg = Mg - d_A Vg$. The equation simplifies to $m - d_A v = M - d_A V$ as long as g remains constant. Taking the density of the sample to be d (equal to m/v) and that of the standard to be D (equal to M/V), it is easily shown that $m = M \times [(1 - d_A/D \div (1 - d_A/d)]$ (Kupper, 1997). This equation illustrates the dependence of a mass measurement on the air density: only when the density of the sample is identical to that of the standard (or when no atmosphere is present at all) is the measured weight representative of the sample's actual mass. To put the issue into perspective, a dry atmosphere at sea level has a density of 0.0012 g/cm^3, while that in Denver, Colorado ($\sim$1 mile above sea level) has a density of 0.00098 g/cm^3 (Kupper, 1997). If we take an extreme example, the measurement of the mass of wood (density 0.373 g/cm^3) against steel (density 8.0 g/cm^3) versus the weight of wood against a steel weight, we find that a 1 g weight of wood measured at sea level corresponds to a 1.003077 mass of wood, whereas a 1 g weight of wood measured in Denver corresponds to a 1.002511 mass of wood. The error in reporting that the weight of wood (neglecting air buoyancy) did not change would then be (1.003077 g $-$ 1.002511 g)/1 g = 0.06%, whereas the error in misreporting the mass of the wood at sea level to be 1 g would be (1.003077 g $-$ 1 g)/1.003077 g = 0.3%. It is better to assume that the variation in weight as a function of air buoyancy is negligible than to assume that the weighed amount is synonymous with the mass (Kupper, 1990).

We have not mentioned the variation in g with altitude, nor as influenced by solar and lunar tidal effects. We have already seen that g is factored out of the balance condition as long as it is held constant, so the problem will not be

encountered unless the balance is moved significantly in altitude and latitude without recalibrating. The calibration of a balance should nevertheless be validated any time it is moved to verify proper function. The effect of tidal variations on g has been determined to be of the order of 0.1 ppm (Kupper, 1997), arguably a negligible quantity considering the tolerance levels available (see discussion of Weight Standards).

Density is a derived unit defined as the mass per unit volume. Obviously, an accurate measure of both mass and volume is necessary to effect a measurement of the density. In metric units, density is typically reported in g/cm^3. A related property is the specific gravity, defined as the weight of a substance divided by the weight of an equal volume of water (the water standard is taken at 4°C, where its density is 1.000 g/cm^3). In metric units, the specific gravity has the same numerical value as the density, but is dimensionless.

In practice, density measurements of solids are made in the laboratory by taking advantage of Archimedes' principle of displacement. A fluid material, usually a liquid or gas, is used as the medium to be displaced by the material whose volume is to be measured. Precise density measurements require the material to be scrupulously clean, perhaps even degassed *in vacuo* to eliminate errors associated with adsorbed or absorbed species. The surface of the material may be porous in nature, so that a certain quantity of the displacement medium actually penetrates into the material. The resulting measured density will be intermediate between the "true" or absolute density of the material and the apparent measured density of the material containing, for example, air in its pores. Mercury is useful for the measurement of volumes of relatively smooth materials as the viscosity of liquid mercury at room temperature precludes the penetration of the liquid into pores smaller than ~5 μm at ambient pressure. On the other hand, liquid helium may be used to obtain a more faithful measurement of the absolute density, as the fluid will more completely penetrate voids in the material through pores of atomic dimension.

The true density of a material may be ascertained from the analysis of the lattice parameters obtained experimentally using diffraction techniques (see Parts X, XI, and XIII). The analysis of x-ray diffraction data elucidates the content of the unit cell in a pure crystalline material by providing lattice parameters that can yield information on vacant lattice sites versus free space in the arrangement of the unit cell. As many metallic crystals are heated, the population of vacant sites in the lattice are known to increase, resulting in a disproportionate decrease in density as the material is heated. Techniques for the measurement of true density has been reported by Feder and Nowick (1957) and by Simmons and Barluffi (1959, 1961, 1962).

WEIGHT STANDARDS

Researchers who make an effort to establish a meaningful mass measurement assurance program quickly become embroiled in a sea of acronyms and jargon. While only certain weight standards are germane to user of the precision laboratory balance, all categories of mass standards may be encountered in the literature and so we briefly list them here. In the United States, the three most common sources of weight standard classifications are NIST (formerly the National Bureau of Standards or NBS), ASTM, and the OIML. A 1954 publication of the NBS (NBS Circular 547) established seven classes of standards: J (covering denominations from 0.05 to 50 mg), M (covering 0.05 mg to 25 kg), S (covering 0.05 mg to 25 kg), S-1 (covering 0.1 mg to 50 kg), P (covering 1 mg to 1000 kg), Q (covering 1 mg to 1000 kg), and T (covering 10 mg to 1000 kg). These classifications were all replaced in 1978 by the ASTM standard E617 (ASTM, 1997), which recognizes the OIML recommendation R 111 (OIML, 1994); this standard was updated in 1997. NIST Handbook 105-1 further establishes class F, covering 1 mg to 5000 kg, primarily for the purpose of setting standards for field standards used in commerce. The ASTM standard E617 establishes eight classes (generally with tighter tolerances in the earlier classes): classes 0, 1, 2, and 3 cover the range from 1 mg to 50 kg, classes 4 and 5 cover the range from 1 mg to 5000 kg, class 6 covers 100 mg to 500 kg, and a special class, class 1.1, covers the range from 1 to 500 mg with the lowest set tolerance level (0.005 mg). The OIML R 111 establishes seven classes (also with more stringent tolerances associated with the earlier classes): E1, E2, F1, F2, and M1 cover the range from 1 mg to 50 kg, M2 covers 200 mg to 50 kg, and M3 covers 1 g to 50 kg. The ASTM classes 1 and 1.1 or OIML classes F1 and E2 are the most relevant to the precision laboratory balance. Only OIML class E1 sets stricter tolerances; this class is applied to primary calibration laboratories for establishing reference standards.

The most common material used in mass standards is stainless steel, with a density of 8.0 g/cm^3. Routine laboratory masses are often made of brass with a density of 8.4 g/cm^3. Aluminum, with a 2.7-g/cm^3 density, is often the material of choice for very small mass standards ($\leq$50 mg). The international mass standard is a 1-kg cylinder made of platinum-iridium (density 21.5 g/cm^3); this cylinder is housed in Sevres, France. Weight standard manufacturers should furnish a certificate that documents the traceability of the standard to the Sevres standard. A separate certificate may be issued that documents the calibration process for the weight, and may include a term to the effect "weights adjusted to an apparent density of 8.0 g/cm^3". These weights will have a true density that may actually be different than 8.0 g/cm^3 depending on the material used, as the specification implies that the weights have been adjusted so as to counterbalance a steel weight in an atmosphere of 0.0012 g/cm^3. In practice, variation of apparent density as a function of local atmospheric density is less than 0.1%, which is lower than the tolerances for all but the most exacting reference standards. Test procedures for weight standards are detailed in annex B of the latest OIML R 111 Committee Draft (1994). The magnetic susceptibility of steel weights, which may affect the calibration of balances based on the electromagnetic force compensation principle, is addressed in these procedures.

A few words about the selection of appropriate weight standards for a calibration routine are in order. A

fundamental consideration is the so-called 3:1 transfer ratio rule, which mandates that the error of the standard should be $<\frac{1}{3}$ the tolerance of the device being tested (ASTM, 1997). Two types of weights are typically used during a calibration routine, test weights and standard weights. Test weights are usually made of brass and have less stringent tolerances. These are useful for repetitive measurements such as those that test repeatability and off-center error (see Types of Balances). Standard weights are usually manufactured from steel and have tight, NIST-traceable tolerances. The standard weights are used to establish the accuracy of a measurement process, and must be handled with meticulous care to avoid unnecessary wear, surface contamination, and damage. Recalibration of weight standards is a somewhat nebulous issue, as no standard intervals are established. The investigator must factor in such considerations as the requirements of the particular program, historical data on the weight set, and the requirements of the measurement assurance program being used (see Mass Measurement Process Assurance).

TYPES OF BALANCES

NIST Handbook 44 (NIST, 1999) defines five classes of weighing device. Class I balances are precision laboratory weighing devices. Class II balances are used for laboratory weighing, precious metal and gem weighing, and grain testing. Class III, III L, and IIII balances are larger-capacity scales used in commerce, including everything from postal scales to highway vehicle-weighing scales. Calibration and verification procedures defined in NIST Handbook 44 have been adopted by all state metrology offices in the U.S.

Laboratory balances are chiefly available in three configurations: dual-pan equal-arm, mechanical single-pan, and top-loading. The equal arm balance is in essence that which is shown schematically in Figure 1. The single-pan balance replaces the second pan with a set of sliders, masses mounted on the lever itself, or in some cases a dial with a coiled spring that applies an adjustable and quantifiable counter-force. The most common laboratory balance is the top-loading balance. These normally employ an internal mechanism by which a series of internal masses (usually in the form of steel rings) or a system of mechanical flexures counter the applied load (Kupper, 1999). However, a spring load may be used in certain routine top-loading balances. Such concerns as changes in force constant of the spring, hysteresis in the material, etc., preclude the use of spring-loaded balances for all but the most routine measurements. On the other extreme are balances that employ electromagnetic force compensation in lieu of internal masses or mechanical flexures. These latter balances are becoming the most common laboratory balance due to their stability and durability, but it is important to note that the magnetic fields of the balance and sample may interact. Standard test methods for evaluating the performance of each of the three types of balance are set forth in ASTM standards E1270 (ASTM, 1988a; for equal-arm balances), E319 (ASTM, 1985; for mechanical single-pan balances), and E898 (ASTM, 1988b, for top-loading direct-reading balances).

A number of salient terms are defined in the ASTM standards; these terms are worth repeating here as they are often associated with the specifications that a balance manufacturer may apply to its products. The application of the principles set forth by these definitions in the establishment of a mass measurement process calibration will be summarized in the next section (see Mass Measurement Process Assurance).

Accuracy. The degree to which a measured value agrees with the true value.

Capacity. The maximum load (mass) that a balance is capable of measuring.

Linearity. The degree to which the measured values of a successive set of standard masses weighed on the balance across the entire operating range of the balance approximates a straight line. Some balances are designed to improve the linearity of a measurement by operating in two or more separately calibrated ranges. The user selects the range of operation before conducting a measurement.

Off-center Error. Any differences in the measured mass as a function of distance from the center of the balance pan.

Hysteresis. Any difference in the measured mass as a function of the history of the balance operation—e.g., a difference in measured mass when the last measured mass was larger than the present measurement versus the measurement when the prior measured mass was smaller.

Repeatability. The closeness of agreement for successive measurements of the same mass.

Reproducibility. The closeness of agreement of measured values when measurements of a given mass are repeated over a period of time (but not necessarily successively). Reproducibility may be affected by, e.g., hysteresis.

Precision. The smallest amount of mass difference that a balance is capable of resolving.

Readability. The value of the smallest mass unit that can be read from the readout without estimation. In the case of digital instruments, the smallest displayed digit does not always have a unit increment. Some balances increment the last digit by two or five, for example. Other balances incorporate a vernier or micrometer to subdivide the smallest scale division. In such cases, the smallest graduation on such devices represents the balance's readability.

Since the most common balance encountered in a research laboratory is of the electronic top-loading type, certain peculiar characteristics of this balance will be highlighted here. Balance manufacturers may refer to two categories of balance: those with versus those without internal calibration capability. In essence, an internal calibration capability indicates that a set of traceable standard masses is integrated into the mechanism of the counterbalance.

Table 1. Typical Types of Balance Available, by Capacity and Divisions

Name	Capacity (range)	Divisions Displayed
Ultramicrobalance	2 g	0.1 μg
Microbalance	3–20 g	1 μg
Semimicrobalance	30–200 g	10 μg
Macroanalytical balance	50–400 g	0.1 mg
Precision balance	100 g–30 kg	0.1 mg–1 g
Industrial balance	30–6000 kg	1 g–0.1 kg

A key choice that must be made in the selection of a balance for the laboratory is that of the operating range. A market survey of commercially available laboratory balances reveals that certain categories of balances are available. Table 1 presents common names applied to balances operating in a variety of weight measurement capacities. The choice of a balance or a set of balances to support a specific project is thus the responsibility of the investigator.

Electronically controlled balances usually include a calibration routine documented in the operation manual. Where they differ, the routine set forth in the relevant ASTM reference (E1270, E319, or E898) should be considered while the investigator identifies the control standard for the measurement process. Another consideration is the comparison of the operating range of the balance with the requirements of the measurement. An improvement in linearity and precision can be realized if the calibration routine is run over a range suitable for the measurement, rather than the entire operating range. However, the integral software of the electronics may not afford the flexibility to do such a limited function calibration. Also, the actual operation of an electronically controlled balance involves the use of a tare setting to offset such weights as that of the container used for a sample measurement. The tare offset necessitates an extended range that is frequently significantly larger than the range of weights to be measured.

MASS MEASUREMENT PROCESS ASSURANCE

An instrument is characterized by its capability to reproducibly deliver a result with a given readability. We often refer to a calibrated instrument; however, in reality there is no such thing as a calibrated balance *per se*. There are weight standards that are used to calibrate a weight measurement procedure, but that procedure can and should include everything from operator behavior patterns to systematic instrumental responses. In other words, it is the measurement process, not the balance itself that must be calibrated. The basic maxims for evaluating and calibrating a mass measurement process are easily translated to any quantitative measurement in the laboratory to which some standards should be attached for purposes of reporting, quality assurance, and reproducibility. While certain industrial, government, and even university programs have established measurement assurance programs [e.g., as required for International Standards Organization (ISO) 9000/9001 certification], the application of established standards to research laboratories is not always well defined. In general, it is the investigator who bears the responsibility for applying standards when making measurements and reporting results. Where no quality assurance programs are mandated, some laboratories may wish to institute a voluntary accreditation program. NIST operates the National Voluntary Laboratory Accreditation Program [NVLAP, telephone number (301) 975-4042] to assist such laboratories in achieving self-imposed accreditation (Harris, 1993).

The ISO Guide on the Expression of Uncertainty in Measurements (1992) identified and recommended a standardized approach for expressing the uncertainty of results. The standard was adopted by NIST and is published in NIST Technical Note 1297 (1994). The NIST publication simplifies the technical aspects of the standard. It establishes two categories of uncertainty, type A and type B. Type A uncertainty contains factors associated with random variability, and is identified solely by statistical analysis of measurement data. Type B uncertainty consists of all other sources of variability; scientific judgment alone quantifies this uncertainty type (Clark, 1994). The process uncertainty under the ISO recommendation is defined as the square root of the sum of the squares of the standard deviations due to all contributing factors. At a minimum, the process variation uncertainty should consist of the standard deviations of the mass standards used (s), the standard deviations of the measurement process (s_P), and the estimated standard deviations due to Type B uncertainty (u_B). Then, the overall process uncertainty (combined standard uncertainty) is $u_c = [(s)^2 + (s_P)^2 + (u_B)^2]^{1/2}$ (Everhart, 1995). This combined uncertainty value is multiplied by the coverage factor (usually 2) to report the expanded uncertainty (U) to the 95% (2-sigma) confidence level. NIST adopted the 2-sigma level for stating uncertainties in January 1994; uncertainty statements from NIST prior to this date were based on the 3-sigma (99%) confidence level. More detailed guidelines for the computation and expression of uncertainty, including a discussion of scatter analysis and error propagation, is provided in NIST Technical Note 1297 (1994). This document has been adopted by the CIPM, and it is available online (see Internet Resources).

J. Everhart (JTI Systems, Inc.) has proposed a process measurement assurance program that affords a powerful, systematic tool for accumulating meaningful data and insight on the uncertainties associated with a measurement procedure, and further helps to improve measurement procedures and data quality by integrating a calibration program with day-to-day measurements (Everhart, 1988). An added advantage of adopting such an approach is that procedural errors, instrument drift or malfunction, or other quality-reducing factors are more likely to be caught quickly.

The essential points of Everhart's program are summarized here.

1. Initial measurements are made by metrology specialists or experts in the measurement using a control standard to establish reference confidence limits.

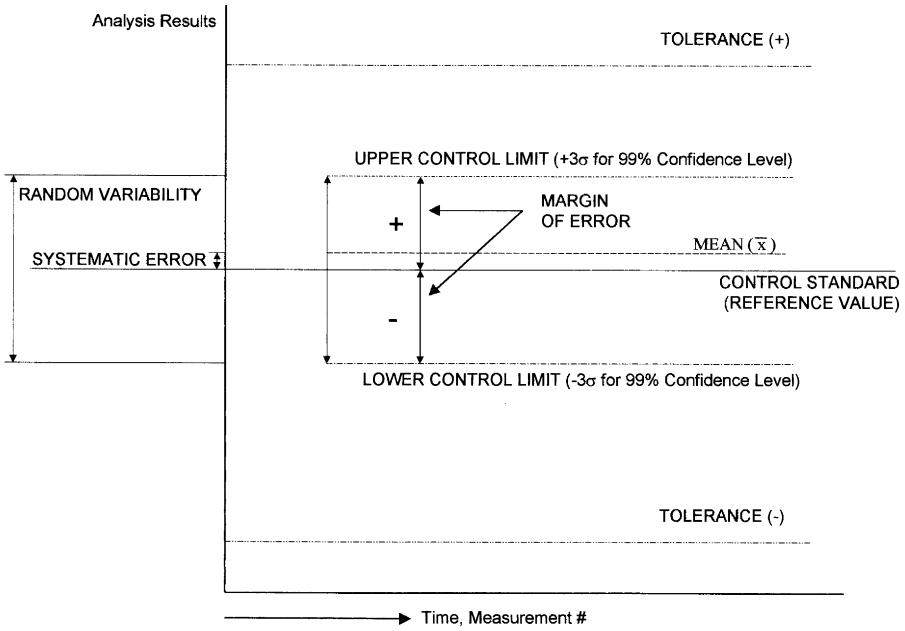


Figure 2. The process measurement assurance program control chart, identifying contributions to the errors associated with any measurement process. (After Everhart, 1988.)

2. Technicians or operators measure the control standard prior to each significant event as determined by the principal investigator (say, an experiment or even a workday).
3. Technicians or operators measure the control standard again after each significant event.
4. The data are recorded and the control measurements checked against the reference confidence limits.

The errors are analyzed and categorized as systematic (bias), random variability, and overall measurement system error. The results are plotted over time to yield a chart like that shown in Figure 2. It is clear how adopting such a discipline and monitoring the charted standard measurement data will quickly identify problems.

Essential practices to institute with any measurement assurance program are to apply an external check on the program (e.g., round robin), to have weight standards recalibrated periodically while surveillance programs are in place, and to maintain a separate calibrated weight standard, which is not used as frequently as the working standards (Harris, 1996). These practices will ensure both accuracy and traceability in the measurement process.

The knowledge of error in measurements and uncertainty estimates can immediately improve a quantitative measurement process. By establishing and implementing standards, higher-quality data and greater confidence in the measurements result. Standards established in industrial or federal settings should be applied in a research environment to improve data quality. The measurement of mass is central to the analysis of materials properties, so the importance of establishing and reporting uncertainties and confidence limits along with the measured results cannot be overstated. Accurate record keeping and data analysis can help investigators identify and correct such problems as bias, operator error, and instrument malfunctions before they do any significant harm.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Georgia Harris of the NIST Office of Weights and Measures for providing information, resources, guidance, and direction in the preparation of this unit and for reviewing the completed manuscript for accuracy. We also wish to thank Dr. John Clark for providing extensive resources that were extremely valuable in preparing the unit.

LITERATURE CITED

ASTM. 1985. Standard Practice for the Evaluation of Single-Pan Mechanical Balances, Standard E319 (reapproved, 1993). American Society for Testing and Materials, West Conshohocken, Pa.

ASTM. 1988a. Standard Test Method for Equal-Arm Balances, Standard E1270 (reapproved, 1993). American Society for Testing Materials, West Conshohocken, Pa.

ASTM. 1988b. Standard Method of Testing Top-Loading, Direct-Reading Laboratory Scales and Balances, Standard E898 (reapproved 1993). American Society for Testing Materials, West Conshohocken, Pa.

ASTM. 1997. Standard Specification for Laboratory Weights and Precision Mass Standards, Standard E617 (originally published, 1978). American Society for Testing and Materials, West Conshohocken, Pa.

ASTM. 1999. Standard Practices for Force Verification of Testing Machines, Standard E4-99. American Society for Testing and Materials, West Conshohocken, Pa.

Baltes, H., Gopel, W., and Hesse, J. (eds.) 1998. Sensors Update, Vol. 4. Wiley-VCH, Weinheim, Germany.

Clark, J. P. 1994. Identifying and managing mass measurement errors. *In* Proceedings of the Weighing, Calibration, and Quality Standards Conference in the 1990s, Sheffield, England, 1994.

Everhart, J. 1988. Process Measurement Assurance Program. JTI Systems, Albuquerque, N. M.

Everhart, J. 1995. Determining mass measurement uncertainty. *Cal. Lab.* May/June 1995.

Feder, R. and Nowick, A. S. 1957. Use of Thermal Expansion Measurements to Detect Lattice Vacancies Near the Melting Point of Pure Lead and Aluminum. *Phys. Rev.*109(6): 1959–1963.

Harris, G. L. 1993. Ensuring accuracy and traceability of weighing instruments. *ASTM Standardization News*, April, 1993.

Harris, G. L. 1996. Answers to commonly asked questions about mass standards. *Cal. Lab.* Nov./Dec. 1996.

Kupper, W. E. 1990. Honest weight—limits of accuracy and practicality. *In* Proceedings of the 1990 Measurement Conference, Anaheim, Calif.

Kupper, W. E. 1997. Laboratory balances. *In* Analytical Instrumentation Handbook, 2nd ed. (G.E. Ewing, ed.). Marcel Dekker, New York.

Kupper, W. E. 1999. Verification of high-accuracy weighing equipment. *In* Proceedings of the 1999 Measurement Science Conference, Anaheim, Calif.

Lide, D. R. 1999. Chemical Rubber Company Handbook of Chemistry and Physics, 80th Edition, CRC Press, Boca Raton, Flor.

NIST. 1999. Specifications, Tolerances, and Other Technical Requirements for Weighing and Measuring Devices, NIST Handbook 44. U. S. Department of Commerce, Gaithersburg, Md.

NIST. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, NIST Technical Note 1297. U. S. Department of Commerce, Gaithersburg, Md.

OIML. 1994. Weights of Classes E_1, E_2, F_1, F_2, M_1, M_2, M_3: Recommendation R111. Edition 1994(E). Bureau International de Metrologie Legale, Paris.

Simmons, R. O. and Barluffi, R. W. 1959. Measurements of Equilibrium Vacancy Concentrations in Aluminum. *Phys. Rev.* 117(1): 52–61.

Simmons, R. O. and Barluffi, R. W. 1961. Measurement of Equilibrium Concentrations of Lattice Vacancies in Gold. *Phys. Rev.* 125(3): 862–872.

Simmons, R. O. and Barluffi, R. W. 1962. Measurement of Equilibrium Concentrations of Vacancies in Copper. *Phys. Rev.* 129(4): 1533–1544.

KEY REFERENCES

ASTM, 1985, 1988a, 1988b (as appropriate for the type of balance used). See above.

These documents delineate the recommended procedure for the actual calibration of balances used in the laboratory.

OIML, 1994. See above.

This document is basis for the establishment of an international standard for metrological control. A draft document designated TC 9/SC 3/N 1 is currently under review for consideration as an international standard for testing weight standards.

Everhart, 1988. See above.

The comprehensive yet easy-to-implement program described in this reference is a valuable suggestion for the implementation of a quality assurance program for everything from laboratory research to industrial production.

INTERNET RESOURCES

http://www.oiml.org

OIML Home Page. General information on the International Organization for Legal Metrology.

http://www.usp.org

United States Pharmacopea Home Page. General information about the program used my many disciplines to establish standards.

http://www.nist.gov/owm

Office of Weights and Measures Home Page. Information on the National Conference on Weights and Measures and laboratory metrology.

http://www.nist.gov/metric

NIST Metric Program Home Page. General information on the metric program including on-line publications.

http://physics.nist.gov/Pubs/guidelines/contents.html

NIST Technical Note 1297: Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results.

http://www.astm.org

American Society for Testing and Materials Home Page. Information on ASTM committees and standards and ASTM publication ordering services.

http://iso.ch

International Standards Organization (ISO) Home Page. Information and calendar on the ISO committee and certification programs.

http://www.ansi.org

American National Standards Institute Home Page. Information on ANSI programs, standards, and committees.

http://www.quality.org/

Quality Resources On-line. Resource for quality-related information and groups.

http://www.fasor.com/~iso25

ISO Guide 25. International list of accreditation bodies, standards organizations, and measurement and testing laboratories.

DAVID DOLLIMORE
The University of Toledo
Toledo, Ohio

ALAN C. SAMUELS
Edgewood Chemical and Biological Center
Aberdeen Proving Ground
Maryland

THERMOMETRY

DEFINITION OF THERMOMETRY AND TEMPERATURE: THE CONCEPT OF TEMPERATURE

Thermometry is the science of measuring temperature, and thermometers are the instruments used to measure temperature. Temperature must be regarded as the scientific measure of "hotness" or "coldness." This unit is concerned with the measurement of temperatures in materials of interest to materials science, and the notion of temperature is thus limited in this discussion to that which applies to materials in the solid, liquid, or gas state (as opposed to the so-called temperature associated with ion

gases and plasmas, which is no longer limited to a measure of the internal kinetic energy of the constituent atoms).

A brief excursion into the history of temperature measurement will reveal that measurement of temperature actually preceded the modern definition of temperature and a temperature scale. Galileo in 1594 is usually credited with the invention of a thermometer in the form that indicated the expansion of air as the environment became hotter (Middleton, 1966). This instrument was called a thermoscope, and consisted of air trapped in a bulb by a column of liquid (Galileo used water) in a long tube attached to the bulb. It can properly be called an air thermometer when a scale is added to measure the expansion, and such an instrument was described by Telioux in 1611. Variation in atmospheric pressure would cause the thermoscope to develop different readings, as the liquid was not sealed into the tube and one surface of the liquid was open to the atmosphere.

The simple expedient of sealing the instrument so that the liquid and gas were contained in the tube really marks the invention of a glass thermometer. By making the diameter of the tube small, so that the volume of the gas was considerably reduced, the liquid dilation in these sealed instruments could be used to indicate the temperature. Such a thermometer was used by Ferdinand II, Grand Duke of Tuscany, about 1654. Fahrenheit eventually substituted mercury for the "spirits of wine" earlier used as the working liquid fluid, because mercury's thermal expansion with temperature is more nearly linear. Temperature scales were then invented using two selected fixed points—usually the ice point and the blood point or the ice point and the boiling point.

THE THERMODYNAMIC TEMPERATURE SCALE

The starting point for the thermodynamic treatment of temperature is to state that it is a property that determines in which direction energy will flow when it is in contact with another object. Heat flows from a higher-temperature object to a lower-temperature object. When two objects have the same temperature, there is no flow of heat between them and the objects are said to be in thermal equilibrium. This forms the basis of the Zeroth Law of thermodynamics. The First Law of thermodynamics stipulates that energy must be conserved during any process. The Second Law introduces the concepts of spontaneity and reversibility—for example, heat flows spontaneously from a higher-temperature system to a lower-temperature one. By considering the direction in which processes occur, the Second Law implicitly demands the passage of time, leading to the definition of entropy. Entropy, S, is defined as the thermodynamic state function of a system where $dS \geq \delta q/T$, where q is the heat and T is the temperature. When the equality holds, the process is said to be reversible, whereas the inequality holds in all known processes (i.e., all known processes occur irreversibly). It should be pointed out that dS (and hence the ratio $\delta q/T$ in a reversible process) is an exact differential, whereas δq is not. The flow of heat in an irreversible process is path dependent. The Third Law defines the absolute zero point of the thermodynamic temperature scale, and further stipulates that no process can reduce the temperature of a macroscopic system to this point. The laws of thermodynamics are defined in detail in all introductory texts on the topic of thermodynamics (e.g., Rock, 1983).

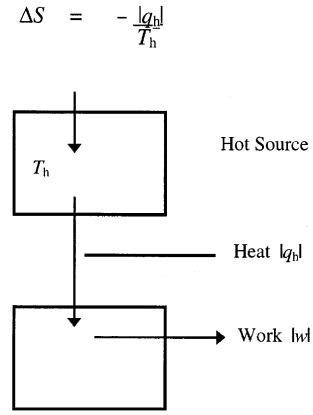


Figure 1. Change in entropy when heat is completely converted into work.

A consideration of the efficiency of heat engines leads to a definition of the thermodynamic temperature scale. A heat engine is a device that converts heat into work. Such a process of producing work in a heat engine must be spontaneous. It is necessary then that the flow of energy from the hot source to the cold sink be accompanied by an overall increase in entropy. Thus in such a hypothetical engine, heat, $|q|$, is extracted from a hot sink of temperature, T_h, and converted completely to work. This is depicted in Figure 1. The change in entropy, ΔS, is then:

$$\Delta S = \frac{-|q|}{T_h} \tag{1}$$

This value of ΔS is negative, and the process is nonspontaneous.

With the addition of a cold sink (see Fig. 2), the removal of the $|q_h|$ from the hot sink changes its entropy by

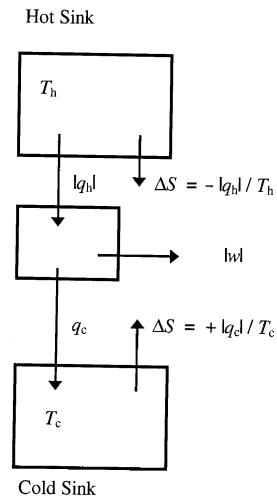


Figure 2. Change in entropy when some heat from the hot sink is converted into work and some into a cold sink.

$-|q_h|/T_h$ and the transfer of $|q_c|$ to the cold sink increases its entropy by $|q_c|/T_c$. The overall entropy change is:

$$\Delta S = \frac{-|q_h|}{T_h} + \frac{|q_c|}{T_c} \tag{2}$$

ΔS is then greater than zero if:

$$|q_c| \geq (T_c/T_h) \cdot |q_h| \tag{3}$$

and the process is spontaneous.

The maximum work of the engine can then be given by:

$$\begin{aligned} |w_{max}| &= |q_h| - |q_{c_{min}}| \\ &= |q_h| - (T_c/T_h) \cdot |q_h| \\ &= [1 - (T_c/T_h)] \cdot |q_h| \end{aligned} \tag{4}$$

The maximum possible efficiency of the engine is:

$$\varepsilon_{rev} = \frac{|w_{max}|}{|q_h|} \tag{5}$$

when by the previous relationship $\varepsilon_{rev} = 1 - (T_c/T_h)$.

If the engine is working reversibly, $\Delta S = 0$ and

$$\frac{|q_c|}{|q_h|} = \frac{T_c}{T_h} \tag{6}$$

Kelvin used this to define the thermodynamic temperature scale using the ratio of the heat withdrawn from the hot sink and the heat supplied to the cold sink. The zero in the thermodynamic temperature scale is the value of T_c at which the Carnot efficiency equals 1, and work output equals the heat supplied (see, e.g., Rock, 1983). Then for $\varepsilon_{rev} = 1$, $T = 0$.

If now a fixed point, such as the triple point of water, is chosen for convenience, and this temperature T_3 is set as 273.16 K to make the Kelvin equivalent to the currently used Celsius degree, then:

$$T_c = (|q_c|/|q_h|) \cdot T_3 \tag{7}$$

The importance of this is that the temperature is defined independent of the working substance.

The perfect-gas temperature scale is independent of the identity of the gas and is identical to the thermodynamic temperature scale. This result is due to the observation of Charles that for a sample of gas subjected to a constant low pressure, the volume, V, varied linearly with temperature whatever the identity of the gas (see, e.g., Rock, 1983; McGee, 1988).

Thus $V =$ constant $(\theta + 273.15^\circ C)$ at constant pressure, where θ denotes the temperature on the Celsius scale and $T\ (\equiv \theta + 273.15)$ is the temperature on the Kelvin, or absolute scale. The volume will approach zero on cooling at $T = -273.15^\circ C$. This is termed the absolute zero. It should be noted that it is not possible to cool real gases to zero volume because they condense to a liquid or a solid before absolute zero is reached.

DEVELOPMENT OF THE INTERNATIONAL TEMPERATURE SCALE OF 1990

In 1927, the International Conference of Weights and Measures approved the establishment of an International Temperature Scale (ITS, 1927). In 1960, the name was changed to the International Practical Temperature Scale (IPTS). Revisions of the scale took place in 1948, 1954, 1960, 1968, and 1990. Six international conferences under the title "Temperature, Its Measurement and Control in Science and Industry" were held in 1941, 1955, 1962, 1972, 1982, and 1992 (Wolfe, 1941; Herzfeld, 1955, 1962; Plumb, 1972; Billing and Quinn, 1975; Schooley, 1982, 1992). The latest revision in 1990 again changed the name of the scale to the International Temperature Scale 1990 (ITS-90).

The necessary features required by a temperature scale are (Hudson, 1982):

1. Definition;
2. Realization;
3. Transfer; and
4. Utilization.

The ITS-90 temperature scale is the best available approximation to the thermodynamic temperature scale. The following points deal with the definition and realization of the IPTS and ITS scale as they were developed over the years.

1. Fixed points are used based on thermodynamic invariant points. These are such points as freezing points and triple points of specific systems. They are classified as (a) defining (primary) and (b) secondary, depending on the system of measurement and/or the precision to which they are known.
2. The instruments used in interpolating temperatures between the fixed points are specified.
3. The equations used to calculate the intermediate temperature between each defining temperature must agree. Such equations should pass through the defining fixed points.

In given applications, use of the agreed IPTS instruments is often not practicable. It is then necessary to transfer the measurement from the IPTS and ITS method to another, more practical temperature-measuring instrument. In such a transfer, accuracy is necessarily reduced, but such transfers are required to allow utilization of the scale.

The IPTS-27 Scale

The IPTS scale as originally set out defined four temperature ranges and specified three instruments for measuring temperatures and provided an interpolating equation for each range. It should be noted that this is now called the IPTS-27 scale even though the 1927 version was called ITS, and two revisions took place before the name was changed to IPTS.

Range I: The oxygen point to the ice point (−182.97 to 0°C);

Range II: The ice point to the aluminum point (0 to 660°C);

Range III: The aluminum point to the gold point (660 to 1063.0°C);

Range IV: Above the gold point (1063.0°C and higher).

The oxygen point is defined as the temperature at which liquid and gaseous oxygen are in equilibrium, and the ice and metal points are defined as the temperature at which the solid and liquid phases of the material are in equilibrium. The platinum resistance thermometer was used for ranges I and II, the platinum versus platinum (90%)/rhodium (10%) thermocouple was used for range III, and the optical pyrometer was used for range IV. In subsequent IPTS scales (and ITS-90), these instruments are still used, but the interpolating equations and limits are modified. The IPTS-27 was based on the ice point and the steam point as true temperatures on the thermodynamic scale.

Subsequent Revision of the Temperature Scales Prior to that of 1990

In 1954, the triple point of water and the absolute zero were used as the two points defining the thermodynamic scale. This followed a proposal originally advocated by Kelvin. In 1975, the Kelvin was adopted as the standard temperature unit. The symbol T represents the thermodynamic temperature with the unit Kelvin. The Kelvin is given the symbol K, and is defined by setting the melting point of water equal to 273.15 K. In practice, for historical reasons, the relationship between T and the Celsius temperature (t) is defined as $t = T - 273.15$ K (the fixed points on the Celsius scale are water's melting point and boiling point). By definition, the degree Celsius (°C) is equal in magnitude to the Kelvin.

The IPTS-68 was amended in 1975 and a set of defining fixed points (involving hydrogen, neon, argon, oxygen, water, tin, zinc, silver, and gold) were listed.

The IPTS-68 had four ranges:

Range I: 13.8 to 273.15 K, measured using a platinum resistance thermometer. This was divided into four parts.

Part A: 13.81 to 17.042 K, determined using the triple point of equilibrium hydrogen and the boiling point of equilibrium hydrogen.

Part B: 17.042 to 54.361 K, determined using the boiling point of equilibrium hydrogen, the boiling point of neon, and the triple point of oxygen.

Part C: 54.361 to 90.188 K, determined using the triple point of oxygen and the boiling point of oxygen.

Part D: 90.188 to 273.15 K, determined using the boiling point of oxygen and the boiling point of water.

Range II: 273.15 to 903.89 K, measured using a platinum resistance thermometer, using the triple point of water, the boiling point of water, the freezing point of tin, and the freezing point of zinc.

Range III: 903.89 to 1337.58 K, measured using a platinum versus platinum (90%)/rhodium (10%) thermocouple, using the antimony point and the freezing points of silver and gold, with cross-reference to the platinum resistance thermometer at the antimony point.

Range IV: All temperatures above 1337.58 K. This is the gold point (at 1064.43°C), but no particular thermal radiation instrument is specified.

It should be noted that temperatures in range IV are defined by fundamental relationships, whereas the other three IPTS defining equations are not fundamental. It must also be stated that the IPTS-68 scale is not defined below the triple point of hydrogen (13.81 K).

The International Temperature Scale of 1990

The latest scale is the International Temperature Scale of 1990 (ITS-90). Figure 3 shows the various temperature ranges set out in ITS-90. T_{90} (meaning the temperature defined according to ITS-90) is stipulated from 0.65 K to the highest temperature using various fixed points and the helium vapor-pressure relations. In Figure 3, these "fixed points" are set out in the diagram to the nearest integer. A review has been provided by Swenson (1992). In ITS-90 an overlap exists between the major ranges for three of the four interpolation instruments, and in addition there are eight overlapping subranges. This represents a change from the IPTS-68. There are alternative definitions of the scale existing for different temperature ranges and types of interpolation instruments. Aspects of the temperature ranges and the interpolating instruments can now be discussed.

Helium Vapor Pressure (0.65 to 5 K). The helium isotopes ^{3}He and ^{4}He have normal boiling points of 3.2 K and 4.2 K respectively, and both remain liquids to $T = 0$. The helium vapor pressure–temperature relationship provides a convenient thermometer in this range.

Interpolating Gas Thermometer (3 to 24.5561 K). An interpolating constant-volume gas thermometer (1 CVGT) using ^{4}He as the gas is suggested with calibrations at three fixed points (the triple point of neon, 24.6 K; the triple point of equilibrium hydrogen, 13.8 K; and the normal boiling point of ^{4}He, 4.2 K).

Platinum Resistance Thermometer (13.8033 K to 961.78°C). It should be noted that temperatures above 0°C are typically recorded in °C and not on the absolute scale. Figure 3 indicates the fixed points and the range over which the platinum resistance thermometer can be used. Swenson (1992) gives some details regarding common practice in using this thermometer. The physical requirement for platinum thermometers that are used at high temperatures and at low temperatures are different, and no single thermometer can be used over the entire range.

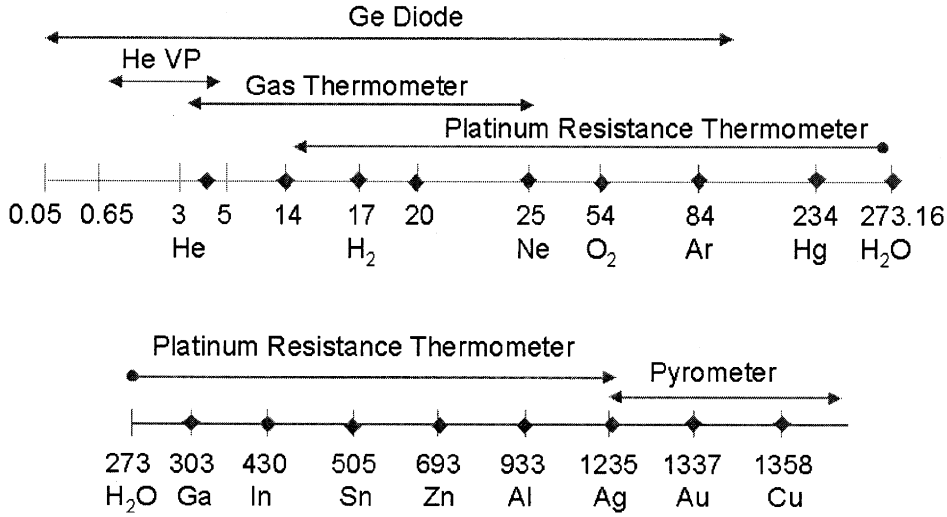


Figure 3. The International Temperature Scale of 1990 with some key temperatures noted. Ge diode range is shown for reference only; it is not defined in the ITS-90 standard. (See text for further details and explanation.)

Optical Pyrometry (above 961.78°C). In this temperature range the silver, gold, or copper freezing point can be used as reference temperatures. The silver point at 961.78°C is at the upper end of the platinum resistance scale, because have thermometers have stability problems (due to such effects as phase changes, changes in heat capacity, and degradation of welds and joints between constituent materials) above this temperature.

TEMPERATURE FIXED POINTS AND DESIRED CHARACTERISTICS OF TEMPERATURE MEASUREMENT PROBES

The Fixed Points

The ITS-90 scale utilized various "defining fixed points." These are given below in order of increasing temperature. (*Note:* all substances except ^{3}He are defined to be of natural isotopic composition.)

- The vapor point of ^{3}He between 3 and 5 K;
- The triple point of equilibrium hydrogen at 13.8033 K (equilibrium hydrogen is defined as the equilibrium concentrations of the *ortho* and *para* forms of the substance);
- An intermediate equilibrium hydrogen vapor point at ≈17 K;
- The normal boiling point of equilibrium hydrogen at ≈20.3 K;
- The triple point of neon at 24.5561 K;
- The triple point of oxygen at 54.3584 K;
- The triple point of argon at 83.8058 K;
- The triple point of mercury at 234.3156 K;
- The triple point of water at 273.16 K (0.01°C);
- The melting point of gallium at 302.9146 K (29.7646°C);
- The freezing point of indium at 429.7485 K (156.5985°C);
- The freezing point of tin at 505.078 K (231.928°C);
- The freezing point of zinc at 692.677 K (419.527°C);
- The freezing point of aluminum at 933.473 K (660.323°C);
- The freezing point of silver at 1234.93 K (961.78°C);
- The freezing point of gold at 1337.33 K (1064.18°C);
- The freezing point of copper at 1357.77 K (1084.62°C).

The reproducibility of the measurement (and/or known difficulties with the occurrence of systematic errors in measurement) dictates the property to be measured in each case on the scale. There remains, however, the question of choosing the temperature measurement probe—in other words, the best thermometer to be used. Each thermometer consists of a temperature sensing device, an interpretation and display device, and a method of connecting one to another.

The Sensing Element of a Thermometer

The desired characteristics for a sensing element always include:

1. *An Unambiguous (Monotonic) Response with Temperature*. This type of response is shown in Figure 4A. Note that the appropriate response need not be linear with temperature. Figure 4B shows an ambiguous response of the sensing property to temperature where there may be more than one temperature at which a particular value of the sensing property occurs.
2. *Sensitivity*. There must be a high sensitivity (δ) of the temperature-sensing property to temperature. Sensitivity is defined as the first derivative of the property (X) with respect to temperature: $\delta = \partial X/\partial T$.
3. *Stability*. It is necessary for the sensing element to remain stable, with the same sensitivity over a long time.

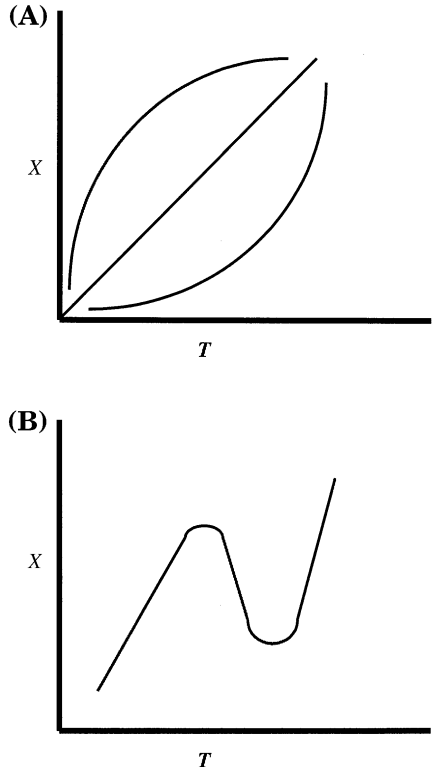


Figure 4. (**A**) An acceptable unambiguous response of temperature-sensing element (*X*) versus temperature (*T*). (**B**) An unacceptable unambiguous response of temperature-sensing element (*X*) versus temperature.

4. *Cost*. A relatively low cost (with respect to the project budget) is desirable.
5. *Range*. A wide range of temperature measurements makes for easy instrumentation.
6. *Size*. The element should be small (i.e., with respect to the sample size, to minimize heat transfer between the sample and the sensor).
7. *Heat Capacity*. A relatively small heat capacity is desirable—i.e., the amount of heat required to change the temperature of the sensor must not be too large.
8. *Response*. A rapid response is required (this is achieved in part by minimizing sensor size and heat capacity).
9. *Usable Output*. A usable output signal is required over the temperature range to be measured (i.e., one should maximize the dynamic range of the signal for optimal temperature resolution).

Naturally, all items are defined relative to the components of the system under study or the nature of the experiment being performed. It is apparent that in any single instrument, compromises must be made.

Readout Interpretation

Resolution of the temperature cannot be improved beyond that of the sensing device. Desirable features on the signal-receiving side include:

1. Sufficient sensitivity for the sensing element;
2. Stability with respect to time;
3. Automatic response to the signal;
4. Possession of data logging and archiving capabilities;
5. Low cost.

TYPES OF THERMOMETER

Liquid-Filled Thermometers

The discussion is limited to practical considerations. It should be noted, however, that the most common type of thermometer in this class is the mercury-filled glass thermometer.

Overall there are two types of liquid-filled systems:

1. Systems filled with a liquid other than mercury
2. Mercury-filled systems.

Both rely on the temperature being indicated by a change in volume. The lower range is dictated by the freezing point of the fill liquid; the upper range must be below the point at which the liquid is unstable, or where the expansion takes place in an unacceptably nonlinear fashion. The nature of the construction material is important. In many instances, the container is a glass vessel with expansion read directly from a scale. In other cases, it is a metal or ceramic holder attached to a capillary, which drives a Bourdon tube (a diaphragm or bellows device). In general organic liquids have a coefficient of expansion some 8 times that of mercury, and temperature spans for accurate work of some 10 to 25°C are dictated by the size of the container bulb for the liquid. Organic fluids are available up to 250°C, while mercury-filled systems can operate up to 650°C. Liquid-filled systems are unaffected by barometric pressure changes. Certificates of performance can generally be obtained from the instrument manufacturers.

Gas-Filled Thermometers

In vapor-pressure thermometers, the container is partially filled with a volatile liquid. Temperature at the bulb is conditioned by the fact that the interface between the liquid and the gas must be located at the point of measurement, and the container must represent the coolest point of the system. Notwithstanding the previous considerations applying to temperature scale, in practice vapor-filled systems can be operated from -40 to ~ 320°C.

A further class of gas-filled systems is one that simply relies on the expansion of a gas. Such systems are based on Charles' Law:

$$P = \frac{KT}{V} \tag{8}$$

where P is the pressure, T is the temperature (Kelvin), and V is the volume. K is a constant. The range of practical application is the widest of any filled systems. The lowest temperature is that at which the gas becomes liquid. The

highest temperature depends on the thermal stability of the gas or the construction material.

Electrical-Resistance Thermometers

A resistance thermometer is dependent upon the electrical resistance of a conducting metal changing with the temperature. In order to minimize the size of the equipment, the resistance of the wire or film should be relatively high so that the resistance can easily be measured. The change in resistance with temperature should also be large. The most common material used is a platinum wire-wound element with a resistance of $\sim$100 Ω at 0°C. Calibration standards are essential (see the previous discussion on the International Temperature Standards). Commercial manufacturers will provide such instruments together with calibration details. Thin-film platinum elements may provide an alternative design feature and are priced competitively with the more common wire-wound elements. Nickel and copper resistance elements are also commercially available.

Thermocouple Thermometers

A thermocouple is an assembly of two wires of unlike metals joined at one end where the temperature is to be measured. If the other end of one of the thermocouple wires leads to a second, similar junction that is kept at a constant reference temperature, then a temperature-dependent voltage develops called the Seebeck voltage (see, e.g., McGee, 1988). The constant-temperature junction is often kept at 0°C and is referred to as the cold junction. Tables are available of the EMF generated versus temperature when one thermocouple is kept as a cold junction at 0°C for specified metal/metal thermocouple junctions. These scales may show a nonlinear variation with temperature, so it is essential that calibration be carried out using phase transitions (i.e., melting points or solid-solid transitions). Typical thermocouple systems are summarized in Table 1.

Table 1. Some Typical Thermocouple Systems[a]

System	Use
Iron-Constantan	Used in dry reducing atmospheres up to 400°C General temperature scale: 0° to 750°C
Copper-Constantan	Used in slightly oxidizing or reducing atmospheres For low-temperature work: −200° to 50°C
Chromel-Alumel	Used only in oxidizing atmospheres Temperature range: 0° to 1250°C
Chromel-Constantan	Not to be used in atmospheres that are strongly reducing atmospheres or contain sulfur compounds Temperature range: −200° to 900°C
Platinum-Rhodium (or suitable alloys)	Can be used as components for thermocouples operating up to 1700°C

[a]Operating conditions and calibration details should always be sought from instrument manufactures.

Thermistors and Semiconductor-Based Thermometers

There are various types of semiconductor-based thermometers. Some semiconductors used for temperature measurements are called thermistors or resistive temperature detectors (RTDs). Materials can be classified as electrical conductors, semiconductors, or insulators depending on their electrical conductivity. Semiconductors have $\sim$10 to 10^6 Ω-cm resistivity. The resistivity changes with temperature, and the logarithm of the resistance plotted against reciprocal of the absolute temperature is often linear. The actual value for the thermistor can be fixed by deliberately introducing impurities. Typical materials used are oxides of nickel, manganese, copper, titanium, and other metals that are sintered at high temperatures. Most thermistors have a negative temperature coefficient, but some are available with a positive temperature coefficient. A typical thermistor with a resistance of 1200 Ω at 40°C will have a 120-Ω resistance at 110°C. This represents a decrease in resistance by a factor of about 2 for every 20°C increase in temperature, which makes it very useful for measuring very small temperature spans. Thermistors are available in a wide variety of styles, such as small beads, discs, washers, or rods, and may be encased in glass or plastic or used bare as required by their intended application. Typical temperature ranges are from −30° to 120°C, with generally a much greater sensitivity than for thermocouples. The germanium diode is an important thermistor due to its responsivity range—germanium has a well-characterized response from 0.05° to 100° Kelvin, making it well-suited to extremely low-temperature measurement applications. Germanium diodes are also employed in emissivity measurements (see the next section) due to their extremely fast response time—nanosecond time resolution has been reported (Xu et al., 1996). Ruthenium oxide RTDs are also suitable for extremely low-temperature measurement and are found in many cryostat applications.

Radiation Thermometers

Radiation incident on matter must be reflected, transmitted, or absorbed to comply with the First Law of Thermodynamics. Thus the reflectance, r, the transmittance, t, and the absorbance, a, sum to unity (the reflectance [transmittance, absorbance] is defined as the ratio of the reflected [transmitted, absorbed] intensity to the incident intensity). This forms the basis of Kirchoff's law of optics. Kirchoff recognized that if an object were a perfect absorber, then in order to conserve energy, the object must also be a perfect emitter. Such a perfect absorber/emitter is called a "black body." Kirchoff further recognized that the absorbance of a black body must equal its emittance, and that a black body would be thus characterized by a certain brightness that depends upon its temperature (Wolfe, 1998). Max Planck identified the quantized nature of the black body's emittance as a function of frequency by treating the emitted radiation as though it were the result of a linear field of oscillators with quantized energy states. Planck's famous black-body law relates the radiant

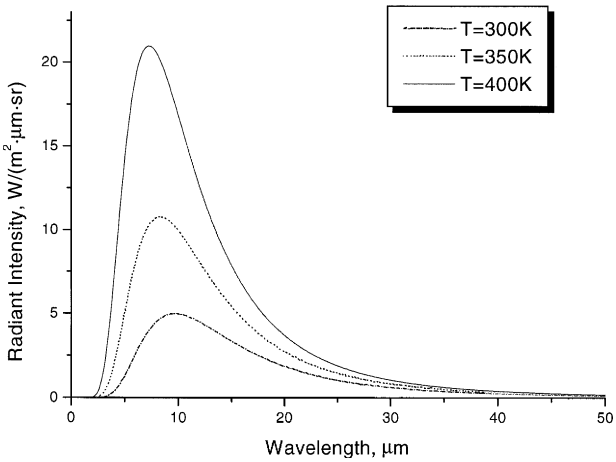


Figure 5. Spectral distribution of radiant intensity as a function of temperature.

intensity to the temperature as follows:

$$I_v = \frac{2hv^3n^2}{c^2}\frac{1}{\exp(hv/kT)-1} \tag{9}$$

where h is Planck's constant, v is the frequency, n is the refractive index of the medium into which the radiation is emitted, c is the velocity of light, k is Boltzmann's constant, and T is the absolute temperature. Planck's law is frequently expressed in terms of the wavelength of the radiation since in practice the wavelength is the measured quantity. Planck's law is then expressed as

$$I_\lambda = \frac{C_1}{\lambda^5(e^{C_2/\lambda T}-1)} \tag{10}$$

where C_1 $(= 2hc^2/n^2)$ and C_2 $(= hc/nk)$ are known as the first and second radiation constant, respectively. A plot of the Planck's law intensity at various temperatures (Fig. 5) demonstrates the principle by which radiation thermometers operate.

The radiative intensity of a black-body surface depends upon the viewing angle according to the Lambert cosine law, $I_\theta = I\cos\theta$. Using the projected area in a given direction given by $dA\cos\theta$, the radiant emission per unit of projected area, or radiance (L), for a black body is given by $L = I\cos\theta/\cos\theta = I$. For real objects, $L \neq I$ because factors such as surface shape, roughness, and composition affect the radiance. An important consequence of this is that emittance is a not an intrinsic materials property. The emissivity, emittance from a perfect material under ideal conditions (pure, atomically smooth and flat surface free of pores or oxide coatings), is a fundamental materials property defined as the ratio of the radiant flux density of the material to that of a black body under the same conditions (likewise, absorptivity and reflectivity are materials properties). Emissivity is extremely difficult to measure accurately, and emittance is often erroneously reported as emissivity in the literature (McGee, 1988). Both emittance and emissivity can be taken at a single wavelength (spectral emittance), over a range of wavelengths (partial emittance), or over all wavelengths (total emittance). It is important to properly determine the emittance of any real material being measured in order convert radiant intensity to temperature. For example, the spectral emittance of polished brass is 0.05, while that of oxidized brass is 0.61 (McGee, 1988). An important ramification of this effect is the fact that it is not generally possible to accurately measure the emissivity of polished, shiny surfaces, especially metallic ones, whose signal is dominated by reflectivity.

Radiation thermometers measure the amount of radiation (in a selected spectral band—see below) emitted by the object whose temperature is to be measured. Such radiation can be measured from a distance, so there is no need for contact between the thermometer and the object. Radiation thermometers are especially suited to the measurement of moving objects, or of objects inside vacuum or pressure vessels. The types of radiation thermometers commonly available are:

broadband thermometers
bandpass thermometers
narrow-band thermometers
ratio thermometers
optical pyrometers and
fiberoptic thermometers.

Broadband thermometers have a response from 0.3 μm optical wavelength to an upper limit of 2.5 to 20 μm, governed by the lens or window material. Bandpass thermometers have lenses or windows selected to view only nine selected portions of the spectrum. Narrow-band thermometers respond to an extremely narrow range of wavelengths. A ratio thermometer measures radiated energy in two narrow bands and calculates the ratio of intensities at the two energies. Optical pyrometers are really a special form of narrow-band thermometer, measuring radiation from a target in a narrow band of visible wavelengths centered at ~0.65 μm in the red portion of the spectrum. A fiberoptic thermometer uses a light guide to guide the radiation from the target to the detector. An important consideration is the so-called "atmospheric window" when making distance measurements. The 8- to 14-μm range is the most common region selected for optical pyrometric temperature measurement. The constituents of the atmosphere are relatively transparent in this region (that is, there are no infrared absorption bands from the most common atmospheric constituents, so no absorption or emission from the atmosphere is observed in this range).

TEMPERATURE CONTROL

It is not necessarily sufficient to measure temperature. In many fields it is also necessary to control the temperature. In an oven, a furnace, or a water bath, a constant temperature may be required. In other uses in analytical instrumentation, a more sophisticated temperature program may be required. This may take the form of a constant heating rate or may be much more complicated.

A temperature controller must:

1. receive a signal from which a temperature can be deduced;
2. compare it with the desired temperature; and
3. produce a means of correcting the actual temperature to move it towards the desired temperature.

The control action can take several forms. The simplest form is an on-off control. Power to a heater is turned on to reach a desired temperature but turned off when a certain temperature limit is reached. This cycling motion of control results in temperatures that oscillate between two set points once the desired temperature has been reached. A proportional control, in which the amount of temperature correction power depends on the magnitude of the "error" signal, provides a better system. This may be based on proportional bandwidth integral control or derivative control. The use of a dedicated computer allows observations to be set (and corrected) at desired intervals and corresponding real-time plots of temperature versus time to be obtained.

The Bureau International des Poids et Mesures (BIPM) ensures worldwide uniformity of measurements and their traceability to the Système Internationale (SI), and carries out measurement-related research. It is the proponent for the ITS-90 and the latest definitions can be found (in French and English) at *http://www.bipm.fr*.

The ITS-90 site, at *http://www.its-90.com*, also has some useful information. The text of the document is reproduced here with the permission of Metrologia (Springer-Verlag).

ACKNOWLEDGMENTS

The series editor gratefully acknowledges Christopher Meyer, of the Thermometry group of the National Institute of Standards and Technology (NIST), for discussions and clarifications concerning the ITS-90 temperature standards.

LITERATURE CITED

Billing, B. F. and Quinn, T. J. 1975. Temperature Measurement, Conference Series No. 26. Institute of Physics, London.

Herzfeld, C. M. (ed.) 1955. Temperature: Its Measurement and Control in Science and Industry, Vol. 2. Reinhold, New York.

Herzfeld, C. M. (ed.) 1962. Temperature: Its Measurement and Control in Science and Industry, Vol. 3. Reinhold, New York.

Hudson, R. P. 1982. *In* Temperature: Its Measurement and Control in Science and Industry, Vol. 5, Part 1 (J.F. Schooley, ed.). Reinhold, New York.

ITS. 1927. International Committee of Weights and Measures. *Conf. Gen. Poids. Mes*. 7:94.

McGee. 1988. Principles and Methods of Temperature Measurement. John Wiley & Sons, New York.

Middleton, W. E. K. 1966. A History of the Thermometer and Its Use in Meteorology. John Hopkins University Press, Baltimore.

Plumb, H. H. (ed.) 1972. Temperature: Its Measurement and Control in Science and Industry, Vol. 4. Instrument Society of America, Pittsburgh.

Rock, P. A. 1983. Chemical Thermodynamics. University Science Books, Mill Valley, Calif.

Schooley, J. F. (ed.) 1982. Temperature: Its Measurement and Control in Science and Industry, Vol. 5. American Institute of Physics, New York.

Schooley, J. F. (ed.) 1992. Temperature: Its Measurement and Control in Science and Industry, Vol. 6. American Institute of Physics, New York.

Swenson, C. A. 1992. *In* Temperature: Its Measurement and Control in Science and Industry, Vol. 6 (J.F. Schooley, ed.). American Institute of Physics, New York.

Wolfe, Q. C. (ed.) 1941. Temperature: Its Measurement and Control in Science and Industry, Vol. 1. Reinhold, New York.

Wolfe, W. L. 1998. Introduction to Radiometry. SPIE Optical Engineering Press, Bellingham, Wash.

Xu, X., Grigoropoulos, C. P., and Russo, R. E. 1996. Nanosecond–time resolution thermal emission measurement during pulsed-excimer. *Appl. Phys. A* 62:51–59.

APPENDIX: TEMPERATURE-MEASUREMENT RESOURCES

Several manufacturers offer a significant level of expertise in the practical aspects of temperature measurement, and can assist researchers in the selection of the most appropriate instrument for their specific task. A noteworthy source for thermometers, thermocouples, thermistors, and pyrometers is Omega Engineering (*www.omega.com*). Omega provides a detailed catalog of products interlaced with descriptive essays of the underlying principles and practical considerations. The Mikron Instrument Company (*www.mikron.com*) manufactures an extensive line of infrared temperature measurement and calibration black-body sources. Graesby is also an excellent resource for extended-area calibration black-body sources. Inframetrics (*www. inframetrics.com*) manufactures a line of highly configurable infrared radiometers. All temperature measurement manufacturers offer calibration services with NIST-traceable certification.

Germanium and ruthenium RTDs are available from most manufacturers specializing in cryostat applications. Representative companies include Quantum Technology (*www.quantum-technology.com*) and Scientific Instruments (*www.scientificinstruments.com*). An extremely useful source for instrument interfacing (for automation and digital data acquisition) is National Instruments (*www. natinst.com*)

DAVID DOLLIMORE
The University of Toledo
Toledo, Ohio

ALAN C. SAMUELS
Edgewood Chemical Biological Center
Aberdeen Proving Ground,
Maryland

SYMMETRY IN CRYSTALLOGRAPHY

INTRODUCTION

The study of crystals has fascinated humanity for centuries, with motivations ranging from scientific curiosity to the belief that they had magical powers. Early crystal science was devoted to descriptive efforts, limited to measuring interfacial angles and determining optical properties. Some investigators, such as Haüy, attempted to deduce the underlying atomic structure from the external morphology. These efforts were successful in determining the symmetry operations relating crystal faces and led to the theory of point groups, the assignment of all known crystals to only seven crystal systems, and extensive compilations of axial ratios and optical indices. Roentgen's discovery of x rays and Laue's subsequent discovery of the scattering of x rays by crystals revolutionized the study of crystallography: crystal structures—i.e., the relative location of atoms in space—could now be determined unequivocally. The benefits derived from this knowledge have enhanced fundamental science, technology, and medicine ever since and have directly contributed to the welfare of human society.

This chapter is designed to introduce those with limited knowledge of space groups to a topic that many find difficult.

SYMMETRY OPERATORS

A crystalline material contains a periodic array of atoms in three dimensions, in contrast to the random arrangement of atoms in an amorphous material such as glass. The periodic repetition of a motif along a given direction in space within a fixed length t parallel to that direction constitutes the most basic symmetry operation. The motif may be a single atom, a simple molecule, or even a large, complex molecule such as a polymer or a protein. The periodic repetition in space along three noncollinear, noncoplanar vectors describes a unit parallelepiped, the unit cell, with periodically repeated lengths **a**, **b**, and **c**, the metric unit cell parameters (Fig. 1). The atomic content of this unit cell is the fundamental building block of the crystal structure. The macroscopic crystalline material results from the periodic repetition of this unit cell. The atoms within a unit cell may be related by additional symmetry operators.

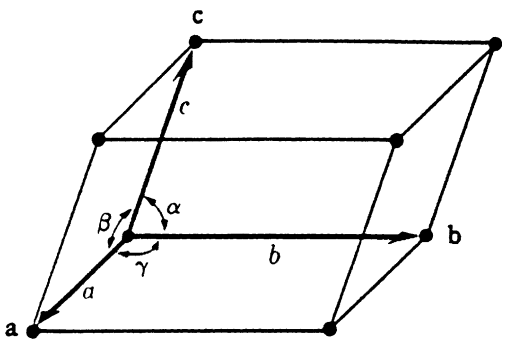


Figure 1. A unit cell.

Proper Rotation Axes

A proper rotation axis, n, repeats an object every $2\pi/n$ radians. Only 1-, 2-, 3-, 4-, and 6-fold axes are consistent with the periodic, space-filling repetition of the unit cell. In contrast, molecular symmetry axes can have any value of n. Figure 2A illustrates the appearance of space that results from the action of a proper rotation axis on a given motif. Note that a 1-fold axis—i.e., rotation by 360°—is a legitimate symmetry operation. These objects retain their handedness. The reversal of the direction of rotation will superimpose the objects without removing them from the plane perpendicular to the rotation axis. After 2π radians the rotated motif superimposes directly on the initial object. The repetition of motifs by a proper rotation axis forms congruent objects.

Improper Rotation Axes

Improper rotation axes are compound symmetry operations consisting of rotation followed by inversion or mirror reflection. Two conventions are used to designate symmetry operators. The International or Hermann-Mauguin symbols are based on rotoinversion operations, and the Schönflies notation is based on rotoreflection operations. The former is the standard in crystallography, while the latter is usually employed in molecular spectroscopy.

Consider the origin of a coordinate system, a b c, and an object located at coordinates $x\ y\ z$. Atomic coordinates are expressed as dimensionless fractions of the three-dimensional periodicities. From the origin draw a vector to every point on the object at $x\ y\ z$, extend this vector the same length through the origin in the opposite direction, and mark off this length. Thus, for every $x\ y\ z$ there will be a $-x$, $-y$, $-z$, ($\bar{x}$, $\bar{y}$, $\bar{z}$ in standard notation). This mapping creates a center of inversion or center of symmetry at the origin. The result of this operation changes the handedness of an object, and the two symmetry-related objects are enantiomorphs. Figure 3A illustrates this operator. It has the International or Hermann-Mauguin symbol $\bar{1}$ read as "one bar." This symbol can be interpreted as a 1-fold rotoinversion axis: i.e., an object is rotated 360° followed by the inversion operation. Similarly, there are $\bar{2}$, $\bar{3}$, $\bar{4}$, and $\bar{6}$ axes (Fig. 2B). Consider the $\bar{2}$ operation: a 2-fold axis perpendicular to the ac plane rotates an object 180° and immediately inverts it through the origin, defined as the point of intersection of the plane and the 2-fold axis. The two objects are related by a mirror and $\bar{2}$ is usually given the special symbol m. The object at $x\ y\ z$ is reproduced at $x\ \bar{y}\ z$ (Fig. 3B). The two objects created by inversion or mirror reflection cannot be superimposed by a proper rotation axis operation. They are related as the right hand is to the left. Such objects are known as enantiomorphs.

The Schönflies notation is based on the compound operation of rotation and reflection, and the operation is designated $\tilde{n}$ or S_n. The subscript n denotes the rotation $2\pi/n$ and S denotes the reflection operation (the German

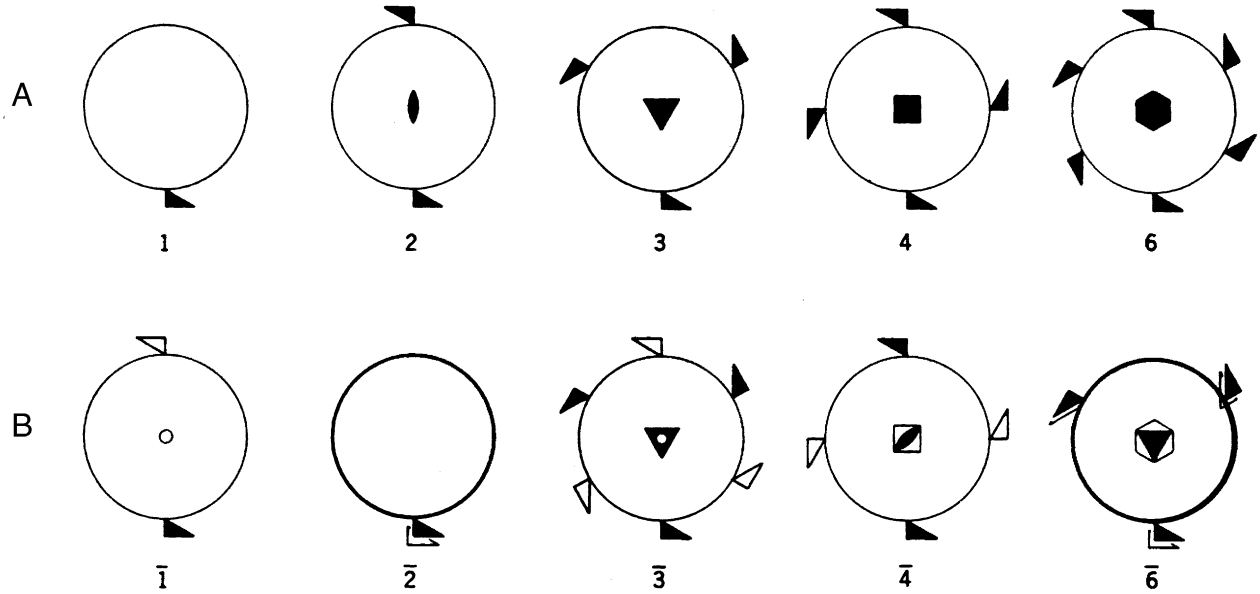


Figure 2. Symmetry of space around the five proper rotation axes giving rise to congruent objects (**A.**), and the five improper rotation axes giving rise to enantiomorphic objects (**B.**). Note the symbols in the center of the circles. The dark circles for $\bar{2}$ and $\bar{6}$ denote mirror planes in the plane of the paper. Filled triangles are above the plane of the paper and open ones below. Courtesy of Buerger (1970).

word for mirror is *spiegel*). Consider a $\tilde{2}$ axis perpendicular to the *ac* plane that contains an object above that plane at *x y z*. Rotate the object 180° and immediately reflect it through the plane. The positional parameters of this object are $\bar{x}, \bar{y}, \bar{z}$. The point of intersection of the 2-fold rotor and the plane is an inversion point and the two objects are enantiomorphs. The special symbol i is assigned to $\tilde{2}$ or S_2, and is equivalent to $\bar{1}$ in the Hermann-Mauguin system. In this discussion only the International (Hermann-Mauguin) symbols will be used.

Screw Axes, Glide Planes

A rotation axis that is combined with translation is called a screw axis and is given the symbol n_t. The subscript t denotes the fractional translation of the periodicity

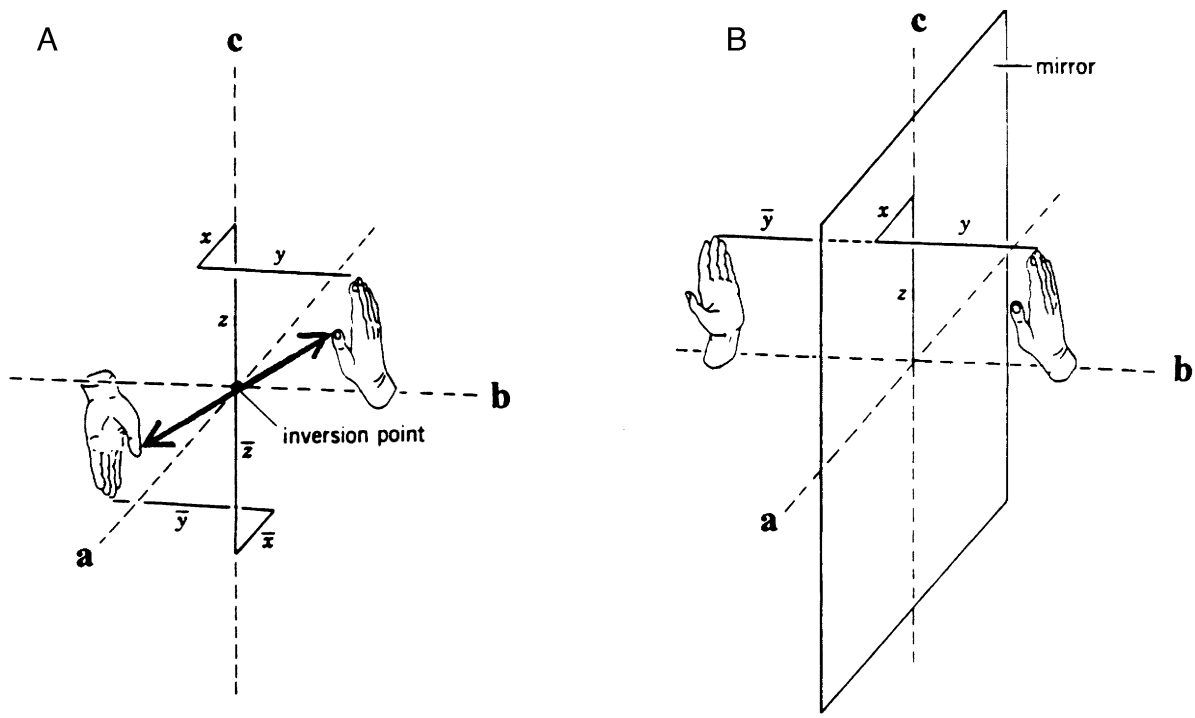


Figure 3. (**A**) A center of symmetry; (**B**) mirror reflection. Courtesy of Buerger (1970).

parallel to the rotation axis n, where $t = m/n$, $m = 1, \ldots, n-1$. Consider a 2-fold screw axis parallel to the b-axis of a coordinate system. The 2-fold rotor acts on an object by rotating it 180° and is immediately followed by a translation, $t/2$, of $\frac{1}{2}$ the b-axis periodicity. An object at $x\,y\,z$ is generated at $\bar{x}, y+\frac{1}{2}, \bar{z}$ by this 2_1 screw axis. All crystallographic symmetry operations must operate on a motif a sufficient number of times so that eventually the motif coincides with the original object. This is not the case at this juncture. This screw operation has to be repeated again, resulting in an object at x, $y+1$, z. Now the object is located one b-axis translation from the original object. Since this constitutes a periodic translation, b, the two objects are identical and the space has the proper symmetry 2_1. The possible screw axes are 2_1, 3_1, 3_2, 4_1, 4_2, 4_3, 6_1, 6_2, 6_3, 6_4, and 6_5 (Fig. 4). Note that the screw axes 3_1 and 3_2 are related as a right-handed thread is to a left-handed one. Similarly, this relationship is present for spaces exhibiting 4_1 and 4_3, 6_1 and 6_5, etc., symmetries. Note the symbols above the axes in Figure 4 that indicate the type of screw axis.

The combination of a mirror plane with translation parallel to the reflecting plane is known as a glide plane. Consider a coordinate system a b c in which the bc plane is a mirror. An object located at $x\,y\,z$ is reflected, which would

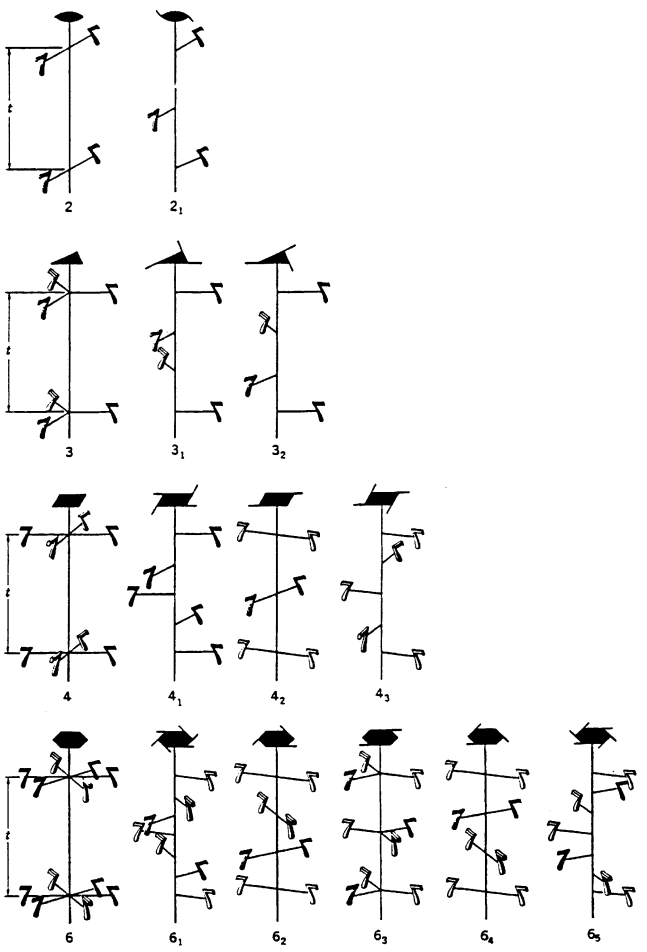


Figure 4. The eleven screw axes. The pure rotation axes are also shown. Note the symbols above the axes. Courtesy of Azaroff (1968).

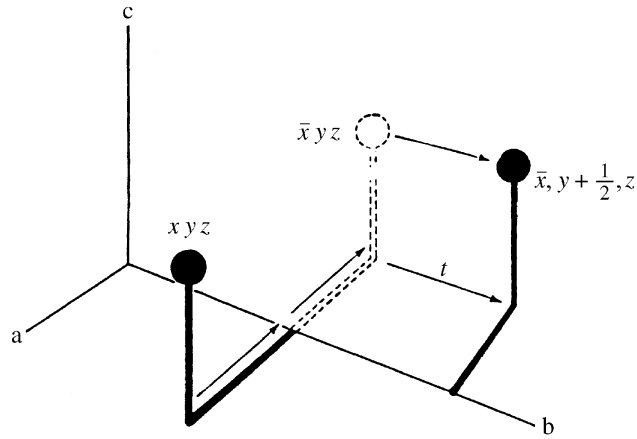


Figure 5. A b glide plane perpendicular to the a-axis.

bring it temporarily to the position $\bar{x}\,y\,z$. However, it does not remain there but is translated by $\frac{1}{2}$ of the b-axis periodicity to the point $\bar{x}, y+\frac{1}{2}, z$ (Fig. 5). This operation must be repeated to satisfy the requirement of periodicity so that the next operation brings the object to x, $y+1$, z, which is identical to the starting position but one b-axis periodicity away. Note that the first operation produces an enantiomorphic object and the second operation reverses this handedness, making it congruent with the initial object. This glide operation is designated as a b glide plane and has the symbol b. We could have moved the object parallel to the c axis by $\frac{1}{2}$ of the c-axis periodicity, as well as by the vector $\frac{1}{2}(\mathbf{b}+\mathbf{c})$. The former symmetry operator is a c glide plane, denoted by c, and the latter is an n glide plane, symbolized by n. Note that in this example an a glide plane operation is meaningless. If the glide plane is perpendicular to the b-axis, then a, c, and $n = \frac{1}{2}(\mathbf{a}+\mathbf{c})$ glide planes can exist. The extension to a glide plane perpendicular to the c-axis is obvious. One other glide operation needs to be described, the diamond glide d. It is characterized by the operation $\frac{1}{4}(\mathbf{a}+\mathbf{b})$, $\frac{1}{4}(\mathbf{a}+\mathbf{c})$, and $\frac{1}{4}(\mathbf{b}+\mathbf{c})$. The diagonal glide with translation $\frac{1}{4}(\mathbf{a}+\mathbf{b}+\mathbf{c})$ can be considered part of the diamond glide operation. All of these operations must be applied repeatedly until the last object that is generated is identical with the object at the origin but one periodicity away.

Symmetry-related positions, or equivalent positions, can be generated from geometrical considerations. However, the operations can be represented by matrices operating on a given position. In general, one can write the matrix equation

$$\mathbf{X}' = R\mathbf{X} + T$$

where $\mathbf{X}'(x'\ y'\ z')$ are the transformed coordinates, R is a rotation operator applied to $\mathbf{X}(x\ y\ z)$, and T is the translation operator. For the $\bar{1}$ operation, $x\ y\ z \Rightarrow \bar{x}\ \bar{y}\ \bar{z}$, and in matrix formulation the transformation becomes

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \bar{1} & 0 & 0 \\ 0 & \bar{1} & 0 \\ 0 & 0 & \bar{1} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \qquad (1)$$

For an a glide plane perpendicular to the c-axis, $x\,y\,z \Rightarrow x+\frac{1}{2}, y, \bar{z}$, or in matrix notation

$$\begin{matrix} x' \\ y' \\ z' \end{matrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \bar{1} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix} \qquad (2)$$

(Hall, 1969; Burns and Glazer, 1978; Stout and Jensen, 1989; Giacovazzo et al., 1992).

POINT GROUPS

The symmetry of space about a point can be described by a collection of symmetry elements that intersect at that point. The point of intersection of the symmetry axes is the origin. The collection of crystallographic symmetry operators constitutes the 32 crystallographic point groups. The external morphology of three-dimensional crystals can be described by one of these 32 crystallographic point groups. Since they describe the interrelationship of faces on a crystal, the symmetry operators cannot contain translational components that refer to atomic-scale relations such as screw axes or glide planes. The point groups can be divided into (1) simple rotation groups and (2) higher symmetry groups. In (1), there exist only 2-fold axes or one unique symmetry axis higher than a 2-fold axis. There are 27 such point groups. In (2), no single unique axis exists but more than one n-fold axis is present, $n > 2$.

The simple rotation groups consist only of one single n-fold axis. Thus, the point groups 1, 2, 3, 4, and 6 constitute the five pure rotation groups (Fig. 2). There are four distinct, unique rotoinversion groups: $\bar{1}$, $\bar{2} = \mathrm{m}$, $\bar{3}$, $\bar{4}$, $\bar{6}$. It has been shown that $\bar{2}$ is equivalent to a mirror, m, which is perpendicular to that axis, and the standard symbol for $\bar{2}$ is m. Group $\bar{6}$ is usually labeled 3/m and is assigned to group n/m. This last symbol will be encountered frequently. It means that there is an n-fold axis parallel to a given direction, and that perpendicular to that direction a mirror plane or some other symmetry plane exists. Next are four unique point groups that contain a mirror perpendicular to the rotation axis: 2/m, $3/\mathrm{m} = \bar{6}$, 4/m, 6/m. There are four groups that contain mirrors parallel to a rotation axis: 2mm, 3m, 4mm, 6mm. An interesting change in notation has occurred. Why is 2mm and not simply 2m used, while 3m is correct? Consider the intersection of two orthogonal mirrors. It is easy to show by geometry that the line of intersection is a 2-fold axis. It is particularly easy with matrix algebra (Stout and Jensen, 1989; Giacovazzo et al., 1992). Let the ab and ac coordinate planes be orthogonal mirror planes. The line of intersection is the a-axis. The multiplication of the respective mirror matrices yields the matrix representation of the 2-fold axis parallel to the a-axis:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \bar{1} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \bar{1} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \bar{1} & 0 \\ 0 & 0 & \bar{1} \end{pmatrix} \qquad (3)$$

Thus, a combination of two intersecting orthogonal mirrors yields a 2-fold axis of symmetry and similarly the combination of a 2-fold axis lying in a mirror plane produces another mirror orthogonal to it.

Let us examine 3m in a similar fashion. Let the 3-fold axis be parallel to the c-axis. A 3-fold symmetry axis demands that the a and b axes must be of equal length and at 120° to each other. Let the mirror plane contain the c-axis and the perpendicular direction to the b-axis. The respective matrices are

$$\begin{pmatrix} 0 & \bar{1} & 0 \\ 1 & \bar{1} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & \bar{1} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \bar{1} & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (4)$$

and the product matrix represents a mirror plane containing the c-axis and the perpendicular direction to the a-axis. These two directions are at 60° to each other. Since 3-fold symmetry requires a mirror every 120°, this is not a new symmetry operator. In general, when n of an n-fold rotor is odd no additional symmetry operators are generated, but when n is even a new symmetry operator comes into existence.

One can combine two symmetry operators in an arbitrary manner with a third symmetry operator. Will this combination be a valid crystallographic point group? This complex problem was solved by Euler (Buerger, 1956; Azaroff, 1968; McKie and McKie, 1986). He derived the relation

$$\cos A = \frac{\cos(\beta/2)\cos(\gamma/2) + \cos(\alpha/2)}{\sin(\beta/2)\sin(\gamma/2)} \qquad (5)$$

where A is the angle between two rotation axes with rotation angles β and γ, and α is the rotation angle of the third axis. Consider the combination of two 2-fold axes with one 3-fold axis. We must determine the angle between a 2-fold and 3-fold axis and the angle between the two 2-fold axes. Let angle A be the angle between the 2-fold and 3-fold axes. Applying the formula yields $\cos A = 0$ or $A = 90°$. Similarly, let B be the angle between the two 2-fold axes. Then $\cos B = \frac{1}{2}$ and $B = 60°$. Thus, the 2-fold axes are orthogonal to the 3-fold axis and 60° to each other, consistent with 3-fold symmetry. The crystallographic point group is 32. Note again that the symbol is 32, while for a 4-fold axis combined with an orthogonal 2-fold axis the symbol is 422.

So far 17 point groups have been derived. The next 10 groups are known as the dihedral point groups. There are four point groups containing n 2-fold axes perpendicular to a principal axis: 222, 32, 422, and 622. (Note that for $n = 3$ only one unique 2-fold axis is shown.) These groups can be combined with diagonal mirrors that bisect the 2-fold axes, yielding the two additional groups $\bar{4}$ 2m and $\bar{3}$m. Four additional groups result from the addition of a mirror perpendicular to the principal axis, $\frac{2}{\mathrm{m}}\frac{2}{\mathrm{m}}\frac{2}{\mathrm{m}}$, $\bar{6}$m2, $\frac{4}{\mathrm{m}}\frac{2}{\mathrm{m}}\frac{2}{\mathrm{m}}$, and $\frac{6}{\mathrm{m}}\frac{2}{\mathrm{m}}\frac{2}{\mathrm{m}}$ making a total of 27.

We now consider the five groups that contain more than one axis of at least 3-fold symmetry: 2 3, $\bar{4}$ 3m, m$\bar{3}$, 4 3 2, and m3m. Note that the position of the 3-fold axis is in the second position of the symbols. This indicates that these point groups belong to the cubic crystal system. The stereographic projections of the 32 point groups are shown in Figure 6 (*International Union of Crystallography*, 1952).

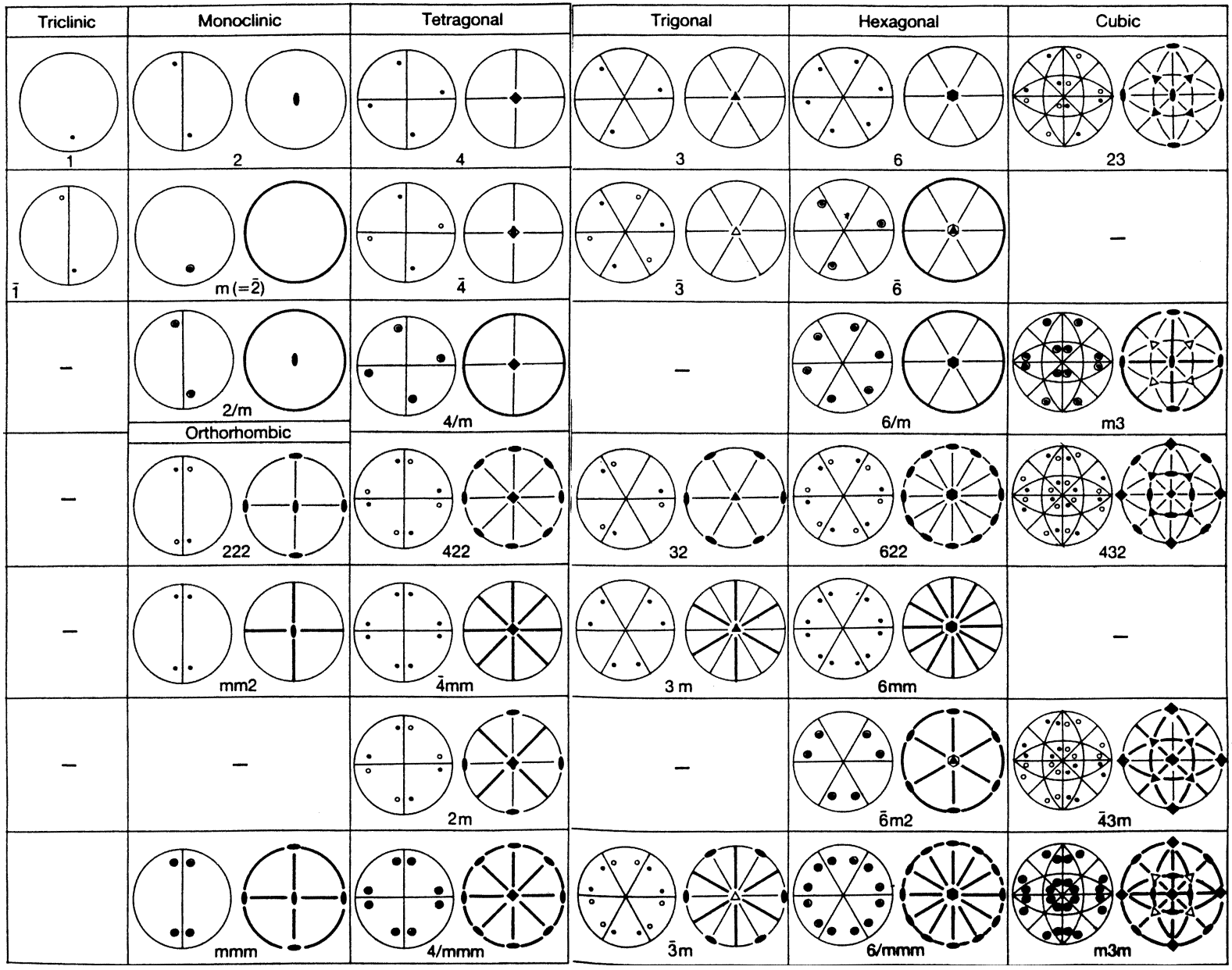


Figure 6. The stereographic projections of the 32 point groups. From the *International Tables for X-Ray Crystallography.* (International Union of Crystallography, 1952).

CRYSTAL SYSTEMS

The presence of a symmetry operator imparts a distinct appearance to a crystal. A 3-fold axis means that crystal faces around the axis must be identical every 120°. A 4-fold axis must show a 90° relationship among faces. On the basis of the appearance of crystals due to symmetry, classical crystallographers could divide them into seven groups as shown in Table 1. The symbol $\neq$ should be read as "not necessarily equal." The relationship among the metric parameters is determined by the presence of the symmetry operators among the atoms of the unit cell, but the metric parameters do not determine the crystal system. Thus, one could have a metrically cubic cell, but if only 1-fold axes are present among the atoms of the unit cell, then the crystal system is triclinic. This is the case for hexamethylbenzene, but, admittedly, this is a rare occurrence. Frequently, the rhombohedral unit cell is reoriented so that it can be described on the basis of a hexagonal unit cell (Azaroff, 1968). It can be considered a subsystem of the hexagonal system, and then one speaks of only six crystal systems.

We can now assign the various point groups to the six crystal systems. Point groups 1 and $\bar{1}$ obviously belong to the triclinic system. All point groups with only one unique 2-fold axis belong to the monoclinic system. Thus, 2, $\bar{2} = m$, and 2/m are monoclinic. Point groups like 222, mmm, etc., are orthorhombic; 32, 6, 6/mmm, etc., are hexagonal (point groups with a 3-fold axis are also labeled trigonal); 4, 4/m, 422, etc., are tetragonal; and 23, m3m, and 432, are cubic. Note the position of the 3-fold axis in the sequence of symbols for the cubic system. The distribution of the 32 point groups among the six crystal systems is shown in Figure 6. The rhombohedral and trigonal systems are not counted separately.

LATTICES

When the unit cell is translated along three periodically repeated noncoplanar, noncollinear vectors, a three-dimensional lattice of points is generated (Fig. 7). When looking at such an array one can select an infinite number of periodically repeated, noncollinear, noncoplanar vectors $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$, in three dimensions, connecting two lattice points, that will constitute the basis vectors of a unit cell for such an array. The choice of a unit cell is one of convenience, but usually the unit cell is chosen to reflect the symmetry operators present. Each lattice point at the eight corners of the unit cell is shared by eight other unit cells. Thus, a unit cell has $8 \times \frac{1}{8} = 1$ lattice point. Such a unit cell is called primitive and is given the symbol *P*. One can choose nonprimitive unit cells that will contain more than one lattice point. The array of atoms around one lattice point is identical to the same array around every other lattice point. This array of atoms may consist either of molecules or of individual atoms, as in NaCl. In the latter case the Na^+ and Cl^- atoms are actually located on lattice points. However, lattice points are usually not occupied by atoms. Confusing lattice points with atoms is a common beginners' error.

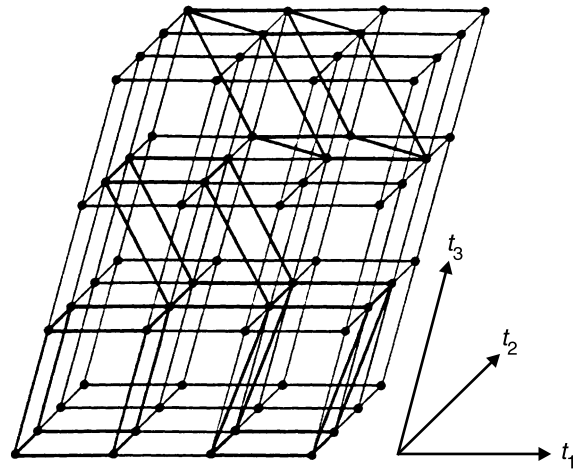


Figure 7. A point lattice with the outline of several possible unit cells.

In Table 1 the unit cells for the various crystal systems are listed with the restrictions on the parameters as a result of the presence of symmetry operators. Let the *b*-axis of a coordinate system be a 2-fold axis. The *ac* coordinate plane can have axes at any angle to each other: i.e., β can have any value. But the b-axis must be perpendicular to the *ac* plane or else the parallelogram defined by the vectors **a** and **b** will not be repeated periodically. The presence of the 2-fold axis imposes the restriction that $\alpha = \gamma = 90°$. Similarly, the presence of three 2-fold axes requires an orthogonal unit cell. Other symmetry operators impose further restrictions, as shown in Table 1.

Consider now a 2-fold *b*-axis perpendicular to a parallelogram lattice *ac*. How can this two-dimensional lattice be

Table 1. The Seven Crystal Systems

Crystal System	Minimum Symmetry	Unit Cell Parameter Relationships
Triclinic (anorthic)	Only a 1-fold axis	$a \neq b \neq c$; $\alpha \neq \beta \neq \gamma$
Monoclinic	One 2-fold axis chosen to be the unique *b*-axis, [010]	$a \neq b \neq c$; $\alpha = \gamma = 90°$, $\beta \neq 90°$
Orthorhombic	Three mutually perpendicular 2-fold axes, [100], [010], [001]	$a \neq b \neq c$; $\alpha = \beta = \gamma = 90°$
Rhombohedral[a]	One 3-fold axis parallel to the long axis of the rhomb, [111]	$a = b = c$; $\alpha = \beta = \gamma \neq 90°$
Tetragonal	One 4-fold axis parallel to the *c*-axis, [001]	$a = b \neq c$; $\alpha = \beta = \gamma = 90°$
Hexagonal[b]	One 6-fold axis parallel to the *c*-axis, [001]	$a = b \neq c$; $\alpha = \beta = 90° \gamma = 120°$
Cubic (isometric)	Four 3-fold axes parallel to the four body diagonals of a cube, [111]	$a = b = c$; $\alpha = \beta = \gamma = 90°$

[a] Usually transformed to a hexagonal unit cell.
[b] Point groups or space groups that are not rhombohedral but contain a 3-fold axis are labeled trigonal.

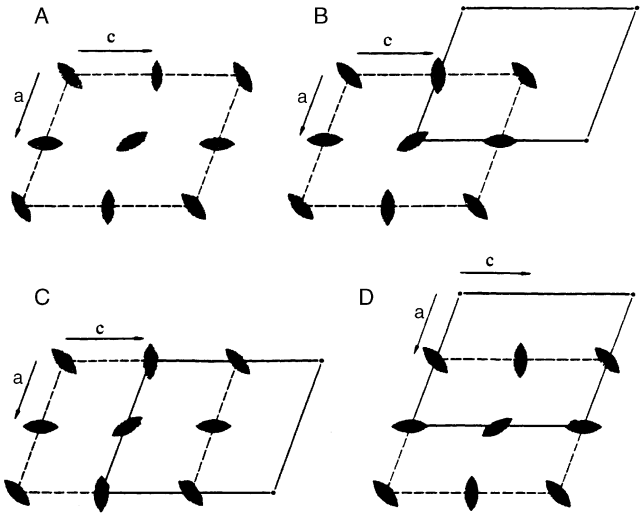


Figure 8. The stacking of parallelogram lattices in the monoclinic system. (**A**) Shifting the zero level up along the 2-fold axis located at the origin of the parallelogram. (**B**) Shifting the zero level up on the 2-fold axis at the center of the parallelogram. (**C**) Shifting the zero level up on the 2-fold axis located at $\frac{1}{2}$c. (**D**) Shifting the zero level up on the 2-fold axis located at $\frac{1}{2}$a. Courtesy of Azaroff (1968).

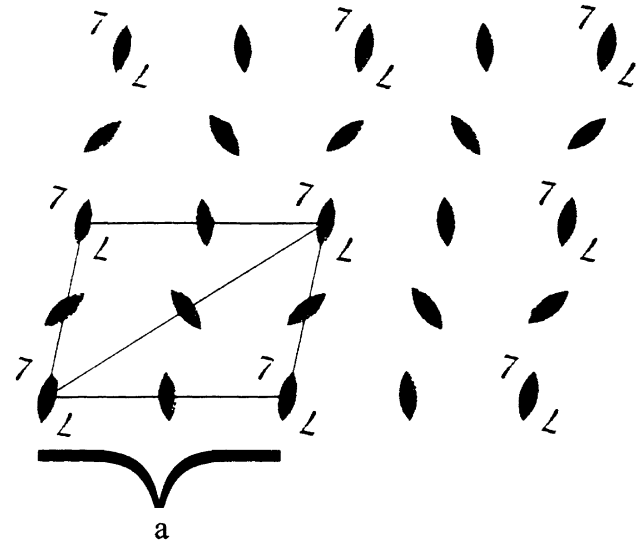


Figure 9. A periodic repetition of a 2-fold axis creates new, crystallographically independent 2-fold axes. Courtesy of Azaroff (1968).

repeated along the third dimension? It cannot be along some arbitrary direction because such a unit cell would violate the 2-fold axis. However, the plane net can be stacked along the *b*-axis at some definite interval to complete the unit cell (Fig. 8A). This symmetry operation produces a primitive cell, *P*. Is this the only possibility? When a 2-fold axis is periodically repeated in space with period **a**, then the 2-fold axis at the origin is repeated at $x = 1, 2, \ldots, n$, but such a repetition also gives rise to new 2-fold axes at $x = \frac{1}{2}, \frac{3}{2}, \ldots, z = \frac{1}{2}, \frac{3}{2}, \ldots$, etc., and along the plane diagonal (Fig. 9). Thus, there are three additional 2-fold axes and three additional stacking possibilities along the 2-fold axes located at $x = \frac{1}{2}, z = 0$; $x = 0, z = \frac{1}{2}$; and $x = \frac{1}{2}, z = \frac{1}{2}$. However, the first layer stacked along the 2-fold axis located at $x = \frac{1}{2}, z = 0$ does not result in a unit cell that incorporates the 2-fold axis. The vector from 0, 0, 0 to that lattice point is not along a 2-fold axis. The stacking sequence has to be repeated once more and now a lattice point on the second parallelogram lattice will lie above the point 0, 0, 0. The vector length from the origin to that point is the periodicity along **b**. An examination of this unit cell shows that there is a lattice point at $\frac{1}{2}, \frac{1}{2}, 0$ so that the *ab* face is centered. Such a unit cell is labeled C-face centered given the symbol C (Fig. 8D), and contains two lattice points: the origin lattice point shared among eight cells and the face-centered point shared between two cells. Stacking along the other 2-fold axes produces an A-face-centered cell given the symbol A (Fig. 8C) and a body-centered cell given the label I. (Fig. 8B). Since every direction in the monoclinic system is a 1-fold axis except for the 2-fold *b*-axis, the labeling of *a* and *c* directions in the plane perpendicular to *b* is arbitrary. Interchanging the *a* and *c* axial labels changes the A-face centering to C-face centering. By convention C-face centering is the standard orientation. Similarly, an I-centered lattice can be reoriented to a C-centered cell by drawing a diagonal in the old cell as a new axis. Thus, there are only two unique Bravais lattices in the monoclinic system, Figure 10.

The systematic investigation of the combinations of the 32 point groups with periodic translation in space generates 14 different space lattices. The 14 Bravais lattices are shown in Figure 10 (Azaroff, 1968; McKie and McKie, 1986).

Miller Indices

To explain x ray scattering from atomic arrays it is convenient to think of atoms as populating planes in the unit cell. The diffraction intensities are considered to arise from reflections of these planes. These planes are labeled by the inverse of their intercepts on the unit cell axes. If a plane intercepts the *a*-axis at $\frac{1}{2}$, the *b*-axis at $\frac{1}{2}$, and the *c*-axis at $\frac{2}{3}$ the distances of their respective periodicities, then the Miller indices are $h = 4$, $k = 4$, $l = 3$, the reciprocals cleared of fractions (Fig. 11), and the plane is denoted as (443). The round brackets enclosing the (*hkl*) indices indicate that this is a plane. Figure 11 illustrates this convention for several planes. Planes that are related by a symmetry operator such as a 4-fold axis in the tetragonal system are part of a common form. Thus, (100), (010), $(\bar{1}00)$ and $(0\bar{1}0)$ are designated by ((100)) or {100}. A direction in the unit cell—e.g., the vector from the origin 0, 0, 0 to the lattice point 1, 1, 1—is represented by square brackets as [111]. Note that the above four planes intersect in a common line, the *c*-axis or the zone axis, which is designated as [001]. A family of zone axes is indicated by angle brackets, $\langle 111 \rangle$. For the tetragonal system, this symbol denotes the symmetry-equivalent directions [111], $[\bar{1}11]$, $[1\bar{1}1]$, and

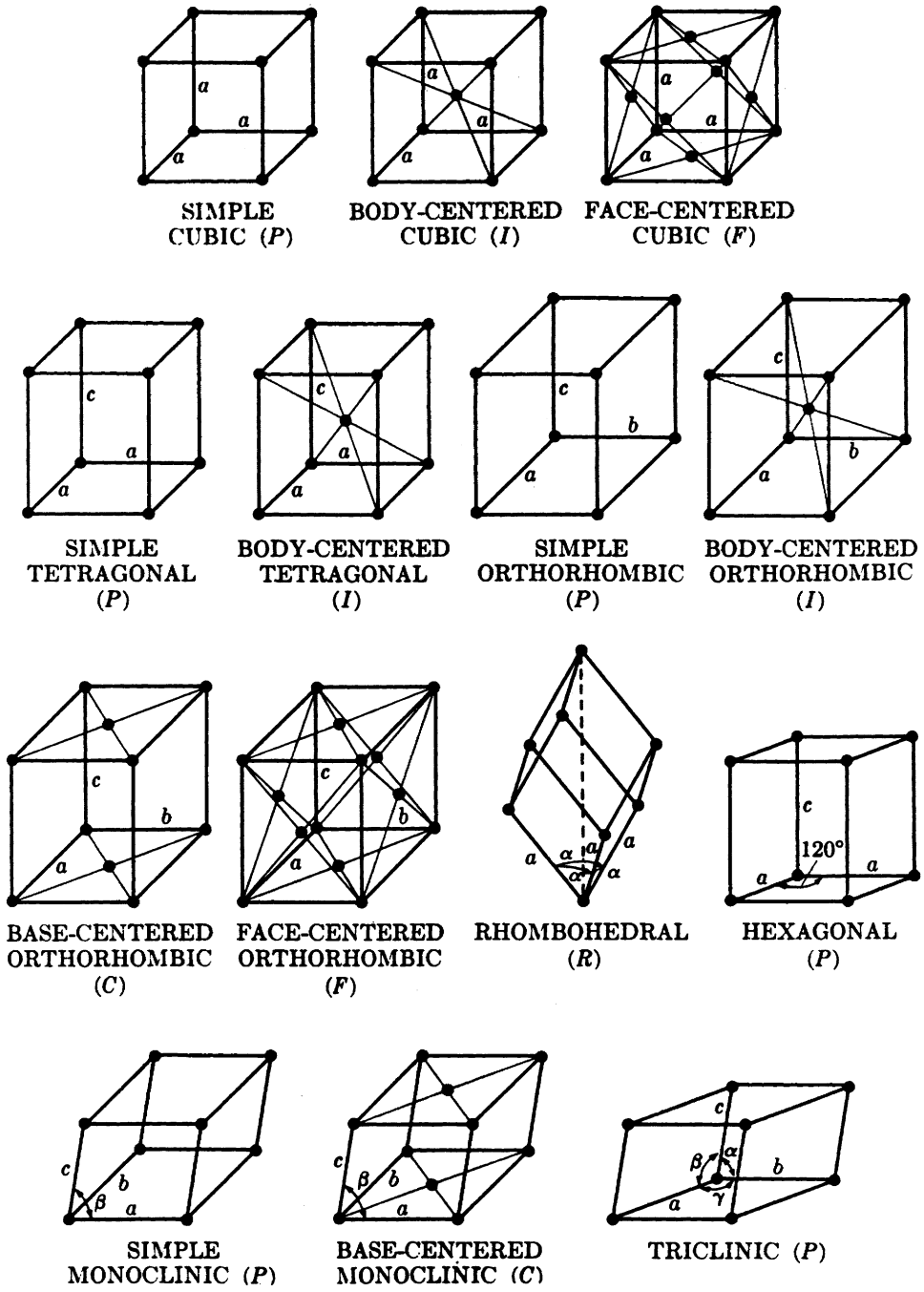


Figure 10. The 14 Bravais lattices. Courtesy of Cullity (1978).

$[\bar{1}\bar{1}1]$. The plane (hkl) and the zone axis $[uvw]$ obey the relationship $hu + kw + lz = 0$.

A complication arises in the hexagonal system. There are three equivalent a-axes due to the presence of a 3-fold or 6-fold symmetry axis perpendicular to the ab plane. If the three axes are the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, then $\mathbf{a}_1 + \mathbf{a}_2 = -\mathbf{a}_3$. To remove any ambiguity about which of the axes are cut by a plane, four symbols are used. They are the Miller-Bravais indices $hkil$, where $i = -(h + k)$ (Azaroff, 1968). Thus, what is ordinarily written as (111) becomes in the hexagonal system $(11\bar{2}1)$ or $(11\cdot1)$, the dot replacing the $\bar{2}$. The Miller indices for the unique directions in the unit cells of the seven crystal systems are listed in Table 1.

SPACE GROUPS

The combination of the 32 crystallographic point groups with the 14 Bravais lattices produces the 230 space groups. The atomic arrangement of every crystalline material displaying 3-dimensional periodicity can be assigned to one of

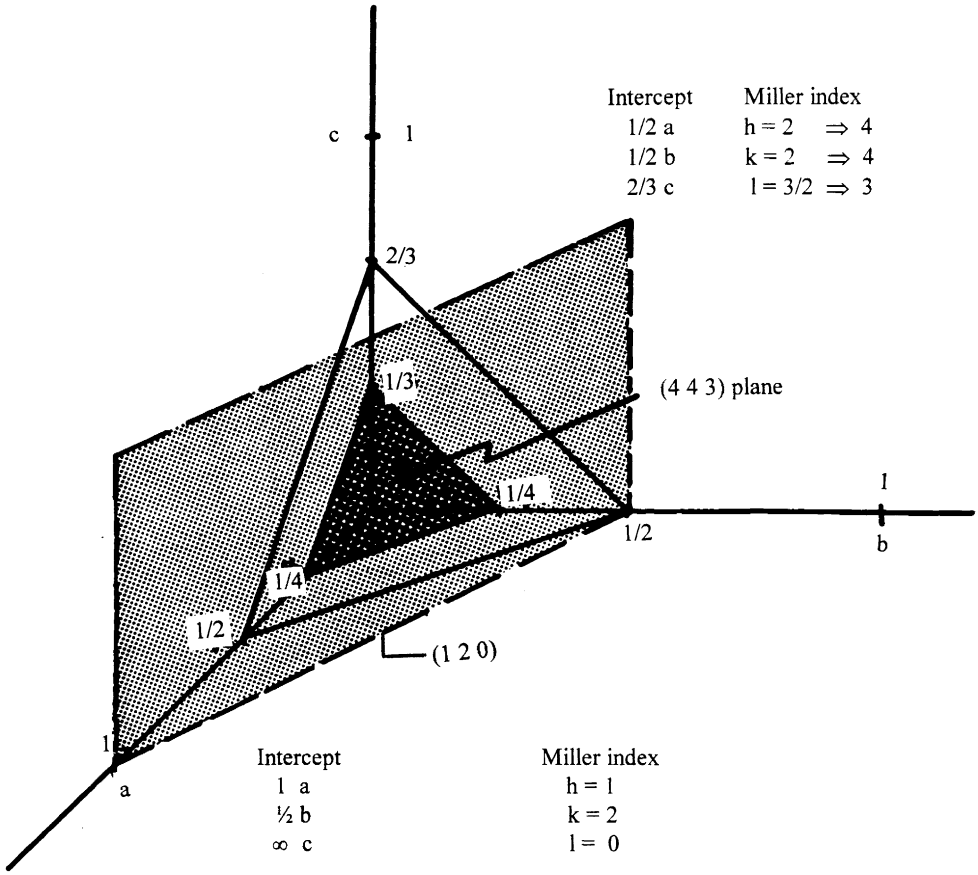


Figure 11. Miller indices of crystallographic planes.

the space groups. Consider the combination of point group 1 with a P lattice. An atom located at position $x\ y\ z$ in the unit cell is periodically repeated in all unit cells. There is only one general position $x\ y\ z$ in this unit cell. Of course, the values for $x\ y\ z$ can vary and there can be many atoms in the unit cell, but usually additional symmetry relations will not exist among them. Now consider the combination of point group $\bar{1}$ with the Bravais lattice P. Every atom at $x\ y\ z$ must have a symmetry-related atom at $\bar{x}\ \bar{y}\ \bar{z}$. Again, these positional parameters can have different values so that many atoms may be present in the unit cell. But for every atom A there must be an identical atom A′ related by a center of symmetry. The two positions are known as equivalent positions and the atom is said to be located in the general position $x\ y\ z$. If an atom is located at the special position 0, 0, 0, then no additional atom is generated. There are eight such special positions, each a center of symmetry, in the unit cell of space group $P\bar{1}$. We have just derived the first two triclinic space groups $P1$ and $P\bar{1}$. The first position in this nomenclature refers to the Bravais lattice. The second refers to a symmetry operator, in this case 1 or $\bar{1}$. The knowledge of symmetry operators relating atomic positions is very helpful when determining crystal structures. As soon as spatial positions $x\ y\ z$ of the atoms of a motif have been determined, e.g., the hand in Figure 3, then all atomic positions of the symmetry related motif(s) are known. The problem of determining all spatial parameters has been halved in the above example. The motif determined by the minimum number of atomic parameters is known as the asymmetric unit of the unit cell.

Let us investigate the combination of point group 2/m with a P Bravais lattice. The presence of one unique 2-fold axis means that this is a monoclinic crystal system and by convention the unique axis is labeled b. The 2-fold axis operating on $x\ y\ z$ generates the symmetry-related position $\bar{x}\ y\ \bar{z}$. The mirror plane perpendicular to the 2-fold axis operating on these two positions generates the additional locations $x\ \bar{y}\ z$ and $\bar{x}\ \bar{y}\ z$ for a total of four general equivalent positions. Note that the combination of 2/m gives rise to a center of symmetry. Special positions such as a location on a mirror $x\ \frac{1}{2}\ z$ permit only the 2-fold axis to produce the related equivalent position $\bar{x}\ \frac{1}{2}\ \bar{z}$. Similarly, the special position at 0, 0, 0 generates no further symmetry-related positions. A total of 14 special positions exist in space group $P2/m$. Again, the symbol shows the presence of only one 2-fold axis; therefore, the space group belongs to the monoclinic crystal system. The Bravais lattice is primitive, the 2-fold axis is parallel to the b-axis, and the mirror plane is perpendicular to the b-axis (International Union of Crystallography, 1983).

Let us consider one more example, the more complicated space group $Cmmm$ (Fig. 12). We notice that the Bravais lattice is C-face centered and that there are three mirrors. We remember that $m = \bar{2}$, so that there are essentially three 2-fold axes present. This makes the crystal system orthorhombic. The three 2-fold axes are orthogonal to

each other. We also know that the line of intersection of two orthogonal mirrors is a 2-fold axis. This space group, therefore, should have as its complete symbol $C2/m\ 2/m\ 2/m$, but the crystallographer knows that the 2-fold axes are there because they are the intersections of the three orthogonal mirrors. It is customary to omit them and write the space group as Cmmm. The three symbols after the Bravais lattice refer to the three orthogonal axes of the unit cell a, b, c. The letters m are really in the denominator so that the three mirrors are located perpendicular to the a-axis, perpendicular to the b-axis, and perpendicular to the c-axis. For the sake of consistency it is wise to consider any numeral as an axis parallel to a direction and any letter as a symmetry operator perpendicular to a direction.

C-face centering means that there is a lattice point at the position $\frac{1}{2}, \frac{1}{2}, 0$ in the unit cell. The atomic environment around any one lattice point is identical to that of any other lattice point. Therefore, as soon as the position $x\ y\ z$ is occupied there must be an identical occupant at $x+\frac{1}{2}$, $y+\frac{1}{2}$, z.

Let us now develop the general equivalent positions, or equipoints, for this space group. The symmetry-related point due to the mirror perpendicular to the a-axis operating on $x\ y\ z$ is $\bar{x}\ y\ z$. The mirror operation on these two equipoints due to the mirror perpendicular to the b-axis yields $x\ \bar{y}\ z$ and $\bar{x}\ \bar{y}\ z$. The mirror perpendicular to the c-axis operates on these four equipoints to yield $x\ y\ \bar{z}$, $\bar{x}\ y\ \bar{z}$, $x\ \bar{y}\ \bar{z}$, and $\bar{x}\ \bar{y}\ \bar{z}$. Note that this space group contains a center

A

Generators selected (1); $t(1,0,0)$; $t(0,1,0)$; $t(0,0,1)$; $t(\frac{1}{2},\frac{1}{2},0)$; (2); (3); (5)

Positions

Multiplicity, Wyckoff letter, Site symmetry			Coordinates				Reflection conditions
			(0,0,0)+	$(\frac{1}{2},\frac{1}{2},0)$+			
							General:
16	r	1	(1) x,y,z	(2) $\bar{x},\bar{y},z$	(3) $\bar{x},y,\bar{z}$	(4) $x,\bar{y},\bar{z}$	hkl : $h+k=2n$
			(5) $\bar{x},\bar{y},\bar{z}$	(6) $x,y,\bar{z}$	(7) $x,\bar{y},z$	(8) $\bar{x},y,z$	$0kl$: $k=2n$
							$h0l$: $h=2n$
							$hk0$: $h+k=2n$
							$h00$: $h=2n$
							$0k0$: $k=2n$
							Special: as above, plus
8	q	..m	$x,y,\frac{1}{4}$	$\bar{x},\bar{y},\frac{1}{4}$	$\bar{x},y,\frac{1}{4}$	$x,\bar{y},\frac{1}{4}$	no extra conditions
8	p	..m	$x,y,0$	$\bar{x},\bar{y},0$	$\bar{x},y,0$	$x,\bar{y},0$	no extra conditions
8	o	.m.	$x,0,z$	$\bar{x},0,z$	$\bar{x},0,\bar{z}$	$x,0,\bar{z}$	no extra conditions
8	n	m..	$0,y,z$	$0,\bar{y},z$	$0,y,\bar{z}$	$0,\bar{y},\bar{z}$	no extra conditions
8	m	..2	$\frac{1}{4},\frac{1}{4},z$	$\frac{3}{4},\frac{3}{4},\bar{z}$	$\frac{3}{4},\frac{3}{4},z$	$\frac{1}{4},\frac{1}{4},\bar{z}$	hkl : $h=2n$
4	l	$mm2$	$0,\frac{1}{2},z$	$0,\frac{1}{2},\bar{z}$			no extra conditions
4	k	$mm2$	$0,0,z$	$0,0,\bar{z}$			no extra conditions
4	j	$m2m$	$0,y,\frac{1}{4}$	$0,\bar{y},\frac{1}{4}$			no extra conditions
4	i	$m2m$	$0,y,0$	$0,\bar{y},0$			no extra conditions
4	h	$2mm$	$x,0,\frac{1}{4}$	$\bar{x},0,\frac{1}{4}$			no extra conditions
4	g	$2mm$	$x,0,0$	$\bar{x},0,0$			no extra conditions
4	f	..$2/m$	$\frac{1}{4},\frac{1}{4},\frac{1}{4}$	$\frac{3}{4},\frac{1}{4},\frac{1}{4}$			hkl : $h=2n$
4	e	..$2/m$	$\frac{1}{4},\frac{1}{4},0$	$\frac{3}{4},\frac{1}{4},0$			hkl : $h=2n$
2	d	mmm	$0,0,\frac{1}{2}$				no extra conditions
2	c	mmm	$\frac{1}{2},0,\frac{1}{2}$				no extra conditions
2	b	mmm	$\frac{1}{2},0,0$				no extra conditions
2	a	mmm	$0,0,0$				no extra conditions

Symmetry of special projections

Along [001] $c\,2mm$
$\boldsymbol{a'=a}$ $\boldsymbol{b'=b}$
Origin at $0,0,z$

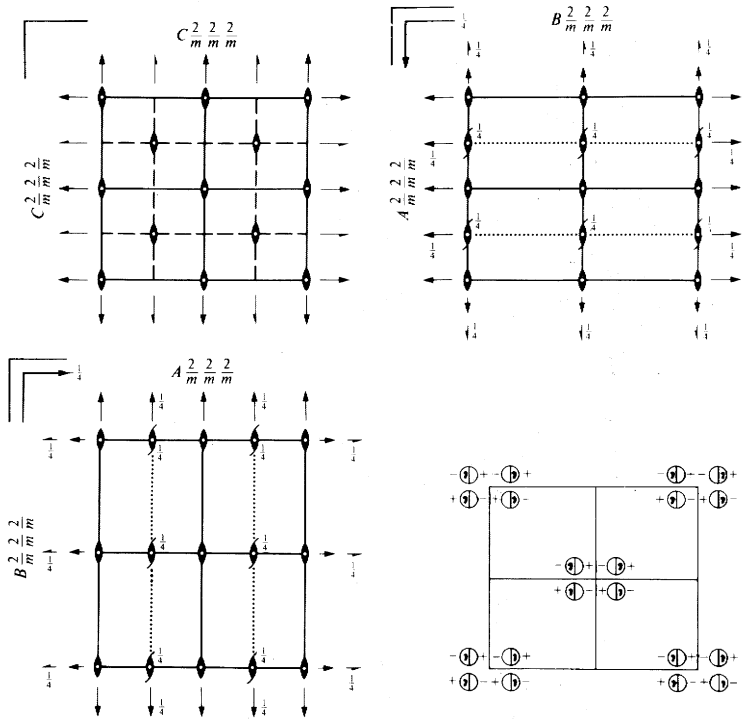
Along [100] $p\,2mm$
$\boldsymbol{a'=\frac{1}{2}b}$ $\boldsymbol{b'=c}$
Origin at $x,0,0$

Along [010] $p\,2mm$
$\boldsymbol{a'=c}$ $\boldsymbol{b'=\frac{1}{2}a}$
Origin at $0,y,0$

Figure 12. The space group C.mmm. (**A**) List of general and special equivalent positions. (**B**) Changes in space groups resulting from relabeling of the coordinate axes. From International Union of Crystallography (1983).

B

C m m m	D_{2h}^{19}	*m m m*	Orthorhombic
No. 65	*C* 2/*m* 2/*m* 2/*m*	Patterson symmetry	*C m m m*



Origin at centre (*m m m*)

Asymmetric unit $0 \le x \le \frac{1}{4}$; $0 \le y \le \frac{1}{4}$; $0 \le z \le \frac{1}{2}$

Symmetry operations

For (0,0,0)+ set

(1) 1	(2) 2 0,0,*z*	(3) 2 0,*y*,0	(4) 2 *x*,0,0
(5) $\bar{1}$ 0,0,0	(6) *m* *x*,*y*,0	(7) *m* *x*,0,*z*	(8) *m* 0,*y*,*z*

For ($\frac{1}{2}$,$\frac{1}{2}$,0)+ set

(1) $t(\frac{1}{2},\frac{1}{2},0)$	(2) 2 $\frac{1}{4},\frac{1}{4},z$	(3) $2(0,\frac{1}{2},0)$ $\frac{1}{4},y,0$	(4) $2(\frac{1}{2},0,0)$ $x,\frac{1}{4},0$
(5) $\bar{1}$ $\frac{1}{4},\frac{1}{4},0$	(6) $n(\frac{1}{2},\frac{1}{2},0)$ *x*,*y*,0	(7) *a* $x,\frac{1}{4},z$	(8) *b* $\frac{1}{4},y,z$

Figure 12 (*Continued*)

of symmetry. To every one of these eight equipoints must be added $\frac{1}{2}$, $\frac{1}{2}$, 0 to take care of C-face centering. This yields a total of 16 general equipoints. When an atom is placed on one of the symmetry operators, the number of equipoints is reduced (Fig. 12A).

Clearly, once one has derived all the positions of a space group there is no point in doing it again. Figure 12 is a copy of the space group information for *C*mmm found in the *International Tables for Crystallography*, Vol. A (International Union of Crystallography, 1983). Note the diagrams in Figure 12B: they represent the changes in the space group symbols as a result of relabeling the unit cell axes. This is permitted in the orthorhombic crystal system since the *a*, *b*, and *c* axes are all 2-fold so that no label is unique. The rectangle in Figure 12B represents the *ab* plane of the unit cell. The origin is in the upper left corner with the *a*-axis pointing down and the *b*-axis pointing to the right; the *c*-axis points out of the plane of the paper. Note the symbols for the symmetry elements and their locations in the unit cell. A complete description of these symbols can be found in the *International Tables for Crystallography*, Vol. A, pages 4–10 (International Union of Crystallography, 1983). In addition to the space group information there is an extensive discussion of many crystallographic topics. No x ray diffraction laboratory should be without this volume. The determination of the space group of a crystalline material is obtained from x ray diffraction data.

Space Group Symbols

In general, space group notations consist of four symbols. The first symbol always refers to the Bravais lattice. Why, then, do we have P 1 or P 2/m? The full symbols are $P111$ and P 1 2/m 1. But in the triclinic system there is no unique direction, since every direction is a 1-fold axis of symmetry. It is therefore sufficient just to write P 1. In the monoclinic system there is only one unique direction—by convention it is the b-axis—and so only the symmetry elements related to that direction need to be specified. In the orthorhombic system there are the three unique 2-fold axes parallel to the lattice parameters a, b, and c. Thus, Pnma means that the crystal system is orthorhombic, the Bravais lattice is P, and there is an n glide plane perpendicular to the a-axis, a mirror perpendicular to the b-axis, and an a glide plane perpendicular to the c-axis. The complete symbol for this space group is $P\ 2_1/n\ 2_1/m\ 2_1/a$. Again, the 2_1 screw axes are a result of the other symmetry operators and are not expressly indicated in the standard symbol. The letter symbols are considered in the denominator and the interpretation is that the operators are perpendicular to the axes. In the tetragonal system there is the unique direction, the 4-fold c-axis. The next unique directions are the equivalent a and b axes, and the third directions are the four equivalent C-face diagonals, the $\langle 110 \rangle$ directions. The symbol $I4cm$ means that the space group is tetragonal, the Bravais lattice is body centered, there is a 4-fold axis parallel to the c-axis, there is are c glide planes perpendicular to the equivalent a and b-axes, and there are mirrors perpendicular to the C-face diagonals. Note that one can say just as well that the symmetry operators c and m are parallel to the directions.

The symbol $P\bar{3}c1$ tells us that the space group belongs to the trigonal system, primitive Bravais lattice, with a 3-fold rotoinversion axis parallel to the c-axis, and a c glide plane perpendicular to the equivalent a and b axes (or parallel to the [210] and [120] directions); the third symbol refers to the face diagonal [110]. Why the 1 in this case? It serves to distinguish this space group from the space group $P\bar{3}1c$, which is different. As before, it is part of the trigonal system. A $\bar{3}$ rotoinversion axis parallel to the c axis is present, but now the c glide plane is perpendicular to the $[1\bar{1}0]$ or parallel to the [110] directions. Since 6-fold symmetry must be maintained there are also c glide planes parallel to the a and b axes. The symbol R denotes the rhombohedral Bravais lattice, but the lattice is usually reoriented so that the unit cell is hexagonal.

The symbol $F\text{m}\bar{3}\text{m}$ full symbol $F\ 4/\text{m}\ \bar{3}\ 2/\text{m}$, tells us that this space group belongs to the cubic system (note the position of $\bar{3}$), the Bravais lattice is all faces centered, and there are mirrors perpendicular to the three equivalent a, b, and c axes, a $\bar{3}$ rotoinversion axis parallel to the four body diagonals of the cube, the $\langle 111 \rangle$ directions, and a mirror perpendicular to the six equivalent face diagonals of the cube, the $\langle 110 \rangle$ directions. In this space group additional symmetry elements are generated, such as 4-fold, 2-fold, and 2_1 axes.

Simple Example of the Use of Crystal Structure Knowledge

Of what use is knowledge of the space group for a crystalline material? The understanding of the physical and chemical properties of a material ultimately depends on the knowledge of the atomic architecture—i.e., the location of every atom in the unit cell with respect to the coordinate axes. The density of a material is $\rho = M/V$, where ρ is density, M the mass, and V the volume (see MASS AND DENSITY MEASUREMENTS). The macroscopic quantities can also be expressed in terms of the atomic content of the unit cell. The mass in the unit cell volume V is the formula weight M multiplied by the number of formula weights z in the unit cell divided by Avogadro's number N. Thus, $\rho = Mz/VN$.

Consider NaCl. It is cubic, the space group is $F\text{m}\bar{3}\text{m}$, the unit cell parameter is $a = 4.872$ Å, its density is 2.165 g/cm^3, and its formula weight is 58.44, so that $z = 4$. There are four Na and four Cl ions in the unit cell. The general position $x\ y\ z$ in space group $F\text{m}\bar{3}\text{m}$ gives rise to a total of 191 additional equivalent positions. Obviously, one cannot place an Na atom into a general position. An examination of the space group table shows that there are two special positions with four equipoints labeled $4a$, at 0, 0, 0 and $4b$, at $\frac{1}{2},\frac{1}{2},\frac{1}{2}$. Remember that F means that $x = \frac{1}{2}$, $y = \frac{1}{2}$, $z = 0$; $x = 0, y = \frac{1}{2}, z = \frac{1}{2}$; and $x = \frac{1}{2}, y = 0, z = \frac{1}{2}$ must be added to the positions. Thus, the 4 Na^+ atoms can be located at the $4a$ position and the 4 Cl^- atoms at the $4b$ position, and since the positional parameters are fixed, the crystal structure of NaCl has been determined. Of course, this is a very simple case. In general, the determination of the space group from x-ray diffraction data is the first essential step in a crystal structure determination.

CONCLUSION

It is hoped that this discussion of symmetry will ease the introduction of the novice to this admittedly arcane topic or serve as a review for those who want to extend their expertise in the area of space groups.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support of the Robert A. Welch Foundation of Houston, Texas.

LITERATURE CITED

Azaroff, L. A. 1968. Elements of X-Ray Crystallography. McGraw-Hill, New York.

Buerger, M. J. 1956. Elementary Crystallography. John Wiley & Sons, New York.

Buerger, M. J. 1970. Contemporary Crystallography. McGraw-Hill, New York.

Burns, G. and Glazer, A. M. 1978. Space Groups for Solid State Scientists. Academic Press, New York.

Cullity, B. D. 1978. Elements of X-Ray Diffraction, 2nd ed. Addison-Wesley, Reading, Mass.

Giacovazzo, C., Monaco, H. L., Viterbo, D., Scordari, F., Gilli, G., Zanotti, G., and Catti, M. 1992. Fundamentals of Crystallography. International Union of Crystallography, Oxford University Press, Oxford.

Hall, L. H. 1969. Group Theory and Symmetry in Chemistry. McGraw-Hill, New York.

International Union of Crystallography (Henry, N. F. M. and Lonsdale, K., eds.). 1952. International Tables for Crystallography, Vol. I: Symmetry Groups. The Kynoch Press, Birmingham, UK.

International Union of Crystallography (Hahn, T., ed.). 1983. International Tables for Crystallography, Vol. A: Space-Group Symmetry. D. Reidel, Dordrecht, The Netherlands.

McKie, D. and McKie, C. 1986. Essentials of Crystallography. Blackwell Scientific Publications, Oxford.

Stout, G. H. and Jensen, L. H. 1989. X-Ray Structure Determination: A Practical Guide, 2nd ed. John Wiley & Sons, New York.

KEY REFERENCES

Burns and Glazer, 1978. See above.

An excellent text for self-study of symmetry operators, point groups, and space groups; makes the International Tables for Crystallography understandable.

Hahn, 1983. See above.

Deals with space groups and related topics, and contains a wealth of crystallographic information.

Stout and Jensen, 1989. See above.

Meets its objective as a "practical guide" to single-crystal x-ray structure determination, and includes introductory chapters on symmetry.

INTERNET RESOURCES

http://www.hwi.buffalo.edu/aca

American Crystallographic Association. Site directory has links to numerous topics including Crystallographic Resources.

http://www.iucr.ac.uk

International Union of Crystallography. Provides links to many data bases and other information about worldwide crystallographic activities.

HUGO STEINFINK
University of Texas
Austin, Texas

PARTICLE SCATTERING

INTRODUCTION

Atomic scattering lies at the heart of numerous materials-analysis techniques, especially those that employ ion beams as probes. The concepts of particle scattering apply quite generally to objects ranging in size from nucleons to billiard balls, at classical as well as relativistic energies, and for both elastic and inelastic events. This unit summarizes two fundamental topics in collision theory: kinematics, which governs energy and momentum transfer, and central-field theory, which accounts for the strength of particle interactions.

For definitions of symbols used throughout this unit, see the Appendix.

KINEMATICS

Binary Collisions

The kinematics of two-body collisions are the key to understanding atomic scattering. It is most convenient to consider such binary collisions as occurring between a moving projectile and an initially stationary target. It is sufficient here to assume only that the particles act upon each other with equal repulsive forces, described by some interaction potential. The form of the interaction potential and its effects are discussed below (see Central-Field Theory). A binary collision results in a change in the projectile's trajectory and energy after it scatters from a target atom. The collision transfers energy to the target atom, which gains energy and recoils away from its rest position. The essential parameters describing a binary collision are defined in Figure 1. These are the masses (m_1 and m_2) and the initial and final velocities (v_0, v_1, and v_2) of the projectile and target, the scattering angle (θ_s) of the projectile, and the recoiling angle (θ_r) of the target. Applying the laws of conservation of energy and momentum establishes fundamental relationships among these parameters.

Elastic Scattering and Recoiling

In an elastic collision, the total kinetic energy of the particles is unchanged. The law of energy conservation dictates that

$$E_0 = E_1 + E_2 \tag{1}$$

where $E = 1/2\,mv^2$ is a particle's kinetic energy. The law of conservation of momentum, in the directions parallel and perpendicular to the incident particle's direction, requires that

$$m_1 v_0 = m_1 v_1 \cos\theta_s + m_2 v_2 \cos\theta_r \tag{2}$$

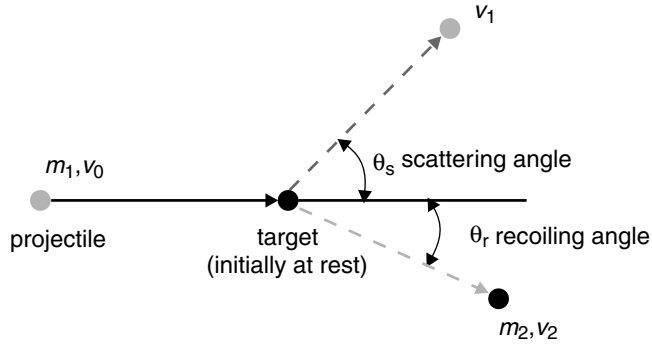


Figure 1. Binary collision diagram in a laboratory reference frame. The initial kinetic energy of the incident projectile is $E_0 = 1/2 m_1 v_0^2$. The initial kinetic energy of the target is assumed to be zero. The final kinetic energy for the scattered projectile is $E_1 = 1/2 m_1 v_1^2$, and for the recoiled particle is $E_2 = 1/2 m_2 v_2^2$. Particle energies (E) are typically expressed in units of electron volts, eV, and velocities (v) in units of m/s. The conversion between these units is $E \approx mv^2/(1.9297 \times 10^8)$, where m is the mass of the particle in amu.

for the parallel direction, and

$$0 = m_1 v_1 \sin\theta_s - m_2 v_2 \sin\theta_r \quad (3)$$

for the perpendicular direction.

Eliminating the recoil angle and target recoil velocity from the above equations yields the fundamental elastic scattering relation for projectiles:

$$2\cos\theta_s = (1+A)v_s + (1-A)/v_s \quad (4)$$

where $A = m_2/m_1$ is the target-to-projectile mass ratio and $v_s = v_1/v_0$ is the normalized final velocity of the scattered particle after the collision.

In a similar manner, eliminating the scattering angle and projectile velocity from Equations 1, 2, and 3 yields the fundamental elastic recoiling relation for targets:

$$2\cos\theta_r = (1+A)v_r \quad (5)$$

where $v_r = v_2/v_0$ is the normalized recoil velocity of the target particle.

Inelastic Scattering and Recoiling

If the internal energy of the particles changes during their interaction, the collision is inelastic. Denoting the change in internal energy by Q, the energy conservation law is stated as:

$$E_0 = E_1 + E_2 + Q \quad (6)$$

It is possible to extend the fundamental elastic scattering and recoiling relations (Equation 4 and Equation 5) to inelastic collisions in a straightforward manner. A kinematic analysis like that given above (see Elastic Scattering and Recoiling) shows the inelastic scattering relation to be:

$$2\cos\theta_s = (1+A)v_s + [1 - A(1-Q_n)]/v_s \quad (7)$$

where $Q_n = Q/E_0$ is the normalized inelastic energy factor. Comparison with Equation 4 shows that incorporating the factor Q_n accounts for the inelasticity in a collision.

When $Q > 0$, it is referred to as an inelastic energy loss; that is, some of the initial kinetic energy E_0 is converted into internal energy of the particles and the total kinetic energy of the system is reduced following the collision. Here Q_n is assumed to have a constant value that is independent of the trajectories of the collision partners, i.e., its value does not depend on θ_s. This is a simplifying assumption, which clearly breaks down if the particles do not collide ($\theta_s = 0$).

The corresponding inelastic recoiling relation is

$$2\cos\theta_r = (1+A)v_r + \frac{Q_n}{Av_r} \quad (8)$$

In this case, inelasticity adds a second term to the elastic recoiling relation (Equation 5).

A common application of the above kinematic relations is in identifying the mass of a target particle by measuring the kinetic energy loss of a scattered probe particle. For example, if the mass and initial velocity of the probe particle are known and its elastic energy loss is measured at a particular scattering angle, then the Equation 4 can be solved in terms of m_2. Or, if both the projectile and target masses are known and the collision is inelastic, Q can be found from Equation 7. A number of useful forms of the fundamental scattering and recoiling relations for both the elastic and inelastic cases are listed in the Appendix at the end of this unit (see Solutions of the Fundamental Scattering and Recoiling Relations in Terms of v, E, θ, A, and Q_n for Nonrelativistic Collisions).

A General Binary Collision Formula

It is possible to collect all the kinematic expressions of the preceding sections and cast them into a single fundamental form that applies to all nonrelativistic, mass-conserving binary collisions. This general formula in which the particles scatter or recoil through the laboratory angle θ is

$$2\cos\theta = (1+A)v_n + h/v_n \quad (9)$$

where $v_n = v/v_0$ and $h = 1 - A(1 - Q_n)$ for scattering and Q_n/A for recoiling. In the above expression, v is a particle's velocity after collision (v_1 or v_2) and the other symbols have their usual meanings. Equation 9 is the essence of binary collision kinematics.

In experimental work, the measured quantity is often the energy of the scattered or recoiled particle, E_1 or E_2. Expressing Equation 9 in terms of energy yields

$$\frac{E}{E_0} = \frac{A}{g}\left[\frac{1}{1+A}(\cos\theta \pm \sqrt{f^2 g^2 - \sin^2\theta})\right]^2 \quad (10)$$

where $f^2 = 1 - Q_n(1+A)/A$ and $g = A$ for scattering and 1 for recoiling. The positive sign is taken when $A > 1$ and both signs are taken when $A < 1$.

Scattering and Recoiling Diagrams

A helpful and instructive way to become familiar with the fundamental scattering and recoiling relations is to look at their geometric representations. The traditional approach is to plot the relations in center-of-mass coordinates, but an especially clear way of depicting these relations, particularly for materials analysis applications, is to use the laboratory frame with a properly scaled polar coordinate system. This approach will be used extensively throughout the remainder of this unit.

The fundamental scattering and recoil relations (Equation 4, Equation 5, Equation 7, and Equation 8) describe circles in polar coordinates, $(\sqrt{E}, \theta)$. The radial coordinate is taken as the square root of normalized energy (E_s or E_r) and the angular coordinate, θ, is the laboratory observation angle (θ_s or θ_r). These curves provide considerable insight into the collision kinematics. Figure 2 shows a typical elastic scattering circle. Here, $\sqrt{E}$ is $\sqrt{(E_s)}$, where $E_s = E_1/E_0$ and θ is θ_s. Note that r is simply v_s, so the circle traces out the velocity/angle relationship for scattering. Projectiles can be viewed as having initial velocity vectors

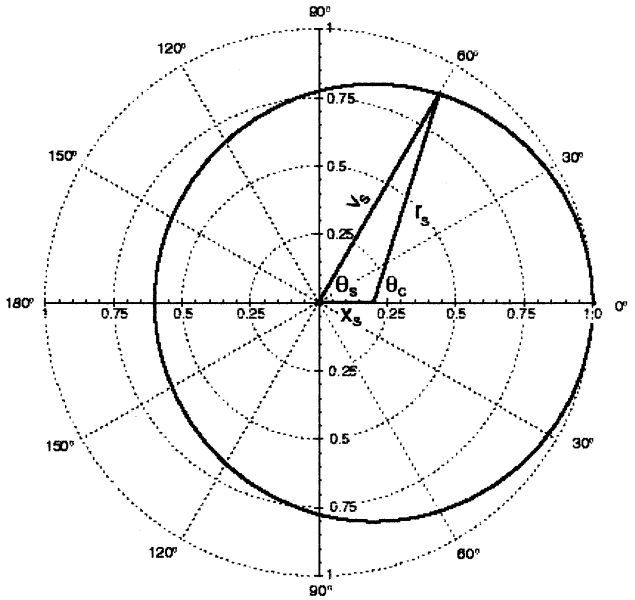


Figure 2. An elastic scattering circle plotted in polar coordinates $(\sqrt{E}, \theta)$ where E is $E_s = E_1/E_0$ and θ is the laboratory scattering angle, θ_s. The length of the line segment from the origin to a point on the circle gives the relative scattered particle velocity, v_s, at that angle. Note that $\sqrt{(E_s)} = v_s = v_1/v_0$. Scattering circles are centered at $(x_s,0^\circ)$, where $x_s = (1 + A)^{-1}$ and $A = m_2/m_1$. All elastic scattering circles pass through the point $(1,0^\circ)$. The circle shown is for the case $A = 4$. The triangle illustrates the relationships $\sin(\theta_c - \theta_s)/\sin(\theta_s) = x_s/r_s = 1/A$.

of unit magnitude traveling from left to right along the horizontal axis, striking the target at the origin, and leaving at angles and energies indicated by the scattering circle. The circle passes through the point $(1,0^\circ)$, corresponding to the situation where no collision occurs. Of course, when there is no scattering ($\theta_s = 0^\circ$), there is no change in the incident particle's energy or velocity ($E_s = 1$ and $v_s = 1$). The maximum energy loss occurs at $\theta = 180^\circ$, when a head-on collision occurs. The circle center and radius are a function of the target-to-projectile mass ratio. The center is located along the 0° direction a distance x_s from the origin, given by

$$x_s = \frac{1}{1+A} \tag{11}$$

while the radius for elastic scattering is

$$r_s = 1 - x_s = A\, x_s = \frac{A}{1+A} \tag{12}$$

For inelastic scattering, the scattering circle center is also given by Equation 11, but the radius is given by

$$r_s = fAx_s = \frac{fA}{1+A} \tag{13}$$

where f is defined as in Equation 10.

Equation 11 and Equation 12 or 13 make it easy to construct the appropriate scattering circle for any given mass ratio A. One simply uses Equation 11 to find the circle center at $(x_s,0^\circ)$ and then draws a circle of radius r_s. The resulting scattering circle can then be used to find the energy of the scattered particle at any scattering angle by drawing a line from the origin to the circle at the selected angle. The length of the line is $\sqrt{(E_s)}$.

Similarly, the polar coordinates for recoiling are $([\sqrt{(E_r)}],\theta_r)$, where $E_r = E_2/E_0$. A typical elastic recoiling circle is shown in Figure 3. The recoiling circle passes through the origin, corresponding to the case where no collision occurs and the target remains at rest. The circle center, x_r, is located at:

$$x_r = \frac{\sqrt{A}}{1+A} \tag{14}$$

and its radius is $r_r = x_r$ for elastic recoiling or

$$r_r = fx_r = \frac{f\sqrt{A}}{1+A} \tag{15}$$

for inelastic recoiling.

Recoiling circles can be readily constructed for any collision partners using the above equations. For elastic collisions ($f = 1$), the construction is trivial, as the recoiling circle radius equals its center distance.

Figure 4 shows elastic scattering and recoiling circles for a variety of mass ratio A values. Since the circles are symmetric about the horizontal (0°) direction, only semicircles are plotted (scattering curves in the upper half plane and recoiling curves in the lower quadrant). Several general properties of binary collisions are evident. First,

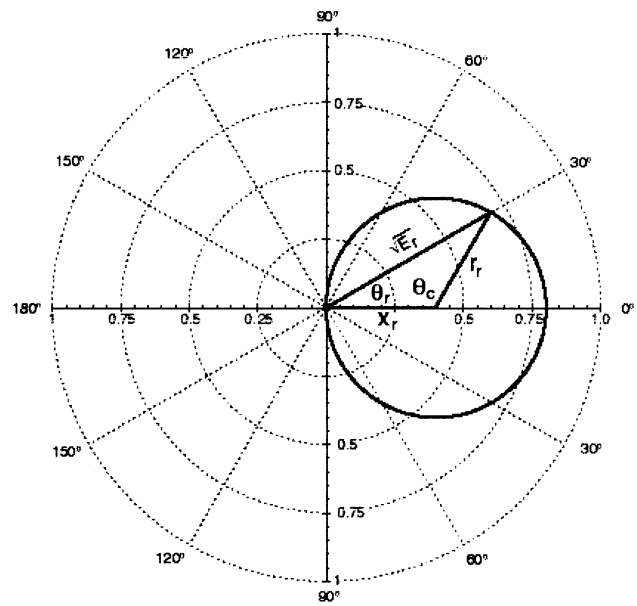


Figure 3. An elastic recoiling circle plotted in polar coordinates $(\sqrt{E},\theta)$ where E is $E_r = E_2/E_0$ and θ is the laboratory recoiling angle, θ_r. The length of the line segment from the origin to a point on the circle gives $\sqrt{(E_r)}$ at that angle. Recoiling circles are centered at $(x_r,0^\circ)$, where $x_r = \sqrt{A}/(1 + A)$. Note that $x_r = r_r$. All elastic recoiling circles pass through the origin. The circle shown is for the case $A = 4$. The triangle illustrates the relationship $\theta_r = (\pi - \theta_c)/2$.

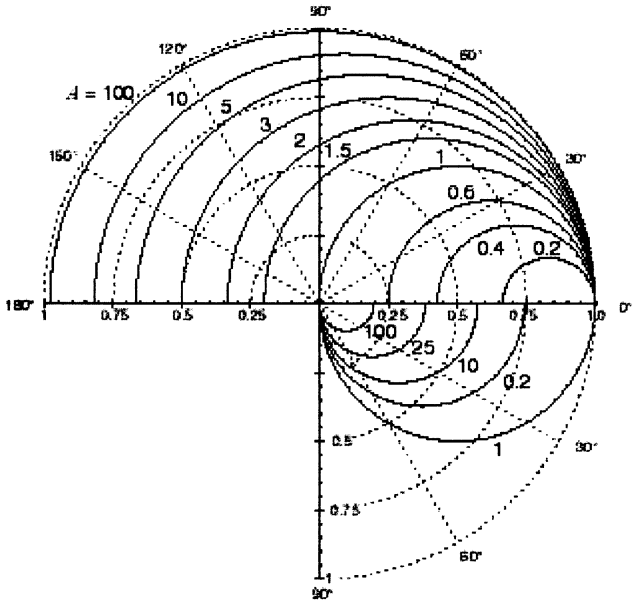


Figure 4. Elastic scattering and recoiling diagram for various values of A. For scattering, $\sqrt{(E_s)}$ versus θ_s is plotted for A values of 0.2, 0.4, 0.6, 1, 1.5, 2, 3, 5, 10, and 100 in the upper half-plane. When $A < 1$, only forward scattering is possible. For recoiling, $\sqrt{(E_r)}$ versus θ_r is plotted for A values of 0.2, 1, 10, 25, and 100 in the lower quadrant. Recoiling particles travel only in the forward direction.

for mass ratio $A > 1$ (i.e., light projectiles striking heavy targets), scattering at all angles $0^\circ < \theta_s \leq 180^\circ$ is permitted. When $A = 1$ (as in billiards, for instance) the scattering and recoil circles are the same. A head-on collision brings the projectile to rest, transferring the full projectile energy to the target. When $A < 1$ (i.e., heavy projectiles striking light targets), only forward scattering is possible and there is a limiting scattering angle, θ_{max}, which is found by drawing a tangent line from the origin to the scattering circle. The value of θ_{max} is arcsin A, because $\theta_s = r_s/x_s = A$. Note that there is a single scattering energy at each scattering angle when $A \geq 1$, but two energies are possible when $A < 1$ and $\theta_s < \theta_{max}$. This is illustrated in Figure 5. For all A, recoiling particles have only one energy and the recoiling angle $\theta_r < 90^\circ$. It is interesting to note that the recoiling circles are the same for A and A^{-1}, so it is not always possible to unambiguously identify the target mass by measuring its recoil energy. For example, using He projectiles, the energies of elastic H and O recoils at any selected recoiling angle are identical.

Center-of-Mass and Relative Coordinates

In some circumstances, such as when calculating collision cross-sections (see Central-Field Theory), it is useful to evaluate the scattering angle in the center of mass reference frame where the total momentum of the system is zero. This is an inertial reference frame with its origin located at the center of mass of the two particles. The center of mass moves in a straight line at constant velocity in the laboratory frame. The relative reference frame is also useful. It is a noninertial frame whose origin is located on the target. The relative frame is equivalent to a situation where a single particle of mass μ interacts with a fixed-point scattering center with the same potential as in the laboratory frame. In both these alternative frames of reference, the two-body collision problem reduces to a one-body problem. The relevant parameters are the reduced mass, μ, the relative energy, E_{rel}, and the center-of-mass scattering angle, θ_c. The term reduced mass originates from the fact that $\mu < m_1 + m_2$. The reduced mass is

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \tag{16}$$

and the relative energy is

$$E_{rel} = E_0\left(\frac{A}{1+A}\right) \tag{17}$$

The scattering angle θ_c is the same in the center-of-mass and relative reference frames. Scattering and recoiling circles show clearly the relationship between laboratory and center-of-mass scattering angles. In fact, the circles can readily be generated by parametric equations involving θ_c. These are simply $x = R\cos\theta_c + C$ and $y = R\sin\theta_c$, where R is the circle radius ($R = r_s$ for scattering, $R = r_r$ for recoiling) and $(C,0^\circ)$ is the location of its center ($C = x_s$ for scattering, $C = x_r$ for recoiling). Figures 2 and 3 illustrate the relationships among θ_s, θ_r, and θ_c. The relationship between θ_c and θ_s can be found by examining the triangle in Figure 2 containing θ_s and having sides of lengths x_s, r_s, and v_s. Applying the law of sines gives, for elastic scattering,

$$\tan\theta_s = \frac{\sin\theta_c}{A^{-1} + \cos\theta_c} \tag{18}$$

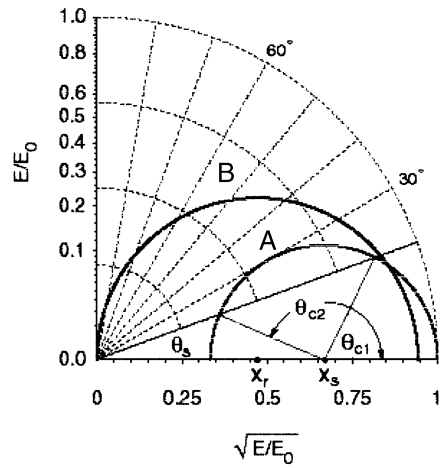


Figure 5. Elastic scattering (**A**) and recoiling (**B**) diagrams for the case $A = 1/2$. Note that in this case scattering occurs only for $\theta_s \leq 30^\circ$. In general, $\theta_s \leq \theta_{max} = \arcsin A$, when $A < 1$. Below θ_{max}, two scattered particle energies are possible at each laboratory observing angle. The relationships among θ_{c1}, θ_{c2}, and θ_s are shown at $\theta_s = 20^\circ$.

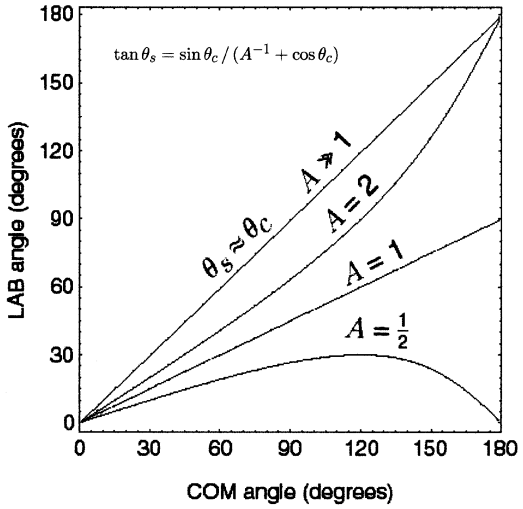


Figure 6. Correspondence between the center-of-mass scattering angle, θ_c, and the laboratory scattering angle, θ_s, for elastic collisions having various values of A: 0.5, 1, 2, and 100. For $A = 1$, $\theta_s = \theta_c/2$. For $A \gg 1$, $\theta_s \approx \theta_c$. When $A < 1$, $\theta_{max} = \arcsin A$.

Inspection of the elastic recoiling circle in Figure 3 shows that

$$\theta_r = \frac{1}{2}(\pi - \theta_c) \tag{19}$$

The relationship is apparent after noting that the triangle including θ_c and θ_r is isosceles. The various conversions between these three angles for elastic and inelastic collisions are listed in Appendix at the end of this unit (see Conversions among θ_s, θ_r, and θ_c for Nonrelativistic Collisions).

Two special cases are worth mentioning. If $A = 1$, then $\theta_c = 2\theta_s$; and as $A \to \infty$, $\theta_c \to \theta_s$. These, along with intermediate cases, are illustrated in Figure 6.

Relativistic Collisions

When the velocity of the projectile is a substantial fraction of the speed of light, relativistic effects occur. The effect most clearly seen as the projectile's velocity increases is distortion of the scattering circle into an ellipse. The relativistic parameter or Lorentz factor, γ, is defined as:

$$\gamma = (1 - \beta^2)^{-1/2} \tag{20}$$

where β, called the reduced velocity, is v_0/c and c is the speed of light. For all atomic projectiles with kinetic energies <1 MeV, γ is effectively 1. But at very high projectile velocities. γ exceeds 1, and the mass of the projectile increases to γm_0, where m_0 is the rest mass of the particle. For example, $\gamma = 1.13$ for a 100-MeV hydrogen projectile. It is when γ is appreciably greater than 1 that the scattering curve becomes elliptical. The center of the ellipse is located at the distance x_s on the horizontal axis in the scattering diagram, given by

$$x_s = \frac{1 + A\gamma}{1 + 2A\gamma + A^2} \tag{21}$$

This definition of x_s is consistent with the earlier, nonrelativistic definition since, when $\gamma = 1$, the center is as given in Equation 11. The major axis of the ellipse for elastic collisions is

$$a = \frac{A(\gamma + A)}{1 + 2A\gamma + A^2} \tag{22}$$

and the minor axis is

$$b = \frac{A}{(1 + 2A\gamma + A^2)^{1/2}} \tag{23}$$

When $\gamma = 1$,

$$a = b = r_s = \frac{A}{1 + A} \tag{24}$$

which indicates that the ellipse turns into the familiar elastic scattering circle under nonrelativistic conditions.

The foci of the ellipse are located at positions $x_s - d$ and $x_s + d$ along the horizontal axis of the scattering diagram, where d is given by

$$d = \frac{A(\gamma^2 - 1)^{1/2}}{1 + 2A\gamma + A^2} \tag{25}$$

The eccentricity of the ellipse, ε, is

$$\varepsilon = \frac{d}{a} = \frac{(\gamma^2 - 1)^{1/2}}{A + \gamma} \tag{26}$$

Examples of relativistic scattering curves are shown in Figure 7.

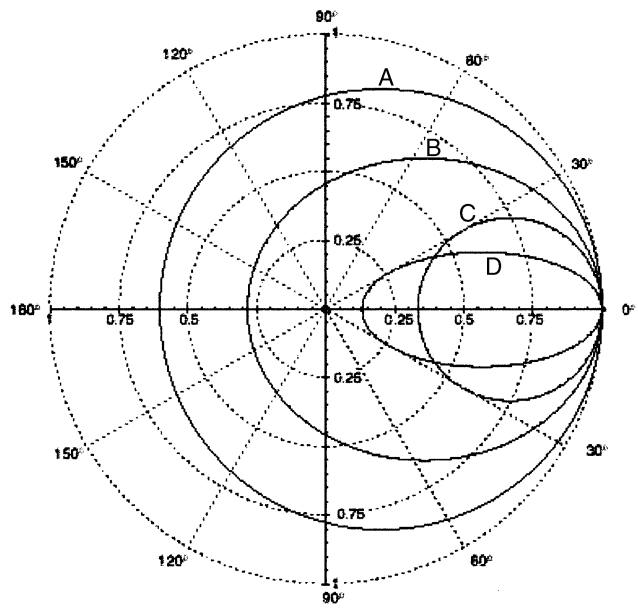


Figure 7. Classical and relativistic scattering diagrams for $A = 1/2$ and $A = 4$. (**A**) $A = 4$ and $\gamma = 1$, (**B**) $A = 4$ and $\gamma = 4.5$, (**C**) $A = 0.5$ and $\gamma = 1$, (**D**) $A = 0.5$ and $\gamma = 4.5$. Note that θ_{max} does not depend on γ.

To understand how the relativistic ellipses are related to the normal scattering circles, consider the positions of the centers. For a given $A > 1$, one finds that $x_s(e) > x_s(c)$, where $x_s(e)$ is the location of the ellipse center and $x_s(c)$ is the location of the circle center. When $A = 1$, then it is always true that $x_s(e) = x_s(c) = 1/2$. And finally, for a given $A < 1$, one finds that $x_s(e) < x_s(c)$. This last inequality has an interesting consequence. As γ increases when $A < 1$, the center of the ellipse moves towards the origin, the ellipse itself becomes more eccentric, and one finds that θ_{max} does not change. The maximum allowed scattering angle when $A < 1$ is always arcsin A. This effect is diagrammed in Figure 7.

For inelastic relativistic collisions, the center of the scattering ellipse remains unchanged from the elastic case (Equation 21). However, the major and minor axes are reduced. A practical way of plotting the ellipse is to use its parametric definition, which is $x = r_s(a)\cos\theta_c + x_s$ and $y = r_s(b)\sin\theta_c$, where

$$r_s(a) = a\sqrt{1 - Q_n/a} \tag{27}$$

and

$$r_s(b) = b\sqrt{1 - Q_n/b} \tag{28}$$

As in the nonrelativistic classical case, the relativistic scattering curves allow one to easily determine the scattered particle velocity and energy at any allowed scattering angle.

In a similar manner, the recoiling curves for relativistic particles can be stated as a straightforward extension of the classical recoiling curves.

Nuclear Reactions

Scattering and recoiling circle diagrams can also depict the kinematics of simple nuclear reactions in which the colliding particles undergo a mass change. A nuclear reaction of the form $\mathbf{A}(\mathbf{b},\mathbf{c})\mathbf{D}$ can be written as $\mathbf{A} + \mathbf{b} \rightarrow \mathbf{c} + \mathbf{D} + Q_{mass}/c^2$, where the mass difference is accounted for by Q_{mass}, usually referred to as the "Q value" for the reaction. The sign of the Q value is conventionally taken as positive for a kinetic energy-producing (exoergic) reaction and negative for a kinetic energy-driven (endoergic) reaction. It is important to distinguish between Q_{mass} and the inelastic energy factor Q introduced in Equation 6. The difference is that Q_{mass} balances the mass in the above equation for the nuclear reaction, while Q balances the kinetic energy in Equation 6. These values are of opposite sign: i.e., $Q = -Q_{mass}$. To illustrate, for an exoergic reaction ($Q_{mass} > 0$), some of the internal energy (e.g., mass) of the reactant particles is converted to kinetic energy. Hence the internal energy of the system is reduced and Q is negative in sign.

For the reaction $\mathbf{A}(\mathbf{b},\mathbf{c})\mathbf{D}$, $\mathbf{A}$ is considered to be the target nucleus, $\mathbf{b}$ the incident particle (projectile), $\mathbf{c}$ the outgoing particle, and $\mathbf{D}$ the recoil nucleus. Let $m_{\mathbf{A}}, m_{\mathbf{b}}, m_{\mathbf{c}}$, and $m_{\mathbf{D}}$ be the corresponding masses and $v_{\mathbf{b}}, v_{\mathbf{c}}$, and $v_{\mathbf{D}}$ be the corresponding velocities ($v_{\mathbf{A}}$ is assumed to be zero). We now define the mass ratios A_1 and A_2 as $A_1 = m_{\mathbf{c}}/m_{\mathbf{b}}$ and $A_2 = m_{\mathbf{D}}/m_{\mathbf{b}}$ and the velocity ratios v_{n1} and v_{n2} as $v_{n1} = v_{\mathbf{c}}/v_{\mathbf{b}}$ and $v_{n2} = v_{\mathbf{D}}/v_{\mathbf{b}}$. Note that A_2 is equivalent to the previously defined target-to-projectile mass ratio A. Applying the energy and momentum conservation laws yields the fundamental kinematic relation for particle $\mathbf{c}$, which is:

$$2\cos\theta_1 = (A_1 + A_2)v_{n1} + \frac{1 - A_2(1 - Q_n)}{A_1 v_{n1}} \tag{29}$$

where θ_1 is the emission angle of particle $\mathbf{c}$ with respect to the incident particle direction. As mentioned above, the normalized inelastic energy factor, Q_n, is Q/E_0, where E_0 is the incident particle kinetic energy. In a similar manner, the fundamental relation for particle $\mathbf{D}$ is found to be

$$2\cos\theta_2 = \frac{(A_1 + A_2)v_{n2}}{\sqrt{A_2}} + \frac{1 - A_1(1 - Q_n)}{\sqrt{A_2}v_{n2}} \tag{30}$$

where θ_2 is the emission angle of particle $\mathbf{D}$.

Equations 29 and 30 can be combined into a single expression for the energy of the products:

$$E = A_i \left[\frac{1}{A_1 + A_2}\left(\cos\theta \pm \sqrt{\cos^2\theta - \frac{(A_1 + A_2)[1 - A_j(1 - Q_n)]}{A_i}}\right)\right]^2 \tag{31}$$

where the variables are assigned according to Table 1.

Equation 31 is a generalization of Equation 10, and its symmetry with respect to the two product particles is noteworthy. The symmetry arises from the common origin of the products at the instance of the collision, at which point they are indistinguishable.

In analogy to the results discussed above (see discussion of Scattering and Recoiling Diagrams), the expressions of Equation 31 describe circles in polar coordinates $(\sqrt{E}, \theta)$. Here the circle center x_1 is given by

$$x_1 = \frac{\sqrt{A_1}}{A_1 + A_2} \tag{32}$$

and the circle center x_2 is given by

$$x_2 = \frac{\sqrt{A_2}}{A_1 + A_2} \tag{33}$$

The circle radius r_1 is

$$r_1 = x_2\sqrt{(A_1 + A_2)(1 - Q_n) - 1} \tag{34}$$

Table 1. Assignment of Variables in Equation 31

	Product Particle	
Variable	**c**	D
E	E_1/E_0	E_2/E_0
θ	q_1	θ_2
A_i	A_1	A_2
A_j	A_2	A_1

and the circle radius r_2 is

$$r_2 = x_1\sqrt{(A_1 + A_2)(1 - Q_n) - 1} \tag{35}$$

Polar ($\sqrt{E}, \theta$) diagrams can be easily constructed using these definitions. In this case, the terms light product and heavy product should be used instead of scattering and recoiling, since the reaction products originate from a compound nucleus and are distinguished only by their mass. Note that Equations 29 and 30 become equivalent to Equations 7 and 8 if no mass change occurs, since if $m_{\mathbf{b}} = m_{\mathbf{c}}$, then $A_1 = 1$. Similarly, Equations. 32 and 33 and Equations 34 and 35 are extensions of Equations. 11, 13, 14, and 16, respectively. It is also worth noting that the initial target mass, $m_{\mathbf{A}}$, does not enter into any of the kinematic expressions, since a body at rest has no kinetic energy or momentum.

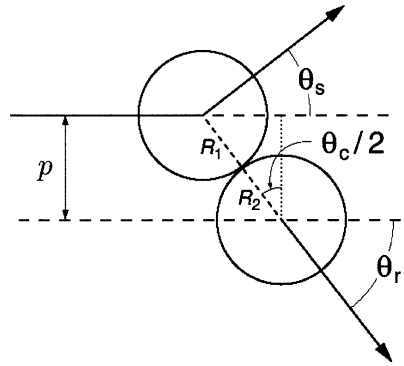


Figure 8. Geometry for hard-sphere collisions in a laboratory reference frame. The spheres have radii R_1 and R_2. The impact parameter, p, is the minimum separation distance between the particles along the projectile's path if no deflection were to occur. The example shown is for $R_1 = R_2$ and $A = 1$ at the moment of impact. From the triangle, it is apparent that $p/D = \cos(\theta_c/2)$, where $D = R_1 + R_2$.

CENTRAL-FIELD THEORY

While kinematics tells us how energy is apportioned between two colliding particles for a given scattering or recoiling angle, it tells us nothing about how the alignment between the collision partners determines their final trajectories. Central-field theory provides this information by considering the interaction potential between the particles. This section begins with a discussion of interaction potentials, then introduces the notion of an impact parameter, which leads to the formulation of the deflection function and the evaluation of collision cross-sections.

Interaction Potentials

The form of the interaction potential is of prime importance to the accurate representation of atomic scattering. The appropriate form depends on the incident kinetic energy of the projectile, E_0, and on the collision partners. When E_0 is on the order of the energy of chemical bonds ($\approx$1 eV), a potential that accounts for chemical interactions is required. Such potentials frequently consist of a repulsive term that operates at short distances and a long-range attractive term. At energies above the thermal and hyperthermal regimes (>100 eV), atomic collisions can be modeled using screened Coulomb potentials, which consider the Coulomb repulsion between nuclei attenuated by electronic screening effects. This energy regime extends up to some tens or hundreds of keV. At higher E_0, the interaction potential becomes practically Coulombic in nature and can be cast in a simple analytic form. At still higher energies, nuclear reactions can occur and must be considered. For particles with very high velocities, relativistic effects can dominate. Table 2 summarizes the potentials commonly used in various energy regimes. In the following, we will consider only central potentials, $V(r)$, which are spherically symmetric and depend only upon the distance between nuclei. In many materials-analysis applications, the energy of the interacting particles is such that pure or screened Coulomb central potentials prove highly useful.

Impact Parameters

The impact parameter is a measure of the alignment of the collision partners and is the distance of closest approach between the two particles in the absence of any forces. Its measure is the perpendicular distance between the projectile's initial direction and the parallel line passing through the target center. The impact parameter p is shown in Figure 8 for a hard-sphere binary collision. The impact parameter can be defined in a similar fashion for any binary collision; the particles can be point-like or have a physical extent. When $p = 0$, the collision is head on. For hard spheres, when p is greater than the sum of the spheres' radii, no collision occurs.

The impact parameter is similarly defined for scattering in the relative reference frame. This is illustrated in

Table 2. Interatomic Potentials Used in Various Energy Regimes[a,b]

Regime	Energy Range	Applicable Potential	Comments
Thermal	<1 eV	Attractive/repulsive	Van der Waals attraction
Hyperthermal	1–100 eV	Many body	Chemical reactions
Low	100 eV–10 keV	Screened Coulomb	Binary collisions
Medium	10 keV–1 MeV	Screened/pure Coulomb	Rutherford scattering
High	1–100 MeV	Coulomb	Nuclear reactions
Relativistic	>100 MeV	Liènard-Wiechert	Pair production

[a] Boundaries between regimes are approximate and depend on the characteristics of the collision partners.
[b] Below the Bohr electron velocity, $e^2/\hbar = 2.2 \times 10^6$ m/s, ionization and neutralization effects can be significant.

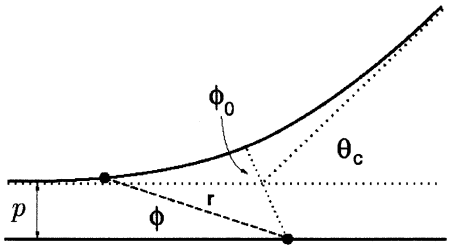


Figure 9. Geometry for scattering in the relative reference frame between a particle of mass μ and a fixed point target with a replusive force acting between them. The impact parameter p is defined as in Figure 8. The actual minimum separation distance is larger than p, and is referred to as the apsis of the collision. Also shown are the orientation angle ϕ and the separation distance r of the projectile as seen by an observer situated on the target particle. The apsis, r_0, occurs at the orientation angle ϕ_0. The relative scattering angle, shown as θ_c, is identical to the center-of-mass scattering angle. The relationship $\theta_c = |\pi - 2\phi_0|$ is apparent by summing the angles around the projectile asymptote at the apsis.

Figure 9 for a collision between two particles with a repulsive central force acting on them. For a given impact parameter, the final trajectory, as defined by θ_c, depends on the strength of the potential field. Also shown in the figure is the actual distance of closest approach, or apsis, which is larger than p for any collision involving a repulsive potential.

Shadow Cones

Knowing the interaction potential, it is straightforward, though perhaps tedious, to calculate the trajectories of a projectile and target during a collision, given the initial state of the system (coordinates and velocities). One does this by solving the equations of motion incrementally. With a sufficiently small value for the time step between calculations and a large number of time steps, the correct trajectory emerges. This is shown in Figure 10, for a representative atomic collision at a number of impact parameters. Note the appearance of a shadow cone, a region inside of which the projectile is excluded regardless of the impact parameter. Many weakly deflected projectile trajectories pass near the shadow cone boundary, leading to a flux-focusing effect. This is a general characteristic of collisions with $A > 1$.

The shape of the shadow cone depends on the incident particle energy and the interaction potential. For a pure Coulomb interaction, the shadow cone (in an axial plane) forms a parabola whose radius, $\hat{r}$ is given by $\hat{r} = 2(\hat{b}\hat{l})^{1/2}$ where $\hat{b} = Z_1 Z_2 e^2/E_0$ and $\hat{l}$ is the distance beyond the target particle. The shadow cone narrows as the energy of the incident particles increases. Approximate expressions for the shadow-cone radius can be used for screened Coulomb interaction potentials, which are useful at lower particle energies. Shadow cones can be utilized by ion-beam analysis methods to determine the surface structure of crystalline solids.

Deflection Functions

A general objective is to establish the relationship between the impact parameter and the scattering and recoiling angles. This relationship is expressed by the deflection function, which gives the center-of-mass scattering angle in terms of the impact parameter. The deflection function is of considerable practical use. It enables one to calculate collision cross-sections and thereby relate the intensity of scattering or recoiling with the interaction potential and the number of particles present in a collision system.

Deflection Function for Hard Spheres

The deflection function can be most simply illustrated for the case of hard-sphere collisions. Hard-sphere collisions have an interaction potential of the form

$$V(r) = \begin{cases} \infty & \text{when} \quad 0 \le p \le D \\ 0 & \text{when} \quad p > D \end{cases} \tag{36}$$

where D, called the collision diameter, is the sum of the projectile and target sphere radii R_1 and R_2. When p is greater than D, no collision occurs.

A diagram of a hard-sphere collision at the moment of impact is shown in Figure 8. From the geometry, it is seen that the deflection function for hard spheres is

$$\theta_c(p) = \begin{cases} 2\arccos(p/D) & \text{when} \quad 0 \le p \le D \\ 0^\circ & \text{when} \quad p > D \end{cases} \tag{37}$$

For elastic billiard ball collisions ($A = 1$), the deflection function expressed in laboratory coordinates using Equations 18 and 19 is particularly simple. For the projectile it is

$$\theta_s = \arccos(p/D) \qquad 0 \le p \le D, \quad A = 1 \tag{38}$$

and for the target

$$\theta_r = \arcsin(p/D) \qquad 0 \le p \le D \tag{39}$$

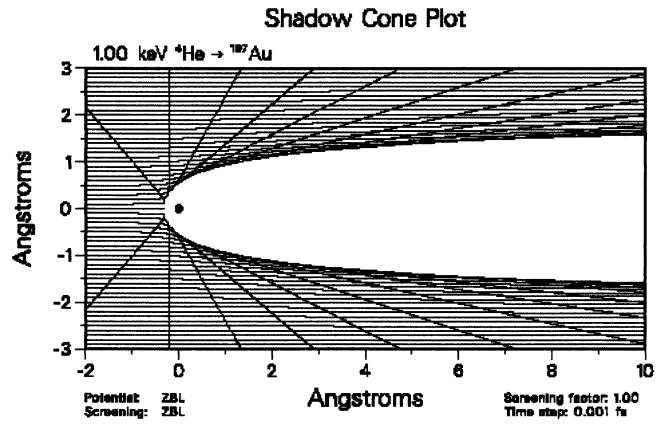


Figure 10. A two-dimensional representation of a shadow cone. The trajectories for a 1-keV ^{4}He atom scattering from a ^{197}Au target atom are shown for impact parameters ranging from +3 to −3 Å in steps of 0.1 Å. The ZBL interaction potential was used. The trajectories lie outside a parabolic shadow region. The full shadow cone is three dimensional and has rotational symmetry about its axis. Trajectories for the target atom are not shown. The dot marks the initial target position, but does not represent the size of the target nucleus, which would appear much smaller at this scale.

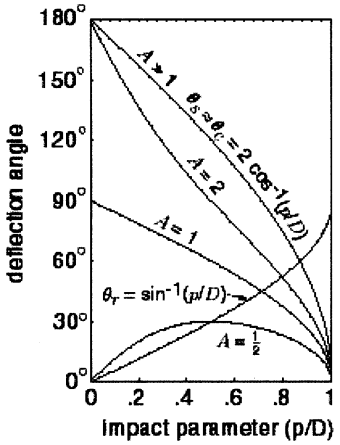


Figure 11. Deflection function for hard-sphere collisions. The center-of-mass deflection angle, θ_c is given by $2\cos^{-1}(p/D)$, where p is the impact parameter (see Fig. 8) and D is the collision diameter (sum of particle radii). The scattering angle in the laboratory frame, θ_s, is given by Equation 37 and is plotted for A values of 0.5, 1, 2, and 100. When $A = 1$, it equals $\cos^{-1}(p/D)$. At large A, it converges to the center-of-mass function. The recoiling angle in the laboratory frame, θ_r, is given by $\sin^{-1}(p/D)$ and does not depend on A.

When $A = 1$, the projectile and target trajectories after the collision are perpendicular.

For $A \neq 1$, the laboratory deflection function for the projectile is not as simple:

$$\theta_s = \arccos\left(\frac{1 - A + 2A(p/D)^2}{\sqrt{1 - 2A + A^2 + 4A(p/D)^2}}\right) \quad 0 \leq p \leq D \tag{40}$$

In the limit, as $A \to \infty$, $\theta_s \to 2\arccos(p/D)$. In contrast, the laboratory deflection function for the target recoil is independent of A and Equation 39 applies for all A. The hard-sphere deflection function is plotted in Figure 11 in both coordinate systems for selected A values.

Deflection Function for Other Central Potentials

The classical deflection function, which gives the center-of-mass scattering angle θ_c as a function of the impact parameter p, is

$$\theta_c(p) = \pi - 2p\int_{r_0}^{\infty} \frac{1}{r^2 f(r)^{1/2}}\,dr \tag{41}$$

where

$$f(r) = 1 - \frac{p^2}{r^2} - \frac{V(r)}{E_{rel}} \tag{42}$$

and $f(r_0) = 0$.

The physical meanings of the variables used in these expressions, for the case of two interacting particles, are as follows:

r: particle separation distance;

r_0: distance at closest approach (turning point or apsis);

$V(r)$: the interaction potential;

E_{rel}: kinetic energy of the particles in the center-of-mass and relative coordinate systems (relative energy).

We will examine how the deflection function can be evaluated for various central potentials. When $V(r)$ is a simple central potential, the deflection function can be evaluated analytically. For example, suppose $V(r) = k/r$, where k is a constant. If $k < 0$, then $V(r)$ represents an attractive potential, such as gravity, and the resulting deflection function is useful in celestial mechanics. For example, in Newtonian gravity, $k = -Gm_1m_2$, where G is the gravitational constant and the masses of the celestial bodies are m_1 and m_2. If $k > 0$, then $V(r)$ represents a repulsive potential, such as the Coulomb field between like-charged atomic particles. For example, in Rutherford scattering, $k = Z_1Z_2e^2$, where Z_1 and Z_2 are the atomic numbers of the nuclei and e is the unit of elementary charge. Then the deflection function is exactly given by

$$\theta_c(p) = \pi - 2\arctan\left(\frac{2pE_{rel}}{k}\right) = 2\arctan\left(\frac{k}{2pE_{rel}}\right) \tag{43}$$

Another central potential for which the deflection function can be exactly solved is the inverse square potential. In this case, $V(r) = k/r^2$, and the corresponding deflection function is:

$$\theta_c(p) = \pi\left(1 - \frac{p}{\sqrt{p^2 + k/E_{rel}}}\right) \tag{44}$$

Although the inverse square potential is not, strictly speaking, encountered in nature, it is a rough approximation to a screened Coloumb field when $k = Z_1Z_2e^2$. Realistic screened Coulomb fields decrease even more strongly with distance than the k/r^2 field.

Approximation of the Deflection Function

In cases where $V(r)$ is a more complicated function, sometimes no analytic solution for the integral exists and the function must be approximated. This is the situation for atomic scattering at intermediate energies, where the appropriate form for $V(r)$ is given by:

$$V(r) = \frac{k}{r}\Phi(r) \tag{45}$$

$F(r)$ is referred to as a screening function. This form for $V(r)$ with $k = Z_1Z_2e^2$ is the screened Coulomb potential. The constant term e^2 has a value of 14.40 eV-Å ($e^2 = c\alpha$, where $= h/2\pi$, h is Planck's constant, and α is the fine-structure constant). Although the screening function is not usually known exactly, several approximations appear to be reasonably accurate. These approximate functions have the form

$$\Phi(r) = \sum_{i=1}^{n} a_i \exp\left(\frac{-b_i r}{l}\right) \tag{46}$$

where a_i, b_i, and l are all constants. Two of the better known approximations are due to Moliére and to Ziegler, Biersack, and Littmark (ZBL). For the Moliére approximation, $n = 3$, with

$$\begin{aligned} a_1 &= 0.35 \qquad b_1 = 0.3 \\ a_2 &= 0.55 \qquad b_2 = 1.2 \\ a_3 &= 0.10 \qquad b_3 = 6.0 \end{aligned} \tag{47}$$

and

$$l = \frac{1}{4}\left[\frac{(3\pi)^2}{2}\right]^{1/3} a_0\left(Z_1^{1/2} + Z_2^{1/2}\right)^{-2/3} \tag{48}$$

where

$$a_0 = \frac{}{m_e e^2} \tag{49}$$

In the above, l is referred to as the screening length (the form shown is the Firsov screening length), a_0 is the Bohr radius, and m_e is the rest mass of the electron.

For the ZBL approximation, $n = 4$, with

$$\begin{aligned} a_1 &= 0.18175 \qquad b_1 = 3.19980 \\ a_2 &= 0.50986 \qquad b_2 = 0.94229 \\ a_3 &= 0.28022 \qquad b_3 = 0.40290 \\ a_4 &= 0.02817 \qquad b_4 = 0.20162 \end{aligned} \tag{50}$$

and

$$l = \frac{1}{4}\left[\frac{(3\pi)^2}{2}\right]^{1/3} a_0\left(Z_1^{0.23} + Z_2^{0.23}\right)^{-1} \tag{51}$$

If no analytic form for the deflection integral exists, two types of approximations are popular. In many cases, analytic approximations can be devised. Otherwise, the function can still be evaluated numerically. Gauss-Mehler quadrature (also called Gauss-Chebyshev quadrature) is useful in such situations. To apply it, the change of variable $x = r_0/r$ is made. This gives

$$\theta_c(p) = \pi - 2\hat{p}\int_0^1 \frac{1}{\sqrt{1 - (\hat{p}x)^2 - \dfrac{V}{E}}}\,dx \tag{52}$$

where $\hat{p} = p/r_0$.

The Gauss-Mehler quadrature relation is

$$\int_{-1}^{1} \frac{g(x)}{\sqrt{(1-x^2)}}\,dx \doteq \sum_{i=1}^{n} w_i g(x_i) \tag{53}$$

where $w_i = \pi/n$ and

$$x_i = \cos\left[\frac{\pi(2i-1)}{2n}\right] \tag{54}$$

Letting

$$g(x) = \hat{p}\sqrt{\frac{1-x^2}{1-(\hat{p}x)^2 - V/E}} \tag{55}$$

it can be shown that

$$\theta_c(p) \doteq \pi\left[1 - \frac{2}{n}\sum_{i=1}^{n/2} g(x_i)\right] \tag{56}$$

This is a useful approximation, as it allows the deflection function for an arbitrary central potential to be calculated to any desired degree of accuracy.

Cross-Sections

The concept of a scattering cross-section is used to relate the number of particles scattered into a particular angle to the number of incident particles. Accordingly, the scattering cross-section is $d\sigma(\theta_c) = dN/n$, where dN is the number of particles scattered per unit time between the angles θ_c and $\theta_c + d\theta_c$, and n is the incident flux of projectiles. With knowledge of the scattering cross-section, it is possible to relate, for a given incident flux, the number of scattered particles to the number of target particles. The value of scattering cross-section depends upon the interaction potential and is expressed most directly using the deflection function.

The differential cross-section for scattering into a differential solid angle $d\Omega$ is

$$\frac{d\sigma(\theta_c)}{d\Omega} = \frac{p}{\sin(\theta_c)}\left|\frac{dp}{d\theta_c}\right| \tag{57}$$

Here the solid and plane angle elements are related by $d\Omega = 2\pi \sin(\theta_c)\, d\theta_c$. Hard-sphere collisions provide a simple example. Using the hard-sphere potential (Equation 36) and deflection function (Equation 37), one obtains $d\sigma(\theta_c)/d\Omega = D^2/4$. Hard-sphere scattering is isotropic in the center-of-mass reference frame and independent of the incident energy. For the case of a Coulomb interaction potential, one obtains the Rutherford formula:

$$\frac{d\sigma(\theta_c)}{d\Omega} = \left(\frac{Z_1 Z_2 e^2}{4E_{rel}}\right)^2 \frac{1}{\sin^4(\theta_c/2)} \tag{58}$$

This formula has proven to be exceptionally useful for ion-beam analysis of materials.

For the inverse square potential (k/r^2), the differential cross-section is given by

$$\frac{d\sigma(\theta_c)}{d\Omega} = \frac{k}{E_{rel}}\frac{\pi^2(\pi - \theta_c)}{\theta_c^2(2\pi - \theta_c)^2}\frac{1}{\sin(\theta_c)} \tag{59}$$

For other potentials, numerical techniques (e.g., Equation 56) are typically used for evaluating collision cross-sections. Equivalent forms of Equation 57, such as

$$\frac{d\sigma(\theta_c)}{d\Omega} = \frac{p}{\sin(\theta_c)}\left|\frac{d\theta_c}{dp}\right|^{-1} = \frac{dp^2}{2d(\cos\theta_c)} \tag{60}$$

show that computing the cross-section can be accomplished by differentiating the deflection function or its inverse.

Cross-sections are converted to laboratory coordinates using Equations 18 and 19. This gives, for elastic collisions,

$$\frac{d\sigma(\theta_s)}{d\omega} = \frac{(1 + 2A\cos\theta_c + A^2)^{3/2}}{A^2|(A + \cos\theta_c)|}\frac{d\sigma(\theta_c)}{d\Omega} \qquad (61)$$

for scattering and

$$\frac{d\sigma(\theta_r)}{d\omega} = 4\sin\left(\frac{\theta_c}{2}\right)\frac{d\sigma(\theta_c)}{d\Omega} \qquad (62)$$

for recoiling. Here, the differential solid angle element in the laboratory reference frame, $d\omega$, is $2\pi\sin(\theta)\,d\theta$ and θ is the laboratory observing angle, θ_s or θ_r. For conversions to laboratory coordinates for inelastic collisions, see Conversions among θ_s, θ_r, and θ_c for Nonrelativistic Collisions, in the Appendix at the end of this unit.

Examples of differential cross-sections in laboratory coordinates for elastic collisions are shown in Figures 12 and 13 as a function of the laboratory observing angle. Some general observations can be made. When $A > 1$, scattering is possible at all angles (0° to 180°) and the scattering cross-sections decrease uniformly as the projectile energy and laboratory scattering angle increase. Elastic recoiling particles are emitted only in the forward direction regardless of the value of A. Recoiling cross-sections decrease as the projectile energy increases, but increase with recoiling angle. When $A < 1$, there are two branches in the scattering cross-section curve. The upper branch (i.e., the one with the larger cross-sections) results from collisions with the larger p. The two branches converge at θ_{max}.

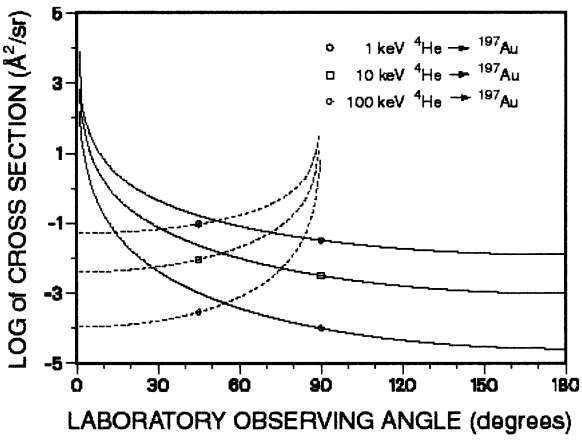


Figure 12. Differential atomic collision cross-sections in the laboratory reference frame for 1-, 10-, and 100-keV ^{4}He projectiles striking ^{197}Au target atoms as a function of the laboratory observing angle. Cross-sections are plotted for both the scattered projectiles (solid lines) and the recoils (dashed lines). The cross-sections were calculated using the ZBL screened Coulomb potential and Gauss-Mehler quadrature of the deflection function.

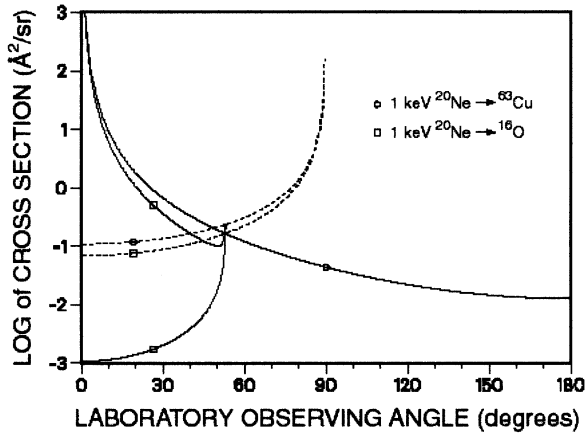


Figure 13. Differential atomic collision cross-sections in the laboratory reference frame for ^{20}Ne projectiles striking ^{63}Cu and ^{16}O target atoms calculated using ZBL interaction potential. Cross-sections are plotted for both the scattered projectiles (solid lines) and the recoils (dashed lines). The limiting angle for ^{20}Ne scattering from ^{16}O is 53.1°.

Applications to Materials Analysis

There are two general ways in which particle scattering theory is utilized in materials analysis. First, kinematics provides the connection between measurements of particle scattering parameters (velocity or energy, and angle) and the identity (mass) of the collision partners. A number of techniques analyze the energy of scattered or recoiled particles in order to deduce the elemental or isotopic identity of a substance. Second, central-field theory enables one to relate the intensity of scattering or recoiling to the amount of a substance present. When combined, kinematics and central-field theory provide exactly the tools needed to accomplish, with the proper measurements, compositional analysis of materials. This is the primary goal of many ion-beam methods, where proper selection of the analysis conditions enables a variety of extremely sensitive and accurate materials-characterization procedures to be conducted. These include elemental and isotopic composition analysis, structural analysis of ordered materials, two- and three-dimensional compositional profiles of materials, and detection of trace quantities of impurities in materials.

KEY REFERENCES

Behrisch, R. (ed). 1981. Sputtering by Particle Bombardment I. Springer-Verlag, Berlin.

Eckstein, W. 1991. Computer Simulation of Ion-Solid Interactions. Springer-Verlag, Berlin.

Eichler, J. and Meyerhof, W. E. 1995. Relativistic Atomic Collisions. Academic Press, San Diego.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. Elsevier Science Publishing, New York.

Goldstein, H. G. 1959. Classical Mechanics. Addison-Wesley, Reading, Mass.

Hagedorn, R. 1963. Relativistic Kinematics. Benjamin/Cummings, Menlo Park, Calif.

Johnson, R. E. 1982. Introduction to Atomic and Molecular Collisions. Plenum, New York.

Landau, L. D. and Lifshitz, E. M. 1976. Mechanics. Pergamon Press, Elmsford, N. Y.

Lehmann, C. 1977. Interaction of Radiation with Solids and Elementary Defect Production. North-Holland Publishing, Amsterdam.

Levine, R. D. and Bernstein, R. B. 1987. Molecular Reaction Dynamics and Chemical Reactivity. Oxford University Press, New York.

Mashkova, E. S. and Molchanov, V. A. 1985. Medium-Energy Ion Reflection from Solids. North-Holland Publishing, Amsterdam.

Parilis, E. S., Kishinevsky, L. M., Turaev, N. Y., Baklitzky, B. E., Umarov, F. F., Verleger, V. K., Nizhnaya, S. L., and Bitensky, I. S. 1993. Atomic Collisions on Solid Surfaces. North-Holland Publishing, Amsterdam.

Robinson, M. T. 1970. Tables of Classical Scattering Integrals. ORNL-4556, UC-34 Physics. Oak Ridge National Laboratory, Oak Ridge, Tenn.

Satchler, G. R. 1990. Introduction to Nuclear Reactions. Oxford University Press, New York.

Sommerfeld, A. 1952. Mechanics. Academic Press, New York.

Ziegler, J. F., Biersack, J. P., and Littmark, U. 1985. The Stopping and Range of Ions in Solids. Pergamon Press, Elmsford, N.Y.

ROBERT BASTASZ
Sandia National Laboratories
Livermore, California

WOLFGANG ECKSTEIN
Max-Planck-Institut für Plasmaphysik
Garching, Germany

APPENDIX

Solutions of Fundamental Scattering and Recoiling Relations in Terms of v, E, θ, A, and Q_n for Nonrelativistic Collisions

For elastic scattering:

$$v_s = \sqrt{E_s} = \frac{\cos\theta_s \pm \sqrt{A^2 - \sin^2\theta_s}}{1+A} \tag{63}$$

$$\theta_s = \arccos\left[\frac{(1+A)v_s}{2} + \frac{1-A}{2v_s}\right] \tag{64}$$

$$A = \frac{2(1 - v_s\cos\theta_s)}{1 - v_s^2} - 1 \tag{65}$$

For inelastic scattering:

$$v_s = \sqrt{E_s} = \frac{\cos\theta_s \pm \sqrt{A^2 f^2 - \sin^2\theta_s}}{1+A} \tag{66}$$

$$\theta_s = \arccos\left[\frac{(1+A)v_s}{2} + \frac{1 - A(1-Q_n)}{2v_s}\right] \tag{67}$$

$$A = \frac{1 + v_s^2 - 2v_s\cos\theta_s}{1 - v_s^2 - Q_n} = \frac{1 + E_s - 2\sqrt{E_s}\cos\theta_s}{1 - E_s - Q_n} \tag{68}$$

$$Q_n = 1 - \frac{1 - v_s[2\cos\theta_s - (1+A)v_s]}{A} \tag{69}$$

In the above relations,

$$f^2 = 1 - \frac{1+A}{A}Q_n \tag{70}$$

and $A = m_2/m_1, v_s = v_1/v_0, E_s = E_1E_0, Q_n = Q/E_0$, and θ_s is the laboratory scattering angle as defined in Figure 1.

For elastic recoiling:

$$v_r = \sqrt{\frac{E_r}{A}} = \frac{2\cos\theta_r}{1+A} \tag{71}$$

$$\theta_r = \arccos\left[\frac{(1+A)v_r}{2}\right] \tag{72}$$

$$A = \frac{2\cos\theta_r}{v_r} - 1 \tag{73}$$

For inelastic recoiling:

$$v_r = \sqrt{\frac{E_r}{A}} = \frac{\cos\theta_r \pm \sqrt{f^2 - \sin^2\theta_r}}{1+A} \tag{74}$$

$$\theta_r = \arccos\left[\frac{(1+A)v_r}{2} + \frac{Q_n}{2Av_r}\right] \tag{75}$$

$$A = \frac{2\cos\theta_r - v_r \pm \sqrt{(2\cos\theta_r - v_r)^2 - 4Q_n}}{2v_r} = \frac{(\cos\theta_r \pm \sqrt{\cos^2\theta_r - E_r - Q_n})^2}{E_r} \tag{76}$$

$$Q_n = Av_r[2\cos\theta_r - (1+A)v_r] \tag{77}$$

In the above relations,

$$f^2 = 1 - \frac{1+A}{A}Q_n \tag{78}$$

and $A = m_2/m_1$, $v_r = v_2/v_0$, $E_r = E_2/E_0$, $Q_n = Q/E_0$, θ_r is the laboratory recoiling angle as defined in Figure 1.

Conversions among θ_s, θ_r, and θ_c for Nonrelativistic Collisions

$$\theta_s = \arctan\left[\frac{\sin 2\theta_r}{(Af)^{-1} - \cos 2\theta_r}\right] = \arctan\left[\frac{\sin\theta_c}{(Af)^{-1} + \cos\theta_c}\right] \tag{79}$$

$$\theta_r = \arctan\left[\frac{\sin\theta_c}{f^{-1} - \cos\theta_c}\right] \tag{80}$$

$$\theta_r = \frac{1}{2}(\pi - \theta_c) \quad \text{for} \quad f = 1 \tag{81}$$

$$\theta_{c_1} = \theta_s + \arcsin\left[(Af)^{-1}\sin\theta_s\right] \tag{82}$$

$$\theta_{c_2} = 2\theta_s - \theta_{c_1} + \pi \qquad \text{for} \quad \sin\theta_s < Af < 1 \tag{83}$$

$$\frac{d\sigma(\theta_s)}{d\omega} = \frac{(1 + 2Af\cos\theta_c + (A^2f^2)^{3/2}}{A^2f^2|(Af + \cos\theta_c)|}\frac{d\sigma(\theta_c)}{d\Omega} \tag{84}$$

$$\frac{d\sigma(\theta_r)}{d\omega} = \frac{(1 - 2f\cos\theta_c + f^2)^{3/2}}{f^2|\cos\theta_c - f|}\frac{d\sigma(\theta_c)}{d\Omega} \tag{85}$$

In the above relations:

$$f = \sqrt{1 - \frac{1+A}{A} Q_n} \tag{86}$$

Note that: (1) $f = 1$ for elastic collisions; (2) when $A < 1$ and $\sin \theta_s$ A, two values of θ_c are possible for each θ_s; and (3) when $A = 1$ and $f = 1$, $(\tan \theta_s)(\tan \theta_r) = 1$.

Glossary of Terms and Symbols

α	Fine-structure constant ($\approx 7.3 \times 10^{-3}$)
A	Target to projectile mass ratio (m_2/m_1)
A_1	Ratio of product **c** mass to projectile **b** ($m_{\mathbf{c}}/m_{\mathbf{b}}$)
A_2	Ratio of product **D** mass to projectile **b** mass ($m_{\mathbf{D}}/m_{\mathbf{b}}$)
a	Major axis of scattering ellipse
a_0	Bohr radius ($\approx 29 \times 10^{-11}$ m)
β	Reduced velocity (v_0/c)
b	Minor axis of scattering ellipse
c	Velocity of light ($\approx 3.0 \times 10^8$ m/s)
d	Distance from scattering ellipse center to focal point
D	Collision diameter
$d\sigma$	Scattering cross-section
ε	Eccentricity of scattering ellipse
e	Unit of elementary charge ($\approx 1.602 \times 10^{-19}$ C)
E_0	Initial kinetic energy of projectile
E_1	Final kinetic energy of scattered projectile or product **c**
E_2	Final kinetic energy of recoiled target or product **D**
E_r	Normalized energy of the recoiled target (E_2/E_0)
E_{rel}	Relative energy
E_s	Normalized energy of the scattered projectile (E_1/E_0)
γ	Relativistic parameter (Lorentz factor)
h	Planck constant (4.136×10^{-15} eV-s)
l	Screening length
μ	Reduced mass ($\mu^{-1} = m_1^{-1} + m_2^{-1}$)
m_1, $m_{\mathbf{b}}$	Mass of projectile
m_2	Mass of recoiling particle (target)
$m_{\mathbf{A}}$	Initial target mass
$m_{\mathbf{c}}$	Light product mass
$m_{\mathbf{D}}$	Heavy product mass
m_e	Electron rest mass ($\approx 9.109 \times 10^{-31}$ kg)
p	Impact parameter
ϕ	Particle orientation angle in the relative reference frame
ϕ_0	Particle orientation angle at the apsis
Φ	Electron screening function
Q	Inelastic energy factor
Q_{mass}	Energy equivalent of particle mass change (Q value)
Q_n	Normalized inelastic energy factor (Q/E_0)
r	Particle separation distance
r_0	Distance of closest approach (apsis or turning point)
r_1	Radius of product **c** circle
r_2	Radius of product **D** circle
r_r	Radius of recoiling circle or ellipse
r_s	Radius of scattering circle or ellipse
R_1	Hard sphere radius of projectile
R_2	Hard sphere radius of target
θ_1	Emission angle of product **c** particle in the laboratory frame
θ_2	Emission angle of product **D** particle in the laboratory frame
θ_c	Center-of-mass scattering angle
θ_{max}	Maximum permitted scattering angle
θ_r	Recoiling angle of target in the laboratory frame
θ_s	Scattering angle of projectile in the laboratory frame
V	Interatomic interaction potential
v_0, $v_{\mathbf{b}}$	Initial velocity of projectile
v_1	Final velocity of scattered projectile
v_2	Final velocity of recoiled target
v_c	Velocity of light product
v_D	Velocity of heavy product
v_{n1}	Normalized final velocity of product **c** particle ($v_{\mathbf{c}}/v_{\mathbf{b}}$)
v_{n2}	Normalized final velocity of product **D** particle ($v_{\mathbf{D}}/v_{\mathbf{b}}$)
v_r	Normalized final velocity of target particle (v_2/v_0)
x_1	Position of product **c** circle or ellipse center
x_2	Position of product **D** circle or ellipse center
x_r	Position of recoiling circle or ellipse center
x_s	Position of scattering circle or ellipse center
Z_1	Atomic number of projectile
Z_2	Atomic number of target

SAMPLE PREPARATION FOR METALLOGRAPHY

INTRODUCTION

Metallography, the study of metal and metallic alloy structure, began at least 150 years ago with early investigations of the science behind metalworking. According to Rhines (1968), the earliest recorded use of metallography was in 1841(Anosov, 1954). Its first systematic use can be traced to Sorby (1864). Since these early beginnings, metallography has come to play a central role in metallurgical studies—a recent (1998) search of the literature revealed over 20,000 references listing metallography as a keyword!

Metallographic sample preparation has evolved from a black art to the highly precise scientific technique it is today. Its principal objective is the preparation of artifact-free representative samples suitable for microstructural examination. The particular choice of a sample preparation procedure depends on the alloy system and also on the focus of the examination, which could include process optimization, quality assurance, alloy design, deformation studies, failure analysis, and reverse engineering. The details of how to make the most appropriate choice and perform the sample preparation are the subject of this unit.

Metallographic sample preparation is divided broadly into two stages. The aim of the first stage is to obtain a planar, specularly reflective surface, where the scale of the artifacts (e.g., scratches, smears, and surface deformation)

is smaller than that of the microstructure. This stage commonly comprises three or four steps: sectioning, mounting (optional), mechanical abrasion, and polishing. The aim of the second stage is to make the microstructure more visible by enhancing the difference between various phases and microstructural features. This is generally accomplished by selective chemical dissolution or film formation—etching. The procedures discussed in this unit are also suitable (with slight modifications) for the preparation of metal and intermetallic matrix composites as well as for semiconductors. The modifications are primarily dictated by the specific applications, e.g., the use of coupled chemical-mechanical polishing for semiconductor junctions.

The basic steps in metallographic sample preparation are straightforward, although for each step there may be several options in terms of the techniques and materials used. Also, depending on the application, one or more of the steps may be elaborated or eliminated. This unit provides guidance on choosing a suitable path for sample preparation, including advice on recognizing and correcting an unsuitable choice.

This discussion assumes access to a laboratory equipped with the requisite equipment for metallographic sample preparation. Listings of typical equipment and supplies (see Table 1) and World Wide Web addresses for major commercial suppliers (see Internet Resources) are provided for readers wishing to start or upgrade a metallography laboratory.

STRATEGIC PLANNING

Before devising a procedure for metallographic sample preparation, it is essential to define the scope and objectives of the metallographic analysis and to determine the requirements of the sample. Clearly defined objectives

Table 1. Typical Equipment and Supplies for Preparation of Metallographic Samples

Application	Items required
Sectioning	Band saw
	Consumable-abrasive cutoff saw
	Low-speed diamond saw or continous-loop wire saw
	Silicon-carbide wheels (coarse and fine grade)
	Alumina wheels (coarse and fine grade)
	Diamond saw blades or wire saw wires
	Abrasive powders for wire saw (Silicon carbide, silicon nitride, boron nitride, alumina)
	Electric-discharge cutter (optional)
Mounting	Hot mounting press
	Epoxy and hardener dispenser
	Vacuum impregnation setup (optional)
	Thermosetting resins
	Thermoplastic resins
	Castable resins
	Special conductive mounting compounds
	Edge-retention additives
	Electroless-nickel plating solutions
Mechanical abrasion and polishing	Belt sander
	Two-wheel mechanical abrasion and polishing station
	Automated polishing head (medium-volume laboratory) or automated grinding and polishing system (high-volume laboratory)
	Vibratory polisher (optional)
	Paper-backed emery and silicon-carbide grinding disks (120, 180, 240, 320, 400, and 600 grit)
	Polishing cloths (napless and with nap)
	Polishing suspensions (15-,9-,6-, and 1-μm diamond; 0.3-μm α-alumina and 0.05-μm γ-alumina; colloidal silica; colloidal magnesia; 1200- and 3200-grit emery)
	Metal and resin-bonded-diamond grinding disks (optional)
Electropolishing (optional)	Commercial electropolisher (recommended)
	Chemicals for electropolishing
Etching	Fume hood and chemical storage cabinets
	Etching chemicals
Miscellaneous	Ultrasonic cleaner
	Stir/heat plates
	Compressed air supply (filtered)
	Specimen dryer
	Multimeter
	Acetone
	Ethyl and methyl alcohol
	First aid kit
	Access to the Internet
	Material safety data sheets for all applicable chemicals
	Appropriate reference books (see Key References)

may help to avoid many frustrating and unrewarding hours of metallography.

It also important to search the literature to see if a sample preparation technique has already been developed for the application of interest. It is usually easier to fine tune an existing procedure than to develop a new one.

Defining the Objectives

Before proceeding with sample preparation, the metallographer should formulate a set of questions, the answers to which will lead to a definition of the objectives. The list below is not exhaustive, but it illustrates the level of detail required.

1. Will the sample be used only for general microstructural evaluation?
2. Will the sample be examined with an electron microscope?
3. Is the sample being prepared for reverse engineering purposes?
4. Will the sample be used to analyze the grain flow pattern that may result from deformation or solidification processing?
5. Is the procedure to be integrated into a new alloy design effort, where phase identification and quantitative microscopy will be used?
6. Is the procedure being developed for quality assurance, where a large number of similar samples will be processed on a regular basis?
7. Will the procedure be used in failure analysis, requiring special techniques for crack preservation?
8. Is there a requirement to evaluate the composition and thickness of any coating or plating?
9. Is the alloy susceptible to deformation-induced damage such as mechanical twinning?

Answers to these and other pertinent questions will indicate the information that is already available and the additional information needed to devise the sample preparation procedure. This leads to the next step, a literature survey.

Surveying the Literature

In preparing a metallographic sample, it is usually easier to fine-tune an existing procedure, particularly in the final polishing and etching steps, than to develop a new one. Moreover, the published literature on metallography is exhaustive, and for a given application there is a high probability that a sample preparation procedure has already been developed; hence, a thorough literature search is essential. References provided later in this unit will be useful for this purpose (see Key References; see Internet Resources).

PROCEDURES

The basic procedures used to prepare samples for metallographic analysis are discussed below. For more detail, see *ASM Metals Handbook, Volume 9: Metallography and Microstructures* (ASM Handbook Committee, 1985) and Vander Voort (1984).

Sectioning

The first step in sample preparation is to remove a small representative section from the bulk piece. Many techniques are available, and they are discussed below in order of increasing mechanical damage:

Cutting with a continuous-loop wire saw causes the least amount of mechanical damage to the sample. The wire may have an embedded abrasive, such as diamond, or may deliver an abrasive slurry, such as alumina, silicon carbide, and boron nitride, to the root of the cut. It is also possible to use a combination of chemical attack and abrasive slurry.

This cutting method does not generate a significant amount of heat, and it can be used with very thin components. Another important advantage is that the correct use of this technique reduces the time needed for mechanical abrasion, as it allows the metallographer to eliminate the first three abrasion steps.

The main drawback is low cutting speed. Also, the proper cutting pressure often must be determined by trial-and-error.

Electric-discharge machining is extremely useful when cutting superhard alloys but can be used with practically any alloy. The damage is typically low and occurs primarily by surface melting. However, the equipment is not commonly available. Moreover, its improper use can result in melted surface layers, microcracking, and a zone of damage several millimeters below the surface.

Cutting with a nonconsumable abrasive wheel, such as a low-speed diamond saw, is a very versatile sectioning technique that results in minimal surface deformation. It can be used for specimens containing constituents with widely differing hardnesses. However, the correct use of an abrasive wheel is a trial-and-error process, as too much pressure can cause seizing and smearing.

Cutting with a consumable abrasive wheel is especially useful when sectioning hard materials. It is important to use copious amounts of coolant. However, when cutting specimens containing constituents with widely differing hardnesses, the softer constituents are likely to undergo selective ablation, which increases the time required for the mechanical abrasion steps.

Sawing is very commonly used and yields satisfactory results in most instances. However, it generates heat, so it is necessary to use copious amounts of cooling fluid when sectioning hard alloys; failure to do so can result in localized "burns" and microstructure alterations. Also, sawing can damage delicate surface coatings and cause "peel-back." It should not be used when an analysis of coated materials or coatings is required.

Shearing is typically used for sheet materials and wires. Although it is a fast procedure, shearing causes extremely heavy deformation, which may result in artifacts. Alternative techniques should be used if possible.

Fracturing, and in particular cleavage fracturing, may be used for certain alloys when it is necessary to examine a

crystallographically specific surface. In general, fracturing is used only as a last resort.

Mounting

After sectioning, the sample may be placed on a plastic mounting material for ease of handling, automated grinding and polishing, edge retention, and selective electropolishing and etching. Several types of plastic mounting materials are available; which type should be used depends on the application and the nature of the sample.

Thermosetting molding resins (e.g., bakelite, diallyl phthalate, and compression-mounting epoxies) are used when ease of mounting is the primary consideration. Thermoplastic molding resins (e.g., methyl methacrylate, PVC, and polystyrene) are used for fragile specimens, as the molding pressure is lower for these resins than for the thermosetting ones. Castable resins (e.g., acrylics, polyesters, and epoxies) are used when good edge retention and resistance to etchants is required. Additives can be included in the castable resins to make the mounts electrically conductive for electropolishing and electron microscopy. Castable resins also facilitate vacuum impregnation, which is sometimes required for powder metallurgical and failure analysis specimens.

Mechanical Abrasion

Mechanical abrasion typically uses abrasive particles bonded to a substrate, such as waterproof paper. Typically, the abrasive paper is placed on a platen that is rotated at 150 to 300 rpm. The particles cut into the specimen surface upon contact, forming a series of "vee" grooves. Successively finer grits (smaller particle sizes) of abrasive material are used to reduce the mechanically damaged layer and produce a surface suitable for polishing. The typical sequence is 120-, 180-, 240-, 320-, 400-, and 600-grit material, corresponding approximately to particle sizes of 106, 75, 52, 34, 22, and 14 μm. The principal abrasive materials are silicon carbide, emery, and diamond.

In metallographic practice, mechanical abrasion is commonly called grinding, although there are distinctions between mechanical abrasion techniques and traditional grinding. Metallographic mechanical abrasion uses considerably lower surface speeds (between 150 and 300 rpm) and a copious amount of fluid, both for lubrication and for removing the grinding debris. Thus, frictional heating of the specimen and surface damage are significantly lower in mechanical abrasion than in conventional grinding.

In mechanical abrasion, the specimen typically is held perpendicular to the grinding platen and moved from the edge to the center. The sample is rotated 90° with each change in grit size, in order to ensure that scratches from the previous operation are completely removed. With each move to finer particle sizes, the rule of thumb is to grind for twice the time used in the previous step. Consequently, it is important to start with the finest possible grit in order to minimize the time required. The size and grinding time for the first grit depends on the sectioning technique used. For semiautomatic or automatic operations, it is best to start with the manufacturer's recommended procedures and fine-tune them as needed.

Mechanical abrasion operations can also be carried out by rubbing the specimen on a series of stationary abrasive strips arranged in increasing fineness. This method is not recommended because of the difficulty in maintaining a flat surface.

Polishing

After mechanical abrasion, a sample is polished so that the surface is specularly reflective and suitable for examination with an optical or scanning electron microscope (SEM). Metallographic polishing is carried out both mechanically and electrolytically. In some cases, where etching is unnecessary or even undesirable—for example, the study of porosity distribution, the detection of cracks, the measurement of plating or coating thickness, and microlevel compositional analysis—polishing is the final step in sample preparation.

Mechanical polishing is essentially an extension of mechanical abrasion; however, in mechanical polishing, the particles are suspended in a liquid within the fibers of a cloth, and the wheel rotation speed is between 150 and 600 rpm. Because of the manner in which the abrasive particles are suspended, less force is exerted on the sample surface, resulting in shallower grooves. The choice of polishing cloth depends on the particular application. When the specimen is particularly susceptible to mechanical damage, a cloth with high nap is preferred. On the other hand, if surface flatness is a concern (e.g., edge retention) or if problems such as second-phase "pullout" are encountered, a napless cloth is the proper choice. Note that in selected applications a high-nap cloth may be used as a "backing" for a napless cloth to provide a limited amount of cushioning and retention of polishing medium. Typically, the sample is rotated continously around the central axis of the wheel, counter to the direction of wheel rotation, and the polishing pressure is held constant until nearly the end, when it is greatly reduced for the finishing touches. The abrasive particles used are typically diamond (6 μm and 1 μm), alumina (0.5 μm and 0.03 μm), and colloidal silica and colloidal magnesia.

When very high quality samples are required, rotating-wheel polishing is usually followed by vibratory polishing. This also uses an abrasive slurry with diamond, alumina, and colloidal silica and magnesia particles. The samples, usually in weighted holders, are placed on a platen which vibrates in such a way that the samples track a circular path. This method can be adapted for chemo-mechanical polishing by adding chemicals either to attack selected constituents or to suppress selective attack.

The end result of vibratory polishing is a specularly reflective surface that is almost free of deformation caused by the previous steps in the sample preparation process. Once the procedure is optimized, vibratory polishing allows a large number of samples to be polished simultaneously with reproducibly excellent quality.

Electrolytic polishing is used on a sample after mechanical abrasion to a 400- or 600-grit finish. It too produces a specularly reflective surface that is nearly free of deformation.

Electropolishing is commonly used for alloys that are hard to prepare or particularly susceptible to deformation artifacts, such as mechanical twinning in Mg, Zr, and Bi. Electropolishing may be used when edge retention is not required or when a large number of similar samples is expected, for example, in process control and alloy development.

The use of electropolishing is not widespread, however, as it (1) has a long development time; (2) requires special equipment; (3) often requires the use of highly corrosive, poisonous, or otherwise dangerous chemicals; and (4) can cause accelerated edge attack, resulting in an enlargement of cracks and porosity, as well as preferential attack of some constituent phases.

In spite these disadvantages, electropolishing may be considered because of its processing speed; once the technique is optimized for a particular application, there is none better or faster.

In electropolishing, the sample is set up as the anode in an electrolytic cell. The cathode material depends on the alloy being polished and the electrolyte: stainless steel, graphite, copper, and aluminum are commonly used. Direct current, usually from a rectified current source, is supplied to the electrolytic cell, which is equipped with an ammeter and voltmeter to monitor electropolishing conditions.

Typically, the voltage-current characteristics of the cell are complex. After an initial rise in current, an "electropolishing plateau" is observed. This plateau results from the formation of a "polishing film," which is a stable, high-resistance viscous layer formed near the anode surface by the dissolution of metal ions. The plateau represents optimum conditions for electropolishing: at lower voltages etching takes place, while at higher voltages there is film breakdown and gas evolution.

The mechanism of electropolishing is not well understood, but is generally believed to occur in two stages: smoothing and brightening. The smoothing stage is characterized by a preferential dissolution of the ridges formed by mechanical abrasion (primarily because the resistance at the peak is lower than in the valley). This results in the formation of the viscous polishing film. The brightening phase is characterized by the elimination of extremely small ridges, on the order of 0.01 μm.

Electropolishing requires the optimization of many parameters, including electrolyte composition, cathode material, current density, bath temperature, bath agitation, anode-to-cathode distance, and anode orientation (horizontal, vertical, etc.). Other factors, such as whether the sample should be removed before or after the current is switched off, must also be considered. During the development of an electropolishing procedure, the microstructure be should first be prepared by more conventional means so that any electropolishing artifacts can be identified.

Etching

After the sample is polished, it may be etched to enhance the contrast between various constituent phases and microstructural features. Chemical, electrochemical, and physical methods are available. Contrast on as-polished surfaces may also be enhanced by nondestructive methods, such as dark-field illumination and backscattered electron imaging (see GENERAL VACCUM TECHNIQUES).

In chemical and electrochemical etching, the desired contrast can be achieved in a number of ways, depending on the technique employed. Contrast-enhancement mechanisms include selective dissolution; formation of a film, whose thickness varies with the crystallographic orientation of grains; formation of etch pits and grooves, whose orientation and density depend on grain orientation; and precipitation etching. A variety of chemical mixtures are used for selective dissolution. Heat tinting—the formation of oxide film—and anodizing both produce films that are sensitive to polarized light.

Physical etching techniques, such as ion etching and thermal etching, depend on the selective removal of atoms.

When developing a particular etching procedure, it is important to determine the "etching limit," below which some microstructural features are masked and above which parts of the microstructure may be removed due to excessive dissolution. For a given etchant, the most important factor is the etching time. Consequently, it is advisable to etch the sample in small time increments and to examine the microstructure between each step. Generally the optimum etching program is evident only after the specimen has been over-etched, so at least one polishing-etching-polishing iteration is usually necessary before a properly etched sample is obtained.

ILLUSTRATIVE EXAMPLES

The particular combination of steps used in metallographic sample preparation depends largely on the application. A thorough literature survey undertaken before beginning sample preparation will reveal techniques used in similar applications. The four examples below illustrate the development of a successful metallographic sample preparation procedure.

General Microstructural Evaluation of 4340 Steel

Samples are to be prepared for the general microstructural evaluation of 4340 steel. Fewer than three samples per day of 1-in. (2.5-cm) material are needed. The general microstructure is expected to be tempered martensite, with a bulk hardness of HRC 40 (hardness on the Rockwell C scale). An evaluation of decarburization is required, but a plating-thickness measurement is not needed.

The following sample preparation procedure is suggested, based on past experience and a survey of the metallographic literature for steel.

Sectioning. An important objective in sectioning is to avoid "burning," which can temper the martensite and cause some decarburization. Based on the hardness and required thickness in this case, sectioning is best accomplished using a 60-grit, rubber-resin-bonded alumina wheel and cutting the section while it is submerged in a coolant. The cutting pressure should be such that the 1-in. samples can be cut in 1 to 2 min.

Mounting. When mounting the sample, the aim is to retain the edge and to facilitate SEM examination. A conducting epoxy mount is suggested, using an appropriate combination of temperature, pressure, and time to ensure that the specimen-mount separation is minimized.

Mechanical Abrasion. To minimize the time needed for mechanical abrasion, a semiautomatic polishing head with a three-sample holder should be used. The wheel speed should be 150 rpm. Grinding would begin with 180-grit silicon carbide, and continue in the sequence 240, 320, 400, and 600 grit. Water should be used as a lubricant, and the sample should be rinsed between each change in grit. Rotation of the sample holder should be in the sense counter to the wheel rotation. This process takes ~35 min.

Mechanical Polishing. The objective is to produce a deformation-free and specularly reflective surface. After mechanical abrasion, the sample-holder assembly should be cleaned in an ultrasonicator. Polishing is done with medium-nap cloths, using a 6-μm diamond abrasive followed by a 1-μm diamond abrasive. The holder should be cleaned in an ultrasonicator between these two steps. A wheel speed of 300 rpm should be used and the specimen should be rotated counter to the wheel rotation. Polishing requires 10 min for the first step and 5 min for the second step. (Duration decreases because successively lighter damage from previous steps requires shorter removal times in subsequent steps.)

Etching. The aim is to reveal the structure of the tempered martensite as well as any evidence of decarburization. Etching should begin with super picral for 30 s. The sample should be examined and then etched for an additional 10 s, if required. In developing this procedure, the samples were found to be over-etched at 50 s.

Measurement of Cadmium Plating Composition and Thickness on 4340 Steel

This is an extension of the previous example. It illustrates the manner in which an existing procedure can be modified slightly to provide a quick and reliable technique for a related application.

The measurement of plating composition and thickness requires a special edge-retention treatment due to the difference in the hardness of the cadmium plating and the bulk specimen. Minor modifications are also required to the polishing procedure due to the possibility of a selective chemical attack.

Based on a literature survey and past experience, the previous sample preparation procedure was modified to accommodate a measurement of plating composition and thickness.

Sectioning. When the sample is cut with an alumina wheel, several millimeters of the cadmium plating will be damaged below the cut. Hand grinding at 120 grit will quickly reestablish a sound layer of cadmium at the surface. An alternative would be to use a diamond saw for sectioning, but this would require a significantly longer cutting time.

After sectioning and before mounting, the sample should be plated with electroless nickel. This will surround the cadmium plating with a hard layer of nickel-sulfur alloy (HRC 60) and eliminate rounding of the cadmium plating during grinding.

Polishing. A buffered solution should be used during polishing to reduce the possibility of selective galvanic attack at the steel-cadmium interface.

Etching. Etching is not required, as the examination will be more exact on an unetched surface. The evaluation, which requires both thickness and compositional measurements, is best carried out with a scanning electron microscope equipped with an energy-dispersive spectroscope (EDS, see SYMMETRY IN CRYSTALLOGRAPHY).

Microstructural Evaluation of 7075-T6 Anodized Aluminum Alloy

Samples are required for the general microstructure evaluation of the aluminum alloy 7075-T6. The bulk hardness is HRB 80 (Rockwell B scale). A single 1/2-in.-thick (1.25-cm) sample will be prepared weekly. The anodized thickness is specified as 1 to 2 μm, and a measurement is required.

The following sample preparation procedure is suggested, based on past experience and a survey of the metallographic literature for aluminum.

Sectioning. The aim is to avoid excessive deformation. Based on the hardness and because the aluminum is anodized, sectioning should be done with a low-speed diamond saw, using copious quantities of coolant. This will take ~20 min.

Mounting. The goal is to retain the edge and to facilitate SEM examination. In order the preserve the thin anodized layer, electroless nickel plating is required before mounting. The anodized surface should be first brushed with an intermediate layer of colloidal silver paint and then plated with electroless nickel for edge retention.

A conducting epoxy mount should be used, with an appropriate combination of temperature, pressure, and time to ensure that the specimen-mount separation is minimized.

Mechanical Abrasion. Manual abrasion is suggested, with water as a lubricant. The wheel should be rotated at 150 rpm, and the specimen should be held perpendicular to the platen and moved from outer edge to center of the grinding paper. Grinding should begin with 320-grit silicon carbide and continue with 400- and 600-grit paper. The sample should be rinsed between each grit and turned 90°. The time needed is ~15 min.

Mechanical Polishing. The aim is to produce a deformation-free and specularly reflective surface. After mechanical abrasion, the holder should be cleaned in an

ultrasonicator. Polishing is accomplished using medium-nap cloths, first with a 0.5-μm α-alumina abrasive and then with a 0.03-μm γ-alumina abrasive. The holder should be cleaned in an ultrasonicator between these two steps. A wheel speed of 300 rpm should be used, and the specimen should be rotated counter to the wheel rotation. Polishing requires 10 min for the first step and 5 min for the second step.

SEM Examination. The objective is to image the anodized layer in backscattered electron mode and measure its thickness. This step is best accomplished using an as-polished surface.

Etching. Etching is required to reveal the microstructure in a T6 state (solution heat treated and artificially aged). Keller's reagent (2 mL 48% HF/3 mL concentrated HCl/5 mL concentrated HNO_3/190 mL H_2O) can be used to distinguish between T4 (solution heat treated and naturally aged to a substantially stable condition) and T6 heat treatment; supplementary electrical conductivity measurements will also aid in distinguishing between T4 and T6. The microstructure should also be checked against standard sources in the literature, however.

Microstructural Evaluation of Deformed High-Purity Aluminum

A sample preparation procedure is needed for a high volume of extremely soft samples that were previously deformed and partially recrystallized. The objective is to produce samples with no artifacts and to reveal the fine substructure associated with the thermomechanical history.

There was no in-house experience and an initial survey of the metallographic literature for high-purity aluminum did not reveal a previously developed technique. A broader literature search that included Ph.D. dissertations uncovered a successful procedure (Connell, 1972). The methodology is sufficiently detailed so that only slight in-house adjustments are needed to develop a fast and highly reliable sample preparation procedure.

Sectioning. The aim is to avoid excessive deformation of the extremely soft samples. A continuous-loop wire saw should be used with a silicon-carbide abrasive slurry. The 1/4-in. (0.6-cm) section will be cut in 10 min.

Mounting. In order to avoid any microstructural recovery effects, the sample should be mounted at room temperature. An electrical contact is required for subsequent electropolishing; an epoxy mount with an embedded electrical contact could be used. Multiple epoxy mounts should be cured overnight in a cool chamber.

Mechanical Abrasion. Semiautomatic abrasion and polishing is suggested. Grinding begins with 600-grit silicon carbide, using water as lubricant. The wheel is rotated at 150 rpm, and the sample is held counter to wheel rotation and rinsed after grinding. This step takes ~5 min.

Mechanical Polishing. The objective is to produce a surface suitable for electropolishing and etching. After mechanical abrasion, and between the two polishing steps, the holder and samples should be cleaned in an ultrasonicator. Polishing is accomplished using medium-nap cloths, first with 1200- and then 3200-mesh emery in soap solution. A wheel speed of 300 rpm should be used, and the holder should be rotated counter to the wheel rotation. Mechanical polishing requires 10 min for the first step and 5 min for the second step.

Electrolytic Polishing and Etching. The aim is to reveal the microstructure without metallographic artifacts. An electrolyte containing 8.2 cm^3 HF, 4.5 g boric acid, and 250 cm^3 deionized water is suggested. A chlorine-free graphite cathode should used, with an anode-cathode spacing of 2.5 cm and low agitation. The open circuit voltage should be 20 V. The time needed for polishing is 30 to 40 s with an additional 15 to 25 s for etching.

COMMENTARY

These examples emphasize two points. The first is the importance of a conducting thorough literature search before developing a new sample preparation procedure. The second is that any attempt to categorize metallographic procedures through a series of simple steps can misrepresent the field. Instead, an attempt has been made to give the reader an overview with selected examples of various complexity. While these metallographic sample preparation procedures were written with the layman in mind, the literature and Internet sources should be useful for practicing metallographers.

LITERATURE CITED

Anosov, P.P. 1954. Collected Works. Akademiya Nauk SSR, Moscow.

ASM Handbook Committee. 1985. ASM Metals Handbook Volume 9: Metallography and Microstructures. ASM International, Metals Park, Ohio.

Connell, R.G. Jr. 1972. The Microstructural Evolution of Aluminum During the Course of High-Temperature Creep. Ph.D. thesis, University of Florida, Gainesville.

Rhines, F.N. 1968. Introduction. *In* Quantitative Microscopy (R.T. DeHoff and F.N. Rhines, eds.) pp. 1-10. McGraw-Hill, New York.

Sorby, H.C. 1864. On a new method of illustrating the structure of various kinds of steel by nature printing. *Sheffield Lit. Phil. Soc.*, Feb. 1964.

Vander Voort, G. 1984. Metallography: Principle and Practice. McGraw-Hill, New York.

KEY REFERENCES

Books

Huppmann, W.J. and Dalal, K. 1986. Metallographic Atlas of Powder Metallurgy. Verlag Schmid. [Order from Metal Powder Industries Foundation, Princeton, N.J.]

Comprehensive compendium of powder metallurgical microstructures.

ASM Handbook Committee, 1985. See above.

The single most complete and authoritative reference on metallography. No metallographic sample preparation laboratory should be without a copy.

Petzow, G. 1978. Metallographic Etching. American Society for Metals, Metals Park, Ohio.

Comprehensive reference for etching recipes.

Samuels, L.E. 1982. Metallographic Polishing by Mechanical Methods, 3rd ed. American Society for Metals, Metals Park, Ohio.

Complete description of mechanical polishing methods.

Smith, C.S. 1960. A History of Metallography. University of Chicago Press, Chicago.

Excellent account of the history of metallography for those desiring a deeper understanding of the field's development.

Vander Voort, 1984. See above.

One of the most popular and thorough books on the subject.

Periodicals

Praktische Metallographie/Practical Metallography (bilingual German-English, monthly). Carl Hanser Verlag, Munich.

Metallography (English, bimonthly). Elsevier, New York.

Structure (English, German, French editions; twice yearly). Struers, Rodovre, Denmark.

Microstructural Science (English, yearly). Elsevier, New York.

INTERNET RESOURCES

NOTE: *Indicates a "must browse" site.

Metallography: General Interest

http://www.metallography.com/ims/info.htm

**International Metallographic Society. Membership information, links to other sites, including the virtual metallography laboratory, gallery of metallographic images, and more.*

http://www.metallography.com/index.htm

**The Virtual Metallography Laboratory. Extremely informative and useful; probably the most important site to visit.*

http://www.kaker.com

**Kaker d.o.o. Database of metallographic etches and excellent links to other sites. Database of vendors of microscopy products.*

http://www.ozelink.com/metallurgy

Metallurgy Books. Good site to search for metallurgy books online.

Standards

http://www.astm.org/COMMIT/e-4.htm

**ASTM E-4 Committee on Metallography. Excellent site for understanding the ASTM metallography committee activities. Good exposition of standards related to metallography and the philosophy behind the standards. Good links to other sites.*

http://www2.arnes.si/~sgszmera1/standard.html#main

**Academic and Research Network of Slovenia. Excellent site for list of worldwide standards related to metallography and microscopy. Good links to other sites.*

Microscopy and Microstructures

http://microstructure.copper.org

Copper Development Association. Excellent site for copper alloy microstructures. Few links to other sites.

http://www.microscopy-online.com

Microscopy Online. Forum for information exchange, links to vendors, and general information on microscopy.

http://www.mwrn.com

MicroWorld Resources and News. Annotated guide to online resources for microscopists and microanalysts.

http://www.precisionimages.com/gatemain.htm

Digital Imaging. Good background information on digital imaging technologies and links to other imaging sites.

http://www.microscopy-online.com

Microscopy Resource. Forum for information exchange, links to vendors, and general information on microscopy.

http://kelvin.seas.virginia.edu/~jaw/mse3101/w4/mse4-0.htm#Objectives

**Optical Metallography of Steel. Excellent exposition of the general concepts, by J.A. Wert*

Commercial Producers of Metallographic Equipment and Supplies

http://www.2spi.com/spihome.html

Structure Probe. Good site for finding out about the latest in electron microscopy supplies, and useful for contacting SPI's technical personnel. Good links to other microscopy sites.

http://www.lamplan.fr/ *or* mmsystem@lamplan.fr

LAM PLAN SA. Good site to search for Lam Plan products.

http://www.struers.com/default2.htm

**Struers. Excellent site with useful resources, online guide to metallography, literature sources, subscriptions, and links.*

http://www.buehlerltd.com/index2.html

Buehler. Good site to locate the latest Buehler products.

http://www.southbaytech.com

Southbay. Excellent site with many links to useful Internet resources, and good search engine for Southbay products.

Archaeometallurgy

http://masca.museum.upenn.edu/sections/met_act.html

Museum Applied Science Center for Archeology, University of Pennsylvania. Fair presentation of archaeometallurgical data. Few links to other sites.

http://users.ox.ac.uk/~salter

**Materials Science–Based Archeology Group, Oxford University. Excellent presentation of archaeometallurgical data, and very good links to other sites.*

ATUL B. GOKHALE
MetConsult, Inc.
New York, New York

COMPUTATION AND THEORETICAL METHODS

INTRODUCTION

Traditionally, the design of new materials has been driven primarily by phenomenology, with theory and computation providing only general guiding principles and, occasionally, the basis for rationalizing and understanding the fundamental principles behind known materials properties. Whereas these are undeniably important contributions to the development of new materials, the direct and systematic application of these general theoretical principles and computational techniques to the investigation of specific materials properties has been less common. However, there is general agreement within the scientific and technological community that modeling and simulation will be of critical importance to the advancement of scientific knowledge in the 21st century, becoming a fundamental pillar of modern science and engineering. In particular, we are currently at the threshold of quantitative and predictive theories of materials that promise to significantly alter the role of theory and computation in materials design. The emerging field of computational materials science is likely to become a crucial factor in almost every aspect of modern society, impacting industrial competitiveness, education, science, and engineering, and significantly accelerating the pace of technological developments.

At present, a number of physical properties, such as cohesive energies, elastic moduli, and expansion coefficients of elemental solids and intermetallic compounds, are routinely calculated from first principles, i.e., by solving the celebrated equations of quantum mechanics: either Schröedinger's equation, or its relativistic version, Dirac's equation, which provide a complete description of electrons in solids. Thus, properties can be predicted using only the atomic numbers of the constituent elements and the crystal structure of the solid as input. These achievements are a direct consequence of a mature theoretical and computational framework in solid-state physics, which, to be sure, has been in place for some time. Furthermore, the ever-increasing availability of midlevel and high-performance computing, high-bandwidth networks, and high-volume data storage and management, has pushed the development of efficient and computationally tractable algorithms to tackle increasingly more complex simulations of materials.

The first-principles computational route is, in general, more readily applicable to solids that can be idealized as having a perfect crystal structure, devoid of grain boundaries, surfaces and other imperfections. The realm of engineering materials, be it for structural, electronics, or other applications, is, however, that of "defective" solids. Defects and their control dictate the properties of real materials. There is, at present, an impressive body of work in materials simulation, which is aimed at understanding properties of real materials. These simulations rely heavily on either a phenomenological or semiempirical description of atomic interactions.

The units in this chapter of *Methods in Materials Research* have been selected to provide the reader with a suite of theoretical and computational tools, albeit at an introductory level, that begins with the microscopic description of electrons in solids and progresses towards the prediction of structural stability, phase equilibrium, and the simulation of microstructural evolution in real materials. The chapter also includes units devoted to the theoretical principles of well established characterization techniques that are best suited to provide exacting tests to the predictions emerging from computation and simulation. It is envisioned that the topics selected for publication will accurately reflect significant and fundamental developments in the field of computational materials science. Due to the nature of the discipline, this chapter is likely to evolve as new algorithms and computational methods are developed, providing not only an up-to-date overview of the field, but also an important record of its evolution.

JUAN M. SANCHEZ

INTRODUCTION TO COMPUTATION

Although the basic laws that govern the atomic interactions and dynamics in materials are conceptually simple and well understood, the remarkable complexity and variety of properties that materials display at the macroscopic level seem unpredictable and are poorly understood. Such a situation of basic well-known governing principles but complex outcomes is highly suited for a computational approach. This ultimate ambition of materials science—to predict macroscopic behavior from microscopic information (e.g., atomic composition)—has driven the impressive development of computational materials science. As is demonstrated by the number and range of articles in this volume, predicting the properties of a material from atomic interactions is by no means an easy task! In many cases it is not obvious how the fundamental laws of physics conspire with the chemical composition and structure of a material to determine a macroscopic property that may be of interest to an engineer. This is not surprising given that on the order of 10^{26} atoms may participate in an observed property. In some cases, properties are simple "averages" over the contributions of these atoms, while for other properties only extreme deviations from the mean may be important. One of the few fields in which a well-defined and justifiable procedure to go from the

atomic level to the macroscopic level exists is the equilibrium thermodynamics of homogeneous materials. In this case, all atoms "participate" in the properties of interest and the macroscopic properties are determined by fairly straightforward averages of microscopic properties. Even with this benefit, the prediction of alloy phase diagrams is still a formidable challenge, as is nicely illustrated in PREDICTION OF PHASE DIAGRAMS. Unfortunately, for many other properties (e.g., fracture), the macroscopic evolution of the material is strongly influenced by singularities in the microscopic distribution of atoms: for instance, a few atoms that surround a void or a cluster of impurity atoms. This dependence of a macroscopic property on small details of the microscopic distribution makes defining a predictive link between the microscopic and macroscopic much more difficult.

Placing some of these difficulties aside, the advantages of computational modeling for the properties that can be determined in this fashion are significant. Computational work tends to be less costly and much more flexible than experimental research. This makes it ideally suited for the initial phase of materials development, where the flexibility of switching between many different materials can be a significant advantage.

However, the ultimate advantage of computing methods, both in basic materials research and in applied materials design, is the level of control one has over the system under study. Whereas in an experimental situation nature is the arbiter of what can be realized, in a computational setting only creativity limits the constraints that can be forced onto a material. A computational model usually offers full and accurate control over structure, composition, and boundary conditions. This allows one to perform computational "experiments" that separate out the influence of a single factor on the property of the material. An interesting example may be taken from this author's research on lithium metal oxides for rechargeable Li batteries. These materials are crystalline oxides that can reversibly absorb and release Li ions through a mechanism called intercalation. Because they can do this at low chemical potential for Li, they are used on the cathode side of a rechargeable Li battery. In the discharge cycle of the battery, Li ions arrive at the cathode and are stored in the crystal structure of the lithium metal oxide. This process is reversed upon charging. One of the key properties of these materials is the electrochemical potential at which they intercalate Li ions, as it directly determines the battery voltage.

Figure 1A shows the potential range at which many transition metal oxides intercalate Li as a function of the number of d electrons in the metal (Ohzuku and Atsushi, 1994). While the graph indicates some upward trend of potential with the number of d electrons, this relation may be perturbed by several other parameters that change as one goes from one material to the other: many of the transition metal oxides in Figure 1A are in different crystal structures, and it is not clear to what extent these structural variations affect the intercalation potential. An added complexity in oxides comes from the small variation in average valence state of the cations, which may result in different oxygen composition, even when the

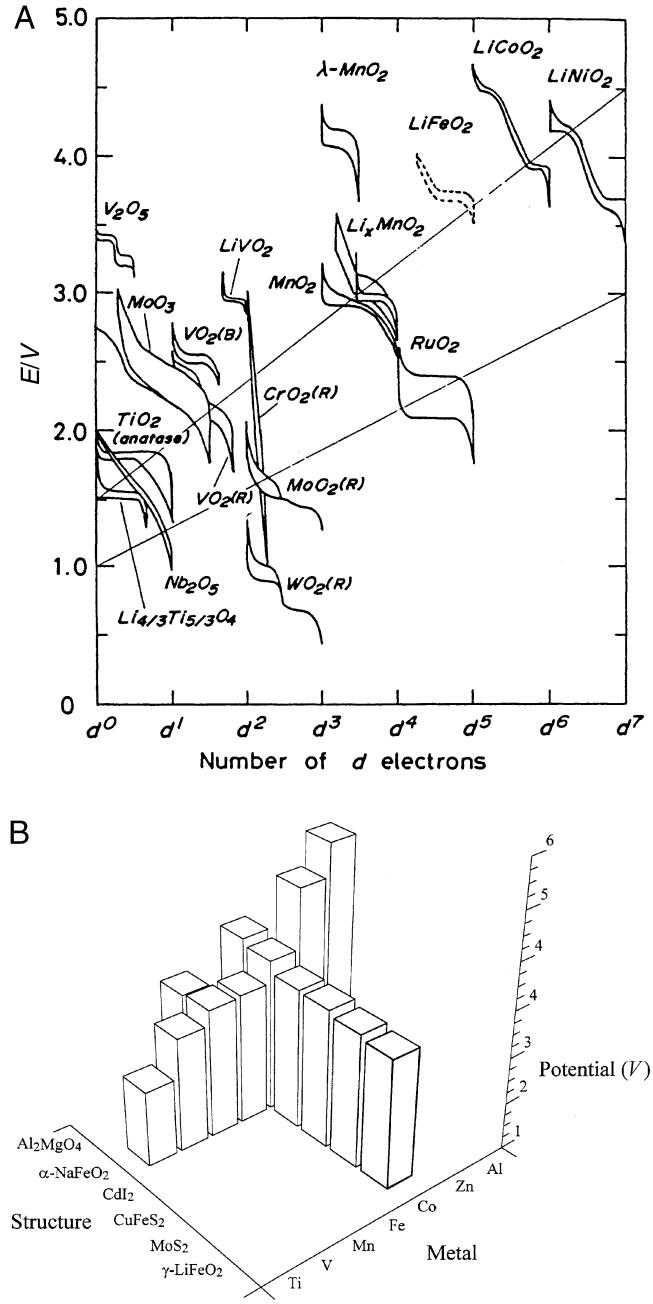


Figure 1. (**A**) Intercalation potential curves for lithium in various metal oxides as a function of the number of d electrons on the transition metal in the compound. (Taken from Ohzuku and Atsushi, 1994.) (**B**) Calculated intercalation potential for lithium in various $LiMO_2$ compounds as a function of the structure of the compound and the choice of metal M. The structures are denoted by their prototype.

chemical formula (based on conventional valences) would indicate the stoichiometry to be the same. These factors convolute the dependence of intercalation potential on the choice of transition metal, making it difficult to separate the roles of each independent factor. Computational methods are better suited to separating the influence of these different factors. Once a method for calculating the intercalation potential has been established, it can be applied to any system, in any crystal structure or oxygen

stoichiometry, whether such conditions correspond to the equilibrium structure of the material or not. By varying only one variable at a time in a calculation of the intercalation potential, a systematic study of each variable (e.g., structure, composition, stoichiometry) can be performed.

Figure 1B, the result of a series of *ab initio* calculations (Aydinol et al., 1997) clearly shows the effect of structure and metal in the oxide independently. Within the 3*d* transition metals, the effect of structure is clearly almost as large as the effect of the number of *d* electrons. Only for the non-*d* metals (Zn, Al) is the effect of metal choice dramatic. The calculation also shows that among the 3*d* metal oxides, $LiCoO_2$ in the spinel structure (Al_2MgO_4) would display the highest potential. Clearly, the advantage of the computational approach is not merely that one can predict the property of interest (in this case the intercalation potential) but also that the factors that may affect it can be controlled systematically.

Whereas the links between atomic-level phenomena and macroscopic properties form the basis for the control and predictive capabilities of computational modeling, they also constitute its disadvantages. The fact that properties must be derived from microscopic energy laws (often quantum mechanics) leads to the predictive characteristics of a method but also holds the potential for substantial errors in the result of the calculation. It is not currently possible to exactly calculate the quantum mechanical energy of a perfect crystalline array of atoms. Any errors in the description of the energetics of a system will ultimately show up in the derived macroscopic results. Many computational models are therefore still not fully quantitative. In some cases, it has not even been possible to identify an explicit link between the microscopic and macroscopic, so quantitative materials studies are not as yet possible.

The units in this chapter deal with a large variety of physical phenomena: for example, prediction of physical properties and phase equilibria, simulation of microstructural evolution, and simulation of chemical engineering processes. Readers may notice that these areas are at different stages in their evolution in applying computational modeling. The most advanced field is probably the prediction of physical properties and phase equilibria in alloys, where a well-developed formalism exists to go from the microscopic to the macroscopic. Combining quantum mechanics and statistical mechanics, a full *ab initio* theory has developed in this field to predict physical properties and phase equilibria, with no more input than the chemical construction of the system (Ducastelle, 1991; Ceder, 1993; de Fontaine, 1994; Zunger, 1994). Such a theory is predictive, and is well suited to the development and study of novel materials for which little or no experimental information is known and to the investigation of materials under extreme conditions.

In many other fields, such as in the study of microstructure or mechanical properties, computational models are still at a stage where they are mainly used to investigate the qualitative behavior of model systems and system-specific results are usually minimal or nonexistent. This lack of an *ab initio* theory reflects the very complex relation between these properties and the behavior of the constituent atoms. An example may be given from the molecular dynamics work on fracture in materials (Abraham, 1997). Typically, such fracture simulations are performed on systems with idealized interactions and under somewhat restrictive boundary conditions. At this time, the value of such modeling techniques is that they can provide complete and detailed information on a well-controlled system and thereby advance the science of fracture in general. Calculations that discern the specific details between different alloys (say Ti-6Al-4V and TiAl) are currently not possible but may be derived from schemes in which the link between the microscopic and the macroscopic is derived more heuristically (Eberhart, 1996).

Many of the mesoscale models (grain growth, film deposition) described in the papers in this chapter are also in this stage of "qualitative modeling." In many cases, however, some agreement with experiments can be obtained for suitable values of the input parameters.

One may expect that many of these computational methods will slowly evolve toward a more predictive nature as methods are linked in a systematic way.

The future of computational modeling in materials science is promising. Many of the trends that have contributed to the rapid growth of this field are likely to continue into the next decade. Figure 2 shows the exponential increase in computational speed over the last 50 years. The true situation is even better than what is depicted in Figure 2 as computer resources have also become less expensive. Over the last 15 years the ratio of computational power to price has increased by a factor of 10^4. Clearly, no other tool in material science and engineering can boast such a dramatic improvement in performance.

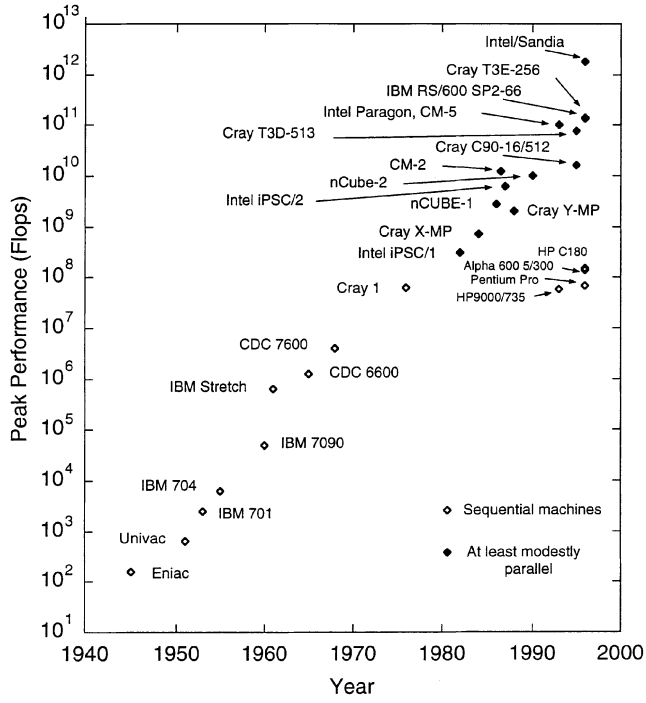


Figure 2. Peak performance of the fastest computers models built as a function of time. The performance is in floating-point operations per second (FLOPS). Data from Fox and Coddington (1993) and from manufacturers' information sheets.

However, it would be unwise to chalk up the rapid progress of computational modeling solely to the availability of cheaper and faster computers. Even more significant for the progress of this field may be the algorithmic development for simulation and quantum mechanical techniques. Highly accurate implementations of the local density approximation (LDA) to quantum mechanics [and its extension to the generalized gradient approximation (GGA)] are now widely available. They are considerably faster and much more accurate now than only a few years ago. The Car-Parrinello method and related algorithms have significantly improved the equilibration of quantum mechanical systems (Car and Parrinello, 1985; Payne et al., 1992). There is no reason to expect this trend to stop, and it is likely that the most significant advances in computational materials science will be realized through novel methods development rather than from ultra-high-performance computing.

Significant challenges remain. In many cases the accuracy of *ab initio* methods is orders of magnitude less than that of experimental methods. For example, in the calculation of phase diagrams an error of 10 meV, not large at all by *ab initio* standards, corresponds to an error of more than 100 K.

The time and size scales over which materials phenomena occur remain the most significant challenge. Although the smallest size scale in a first-principles method is always that of the atom and electron, the largest size scale at which individual features matter for a macroscopic property may be many orders of magnitude larger. For example, microstructure formation ultimately originates from atomic displacements, but the system becomes inhomogeneous on the scale of micrometers through sporadic nucleation and growth of distinct crystal orientations or phases. Whereas statistical mechanics provide guidance on how to obtain macroscopic averages for properties in homogeneous systems, there is no theory for coarse-grain (average) inhomogeneous materials. Unfortunately, most real materials are inhomogeneous.

Finally, all the power of computational materials science is worth little without a general understanding of its basic methods by all materials researchers. The rapid development of computational modeling has not been paralleled by its integration into educational curricula. Few undergraduate or even graduate programs incorporate computational methods into their curriculum, and their absence from traditional textbooks in materials science and engineering is noticeable. As a result, modeling is still a highly undervalued tool that so far has gone largely unnoticed by much of the materials science and engineering community in universities and industry. Given its potential, however, computational modeling may be expected to become an efficient and powerful research tool in materials science and engineering.

LITERATURE CITED

Abraham, F. F. 1997. On the transition from brittle to plastic failure in breaking a nanocrystal under tension (NUT). *Europhys. Lett.* 38:103–106.

Aydinol, M. K., Kohan, A. F., Ceder, G., Cho, K., and Joannopoulos, J. 1997. Ab-initio study of litihum intercalation in metal oxides and metal dichalcogenides. *Phys. Rev. B* 56:1354–1365.

Car, R. and Parrinello, M. 1985. Unified approach for molecular dynamics and density functional theory. *Phys. Rev. Lett.* 55:2471–2474.

Ceder, G. 1993. A derivation of the Ising model for the computation of phase diagrams. *Computat. Mater. Sci.* 1:144–150.

de Fontaine, D. 1994. Cluster approach to order-disorder transformations in alloys. *In* Solid State Physics (H. Ehrenreich and D. Turnbull, eds.). pp. 33–176. Academic Press, San Diego.

Ducastelle, F. 1991. Order and Phase Stability in Alloys. North-Holland Publishing, Amsterdam.

Eberhart, M. E. 1996. A chemical approach to ductile versus brittle phenomena. *Philos. Mag. A* 73:47–60.

Fox, G. C. and Coddington, P. D. 1993. An overview of high performance computing for the physical sciences. *In* High Performance Computing and Its Applications in the Physical Sciences: Proceedings of the Mardi Gras '93 Conference (D. A. Browne et al., eds.). pp. 1–21. World Scientific, Louisiana State University.

Ohzuku, T. and Atsushi, U. 1994. Why transitional metal (di) oxides are the most attractive materials for batteries. *Solid State Ionics* 69:201–211.

Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A., and Joannopoulos, J. D. 1992. Iterative minimization techniques for ab-initio total energy calculations: Molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* 64:1045.

Zunger, A. 1994. First-principles statistical mechanics of semiconductor alloys and intermetallic compounds. *In* Statics and Dynamics of Alloy Phase Transformations (P. E. A. Turchi and A. Gonis, eds.). pp. 361–419. Plenum, New York.

GERBRAND CEDER
Massachusetts Institute of Technology
Cambridge, Massachusetts

SUMMARY OF ELECTRONIC STRUCTURE METHODS

INTRODUCTION

Most physical properties of interest in the solid state are governed by the electronic structure—that is, by the Coulombic interactions of the electrons with themselves and with the nuclei. Because the nuclei are much heavier, it is usually sufficient to treat them as fixed. Under this Born-Oppenheimer approximation, the Schrödinger equation reduces to an equation of motion for the electrons in a fixed external potential, namely, the electrostatic potential of the nuclei (additional interactions, such as an external magnetic field, may be added).

Once the Schrödinger equation has been solved for a given system, many kinds of materials properties can be calculated. Ground-state properties include the cohesive energy, or heats of compound formation, elastic constants or phonon frequencies (Giannozzi and de Gironcoli, 1991), atomic and crystalline structure, defect formation energies, diffusion and catalysis barriers (Blöchl et al., 1993) and even nuclear tunneling rates (Katsnelson et al.,

1995), magnetic structure (van Schilfgaarde et al., 1996), work functions (Methfessel et al., 1992), and the dielectric response (Gonze et al., 1992). Excited-state properties are accessible as well; however, the reliability of the properties tends to degrade—or requires more sophisticated approaches—the larger the perturbing excitation.

Because of the obvious advantage in being able to calculate a wide range of materials properties, there has been an intense effort to develop general techniques that solve the Schrödinger equation from "first principles" for much of the periodic table. An exact, or nearly exact, theory of the ground state in condensed matter is immensely complicated by the correlated behavior of the electrons. Unlike Newton's equation, the Schrödinger equation is a field equation; its solution is equivalent to solving Newton's equation along all paths, not just the classical path of minimum action. For materials with wide-band or itinerant electronic motion, a one-electron picture is adequate, meaning that to a good approximation the electrons (or quasiparticles) may be treated as independent particles moving in a fixed effective external field. The effective field consists of the electrostatic interaction of electrons plus nuclei, plus an additional effective (mean-field) potential that originates in the fact that by correlating their motion, electrons can avoid each other and thereby lower their energy. The effective potential must be calculated self-consistently, such that the effective one-electron potential created from the electron density generates the same charge density through the eigenvectors of the corresponding one-electron Hamiltonian.

The other possibility is to adopt a model approach that assumes some model form for the Hamiltonian and has one or more adjustable parameters, which are typically determined by a fit to some experimental property such as the optical spectrum. Today such Hamiltonians are particularly useful in cases beyond the reach of first-principles approaches, such as calculations of systems with large numbers of atoms, or for strongly correlated materials, for which the (approximate) first-principles approaches do not adequately describe the electronic structure. In this unit, the discussion will be limited to the first-principles approaches.

Summaries of Approaches

The local-density approximation (LDA) is the "standard" solid-state technique, because of its good reliability and relative simplicity. There are many implementations and extensions of the LDA. As shown below (see discussion of The Local Density Approximation) it does a good job in predicting ground-state properties of wide-band materials where the electrons are itinerant and only weakly correlated. Its performance is not as good for narrow-band materials where the electron correlation effects are large, such as the actinide metals, or the late-period transition-metal oxides.

Hartree-Fock (HF) theory is one of the oldest approaches. Because it is much more cumbersome than the LDA, and its accuracy much worse for solids, it is used mostly in chemistry. The electrostatic interaction is called the "Hartree" term, and the Fock contribution that approximates the correlated motion of the electrons is called "exchange." For historic reasons, the additional energy beyond the HF exchange energy is often called "correlation" energy. As we show below (see discussion of Hartree-Fock Theory), the principal failing of Hartree-Fock theory stems from the fact that the potential entering into the exchange interaction should be screened out by the other electrons. For narrow-band systems, where the electrons reside in atomic-like orbitals, Hartree-Fock theory has some important advantages over the LDA. Its nonlocal exchange serves as a better starting point for more sophisticated approaches.

Configuration-interaction theory is an extension of the HF approach that attempts to solve the Schrödinger equation with high accuracy. Computationally, it is very expensive and is feasible only for small molecules with $\sim$10 atoms or fewer. Because it is only applied to solids in the context of model calculations (Grant and McMahan, 1992), it is not considered further here.

The so-called *GW* approximation may be thought of as an extension to Hartree-Fock theory, as described below (see discussion under Dielectric Screening, the Random-Phase, *GW*, and SX Approximations). The *GW* method incorporates a representation of the Green's function (G) and the Coulomb interaction (W). It is a Hartree-Fock-like theory for which the exchange interaction is properly screened. *GW* theory is computationally very demanding, but it has been quite successful in predicting, for example, bandgaps in semiconductors. To date, it has been only possible to apply the theory to optical properties, because of difficulties in reliably integrating the self-energy to obtain a total energy.

The LDA + U theory is a hybrid approach that uses the LDA for the "itinerant" part and Hartree-Fock theory for the "local" part. It has been quite successful in calculating both ground-state and excited-state properties in a number of correlated systems. One criticism of this theory is that there exists no unique prescription to renormalize the Coulomb interaction between the local orbitals, as will be described below. Thus, while the method is *ab initio*, it retains the flavor of a model approach.

The self-interaction correction (Svane and Gunnarsson, 1990) is similar to LDA + U theory, in that a subset of the orbitals (such as the *f*-shell orbitals) are partitioned off and treated in a HF-like manner. It offers a unique and well-defined functional, but tends to be less accurate than the LDA + U theory, because it does not screen the local orbitals.

The quantum Monte Carlo approach is not a mean-field approach. It is an ostensibly exact, or nearly exact, approach to determine the ground-state total energy. In practice, some approximations are needed, as described below (see discussion of Quantum Monte Carlo). The basic idea is to evaluate the Schrödinger equation by brute force, using a Monte Carlo approach. While applications to real materials so far have been limited, because of the immense computational requirements, this approach holds much promise with the advent of faster computers.

Implementation

Apart from deciding what kind of mean-field (or other) approximation to use, there remains the problem of

implementation in some kind of practical method. Many different approaches have been employed, especially for the LDA. Both the single-particle orbitals and the electron density and potential are invariably expanded in some basis set, and the various methods differ in the basis set employed. Figure 1 depicts schematically the general types of approaches commonly used. One principal distinction is whether a method employs plane waves for a basis, or atom-centered orbitals. The other primary distinction among methods is the treatment of the core. Valence electrons must be orthogonalized to the inert core states. The various methods address this by (1) replacing the core with an effective (pseudo)potential, so that the (pseudo)wave functions near the core are smooth and nodeless, or (2) by "augmenting" the wave functions near the nuclei with numerical solutions of the radial Schrödinger equation. It turns out that there is a connection between "pseudizing" or augmenting the core; some of the recently developed methods such as the Planar Augmented Wave method of Blöchl (1994), and the pseudopotential method of Vanderbilt (1990) may be thought of as a kind of hybrid of the two (Dong, 1998).

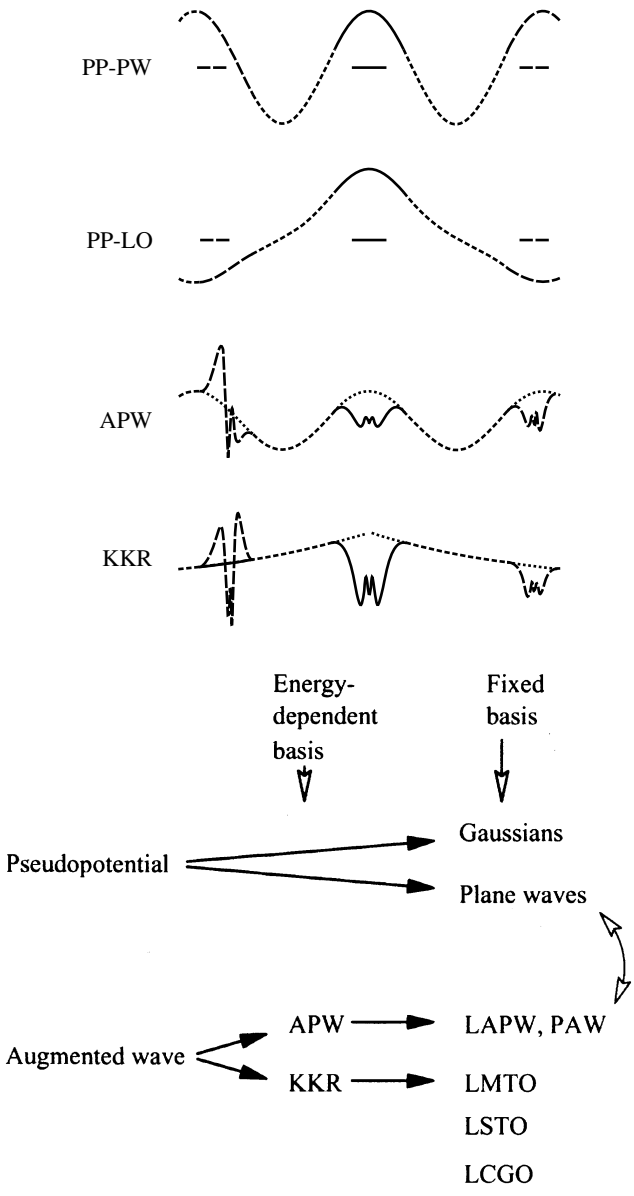


Figure 1. Illustration of different methods, as described in the text. The pseudopotential (PP) approaches can employ either plane waves (PW) or local atom-centered orbitals; similarly the augmented-wave approach employing PW becomes APW or LAPW; using atom-centered Hankel functions it is the KKR method or the method of linear muffin-tin orbitals (LMTO). The PAW (Blöchl, 1994) is a variant of the APW method, as described in the text. LMTO, LSTO and LCGO are atom-centered augmented-wave approaches with Hankel, Slater, and Gaussian orbitals, respectively, used for the envelope functions.

The augmented-wave basis sets are "intelligently" chosen in that they are tailored to solutions of the Schrödinger equation for a "muffin-tin" potential. A muffin-tin potential is flat in the interstitial region, and then spherically symmetric inside nonoverlapping spheres centered at each nucleus, and, for close-packed systems, is a fairly good representation of the true potential. But because the resulting Hamiltonian is energy dependent, both the augmented plane-wave (APW) and augmented atom-centered (Korringa, Kohn, Rostoker; KKR) methods result in a nonlinear algebraic eigenvalue problem. Andersen and Jepsen (1984 also see Andersen, 1975) showed how to linearize the augmented-wave Hamiltonian, and both the APW (now LAPW) and KKR—renamed linear muffin-tin orbitals (LMTO)—methods are vastly more efficient.

The choice of implementation introduces further approximations, though some techniques have enough machinery now to solve a given one-electron Hamiltonian nearly exactly. Today the LAPW method is regarded as the "industry standard" high-precision method, though some implementations of the LMTO method produces a corresponding accuracy, as does the plane-wave pseudopotential approach, provided the core states are sufficiently deep and enough plane waves are chosen to make the basis reasonably complete. It is not always feasible to generate a well-converged pseudopotential; for example, the high-lying *d* cores in Ga can be a little too shallow to be "pseudized" out, but are difficult to treat explicitly in the valence band using plane waves. Traditionally the augmented-wave approaches have introduced shape approximations to the potential, "spheridizing" the potential inside the augmentation spheres. This is often still done today; the approximation tends usually to be adequate for energy bands in reasonably close-packed systems, and relatively coarse total energy differences. This approximation, combined with enlarging the augmentation spheres and overlapping them so that their volume equals the unit cell volume, is known as the atomic spheres approximation (ASA). Extensions, such as retaining the correction to the spherical part of the electrostatic potential from the nonspherical part of the density (Skriver and Rosengaard, 1991) eliminate most of the errors in the ASA.

Extensions

The "standard" implementations of, for example, the LDA, generate electron eigenstates through diagonalization of the one-electron wave function. As noted before, the

one-electron potential itself must be determined self-consistently, so that the eigenstates generate the same potential that creates them.

Some information, such as the total energy and internuclear forces, can be directly calculated as a byproduct of the standard self-consistency cycle. There have been many other properties that require extensions of the "standard" approach. Linear-response techniques (Baroni et al., 1987; Savrasov et al., 1994) have proven particularly fruitful for calculation of a number of properties, such as phonon frequencies (Giannozzi and de Gironcoli, 1991), dielectric response (Gonze et al., 1992), and even alloy heats of formation (de Gironcoli et al., 1991). Linear response can also be used to calculate exchange interactions and spin-wave spectra in magnetic systems (Antropov et al., unpub. observ.). Often the LDA is used as a parameter generator for other methods. Structural energies for phase diagrams are one prime example. Another recent example is the use of precise energy band structures in GaN, where small details in the band structure are critical to how the material behaves under high-field conditions (Krishnamurthy et al., 1997).

Numerous techniques have been developed to solve the one-electron problem more efficiently, thus making it accessible to larger-scale problems. Iterative diagonalization techniques have become indispensable to the plane wave basis. Though it was not described this way in their original paper, the most important contribution from Carr and Parrinello's (1985) seminal work was their demonstration that special features of the plane-wave basis can be exploited to render a very efficient iterative diagonalization scheme. For layered systems, both the eigenstates and the Green's function (Skriver and Rosengaard, 1991) can be calculated in $O(N)$ time, with N being the number of layers (the computational effort in a straightforward diagonalization technique scales as cube of the size of the basis). Highly efficient techniques for layered systems are possible in this way. Several other general-purpose $O(N)$ methods have been proposed. A recent class of these methods computes the ground-state energy in terms of the density matrix, but not spectral information (Ordejón et al., 1995). This class of approaches has important advantages for large-scale calculations involving 100 or more atoms, and a recent implementation using the LDA has been reported (Ordejón et al., 1996); however, they are mainly useful for insulators. A Green's function approach suitable for metals has been proposed (Wang et al., 1995), and a variant of it (Abrikosov et al., 1996) has proven to be very efficient to study metallic systems with several hundred atoms.

HARTREE-FOCK THEORY

In Hartree-Fock theory, one constructs a Slater determinant of one-electron orbitals ψ_j. Such a construct makes the total wave function antisymmetric and better enables the electrons to avoid one another, which leads to a lowering of total energy. The additional lowering is reflected in the emergence of an additional effective (exchange) potential v^x (Ashcroft and Mermin, 1976). The resulting one-electron Hamiltonian has a local part from the direct electrostatic (Hartree) interaction v^H and external (nuclear) potential v^{ext}, and a nonlocal part from v^x

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v^{\text{ext}}(\mathbf{r}) + v^H(\mathbf{r})\right]\psi_i(\mathbf{r}) + \int d^3\mathbf{r}' v^x(\mathbf{r},\mathbf{r}')\psi(\mathbf{r}') = \varepsilon_i\psi_i(\mathbf{r}) \quad (1)$$

$$v^H(\mathbf{r}) = \int \frac{e^2}{|\mathbf{r}-\mathbf{r}'|} n(\mathbf{r}') \quad (2)$$

$$v_x(\mathbf{r},\mathbf{r}') = -\sum_j \frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}) \quad (3)$$

where e is the electronic charge and $n(\mathbf{r})$ is the electron density.

Thanks to Koopman's theorem, the change in energy from one state to another is simply the difference between the Hartree-Fock parameters ε in two states. This provides a basis to interpret the ε in solids as energy bands. In comparison to the LDA (see discussion of The Local Density Approximation), Hartree-Fock theory is much more cumbersome to implement, because of the nonlocal exchange potential $v^x(\mathbf{r},\mathbf{r}')$ which requires a convolution of v^x and ψ. Moreover, the neglect of correlations beyond the exchange renders it a much poorer approximation to the ground state than the LDA. Hartree-Fock theory also usually describes the optical properties of solids rather poorly. For example, it rather badly overestimates the bandgap in semiconductors. The Hartree-Fock gaps in Si and GaAs are both $\sim$5 eV (Hott, 1991), in comparison to the observed 1.1 and 1.5 eV, respectively.

THE LOCAL-DENSITY APPROXIMATION

The LDA actually originates in the X-α method of Slater (1951), who sought a simplifying approximation to the HF exchange potential. By assuming that the exchange varied in proportion to $n^{1/3}$, with n the electron density, the HF exchange becomes local and vastly simplifies the computational effort. Thus, as it was envisioned by Slater, the LDA is an approximation to Hartree-Fock theory, because the exact exchange is approximated by a simple functional of the density n, essentially proportional to $n^{1/3}$. Modern functionals go beyond Hartree-Fock theory because they include correlation energy as well.

Slater's X-α method was put on a firm foundation with the advent of density-functional theory (Hohenberg and Kohn, 1964). It established that the ground-state energy is strictly a functional of the total density. But the energy functional, while formally exact, is unknown. The LDA (Kohn and Sham, 1965) assumes that the exchange plus correlation part of the energy E_{xc} is a strictly local functional of the density:

$$E_{\text{xc}}[n] \approx \int d^3\mathbf{r} n(\mathbf{r})\varepsilon_{\text{xc}}[n(\mathbf{r})] \quad (4)$$

This *ansatz* leads, as in the Hartree-Fock case, to an equation of motion for electrons moving independently in

an effective field, except that now the potential is strictly local:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v^{\text{ext}}(\mathbf{r}) + v^H(\mathbf{r}) + v^{\text{xc}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}) \quad (5)$$

This one-electron equation follows directly from a functional derivative of the total energy. In particular, $v^{\text{xc}}(\mathbf{r})$ is the functional derivative of E_{xc}:

$$v^{\text{xc}}(\mathbf{r}) \equiv \frac{\delta E_{\text{xc}}}{\delta n} \quad (6)$$

$$\overset{\text{LDA}}{=} \frac{\delta}{\delta n}\int d^3\mathbf{r} n(\mathbf{r})\varepsilon_{\text{xc}}[n(\mathbf{r})] \quad (7)$$

$$= \varepsilon_{\text{xc}}[n(\mathbf{r})] + n(\mathbf{r})\frac{d}{dn}\varepsilon_{\text{xc}}[n(\mathbf{r})] \quad (8)$$

Both exchange and correlation are calculated by evaluating the exact ground state for a jellium (in which the discrete nuclear charge is smeared out into a constant background). This is accomplished either by Monte Carlo techniques (Ceperley and Alder, 1980) or by an expansion in the random-phase approximation (von Barth and Hedin, 1972). When the exchange is calculated exactly, the self-interaction terms (the interaction of the electron with itself) in the exchange and direct Coulomb terms cancel exactly. Approximation of the exact exchange by a local density functional means that this is no longer so, and this is one key source of error in the LD approach. For example, near surfaces, or for molecules, the asymptotic decay of the electron potential is exponential, whereas it should decay as 1/r, where r is the distance to the nucleus. Thus, molecules are less well described in the LDA than are solids.

In Hartree-Fock theory, the opposite is the case. The self-interaction terms cancel exactly, but the operator $1/|\mathbf{r}-\mathbf{r}'|$ entering into the exchange should in effect be screened out. Thus, Hartree-Fock theory does a reasonable job in small molecules, where the screening is less important, while for solids it fails rather badly. Thus, the LDA generates much better total energies in solids than Hartree-Fock theory. Indeed, on the whole, the LDA predicts, with rather good accuracy, ground-state properties, such as crystal structures and phonon frequencies in itinerant materials and even in many correlated materials.

Gradient Corrections

Gradient corrections extend slightly the *ansatz* of the local density approximation. The idea is to assume that E_{xc} is not only a local functional of the density, but a functional of the density and its Laplacian. It turns out that the leading correction term can be obtained exactly in the limit of a small, slowly varying density, but it is divergent. To render the approach practicable, a wave-vector analysis is carried out and the divergent, low wave-vector part of the functional is cut off; these are called "generalized gradient approximations" (GGAs). Calculations using gradient corrections have produced mixed results. It was hoped that since the LDA does quite well in predicting many ground-state properties, gradient corrections would introduce the small corrections needed, particularly in systems in which the density is slowly varying. On the whole, the GGA tends to improve some properties, though not consistently so. This is probably not surprising, since the main ingredients missing in the LDA, (e.g., inexact cancellation of the self-interaction and nonlocal potentials) are also missing for gradient-corrected functionals. One of the first approximations was that of Langreth and Mehl (1981). Many of the results in the next section were produced with their functional. Some newer functionals, most notably the so-called "PBE" (named after Perdew, Burke, Enzerhof) functional (Perdew, 1997) improve results for some properties of solids, while worsening others.

Table 1. Heats of Formation, in eV, for the Hydroxyl ($OH + H_2 \rightarrow H_2O + H$), Tetrazine ($H_2C_2N_4 \rightarrow 2\,HCN + N_2$), and Vinyl Alcohol ($C_2OH_3 \rightarrow$ Acetaldehyde) Reactions, Calculated with Different Methods[a]

Method	Hydroxyl	Tetrazine	Vinyl Alcohol
HF	−0.07	−3.41	−0.54
LDA	−0.69	−1.99	−0.34
Becke	−0.44	−2.15	−0.45
PW-91	−0.64	−1.73	−0.45
QMC	−0.65	−2.65	−0.43
Expt	−0.63	—	−0.42

[a]Abbreviations: HF, Hartree-Fock; LDA, local-density approximation; Becke, GGA functional (Becke, 1993); PW-91, GGA functional (Perdew, 1997); QMC, quantum Monte Carlo. Calculations by Grossman and Mitas (1997).

One recent calculation of the heat of formation for three molecular reactions offers a detailed comparison of the HF, LDA, and GGA to (nearly) exact quantum Monte Carlo results. As shown in Table 1, all of the different mean-field approaches have approximately similar accuracy in these small molecules.

Excited-state properties, such as the energy bands in itinerant or correlated systems, are generally not improved at all with gradient corrections. Again, this is to be expected since the gradient corrections do not redress the essential ingredients missing from the LDA, namely, the cancellation of the self-interaction or a proper treatment of the nonlocal exchange.

LDA Structural Properties

Figure 2 compares predicted atomic volumes for the elemental transition metals and some *sp* bonded semiconductors to corresponding experimental values. The errors shown are typical for the LDA, underestimating the volume by ∼0% to 5% for *sp* bonded systems, by ∼0% to 10% for *d*-bonded systems with the worst agreement in the 3*d* series, and somewhat more for *f*-shell metals (not shown). The error also tends to be rather severe for the extremely soft, weakly bound alkali metals.

The crystal structure of Se and Te poses a more difficult test for the LDA. These elements form an open, low-symmetry crystal with ∼90° bond angles. The electronic structure is approximately described by pure atomic *p* orbitals linked together in one-dimensional chains, with a weak interaction between the chains. The weak

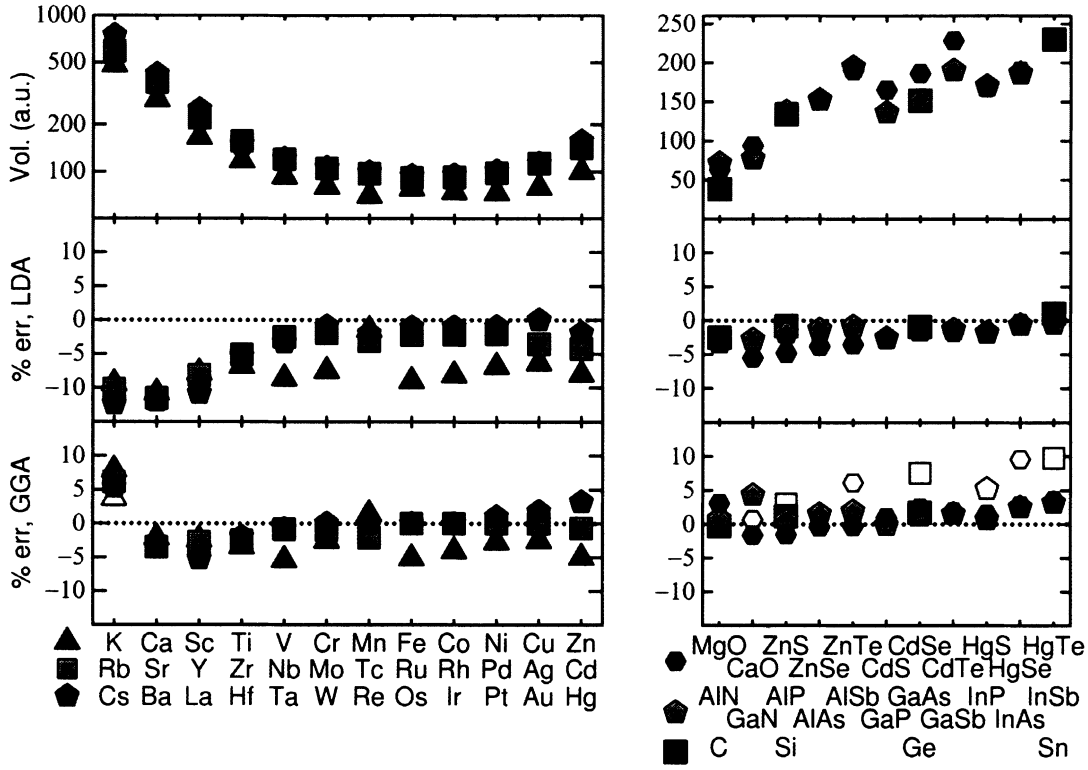


Figure 2. Unit cell volume for the elemental transition metals (left) and semiconductors (right). Left: triangles, squares, and pentagons refer to 3-, 4-, and 5-*d* metals, respectively. Right: squares, pentagons, and hexagons refer to group IV, III-V, and II-VI compounds. Upper panel: volume per unit cell; middle panel: relative error predicted by the LDA; lower panel: relative error predicted by the LDA + GGA of Langreth and Mehl (1981), except for light symbols, which are errors in the PBE functional (Perdew et al., 1996, 1997).

inter-chain interaction combined with the low symmetry and open structure make a difficult test for the local-density approximation. The crystal structure of Se and Te is hexagonal with three atoms per unit cell, and may be specified by the a and c parameters of the hexagonal cell, and one internal displacement parameter, u. Table 2 shows that the LDA predicts rather well the strong intra-chain bond length, but rather poorly reproduces the inter-chain bond length.

One of the largest effects of gradient-corrected functionals is to increase systematically and on the average improve, the equilibrium bond lengths (Fig. 2). The GGA of Langreth and Mehl (1981) significantly improves on the transition metal lattice constants; they similarly significantly improve on the predicted inter-chain bond length in Se (Table 2). In the case of the semiconductors, there is a tendency to overcorrect for the heavier elements. The newer GGA of Perdew et al. (1996, 1997) rather badly overestimates lattice constants in the heavy semiconductors.

Table 2. Crystal Structure of Se, Comparing the LDA to GGA Results (Perdew, 1991), as taken from Dal Corso and Resta (1994)[a]

	a	c	u	d_1	d_2
LDA	7.45	9.68	0.256	4.61	5.84
GGA	8.29	9.78	0.224	4.57	6.60
Expt	8.23	9.37	0.228	4.51	6.45

[a]Lattice parameters a and c are in atomic units (i.e., units of the Bohr radius a_0), as are intra-chain bond length d_1 and inter-chain bond length d_2. The parameter u is an internal displacement parameter as described in Dal Corso and Resta (1994).

LDA Heats of Formation and Cohesive Energies

One of the largest systematic errors in the LDA is the cohesive energy, i.e., the energy of formation of the crystal from the separated elements. Unlike Hartree-Fock theory, the LD functional has no variational principle that guarantees its ground-state energy is less than the true one. The LDA usually overestimates binding energies. As expected, and as Figure 3 illustrates, the errors tend to be greater for transition metals than for *sp* bonded compounds. Much of the error in the transition metals can be traced to errors in the spin multiplet structure in the atom (Jones and Gunnarsson, 1989); thus, the sudden change in the average overbinding for elements to the left of Cr and the right of Mn.

For reasons explained above, errors in the heats of formation between molecules and solid phases, or between different solid phases (Pettifor and Varma, 1979) tend to be much smaller than those of the cohesive energies. This is especially true when the formation involves atoms arranged on similar lattices. Figure 4 shows the errors typically encountered in solid-solid reactions between a

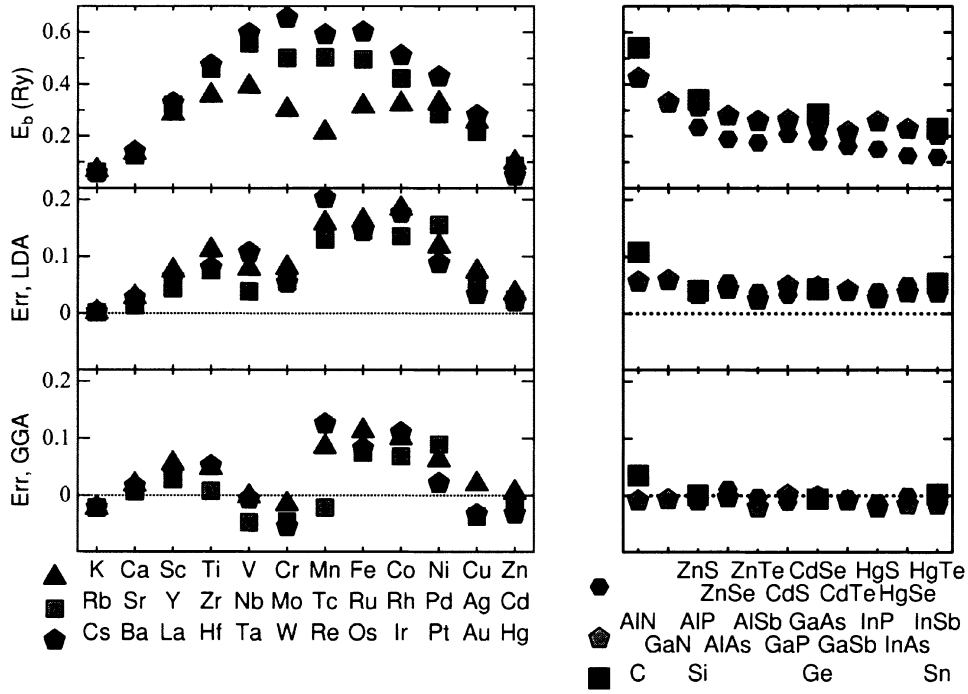


Figure 3. Heats of formation for elemental transition metals (left) and semiconductors (right). Left: triangles, squares, and pentagons refer to 3-, 4-, and 5-*d* metals, respectively. Right: squares, pentagons, and hexagons refer to group IV, III-V, and II-VI compounds. Upper panel: heat of formation per atom (Ry); middle panel: error predicted by the LDA; lower panel: error predicted by the LDA + GGA of Langreth and Mehl (1981).

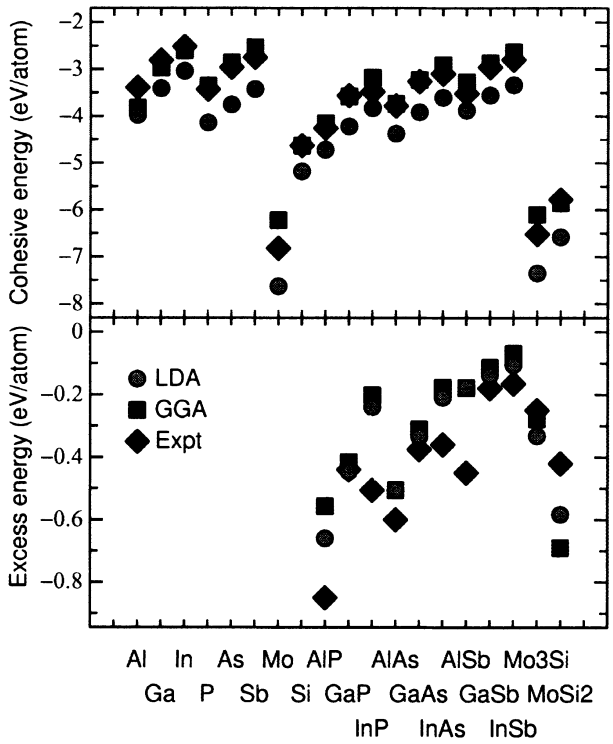


Figure 4. Cohesive energies (top), and heats of formation (bottom) of compounds from the elemental solids. The MoSi data are taken from McMahan et al. (1994). The other data were calculated by Berding and van Schilfgaarde using the FP-LMTO method (unpub. observ.).

wide range of dissimilar phases. Al is face-centered cubic (fcc), P and Si are open structures, and the other elements form a range of structures intermediate in their packing densities. This figure also encapsulates the relative merits of the LDA and GGA as predictors of binding energies. The GGA generally predicts the cohesive energies significantly better than the LDA, because the cohesive energies involve free atoms. But when solid-solid reactions are considered, the improvement disappears. Compounds of Mo and Si make an interesting test case for the GGA (McMahan et al., 1994). The LDA tends to overbind; but the GGA of Langreth and Mehl (1981) actually fares considerably worse, because the amount of overbinding is less systematic, leading to a prediction of the wrong ground state for some parts of the phase diagram. When recalculated using Perdew's PBE functional, the difficulty disappears (J. Klepis, pers. comm.).

The uncertainties are further reduced when reactions involve atoms rearranged on similar or the same crystal structures. One testimony to this is the calculation of structural energy differences of elemental transitional metals in different crystal structures. Figure 5 compares the local density hexagonal close packed–body centered cubic (hcp-bcc) and fcc-bcc energy differences in the 3-*d* transition metals, calculated nonmagnetically (Paxton et al., 1990; Skriver, 1985; Hirai, 1997). As the figure shows, there is a trend to stabilize bcc for elements with the *d* bands less than half-full, and to stabilize a close-packed structure for the late transition metals. This trend can be attributed to the two-peaked structure in the bcc *d* contribution to the density of states, which gains energy

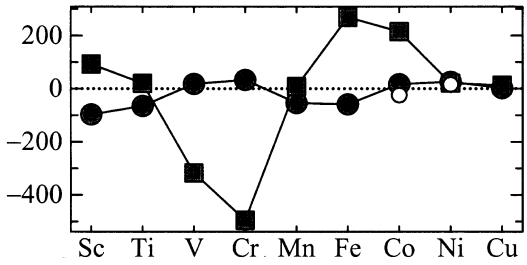


Figure 5. Hexagonal close packed–face centered cubic (hcp-fcc; circles) and body centered cubic–face centered cubic (bcc-fcc; squares) structural energy differences, in meV, for the 3-*d* transition metals, as calculated in the LDA, using the full-potential LMTO method. Original calculations are nonmagnetic (Paxton et al., 1990; Skriver, 1985); white circles are recalculations by the present author, with spin-polarization included.

when the lower (bonding) portion is filled and the upper (antibonding) portion is empty. Except for Fe (and with the mild exception of Mn, which has a complex structure with noncollinear magnetic moments and was not considered here), each structure is correctly predicted, including resolution of the experimentally observed sign of the hcp-fcc energy difference. Even when calculated magnetically, Fe is incorrectly predicted to be fcc. The GGAs of Langreth and Mehl (1981) and Perdew and Wang (Perdew, 1991) rectify this error (Bagno et al., 1989), possibly because the bcc magnetic moment, and thus the magnetic exchange energy, is overestimated for those functionals. In an early calculation of structural energy differences (Pettifor, 1970), Pettifor compared his results to inferences of the differences by Kaufman who used "judicious use of thermodynamic data and observations of phase equilibria in binary systems." Pettifor found that his calculated differences are two to three times larger than what Kaufman inferred (Paxton's newer calculations produces still larger discrepancies). There is no easy way to determine which is more correct.

Figure 6 shows some calculated heats of formation for the TiAl alloy. From the point of view of the electronic structure, the alloy potential may be thought of as a rather weak perturbation to the crystalline one, namely, a permutation of nuclear charges into different arrangements on the same lattice (and additionally some small distortions about the ideal lattice positions). The deviation from the regular solution model is properly reproduced by the LDA, but there is a tendency to overbind, which leads to an overestimate of the critical temperatures in the alloy phase diagram (Asta et al., 1992).

LDA Elastic Constants

Because of the strong volume dependence of the elastic constants, the accuracy to which LDA predicts them depends on whether they are evaluated at the observed volume or the LDA volume. Figure 7 shows both for the elemental transition metals and some *sp*-bonded compounds. Overall, the GGA of Langreth and Mehl (1981) improves on the LDA; how much improvement depends on which lattice constant one takes. The accuracy of other

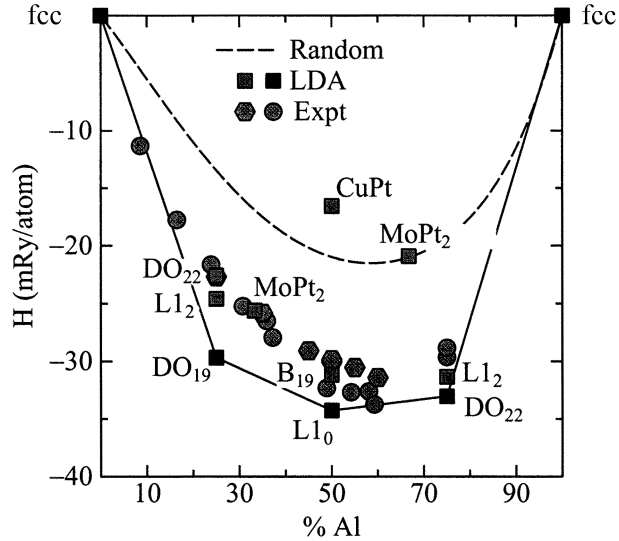


Figure 6. Heat of formation of compounds of Ti and Al from the fcc elemental states. Circles and hexagons are experimental data, taken from Kubaschewski and Dench (1955) and Kubaschewski and Heymer (1960). Light squares are heats of formation of compounds from the fcc elemental solids, as calculated from the LDA. Dark squares are the minimum-energy structures and correspond to experimentally observed phases. Dashed line is the estimated heat formation of a random alloy. Calculated values are taken from Asta et al. (1992).

elastic constants and phonon frequencies are similar (Baroni et al., 1987; Savrosov et al., 1994); typically they are predicted to within ~20% for *d*-shell metals and somewhat better than that for *sp*-bonded compounds. See Figure 8 for a comparison of c_{44}, or its hexagonal analog.

LDA Magnetic Properties

Magnetic moments in the itinerant magnets (e.g., the 3*d* transition metals) are generally well predicted by the LDA. The upper right panel of Figure 9 compares the LDA moments to experiment both at the LDA minimum-energy volume and at the observed volume. For magnetic properties, it is most sensible to fix the volume to experiment, since for the magnetic structure the nuclei may by viewed as an external potential. The classical GGA functionals of Langreth and Mehl (1981) and Perdew and Wang (Perdew, 1991) tend to overestimate the moments and worsen agreement with experiment. This is less the case with the recent PBE functional, however, as Figure 9 shows.

Cr is an interesting case because it is antiferromagnetic along the [001] direction, with a spin-density wave, as Figure 9 shows. It originates as a consequence of a nesting vector in the Cr Fermi surface (also shown in the figure), which is incommensurate with the lattice. The half-period is approximately the reciprocal of the difference in the length of the nesting vector in the figure and the half-width of the Brillouin zone. It is experimentally ~21.2 monolayers (ML) (Fawcett, 1988), corresponding to a nesting vector $q = 1.047$. Recently, Hirai (1997) calculated the

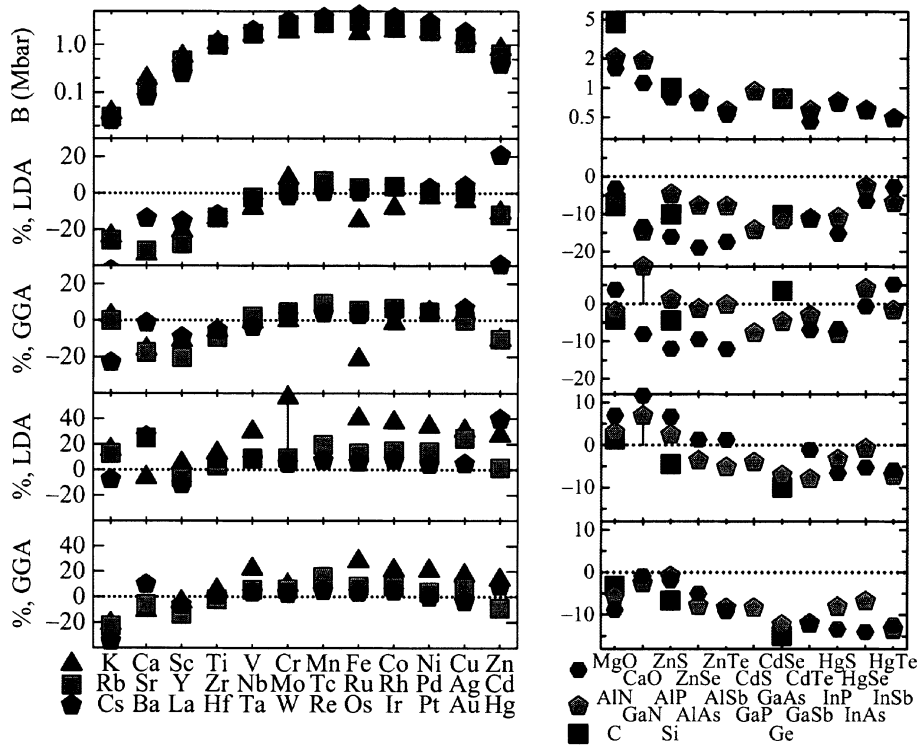


Figure 7. Bulk modulus for the elemental transition metals (left) and semiconductors (right). Left: triangles, squares, and pentagons refer, to 3-, 4-, and 5-*d* metals, respectively. Right: squares, pentagons, and hexagons refer to group IV, III-V, and II-VI compounds. Top panels: bulk modulus; second panel from top: relative error predicted by the LDA at the observed volume; third panel from top: same, but for the LDA + GGA of Langreth and Mehl (1981) except for light symbols, which are errors in the PBE functional (Perdew et al., 1996, 1997); fourth and fifth panels from top: same as the second and third panels but evaluated at the minimum-energy volume.

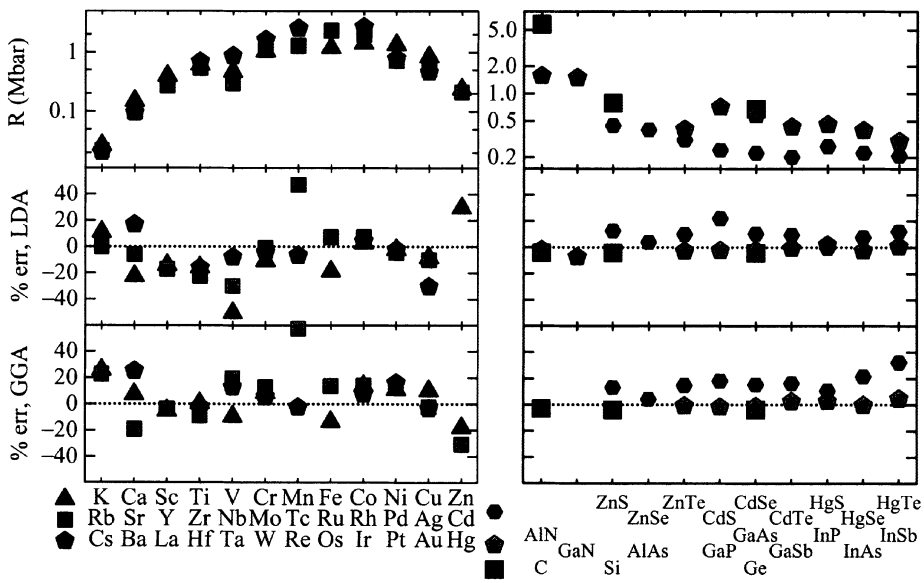


Figure 8. Elastic constant (R) for the elemental transition metals (left) and semiconductors (right), and the experimental atomic volume. For cubic structures, $R = c_{44}$. For hexagonal structures, $R = (c_{11} + 2c_{33} + c_{12}\text{-}4c_{13})/6$ and is analogous to c_{44}. Left: triangles, squares, and pentagons refer to 3-, 4-, and 5-*d* metals, respectively. Right: squares, pentagons, and hexagons refer to group IV, III-V and II-VI compounds. Upper panel: volume per unit cell; middle panel: relative error predicted by the LDA; lower panel: relative error predicted by the LDA + GGA of Langreth and Mehl (1981) except for light symbols, which are errors in the PBE functional (Perdew et al., 1996, 1997).

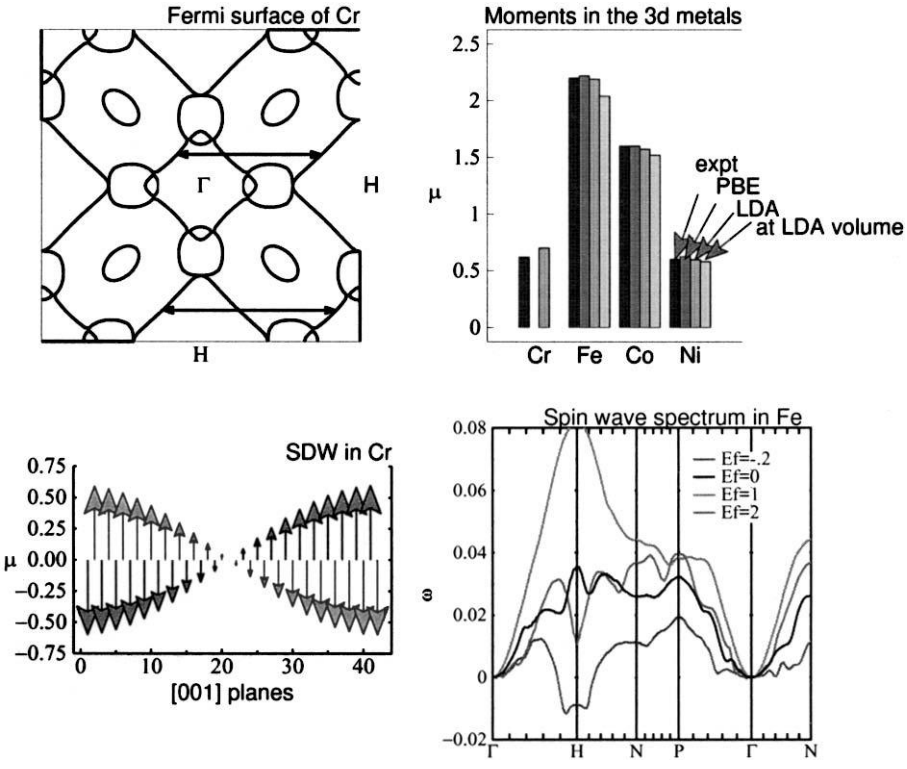


Figure 9. Upper left: LDA Fermi surface of nonmagnetic Cr. Arrows mark the nesting vectors connecting large, nearly parallel sheets in the Brillouin zone. Upper right: magnetic moments of the 3-*d* transition metals, in Bohr magnetons, calculated in the LDA at the LDA volume, at the observed volume, and using the PBE (Perdew et al., 1996, 1997) GGA at the observed volume. The Cr data is taken from Hirai (1997). Lower left: the magnetic moments in successive atomic layers along [001] in Cr, showing the antiferromagnetic spin-density wave. The observed period is ~21.2 lattice spacings. Lower right: spin-wave spectrum in Fe, in meV, calculated in the LDA (Antropov et al., unpub. observ.) for different band fillings, as discussed in the text.

period in Cr by constructing long supercells and evaluating the total energy using a layer KKR technique as a function of the cell dimensions. The calculated moment amplitude (Fig. 9) and period were both in good agreement with experiment. Hirai's calculated period was ~20.8 ML, in perhaps fortuitously good agreement with experiment. This offers an especially rigorous test of the LDA, because small inaccuracies in the Fermi surface are greatly magnified by errors in the period.

Finally, Figure 9 shows the spin-wave spectrum in Fe as calculated in the LDA using the atomic spheres approximation and a Green's function technique, plotted along high-symmetry lines in the Brillouin zone. The spin stiffness D is the curvature of ω at Γ, and is calculated to be 330 meV-A^2, in good agreement with the measured 280–310 meV-A^2. The four lines also show how the spectrum would change with different band fillings (as defined in the legend)—this is a "rigid band" approximation to alloying of Fe with Mn ($E_F < 0$) or Co ($E_F > 0$). It is seen that ω is positive everywhere for the normal Fe case (black line), and this represents a triumph for the LDA, since it demonstrates that the global ground state of bcc Fe is the ferromagnetic one. ω remains positive by shifting the Fermi level in the "Co alloy" direction, as is observed experimentally. However, changing the filling by only -0.2 eV ("Mn alloy") is sufficient to produce an instability at H, thus driving it to an antiferromagnetic structure in the [001] direction, as is experimentally observed.

Optical Properties

In the LDA, in contradistinction to Hartree-Fock theory, there is no formal justification for associating the eigenvalues ε of Equation 5 with energy bands. However, because the LDA is related to Hartree-Fock theory, it is reasonable to expect that the LDA eigenvalues ε bear a close resemblance to energy bands, and they are widely interpreted that way. There have been a few "proper" local-density calculations of energy gaps, calculated by the total energy difference of a neutral and a singly charged molecule; see, for example, Cappellini et al. (1997) for such a calculation in C_{60} and Na_4. The LDA systematically underestimates bandgaps by ~1 to 2 eV in the itinerant semiconductors; the situation dramatically worsens in more correlated materials, notably *f*-shell metals and some of the late-transition-metal oxides. In Hartree-Fock theory, the nonlocal exchange potential is too large because it neglects the ability of the host to screen out the bare Coulomb interaction $1/|\mathbf{r} - \mathbf{r}'|$. In the LDA, the nonlocal character of the interaction is simply missing. In semiconductors, the long-ranged part of this interaction should be present but screened by the dielectric constant ε_∞. Since $\varepsilon_\infty \gg 1$, the LDA does better by ignoring the nonlocal interaction altogether than does Hartree-Fock theory by putting it in unscreened.

Harrison's model of the gap underestimate provides us with a clear physical picture of the missing ingredient in the LDA and a semiquantitative estimate for the correction (Harrison, 1985). The LDA uses a fixed one-electron potential for all the energy bands; that is, the effective one-electron potential is unchanged for an electron excited across the gap. Thus, it neglects the electrostatic energy cost associated with the separation of electron and hole for such an excitation. This was modeled by Harrison by noting a Coulombic repulsion U between the local excess charge and the excited electron. An estimate of this

Coulombic repulsion U can be made from the difference between the ionization potential and electron affinity of the free atom; (Harrison, 1985) it is ~10 eV. U is screened by the surrounding medium so that an estimate for the additional energy cost, and therefore a rigid shift for the entire conduction band including a correction to the band-gap, is U/ε_∞. For a dielectric constant of 10, one obtains a constant shift to the LDA conduction bands of ~1 eV, with the correction larger for wider gap, smaller ε materials.

DIELECTRIC SCREENING, THE RANDOM-PHASE, *GW*, AND SX APPROXIMATIONS

The way in which screening affects the Coulomb interaction in the Fock exchange operator is similar to the screening of an external test charge. Let us then consider a simple model of static screening of a test charge in the random-phase approximation. Consider a lattice of points (spheres), with the electron density in equilibrium. We wish to calculate the screening response, i.e., the electron charge δ_{qj} at site j induced by the addition of a small external potential δV_i^0 at site i. Supposing the screening charge did not interact with itself—let us call this the noninteracting screening charge δq_j^0. This quantity is related to δV_j^0 by the noninteracting response function P_{ij}^0:

$$\delta q_k^0 = \sum_j P_{kj}^0 \delta V_j^0 \tag{9}$$

P_{kj}^0 can be calculated directly in first-order perturbation theory from the eigenvectors of the one-electron Schrödinger equation (see Equation 5 under discussion of The Local Density Approximation), or directly from the induced change in the Green's function, G^0, calculated from the one-electron Hamiltonian. By linearizing the Dyson's equation, one obtains an explicit representation of P_{ij}^0 in terms of G^0:

$$\delta G = G^0\, \delta V^0 G \approx G^0\, \delta V^0 G^0 \tag{10}$$

$$\delta q_{\mathrm{k}}^0 = \frac{1}{\pi}\,\mathrm{Im}\int_{-\infty}^{E_F} dz\, \delta G_{kk} \tag{11}$$

$$= \left[\frac{1}{\pi}\mathrm{Im}\sum_k \int_{-\infty}^{E_F} dz G_{kj}^0 G_{jk}^0\right]\delta V_j^0 \tag{12}$$

$$= \sum_j P_{kj}^0\, \delta V_j^0 \tag{13}$$

It is straightforward to see how the full screening proceeds in the random-phase approximation (RPA). The RPA assumes that the screening charge does not induce further correlations; that is, the potential induced by the screening charge is simply the classical electrostatic potential corresponding to the screening charge. Thus, δq_j^0 induces a new electrostatic potential δV_i^1 In the discrete lattice model we consider here, the electrostatic potential is linearly related to a collection of charges by some matrix M, i.e.,

$$\delta V_k^1 = \sum_j M_{jk}\delta q_k^0 \quad \delta V_i^1 = \sum_k M_{ik}\delta q_k^0 \tag{14}$$

If the q_k correspond to spherical charges on a discrete lattice, M_{ij} is $e^2/|\mathbf{r}_i - \mathbf{r}_j|$ (omitting the on-site term), or given periodic boundary conditions, M is the Madelung matrix (Slater, 1967). Equation 14 can be Fourier transformed, and that is typically done when the q_k are occupations of plane waves. In that case, $M_{jk} = 4\pi e^2 V^{-1}/k^2\,\delta_{jk}$.

Now δV^1 induces a corresponding additional screening charge δq_j^1, which induces another screening charge δq_j^2, and so on. The total perturbing potential is the sum of the external potential and the screening potentials, and the total screening charge is the sum of the δq^n. Carrying out the sum, one arrives at the screened charge, potential, and an explicit representation for the dielectric constant ε

$$\delta q = \sum_n \delta q^n = (1 - MP^0)^{-1} P^0\, \delta V^0 \tag{15}$$

$$\delta V = \delta V^0 + \delta V^{scr} = \sum_n \delta V^n \tag{16}$$

$$= (1 - MP^0)^{-1}\delta V^0 \tag{17}$$

$$= \varepsilon^{-1}\delta V^0 \tag{18}$$

In practical implementations for crystals with periodic boundary conditions, ε is computed in reciprocal space. The formulas above assumed a static screening, but the generalization is obvious if the screening is dynamic, that is, P^0 and ε are functions of energy.

The screened Coulomb interaction, W, proceeds just as in the screening of an external test charge. In the lattice model, the continuous variable $\mathbf{r}$ is replaced with the matrix M_{ij} connecting discrete lattice points; it is the Madelung matrix for a lattice with periodic boundary conditions. Then:

$$W_{ij}(E) = [\varepsilon^{-1}(E)M]_{ij} \tag{19}$$

The *GW* Approximation

Formally, the *GW* approximation is the first term in the series expansion of the self-energy in the screened Coulomb interaction, W. However, the series is not necessarily convergent, and in any case such a viewpoint offers little insight. It is more useful to think of the *GW* approximation as being a generalization of Hartree-Fock theory, with an energy-dependent, nonlocal screened interaction, W, replacing the bare coulomb interaction M entering into the exchange (see Equation 3). The one-electron equation may be written generally in terms of the self-energy Σ:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v^{\mathrm{ext}}(\mathbf{r}) + v^H(\mathbf{r})\right]\psi_i(\mathbf{r}) + \int d^3\mathbf{r}'\Sigma(\mathbf{r},\mathbf{r}';E_i)\psi(\mathbf{r}') = E_i\psi_i(\mathbf{r}) \tag{20}$$

In the *GW* approximation,

$$\Sigma^{\mathrm{GW}}(\mathbf{r},\mathbf{r}';E) = \frac{i}{2\pi}\int_{-\infty}^{\infty} d\omega e^{+i0\omega} G(\mathbf{r},\mathbf{r}';E+\omega)W(\mathbf{r},\mathbf{r}';-\omega) \tag{21}$$

The connection between *GW* theory (see Equations 20 and 21), and HF theory see Equations 1 and 3, is obvious once we make the identification of v^x with the self-energy Σ. This we can do by expressing the density matrix in terms of the Green's function:

$$\sum_j \psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}) = -\frac{1}{\pi}\int_{-\infty}^{E_F} d\omega \ \mathrm{Im}\ G(\mathbf{r}', \mathbf{r}; \omega) \tag{22}$$

$$\Sigma^{\mathrm{HF}}(\mathbf{r}, \mathbf{r}'; E) = v^x(\mathbf{r}, \mathbf{r}') = -\frac{1}{\pi}\int_{-\infty}^{E_F} d\omega \ [\mathrm{Im}\ G(\mathbf{r}, \mathbf{r}'; \omega)]\ M(\mathbf{r}, \mathbf{r}') \tag{23}$$

Comparison of Equations 21 and 23 show immediately that the *GW* approximation is a Hartree-Fock-like theory, but with the bare Coulomb interaction replaced by an energy-dependent screened interaction *W*. Also, note that in HF theory, Σ is calculated from occupied states only, while in *GW* theory, the quasiparticle spectrum requires a summation over unoccupied states as well.

GW calculations proceed essentially along these lines; in practice G, ε^{-1}, W and Σ are generated in Fourier space. The LDA is used to create the starting wave functions that generate them; however, once they are made the LDA does not enter into the Hamiltonian. Usually in semiconductors ε^{-1} is calculated only for $\omega = 0$, and the ω dependence is taken from a plasmon–pole approximation. The latter is not adequate for metals (Quong and Eguiluz, 1993).

The *GW* approximation has been used with excellent results in the calculation of optical excitations, such as the calculation of energy gaps. It is difficult to say at this time precisely how accurate the *GW* approximation is in semiconductors, because only recently has a proper treatment of the semicore states been formulated (Shirley et al., 1997). Table 3 compares some excitation energies of a few semiconductors to experiment and to the LDA.

Because of the numerical difficulty in working with products of four wave functions, nearly all *GW* calculations are carried out using plane waves. There has been, however, an all-electron *GW* method developed (Aryasetiawan and Gunnarsson, 1994), in the spirit of the augmented wave. This implementation permits the *GW* calculation of narrow-band systems. One early application to Ni showed that it narrowed the valence *d* band by ~1 eV relative to the LDA, in agreement with experiment.

The *GW* approximation is structurally relatively simple; as mentioned above, it assumes a generalized HF form. It does not possess higher-order (most notably, vertex) corrections. These are needed, for example, to reproduce the multiple plasmon satellites in the photoemission of the alkali metals. Recently, Aryasetiawan and coworkers introduced a beyond-*GW* "cumulant expansion" (Aryasetiawan et al., 1996), and very recently, an *ab initio* *T*-matrix technique (Springer et al., 1998) that they needed to account for the spectra in Ni.

Usually *GW* calculations to date use the LDA to generate *G*, *W*, etc. The procedure can be made self-consistent, i.e., *G* and *W* remade with the *GW* self-energy; in fact this was essential in the highly correlated case of NiO (Aryasetiawan and Gunnarson, 1995). Recently, Holm and von Barth (1998) investigated properties of the homogeneous electron gas with a *G* and *W* calculated self-consistently, i.e., from a *GW* potential. Remarkably, they found the self-consistency worsened the optical properties with respect to experiment, though the total energy did improve. Comparison of data in Table 3 shows that the self-consistency procedure overcorrected the gap widening in the semiconductors as well.

It may be possible in principle to calculate ground-state properties in the *GW*, but this is extremely difficult in practice, and there has been no successful attempt to date for real materials. Thus, the LDA remains the "industry standard" for total-energy calculations.

Table 3. Energy Bandgaps in the LDA, the *GW* Approximation with the Core Treated in the LDA, and the *GW* Approximation for Both Valence and Core[a]

	Expt	LDA	*GW*+ LDA Core	*GW*+ QP Core	SX	SX+ P^0(SX)
Si						
$\Gamma_{8v}\rightarrow\Gamma_{6c}$	3.45	2.55	3.31	3.28	3.59	3.82
$\Gamma_{8v}\rightarrow$X	1.32	0.65	1.44	1.31	1.34	1.54
$\Gamma_{8v}\rightarrow$L	2.1, 2.4	1.43	2.33	2.11	2.25	2.36
E_g	1.17	0.52	1.26	1.13	1.25	1.45
Ge						
$\Gamma_{8v}\rightarrow\Gamma_{6c}$	0.89	0.26	0.53	0.85	0.68	0.73
$\Gamma_{8v}\rightarrow$X	1.10	0.55	1.28	1.09	1.19	1.21
$\Gamma_{8v}\rightarrow$L	0.74	0.05	0.70	0.73	0.77	0.83
GaAs						
$\Gamma_{8v}\rightarrow\Gamma_{6c}$	1.52	0.13	1.02	1.42	1.22	1.39
$\Gamma_{8v}\rightarrow$X	2.01	1.21	2.07	1.95	2.08	2.21
$\Gamma_{8v}\rightarrow$L	1.84	0.70	1.56	1.75	1.74	1.90
AlAs						
$\Gamma_{8v}\rightarrow\Gamma_{6c}$	3.13	1.76	2.74	2.93	2.82	3.03
$\Gamma_{8v}\rightarrow$X	2.24	1.22	2.09	2.03	2.15	2.32
$\Gamma_{8v}\rightarrow$L		1.91	2.80	2.91	2.99	3.14

[a] After Shirley et al. (1997). SX calculations are by the present author, using either the LDA *G* and *W*, or by recalculating *G* and *W* with the LDA + SX potential.

The SX Approximation

Because the calculations are very heavy and unsuited to calculations of complex systems, there have been several attempts to introduce approximations to the *GW* theory. Very recently, Rücker (unpub. observ.) introduced a generalization of the LDA functional to account for excitations (van Schilfgaarde et al., 1997). His approach, which he calls the "screened exchange" (SX) theory, differs from the usual *GW* approach in that the latter does not use the LDA at all except to generate trial wave functions needed to make the quantities such as G, ε^{-1}, and W. His scheme was implemented in the LMTO–atomic spheres approximation (LMTO-ASA; see the Appendix), and promises to be extremely efficient for the calculation of excited-state properties, with an accuracy approaching

that of the *GW* theory. The principal idea is to calculate the difference between the screened exchange and the contribution to the screened exchange from the local part of the response function. The difference in Σ may be similarly calculated:

$$\delta W = W[P^0] - W[P^{0,\mathrm{LDA}}] \tag{24}$$

$$\delta W = G\ \delta W \tag{25}$$

$$\equiv \delta v^{\mathrm{SX}} \tag{26}$$

The energy bands are generated like in *GW* theory, except that the (small) correction $\delta\Sigma$ is added to the local v^{XC}, instead of Σ being substituted for v^{XC}. The analog with *GW* theory is that

$$\Sigma(\mathbf{r}, \mathbf{r}', E) = v^{\mathrm{SX}}(\mathbf{r}, \mathbf{r}') + \delta(\mathbf{r} - \mathbf{r}')[v^{xc}(\mathbf{r}) - v^{\mathrm{sx,DFT}}(\mathbf{r})] \tag{27}$$

Although it is not essential to the theory, Rücker's implementation uses only the static response function, so that the one-electron equations have the Hartree-Fock form (see Equation 1). The theory is formulated in terms of a generalization of the LDA functional, so that the *N*-particle LDA ground state is exactly reproduced, and also the $(N+1)$-particle ground state is generated with a corresponding accuracy, provided interaction of the additional electron and the *N* particle ground state is correctly depicted by $v^{XC} + \delta\Sigma$. In some sense, Rücker's approach is a formal and more rigorous embodiment of Harrison's model. Some results using this theory are shown in Table 3 along with Shirley's results. The closest points of comparison are the *GW* calculations marked "*GW* + LDA core" and "SX"; these both use the LDA to generate the *GW* self-energy Σ. Also shown is the result of a partially self-consistent calculation, in which the *G* and *W*, Σ were remade using the LDA + SX potential. It is seen that self-consistency widens the gaps, as found by Holm and von Barth (1998) for a jellium.

LDA + *U*

As an approximation, the LDA is most successful for wide-band, itinerant electrons, and begins to fail where the assumption of a single, universal, orbital-independent potential breaks down. The LDA + *U* functional attempts to bridge the gap between the LDA and model Hamiltonian approaches that attempt to account for fluctuations in the orbital occupations. It partitions the electrons into an itinerant subsystem, which is treated in the LDA, and a subsystem of localized (*d* or *f*) orbitals. For the latter, the LDA contribution to the energy functional is subtracted out, and a Hubbard-like contribution is added. This is done in the same spirit as the SX theory described above, except that here, the LDA + *U* theory attempts to properly account for correlations between atomic-like orbitals on a single site, as opposed to the long-range interaction when the components of an electron-hole pair are infinitely separated.

A motivation for the LDA + *U* functional can already be seen in the humble H atom. The total energy of the H^+ ion is 0, the energy of the H atom is 1 Ry, and the H^- ion is barely bound; thus its total energy is also ~1 Ry. Let us assume the LDA correctly predicts these total energies (as it does in practice; the LDA binding of H is ~0.97 Ry). In the LDA, the number of electrons, *N*, is a continuous variable, and the one-electron term value is, by definition, $\varepsilon \cong \delta E/\delta N$. Drawing a parabola through these three energies, it is evident that $\varepsilon \sim 0.5$ Ry in the LDA. By interpreting ε as the energy needed to ionize H (this is the atomic analog of using energy bands in a solid for the excitation spectrum), one obtains a factor-of-two error. On the other hand, using the LDA total energy difference $E(1) - E(0) \sim 0.97$ Ry predicts the ionization of the H atom rather well. Essentially the same point was made in the discussion of Opical Properties, above. In the solid, this difficulty persists, but the error is much reduced because the other electrons that are present will screen out much of the effect, and the error will depend on the context. In semiconductors, bandgap is underestimated because the LDA misses the Coulomb repulsion associated with separating an electron-hole pair (this is almost exactly analogous to the ionization of H). The cost of separating an electron-hole pair would be <~1 Ry, except that it is reduced by the dielectric screening to ~1 Ry/10 or ~1 to 2 eV.

For narrow-band systems, there is a related difficulty. Here the problem is that as the localized orbitals change their occupations, the effective potentials on these orbitals should also change. Let us consider first a simple form of the LDA + *U* functional that corrects for the error in the atomic limit as indicated for the H atom. Consider a collection of localized *d* orbitals with occupation numbers n_i and the total number of $N_d = \sum_i n_i$. In HF theory, the Coulomb interaction between the N_d electrons is $E = UN_d(N_d - 1)/2$. It is assumed that the LDA closely approximates HF theory for this term, and so this contribution is subtracted from the LDA functional, and replaced by the occupation-dependent Hubbard term:

$$E = \frac{U}{2}\sum_{i \neq j} n_i n_j \tag{28}$$

The resulting functional (Anisimov et al., 1993) is

$$E^{\mathrm{LDA+U}} = E^{\mathrm{LDA}} + \frac{U}{2}\sum_{i \neq j} n_i n_j - \frac{U}{2}N_d(N_d - 1) \tag{29}$$

U is determined from

$$U = \frac{\partial^2 E}{\partial N_d^2} \tag{30}$$

The orbital energies are now properly occupation-dependent:

$$\varepsilon_d = \frac{\partial E}{\partial N_d} = \varepsilon_{\mathrm{LDA}} + U\left(\frac{1}{2} - n_i\right) \tag{31}$$

For a fully occupied state, $\varepsilon = \varepsilon_{\mathrm{LDA}} - U/2$, and for an unoccupied state, $\varepsilon = \varepsilon_{\mathrm{LDA}} + U/2$. It is immediately evident that this functional corrects the error in the H ionization potential. For H, $U \sim 1$ Ry, so for the neutral atom $\varepsilon = \varepsilon_{\mathrm{LDA}} - U/2 \sim -1$ Ry, while for the H^+ ion, $\varepsilon = \varepsilon_{\mathrm{LDA}} + U/2 \sim 0$.

A complete formulation of the functional must account for the exchange, and should also permit the U's to be orbital-dependent. Liechtenstein et al. (1995) generalized the functional of Anisimov to be rotationally invariant. In this formulation, the orbital occupations must be generalized to the density matrices n_{ij}:

$$E^{\mathrm{LDA}+U} = E^{\mathrm{LDA}} + \frac{1}{2}\sum_{ijkl} U_{ijkl}\, n_{ij}\, n_{kl} - E_{dc}$$

$$E_{dc} = \frac{U}{2} N_d(N_d - 1) - \frac{J}{2}[n_d^{\uparrow}(N_d^{\uparrow} - 1) + n_d^{\downarrow}(n_d^{\downarrow} - 1)] \tag{32}$$

The terms proportional to J represent the exchange contribution to the LDA energy. The up and down arrows in Equation 32 represent the relative direction of electronic spin.

One of the most problematic aspects to this functional is the determination of the U's. Calculation of U from the bare Coulomb interaction is straightforward, and correct for isolated free atom, but in the solid the screening renormalizes U by a factor of $\sim$2 to 5. Several prescriptions have been proposed to generate the screened U. Perhaps the most popular is the "constrained density-functional" approach, in which a supercell is constructed and one atom is singled out as a defect. The LDA total energy is calculated as a function of the occupation number of a given orbital (the occupation number must be constrained artificially; thus, the name "constrained density functional"), and U is obtained essentially from Equation 30 (Kanamori, 1963; Anisimov and Gunnarsson, 1991). Since U is the static limit of the screened Coulomb interaction, W, it may also be calculated with the RPA as is done in the GW approximation; see Equation 20. Very recently, W was calculated in this way for Ni (Springer and Aryasetiawan, 1998), and was found to be somewhat lower (2.2 eV) than results obtained from the constrained LDA approach (3.7 to 5.4 eV).

Results from LDA + *U*

To date, the LDA + U approach has been primarily used in rather specialized areas, to attack problems beyond the reach of the LDA, but still in an *ab initio* way. It was included in this review because it establishes a physically transparent picture of the limitations of the LDA, and a prescription for going beyond it. It seems likely that the LDA + U functional, or some close descendent, will find its way into "standard" electronic structure programs in the future, for making the small but important corrections to the LDA ground-state energies in simpler itinerant systems.

The reader is referred to an excellent review article (Anisimov et al., 1997) that provides several illustrations of the method, including a description of electronic structure of the f-shell compounds CeSb and $PrBa_2Cu_3O_7$, and some late-transition metal oxides.

Relationship between LDA + *U* and *GW* Methods

As indicated above and shown in detail in Anisimov et al. (1997), the LDA + U approach is in one sense an approximation to the GW theory, or more accurately a hybrid between the LDA and a statically screened GW for the localized orbitals. One drawback to this approach is the *ad hoc* partitioning of the electrons. The practitioner must decide which orbitals are atomic-like and need to be split off. The method has so far only been implemented in basis sets that permit partitioning the orbital occupations into the localized and itinerant states. For the materials that have been investigated so far (e.g., f-shell metals, NiO), this is not a serious problem, because which orbitals are localized is fairly obvious. The other main drawback, namely, the difficulty in determining U, does not seem to be one of principle (Springer and Aryasetiawan, 1998).

On the other hand, the LDA + U approach offers an important advantage: it is a total-energy method, and can deal with both ground-state and excited-state properties.

QUANTUM MONTE CARLO

It was noted that the Hartree-Fock approximation is inadequate for solids, and that the LDA, while faring much better, has in the LD *ansatz* an uncontrolled approximation. Quantum Monte Carlo (QMC) methods are ostensibly a way to solve the Schrödinger equation to arbitrary accuracy. Thus, there is no mean-field approximation as in the other approaches discussed here. In practice, the computational effort entailed in such calculations is formidable, and for the amount of computer time available today exact solutions are not feasible for realistic systems. The number of Monte Carlo calculations today is still limited, and for the most part confined to calculations of model systems. However, as computer power continues to increase, it is likely that such an approach will be used increasingly.

The idea is conceptually simple, and there are two most common approaches. The simplest (and approximate) is the *variational* Monte Carlo (VMC) method. A more sophisticated (and in principle, potentially exact) approach is known as the Green's function, or diffusion Monte Carlo

(GFMC) method. The reader is referred to Ceperley and Kalos (1979) and Ceperley (1999)for an introduction to both of them. In the VMC approach, some functional form for the wave function is assumed, and one simply evaluates by brute force the expectation value of the many-body Hamiltonian. Since this involves integrals over each of the many electronic degrees of freedom, Monte Carlo techniques are used to evaluate them. The assumed form of the wave functions are typically adapted from LDA or HF theory, with some extra degrees of freedom (usually a Jastrow factor; see discussion of Variational Monte Carlo, below) added to incorporate beyond-LDA effects. The free parameters are determined variationally by minimizing the energy.

This is a rapidly evolving subject, still in the relatively early stages insofar as realistic solutions of real materials are concerned. In this brief review, the main ideas are outlined, but the reader is cautioned that while they appear quite simple, there are many important details that need be mastered before these techniques can be reliably prosecuted.

Variational Monte Carlo

Given a trial wave function Ψ_T, the expectation value of any operator $O(\mathbf{R})$ of all the electron coordinates, $\mathbf{R} = r_1 r_2 r_3 \ldots$, is

$$\langle O \rangle = \frac{\int d\mathbf{R} \Psi_T^*(\mathbf{R}) O(\mathbf{R}) \Psi_T(\mathbf{R})}{\int d\mathbf{R} \Psi_T^*(\mathbf{R}) \Psi_T(\mathbf{R})} \tag{33}$$

and in particular the total energy is the expectation value of the Hamiltonian H. Because of the variational principle, it is a minimum when Ψ_T is the ground-state wave function. The variational approach typically guesses a wave function whose form can be freely chosen. In practice, assumed wave functions are of the Jastrow type (Malatesta et al., 1997), such as

$$\Psi_T = D^{\uparrow}(\mathbf{R}) D^{\downarrow}(\mathbf{R}) \exp\left[\sum_i \chi(\mathbf{r}_i) + \sum_{ij} u(\mathbf{r}_i - \mathbf{r}_j)\right] \tag{34}$$

$D^{\uparrow}(\mathbf{R})$ and $D^{\downarrow}(\mathbf{R})$ are the spin-up and spin-down Slater determinants of single-particle wave functions obtained from, for example, a Hartree-Fock or local-density calculation. χ and u are the one-body and two-body pairwise terms in the Jastrow factor, with typically 10 or 20 parameters to be determined variationally.

Integration of the expectation values of interest (usually the total energy) proceeds by integration of Equation 32 using the Metropolis algorithm (Metropolis et al., 1953), which is a powerful approach to compute integrals of many dimensions. Suppose we can generate a collection of electron configurations $\{\mathbf{R}_i\}$ such that the probability of finding configuration

$$p(\mathbf{R}_i) = \frac{|\Psi_T(\mathbf{R}_i)|^2}{\int d\mathbf{R} |\Psi_T(\mathbf{R})|^2} \tag{35}$$

Then by the central limit theorem, for such a collection of points the expectation value of any operator, in particular the Hamiltonian is

$$E = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \psi_T^{-1}(\mathbf{R}_i) H(\mathbf{R}_i) \psi_T(\mathbf{R}_i) \tag{36}$$

How does one arrive at a collection of the $\{\mathbf{R}_i\}$ with the required probability distribution? This one obtains automatically from the Metropolis algorithm, and it is why Metropolis is so powerful. Each particle is randomly moved from an old position to a new position uniformly distributed. If $|\Psi_T(\mathbf{R}_{\text{new}})|^2 \geq |\Psi_T(\mathbf{R}_{\text{old}})|^2$, $\mathbf{R}_{\text{new}}$ is accepted.

Otherwise $\mathbf{R}_{\text{new}}$ is accepted with probability

$$\frac{|\psi_T(\mathbf{R}_{\text{new}})|^2}{|\psi_T(\mathbf{R}_{\text{old}})|^2} \tag{37}$$

Green's Function Monte Carlo

In the GFMC approach, one does not impose the form of the wave function, but it is obtained directly from the Schrödinger equation. The GFMC is one practical realization of the observation that the Schrödinger equation is a diffusion equation in imaginary time, $\tau = it$:

$$H\Psi = \frac{\partial \Psi}{\partial \tau} \tag{38}$$

As $\tau \to \infty$, Ψ evolves to the ground state. To see this, we express the formal solution of Equation 38 in terms of the eigenstates Ψ_n and eigenvalues E_n of H:

$$\Psi(\mathbf{R}, \tau) = e^{-(H + \text{const})\tau} \Psi(\mathbf{R}, 0) = \sum_n e^{-(E_n + \text{const})\tau} \Psi_n(\mathbf{R}) \tag{39}$$

Provided $\Psi(\mathbf{R}, 0)$ is not orthogonal to the ground state, and that the latter is sufficiently distinct in energy from the first excited state, only the ground state will survive as $\tau \to \infty$.

In the GFMC, Ψ is evolved from an integral form for the Schrödinger equation (Ceperley and Kalos, 1979):

$$\Psi(\mathbf{R}) = (E + \text{const}) \int d\mathbf{R}' G(\mathbf{R}, \mathbf{R}') \Psi(\mathbf{R}') \tag{40}$$

One defines a succession of functions $\Psi^{(n)}(\mathbf{R})$

$$\Psi^{(n+1)}(\mathbf{R}) = (E + \text{const}) \int d\mathbf{R}' G(\mathbf{R}, \mathbf{R}') \Psi^{(n)}(\mathbf{R}') \tag{41}$$

and the ground state emerges in the $n \to \infty$ limit.

One begins with a guess for the ground state, $\Psi^{(0)}(\{\mathbf{R}_i\})$, taken, for example, from a VMC calculation, and draws a set of configurations $\{\mathbf{R}_i\}$ randomly, with a probability distribution $\Psi^{(0)}(\mathbf{R})$. One estimates $\Psi^{(1)}(\mathbf{R}_i)$ by integrating Equation 41 for one time step using each of the $\{\mathbf{R}_i\}$ values. This procedure is repeated, with the collection $\{\mathbf{R}_i\}$ at step n drawn from the probability distribution $\Psi^{(n)}$.

The above prescription requires that G be known, which it is not. The key innovation of the method stems

from the observation that one can sample $G(\mathbf{R}, \mathbf{R}')\ \Psi^{(n)}(\mathbf{R}')$ without knowing G explicitly. The reader is referred to Ceperley and Kalos (1979) for details.

There is one very serious obstacle to the practical realization of the GFMC technique. Fermion wave functions necessarily have nodes, and it turns out that there is large cancellation in the statistical averaging between the positive and negative regions. The desired information gets drowned out in the noise after some number of iterations. This difficulty was initially resolved—and still is largely to this day, by making a fixed-node approximation. If the nodes are known, one can partition the GFMC into positive definite regions, and carry out Monte Carlo simulations in each region separately. The nodes are obtained in practice typically by an initial variational QMC calculation.

Applications of the Monte Carlo Method

To date, the vast majority of QMC calculations are carried out for model systems. There have been a few VMC calculations of realistic solids. The first "realistic" calculation was carried out by Fahy et al. (1990) on diamond; recently, nearly the same technique was used to calculate the heat of formation of BN with considerable success (Malatesta et al., 1997). One recent example, albeit for molecules, is the VMC study of heats of formation and barrier heights calculated for three small molecules; see Grossman and Mitás (1997) and Table 1.

Recent fixed-node GFMC have been reported for a number of small molecules, including dissociation energies for the first-row hydrides (LiH to FH) and some silicon hydrides (Lüchow and Anderson, 1996; Greeff and Lester, 1997), and the calculations are in agreement with experiment to within ~20 meV. Grossman and Mitás (1995) have reported a GFMC study of small Si clusters. There are few GFMC calculations in solids, largely because of finite-size effects. GFMC calculations are only possible in finite systems; solids must be approximated with small simulation cells subject to periodic boundary conditions. See Fraser et al. (1996) for a discussion of ways to accelerate convergence of errors resulting from such effects. As early as 1991, Li et al. (1991) published a VMC and GFMC study of the binding in Si. Their VMC and GFMC binding energies were 4.38 and 4.51 eV, respectively, in comparison to the LDA 5.30 eV using the Ceperley-Alder functional (Ceperley and Alder, 1980), and 5.15 eV for the von Barth-Hedin functional (von Barth and Hedin, 1972), and the experimental 4.64 eV. Rajagopal et al. (1995) have reported VMC and GFMC calculations on Ge, using a local pseudopotential for the ion cores.

From the dearth of reported results, it is clear that Monte Carlo methods are only beginning to make their way into the literature. But because the main constraint, the finite capacity of modern-day computers, is diminishing, this approach is becoming a promising prospect for the future.

SUMMARY AND OUTLOOK

This unit reviewed the status of modern electronic structure theory. It is seen that the local-density approximation, while a powerful and general purpose technique, has certain limitations. By comparing the relationships between it and other approaches, the relative strengths and weaknesses were elucidated. This also provides a motivation for reviewing the potential of the most promising of the newer, next-generation techniques.

It is not clear yet which approaches will predominate in the next decade or two. The QMC is yet in its infancy, impeded by the vast amount of computer resources it requires. But obviously, this problem will lessen in time with the advent of ever faster machines. Screened Hartree-Fock-like approaches also show much promise. They have the advantage that they are more mature and can provide physical insight more naturally than the QMC "heavy hammer." A general-purpose HF-like theory is already realized in the *GW* approximation, and some extensions of *GW* theory have been put forward. Whether a corresponding, all-purpose total-energy method will emerge is yet to be seen.

ACKNOWLEDGMENTS

The author thanks Dr. Paxton for a critical reading of the manuscript, and for pointing out Pettifor's work on the structural energy differences in the transition metals. This work was supported under Office of Naval Research contract number N00014-96-C-0183.

LITERATURE CITED

Abrikosov, I. A., Niklasson, A. M. N., Simak, S. I., Johansson, B., Ruban, A. V., and Skriver, H. L. 1996. *Phys. Rev. Lett.* 76:4203.

Andersen, O. K. 1975. *Phys. Rev. B* 12:3060.

Andersen, O. K. and Jepsen, O. 1984. *Phys. Rev. Lett.* 53:2571.

Anisimov, V. I. and Gunnarsson, O. 1991. *Phys. Rev. B* 43:7570.

Anisimov, V. I., Solovyev, I. V., Korotin, M. A., Czyzyk, M. T., and Sawatzky, G. A. 1993. *Phys. Rev. B* 48:16929.

Anisimov, V. I., Aryasetiawan F., Lichtenstein, A. I., von Barth U., and Hedin, L. 1997. *J. Phys. C* 9:767.

Aryasetiawan, F. 1995. *Phys. Rev. B* 52:13051.

Aryasetiawan, F. and Gunnarsson, O. 1994. *Phys. Rev. B* 54: 16214.

Aryasetiawan, F. and Gunnarsson, O. 1995. *Phys. Rev. Lett.* 74: 3221.

Aryasetiawan, F., Hedin, L., and Karlsson, K. 1996. *Phys. Rev. Lett.* 77:2268.

Ashcroft, N. W. and Mermin, D. 1976. Solid State Physics. Holt, Rinehart and Winston, New York.

Asta, M., de Fontaine, D., van Schilfgaarde, M., Sluiter, M., and Methfessel, M. 1992. First-principles phase stability study of FCC alloys in the Ti-Al system. *Phys. Rev. B* 46:5055.

Bagno, P., Jepsen, O., and Gunnarsson, O. 1989. *Phys. Rev. B* 40:1997.

Baroni, S., Giannozzi, P., and Testa, A. 1987. *Phys. Rev. Lett.* 58:1861.

Becke, A. D. 1993. *J. Chem. Phys.* 98:5648.

Blöchl, P. E. 1994. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* 50:17953.

Blöchl, P. E., Smargiassi, E., Car, R., Laks, D. B., Andreoni, W., and Pantelides, S. T. 1993. *Phys. Rev. Lett.* 70:2435.

Cappellini, G., Casula, F., and Yang, J. 1997. *Phys. Rev. B* 56:3628.

Carr, R. and Parrinello, M. 1985. *Phys. Rev. Lett.* 55:2471.

Ceperley, D. M. 1999. Microscopic simulations in physics. *Rev. Mod. Phys.* 71:S438.

Ceperley, D. M. and Alder, B. J. 1980. *Phys. Rev. Lett*. 45:566.

Ceperley, D. M. and Kalos, M. H. 1979. *In* Monte Carlo Methods in Statistical Physics (K. Binder, ed.) p. 145. Springer-Verlag, Heidelberg.

Dal Corso, A. and Resta, R. 1994. *Phys. Rev. B* 50:4327.

de Gironcoli, S., Giannozzi, P., and Baroni, S. 1991. *Phys. Rev. Lett.* 66:2116.

Dong, W. 1998. *Phys. Rev. B* 57:4304.

Fahy, S., Wang, X. W., and Louie, S. G. 1990. *Phys. Rev. B* 42:3503.

Fawcett, E. 1988. *Rev. Mod. Phys.* 60:209.

Fraser, L. M., Foulkes, W. M. C., Rajagopal, G., Needs, R. J., Kenny, S. D., and Williamson, A. J. 1996. *Phys. Rev. B* 53:1814.

Giannozzi, P. and de Gironcoli, S. 1991. *Phys. Rev. B* 43:7231.

Gonze, X., Allan, D. C., and Teter, M. P. 1992. *Phys. Rev. Lett.* 68:3603.

Grant, J. B. and McMahan, A. K. 1992. *Phys. Rev. B* 46:8440.

Greeff, C. W. and Lester, W. A. Jr. 1997. *J. Chem. Phys.* 106:6412.

Grossman, J. C. and Mitas, L. 1995. *Phys. Rev. Lett.* 74:1323.

Grossman, J. C. and Mitas, L. 1997. *Phys. Rev. Lett.* 79:4353.

Harrison, W. A. 1985. *Phys. Rev. B* 31:2121.

Hirai, K. 1997. *J. Phys. Soc. Jpn.* 66:560–563.

Hohenberg, P. and Kohn, W. 1964. *Phys. Rev. B* 136:864.

Holm, B. and von Barth, U. 1998. *Phys. Rev. B*57:2108.

Hott, R. 1991. *Phys. Rev.* B44:1057.

Jones, R. and Gunnarsson, O. 1989. *Rev. Mod. Phys.* 61:689.

Kanamori, J. 1963. *Prog. Theor. Phys.* 30:275.

Katsnelson, M. I., Antropov, V. P., van Schilfgaarde M., and Harmon, B.N. 1995. *JETP Lett.* 62:439.

Kohn, W. and Sham, J. L. 1965. *Phys. Rev.* 140:A1133.

Krishnamurthy, S., van Schilfgaarde, M., Sher, A., and Chen, A.-B. 1997. *Appl. Phys. Lett.* 71:1999.

Kubaschewski, O. and Dench, W. A. 1955. *Acta Met.* 3:339.

Kubaschewski, O. and Heymer, G. 1960. *Trans. Faraday Soc.* 56:473.

Langreth, D. C. and Mehl, M. J. 1981. *Phys. Rev. Lett.* 47:446.

Li, X.-P., Ceperley, D. M., and Martin, R. M. 1991. *Phys. Rev. B* 44:10929.

Liechtenstein, A. I., Anisimov, V. I., and Zaanen, J. 1995. *Phys. Rev. B* 52:R5467.

Lüchow, A. and Anderson, J. B. 1996. *J. Chem. Phys.* 105:7573.

Malatesta, A., Fahy, S., and Bachelet, G. B. 1997. *Phys. Rev. B* 56:12201.

McMahan, A. K., Klepeis, J. E., van Schilfgaarde M., and Methfessel, M. 1994. *Phys. Rev. B* 50:10742.

Methfessel, M., Hennig, D., and Scheffler, M. 1992. *Phys. Rev. B* 46:4816.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, N. N., Teller A. M., and Teller, E. 1953. *J. Chem. Phys.* 21:1087.

Ordejón, P., Drabold, D. A., Martin, R. M., and Grumbach, M. P. 1995. *Phys. Rev. B* 51:1456.

Ordejón, P., Artacho, E., and Soler, J. M. 1996. *Phys. Rev. B* 53:R10441.

Paxton, A. T., Methfessel, M., and Polatoglou, H. M. 1990. *Phys. Rev. B* 41:8127.

Perdew, J. P. 1991. *In* Electronic Structure of Solids (P. Ziesche and H. Eschrig, eds.), p. 11, Akademie-Verlag, Berlin.

Perdew, J. P., Burke, K., and Ernzerhof, M. 1996. *Phys. Rev. Lett.* 77:3865.

Perdew, J. P., Burke, K., and Ernzerfhof, M. 1997. *Phys. Rev. Lett.* 78:1396.

Pettifor, D. 1970. *J. Phys. C* 3:367.

Pettifor, D and Varma, J. 1979. *J. Phys. C* 12:L253.

Quong, A. A. and Eguiluz, A. G. 1993. *Phys. Rev. Lett.* 70:3955.

Rajagopal, G., Needs, R. J., James, A., Kenny, S. D., and Foulkes, W. M. C. 1995. *Phys. Rev. B* 51:10591.

Savrasov, S. Y., Savrasov, D. Y., and Andersen, O. K. 1994. *Phys. Rev. Lett.* 72:372.

Shirley, E. L., Zhu, X., and Louie, S. G. 1997. *Phys. Rev. B* 56:6648.

Skriver, H. 1985. *Phys. Rev. B* 31:1909.

Skriver, H. L. and Rosengaard, N. M. 1991. *Phys. Rev. B* 43:9538.

Slater, J. C. 1951. *Phys. Rev.* 81:385.

Slater, J. C. 1967. *In* Insulators, Semiconductors, and Metals, pp. 215–20. McGraw-Hill, New York.

Springer M. and Aryasetiawan, F. 1998. *Phys. Rev. B* 57:4364.

Springer, M., Aryasetiawan, F., and Karlsson, K. 1998. *Phys. Rev. Lett.* 80:2389.

Svane, A. and Gunnarsson, O. 1990. *Phys. Rev. Lett.* 65:1148.

van Schilfgaarde, M., Antropov V., and Harmon, B. 1996. *J. Appl. Phys.* 79:4799.

van Schilfgaarde, M., Sher, A., and Chen, A.-B. 1997. *J. Crys. Growth*. 178:8.

Vanderbilt, D. 1990. *Phys. Rev. B* 41:7892.

von Barth, U. and Hedin, L. 1972. *J. Phys. C* 5:1629.

Wang, Y., Stocks, G. M., Shelton, W. A. Nicholson, D. M. C., Szotek, Z., and Temmerman, W. M. 1995. *Phys. Rev. Lett.* 75:2867.

MARK VAN SCHILFGAARDE
SRI International
Menlo Park, California

PREDICTION OF PHASE DIAGRAMS

A phase diagram is a graphical object, usually determined experimentally, indicating phase relationships in thermodynamic space. Usually, one coordinate axis represents temperature; the others may represent pressure, volume, concentrations of various components, and so on. This unit is concerned only with temperature-concentration diagrams, limited to binary (two-component) and ternary (three-component) systems. Since more than one component will be considered, the relevant thermodynamic systems will be alloys, by definition, of metallic, ceramic, or semiconductor materials. The emphasis here will be placed primarily on metallic alloys; oxides and semiconductors are covered more extensively elsewhere in this chapter (in SUMMARY OF ELECTRONIC STRUCTURE METHODS, BONDING IN METALS, and BINARY AND MULTICOMPONENT DIFFUSION). The topic of bonding in metals, highly pertinent to this unit as well, is discussed in BONDING IN METALS.

Phase diagrams can be classified broadly into two main categories: experimentally and theoretically determined. The object of the present unit is the theoretical determina-

tion—i.e., the calculation of phase diagrams, meaning ultimately their prediction. But calculation of phase diagrams can mean different things: there are prototype, fitted, and first-principles approaches. Prototype diagrams are calculated under the assumption that energy parameters are known *a priori* or given arbitrarily. Fitted diagrams are those whose energy parameters are fitted to known, experimentally determined diagrams or to empirical thermodynamic data. First-principles diagrams are calculated on the basis of energy parameters calculated from essentially only the knowledge of the atomic numbers of the constituents, hence by actually solving the relevant Schrödinger equation. This is the "Holy Grail" of alloy theory, an objective not yet fully attained, although recently great strides have been made in that direction.

Theory also enters in the experimental determination of phase diagrams, as these diagrams not only indicate the location in thermodynamic space of existing phases but also must conform to rigorous rules of thermodynamic equilibrium (stable or metastable). The fundamental rule of equality of chemical potentials imposes severe constraints on the graphical representation of phase diagrams, while also permitting an extraordinary variety of forms and shapes of phase diagrams to exist, even for binary systems.

That is one of the attractions of the study of phase diagrams, experimental or theoretical: their great topological diversity subject to strict thermodynamic constraints. In addition, phase diagrams provide essential information for the understanding and designing of materials, and so are of vital importance to materials scientists. For theoreticians, first-principles (or *ab initio*) calculations of phase diagrams provide enormous challenges, requiring the use of advanced techniques of quantum and statistical mechanics.

BASIC PRINCIPLES

The thermodynamics of phase equilibrium were laid down by Gibbs over 100 years ago (Gibbs, 1875–1878). His was strictly a "black-box" thermodynamics in the sense that each phase was considered as a uniform continuum whose internal structure did not have to be specified. If the black boxes were very small, then interfaces had to be considered; Gibbs treated these as well, still without having to describe their structure. The thermodynamic functions, internal energy E, enthalpy H ($= E + PV$, where P is pressure and V volume), (Helmholtz) free energy F ($= E - TS$, where T is absolute temperature and S is entropy), and Gibbs free energy G ($= H - TS = F + PV$), were assumed to be known. Unfortunately, these functions (of P, V, T, say)—even for the case of fluids or hydrostatically stressed solids, as considered here—are generally not known. Analytical functions have to be "invented" and their parameters obtained from experiment. If suitable free energy functions can be obtained, then corresponding phase diagrams can be constructed from the law of equality of chemical potentials. If, however, the functions themselves are unreliable and/or the values of the parameters not determined over a sufficiently large region of thermodynamic space, then one must resort to studying in detail the physics of the thermodynamic system under consideration, investigating its internal structure, and relating microscopic quantities to macroscopic ones by the techniques of statistical mechanics. This unit will very briefly review some of the methods used, at various stages of sophistication, to arrive at the "calculation" of phase diagrams. Space does not allow the elaboration of mathematical details, much of which may be found in the author's two *Solid State Physics* review articles (de Fontaine, 1979, 1994) and references cited therein. An historical approach is given in the author's *MRS Turnbull* lecture (de Fontaine, 1996).

Classical Approach

The condition of phase equilibrium is given by the following set of equations (Gibbs, 1875–1878):

$$\mu_I^{\alpha} = \mu_I^{\beta} = \cdots = \mu_I^{\lambda} \tag{1}$$

designating the equality of chemical potentials (μ) for component I ($= 1, \ldots, n$) in phases α, β, and γ. A convenient graphical description of Equations 1 is provided in free energy-concentration space by the common tangent rule, in binary systems, or the common tangent hyperplane in multicomponent systems. A very complete account of multicomponent phase equilibrium and its graphical interpretation can be found in Palatnik and Landau (1964) and is summarized in more readable form by Prince (1966). The reason that Equations 1 lead to a common tangent construction rather than the simple search for the minima of free energy surfaces is that the equilibria represented by those equations are constrained minima, with constraint given by

$$\sum_{I=1}^{n} x_I = 1 \tag{2}$$

where $0 \leq x_I \leq 1$ designates the concentration of component I. By the properties of the x_I values, it follows that a convenient representation of concentration space for an n-component system is that of a regular simplex in $(n - 1)$-dimensional space: a straight line segment of length 1 for binary systems, an equilateral triangle for ternaries (the so-called Gibbs triangle), a regular tetrahedron for quaternaries, and so on (Palatnik and Landau, 1964; Prince, 1966). The temperature axis is then constructed orthogonal to the simplex. The phase diagram space thus has dimensions equal to the number of (nonindependent) components. Although heroic efforts have been made in that direction (Cayron, 1960; Prince, 1966), it is clear that the full graphical representation of temperature-composition phase diagrams is practically impossible for anything beyond ternary systems and awkward even beyond binaries. One then resorts to constructing two-dimensional sections, isotherms, and isopliths (constant ratio of components).

The variance f (or number of thermodynamic degrees of freedom) of a phase region where ϕ (≥ 1) phases are in equilibrium is given by the famous Gibbs phase rule, derived from Equations 1 and 2,

$$f = n - \phi + 1 \quad \text{(at constant pressure)} \tag{3}$$

It follows from this rule that when the number of phases is equal to the number of components plus 1, the equilibrium is invariant. Thus, at a given pressure, there is only one, or a finite number, of discrete temperatures at which $n + 1$ phases may coexist. Let us say that the set of $n + 1$ phases in equilibrium is $\Sigma = \{\alpha, \beta, \ldots, \gamma\}$. Then just above the invariant temperature, a particular subset of Σ will be found in the phase diagram, and just below, a different subset.

It is simpler to illustrate these concepts with a few examples (see Figs. 1 and 2). In isobaric binary systems, three coexisting phases constitute an invariant phase equilibrium, and only two topologically distinct cases are permitted: the eutectic and peritectic type, as illustrated in Figure 1. These diagrams represent schematically regions in temperature-concentration space in a narrow temperature interval just above and just below the three-phase coexistence line. Two-phase monovariant regions ($f = 1$ in Equation 3) are represented by sets of horizontal (isothermal) lines, the tie lines (or conodes), whose extremities indicate the equilibrium concentrations of the coexisting phases at a given temperature. In the eutectic case (Fig. 1A), the high-temperature phase γ (often the liquid phase) ceases to exist just below the invariant temperature, leaving only the $\alpha + \beta$ two-phase equilibrium. In the peritectic case (Fig. 1B), a new phase, β, makes its appearance while the relative amount of $\alpha + \gamma$ two-phase material decreases. It can be seen from Figure 1 that, topologically speaking, these two cases are mirror images of each other, and there can be no other possibility. How the full-phase diagram can be continued will be shown in the next section (Mean-Field Approach) in the eutectic case calculated by means of "regular solution" free energy curves.

Ternary systems exhibit three topologically distinct possibilities: eutectic type, peritectic type, and "intermediate" (Fig. 2B, C, and D, respectively). Figure 2A shows the phase regions in a typical ternary eutectic isothermal section at a temperature somewhat higher than the four-phase invariant ($f = 0$) equilibrium $\gamma + \alpha + \beta + \gamma$. In that case, three three-phase triangles come together to form a larger triangle at the invariant temperature, just below which the γ phase disappears, leaving only the $\alpha + \beta + \gamma$ monovariant equilibrium (see Fig. 2B). For simplicity, two-phase regions with their tie lines have been omitted. As for binary systems, the peritectic case is the topological mirror image of the eutectic case: the $\alpha + \beta + \gamma$ triangle splits into three three-phase triangles, as indicated in Figure 2C. In the "intermediate" case (Fig. 2D), the invariant is no longer a triangle but a quadrilateral. Just above the invariant temperature, two triangles meet along one of the diagonals of the quadrilateral; just below, it splits into two triangles along the other diagonal.

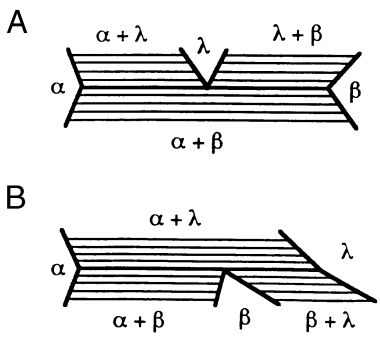


Figure 1. Binary eutectic (**A**) and peritectic (**B**) topology near the three-phase invariant equilibrium.

For quaternary systems, graphical representation is more difficult, since the possible invariant equilibria—of which there are four: eutectic type, peritectic type, and two intermediate cases—involve the merging and splitting of tetrahedra along straight lines or planes of four- or five-vertex (invariant) figures (Cayron, 1960; Prince, 1966). For even higher-component systems, full graphical representation is just about hopeless (Cayron, 1960). The Stockholm-based commercial firm Thermocalc (Jansson et al., 1993) does, however, provide software that constructs two-dimensional sections through multidimensional phase diagrams; these diagrams themselves are constructed analytically by means of Equations 1 applied to semiempirical multicomponent free energy functions.

The reason for dwelling on the universal rules of invariant equilibria is that they play a fundamental role in phase diagram determination. Regardless of the method employed—experimental or computational; prototype, fitted, or *ab initio*—these rules must be respected and often give a clue as to how certain features of the phase diagram must appear. These rules are not always respected even in published phase diagrams. They are, of course, followed in, for example, the three-volume compilation of experimentally determined binary phase diagrams published by the American Society for Metals (Mallaski, 1990).

Mean-Field Approach

The diagrams of Figures 1 and 2 are not even "prototypes"; they are schematic. The only requirements imposed are that they represent the basic types of phase equilibria in binary and ternary systems and that the free-hand-drawn phase regions obey the phase rule. For a slightly more quantitative approach, it is useful to look at the construction of a prototype binary eutectic diagram based on the simplest analytical free energy model, that of the regular solution, beginning with some useful general definitions.

Consider some extensive thermodynamic quantity, such as energy, entropy, enthalpy, or free energy F. For condensed matter, it is customary to equate the Gibbs and Helmholtz free energies. That is not to say that equilibrium will be calculated at constant volume, merely that the Gibbs free energy is evaluated at zero pressure; at reasonably low pressures, at or below atmospheric pressure, the internal states of condensed phases change very little with pressure. The total free energy of a binary solution AB for a given phase is then the sum of an "unmixed" contribution, which is the concentration-weighted average of the free energy of the pure constituents A and B, and of a mixing term that takes into account the mutual interaction of the constituents. This gives the equation

$$F = F_{\text{lin}} + F_{\text{mix}} \tag{4}$$

with

$$F_{\text{lin}} = x_{\text{A}}F_{\text{A}} + x_{\text{B}}F_{\text{B}} \tag{5}$$

where F_{lin}, known as the "linear term" because of its linear dependence on concentration, contains thermodynamic information pertaining only to the pure elements. The mixing term is the difficult one to calculate, as it also contains the important configurational entropy-of-mixing contribution. The simplest mean-field (MF) approximation regards F_{mix} as a slight deviation from the free energy of a completely random solution—i.e., one that would obtain for noninteracting particles (atoms or molecules). In the random case, the free energy is the ideal one, consisting only of the ideal free energy of mixing:

$$F_{\text{id}} = -TS_{\text{id}} \tag{6}$$

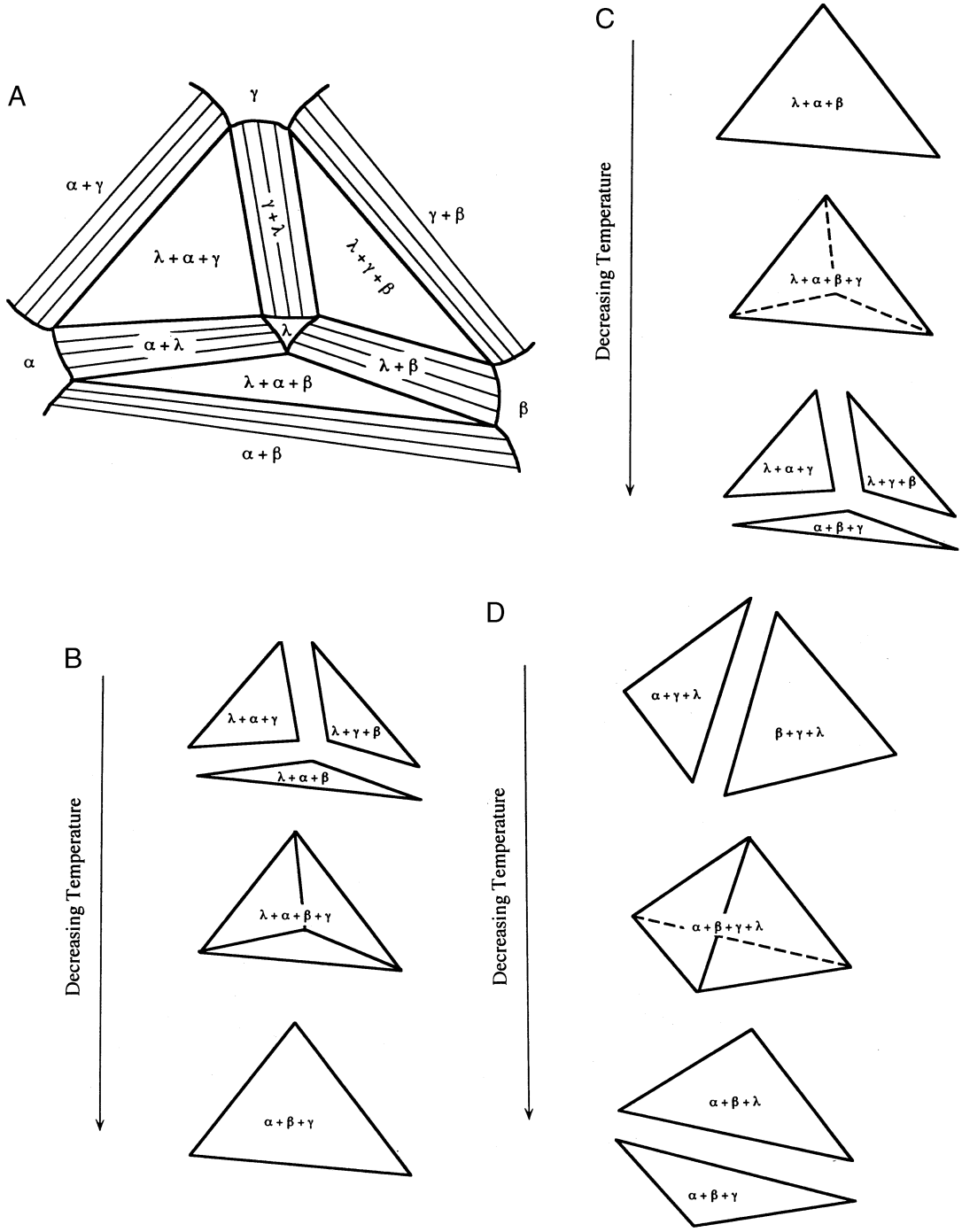


Figure 2. (**A**) Isothermal section just above the $\alpha + \beta + \gamma + \lambda$ invariant ternary eutectic temperature. (**B**) Ternary eutectic case (schematic). (**C**) Ternary peritectic case (schematic). (**D**) Ternary intermediate case (schematic).

since the energy (or enthalpy) of noninteracting particles is zero. The ideal configurational entropy of mixing per particle is given by the universal function

$$S_{\mathrm{id}} = -Nk_{\mathrm{B}}(x_{\mathrm{A}} \log x_{\mathrm{A}} + x_{\mathrm{B}} \log x_{\mathrm{B}}) \tag{7}$$

where N is the number of particles and k_{B} is Boltzmann's constant. Expression 7 is easily generalized to multicomponent systems.

What is left over in the general free energy is, by definition, the excess term, in the MF approximation taken to be a polynomial in concentration and temperature $F_{xs}(x, T)$. Here, since there is only one independent concentration because of Equation 2, one writes $x = x_{\mathrm{B}}$, the concentration of the second constituent, and $x_{\mathrm{A}} = 1 - x$ in the argument of F. In the regular solution model, the simplest formulation of the MF approximation, the simplifying assumption is also made that there is no temperature dependence of the excess entropy and the excess enthalpy depends quadratically on concentration. This gives

$$F_{xs} = x(1-x)W \tag{8}$$

where W is some generalized (concentration- and temperature-independent) interaction parameter.

For the purpose of constructing a prototype binary phase diagram in the simplest way, based on the regular solution model for the solid (σ) and liquid (λ) phases, one writes the pair of free energy expressions

$$f_\sigma = x(1-x)w_\sigma = k_{\mathrm{B}}T[x \log x + (1-x) \log (1-x)] \tag{9a}$$

and

$$f_\lambda = \Delta h - T\,\Delta s + x(1-x)w_\lambda + k_{\mathrm{B}}T[x \log x + (1-x) \log (1-x)] \tag{9b}$$

These equations are arrived at by subtracting the linear term (Equation 5) of the solid from the free energy functions of both liquid and solid (hence, the Δs in Equation 9b) and by dividing through by N, the number of particles. The lowercase letter symbols in Equations 9 indicate that all extensive thermodynamic quantities have been normalized to a single particle. The (linear) Δh term was considered to be temperature independent, and the Δs term was considered to be both temperature and concentration independent, assuming that the entropies of melting the two constituents were the same.

Free energy curves for the solid only (Equation 9a) are shown at the top portion of Figure 3 for a set of normalized temperatures ($t = k_{\mathrm{B}}T/|w|$), with $w_\sigma > 0$. The locus of common tangents, which in this case of symmetric free energy curves are horizontal, gives the miscibility gap (MG) type of phase diagram shown at the bottom portion of Figure 3. Again, tie lines are not shown in the two-phase region $\alpha + \beta$. Above the critical point ($t = 0.5$), the two phases α and β are structurally indistinguishable. Here, Δs and the constants a and b in $\Delta h = a + bx$ were fixed so that the free energy curves of the solid and liquid would intersect in a reasonable temperature interval. Within this interval, and above

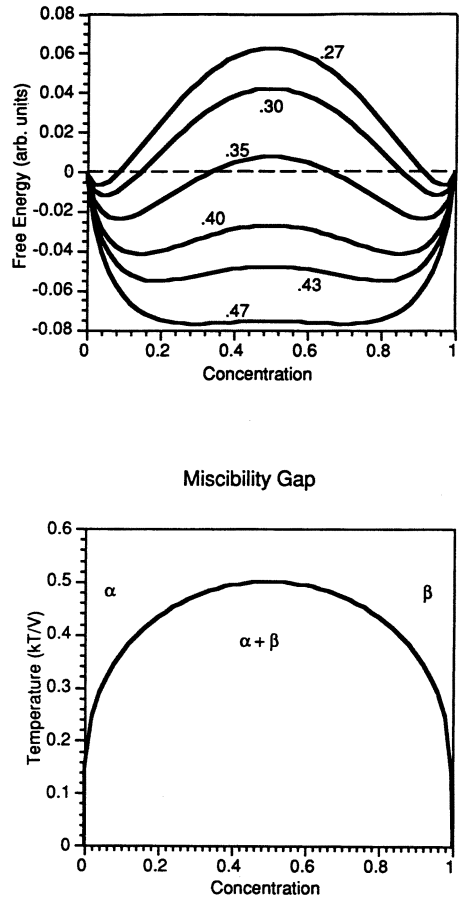


Figure 3. Regular solution free energy curves (at the indicated normalized temperatures) and resulting miscibility gap.

the eutectic ($\alpha + \beta + \gamma$) temperature, common tangency involves both free energy curves, as shown in Figure 4: the full straight-line segments are the equilibrium common tangents (filled-circle contact points) between solid (full curve) and liquid (dashed curve) free energies. For

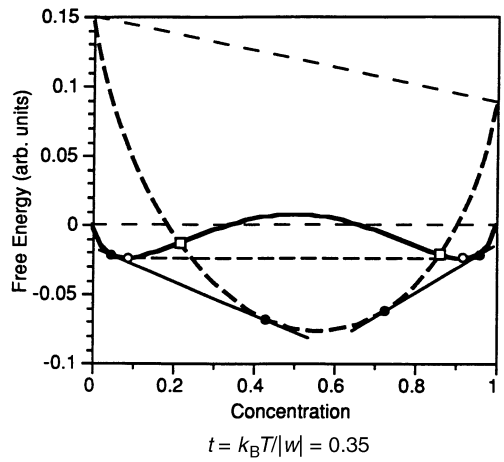


Figure 4. Free energy curves for the solid (full line) and liquid (dashed line) according to Equations 9a and 9b.

the liquid, it is assumed that $|w_\lambda| < w_\sigma$. The dashed common tangent (open circles) represents a metastable MG equilibrium above the eutectic. Open square symbols in Figure 4 mark the intersections of the two free energy curves. The loci of these intersections on a phase diagram trace out the so-called T_0 lines where the (metastable) disordered-state free energies are equal. Sets of both liquid and solid free energy curves are shown in Figure 5 for the same reduced temperatures (t) as were given in Figure 3. The phase diagram resulting from the locus of lowest common tangents (not shown) to curves such as those given in Figure 5 is shown in Figure 6. The melting points of pure A and B are, respectively, at reduced temperatures 0.47 and 0.43; the eutectic is at ~0.32.

This simple example suffices to demonstrate how phase diagrams can be generated by applying the equilibrium conditions (Equation 1) to model free energy functions. Of course, actual phase diagrams are much more complicated, but the general principles are always the same. Note that in these types of calculations the phase rule does not have to be imposed "externally"; it is built into the equilibrium conditions, and indeed follows from them. Thus, for example, the phase diagram of Figure 6 clearly belongs to the eutectic class illustrated schematically in Figure 1A. Peritectic systems are often encountered in cases where the difference of A and B melting temperatures is large with respect to the average normalized melting temperature $\bar{t}$. The general procedure outlined above may of course be extended to multicomponent systems.

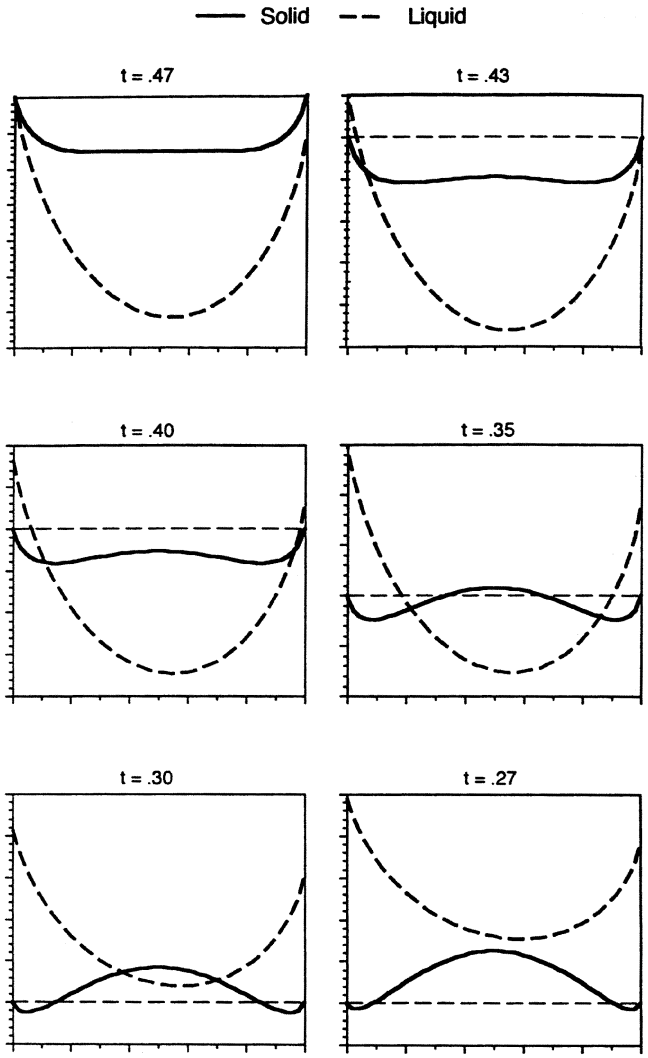


Figure 5. Free energy curves, as in Figure 4, but at the indicated reduced temperature t, generating the phase diagram of Figure 6.

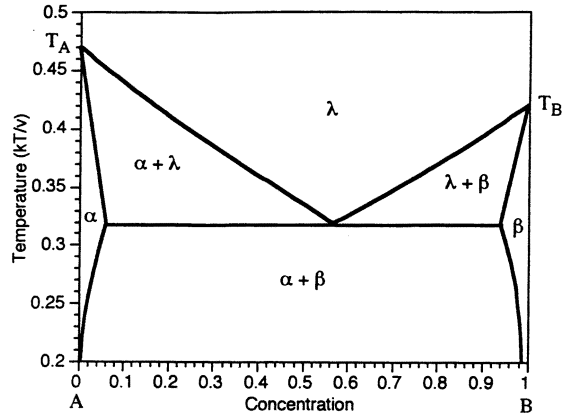


Figure 6. Eutectic phase-diagram generated from curves such as those of Figure 5.

If the interaction parameter W (Equation 8) were negative, then physically this means that unlike (A-B) near-neighbor (nn) "bonds" are favored over the average of like (A-A, B-B) "bonds." The regular solution model cannot do justice to this case, so it is necessary to introduce sublattices, one or more of which will be preferentially occupied by a certain constituent. In a sense, this atomic ordering situation is not very different from that of phase separation, illustrated in Figures 3 and 6, where the two types of atoms separate out in two distinct regions of space; here the separation is between two interpenetrating sublattices. Let there be two distinct sublattices α and β (the same notation is purposely used as in the phase separation case). The pertinent concentration variables (only the B concentration need be specified) are then x^α and x^β, which may each be written as a linear combination of x, the overall concentration of B, and η, an appropriate long-range order (LRO) parameter. In a binary "ordering" system with two sublattices, there are now two independent composition variables: x and η. In general, there is one less order parameter than there are sublattices introduced. When the ideal entropy (de Fontaine, 1979), also containing linear combinations of x and η, multiplied by $-T$ is added on, this gives so-called (generalized) Gorsky-Bragg-Williams (GBW) free energies. To construct a prototype phase diagram, possibly featuring a variety of ordered phases, it is now necessary to minimize these free energy functions with respect to the LRO parameter(s) at given x and T and then to construct common tangents.

Generalization to ordering is very important as it allows treatment of (ordered) compounds with extensive or narrow ranges of solubility. The GBW method is very convenient because it requires little in the way of computational machinery. When the GBW free energy is

expanded to second order, then Fourier transformed, it leads to the "method of concentration waves," discussed in detail elsewhere (de Fontaine, 1979; Khachaturyan, 1983), with which it is convenient to study general properties of ordering in solid solutions, particularly ordering instabilities. However, to construct fitted phase diagrams that resemble those determined experimentally, it is found that an excess free energy represented by a quadratic form will not provide enough flexibility for an adequate fit to thermodynamic data. It is then necessary to write F_{xs} as a polynomial of degree higher than the second in the relevant concentration variables: the temperature T also appears, usually in low powers.

The extension of the GBW method to include higher-degree polynomials in the excess free energy is the technique favored by the CALPHAD group (see *CALPHAD* journal published by Elsevier), a group devoted to the very useful task of collecting and assessing experimental thermodynamic data, and producing calculated phase diagrams that agree as closely as possible with experimental evidence. In a sense, what this group is doing is "thermodynamic modeling," i.e., storing thermodynamic data in the form of mathematical functions with known parameters. One may ask, if the phase diagram is already known, why calculate it? The answer is that phase diagrams have often not been obtained in their entirety, even binary ones, and that calculated free energy functions allow some extrapolation into the unknown. Also, knowing free energies explicitly allows one to extract many other useful thermodynamic functions instead of only the phase diagram itself as a graphical object. Another useful though uncertain aspect is the extrapolation of known lower-dimensional diagrams into unknown higher-dimensional ones. Finally, appropriate software can plot out two-dimensional sections through multicomponents. Such software is available commercially from Thermocalc (Jansson et al., 1993). An example of a fitted phase diagram calculated in this fashion is given below (see Examples of Applications).

It is important to note, however, that the generalizations just described—sublattices, Fourier transforms, polynomials—do not alter the degree of sophistication of the approximation, which remains decidedly MF, i.e., which replaces the averages of products (of concentrations) by products of averages. Such a procedure can give rise to certain unacceptable behavior of calculated diagrams, as explained below (see discussion of Cluster-Expansion Free Energy).

Cluster Approach

Thus far, the treatment has been restricted to a black-box approach: each phase is considered as a thermodynamic continuum, possibly in equilibrium with other phases. Even in the case of ordering, the approach is still macroscopic, each sublattice also being considered as a thermodynamic continuum, possibly interacting with other sublattices. It is not often appreciated that, even though the crystal structure is taken into consideration in setting up the GBW model, the resulting free energy function still does not contain any geometrical information. Each sublattice could be located anywhere, and such things as coordination numbers can be readily incorporated in the effective interactions W. In fact, in the "MF-plus-ideal-entropy" formulation, solids and liquids are treated in the same way: terminal solid solutions, compounds, and liquids are represented by similar black-box free energy curves, and the lowest set of common tangents are constructed at each temperature to generate the phase diagram.

With such a simple technique available, which generally produces formally satisfactory results, why go any further?

There are several reasons to perform more extensive calculations: (1) the MF method often overestimates the configurational entropy and enthalpy contributions; (2) the method often produces qualitatively incorrect results, particularly where ordering phenomena are concerned (de Fontaine, 1979); (3) short-range order (SRO) cannot be taken into account; and (4) the CALPHAD method cannot make thermodynamic predictions based on "atomistics." Hence, a much more elaborate microscopic theory is required for the purpose of calculating *ab initio* phase diagrams.

To set up such a formalism, on the atomic scale, one must first define a reference lattice or structure: face-centered cubic (fcc), body-centered cubic (bcc), or hexagonal close-packed (hcp), for example. The much more difficult case of liquids has not yet been worked out; also, for simplicity, only binary alloys (A-B) will be considered here. At each lattice site (p), define a spinlike configurational operator σ_p equal to +1 if an atom of type A is associated with it, -1 if B. At each site, also attach a (three-dimensional) vector $\mathbf{u}_p$ that describes the displacement (static or dynamic) of the atom from its ideal lattice site. For given configuration $\boldsymbol{\sigma}$ (a boldfaced $\boldsymbol{\sigma}$ indicates a vector of N $\sigma_p = \pm 1$ components, representing the given configuration in a supercell of N lattice sites), an appropriate Hamiltonian is then constructed, and the energy E is calculated at $T = P = 0$ by quantum mechanics. In principle, this program has to be carried out for a variety of supercell configurations, lattice parameters, and internal atomic displacements; then configurational averages are taken at given T and P to obtain expectation values of appropriate thermodynamic functions, in particular the free energy. Such calculations must be repeated for all competing crystalline phases, ordered or disordered, as a function of average concentration x, and lowest tangents constructed at various temperatures and (usually) at $P = 0$. As described, this task is an impossible one, so that rather drastic approximations must be introduced, such as using a variational principle instead of the correct partition function procedure to calculate the free energy and/or using a Hamiltonian based on semiempirical potentials. The first of these approximation methods—derivation of a variational method for the free energy, which is at the heart of the cluster variation method (CVM; Kikuchi, 1951)—is presented here.

The exact free energy F of a thermodynamic system is given by

$$f = -k_B T \ln Z \tag{10}$$

with partition function

$$Z = \sum_{\text{states}} e^{-E(\text{state})/k_B T} \tag{11}$$

The sum over states in Equation 11 may be partially decoupled (Ceder, 1993):

$$Z \Rightarrow \sum_{\{\sigma\}} e^{-E_{\text{stat}}(\sigma)/k_B T} \sum_{\text{dyn}} e^{-E_{\text{dyn}}(\sigma)/k_B T} \tag{12}$$

where

$$E_{\text{stat}}(\sigma) = E_{\text{repl}}(\sigma) + E_{\text{displ}}(\sigma) \tag{13}$$

represents the sum of a replacive and a displacive energy. The former is the energy that results from the rearrangements of types of atoms, still centered on their associated lattice sites; this energy contribution has also been called "ordering," "ideal," "configurational," or "chemical." The displacive energy is associated with static atomic displacements. These displacements themselves may produce volume effects, or change in the volume of the unit cell; cell-external effects, or change in the shape of the unit cell at constant volume; and cell-internal effects, or relaxation of atoms inside the unit cell away from their ideal lattice positions (Zunger, 1994). The sums over dynamical degrees of freedom may be replaced by Boltzmann factors of the type $F_{\text{dyn}}(\sigma) = E_{\text{dyn}}(\sigma) - TS_{\text{dyn}}$ to obtain finally a new partition function

$$Z = \sum_{\{\sigma\}} e^{-\Psi(\sigma)/k_B T} \tag{14}$$

with "hybrid" energy

$$\Psi(\sigma) = E_{\text{repl}}(\sigma) + E_{\text{displ}}(\sigma) + F_{\text{dyn}}(\sigma) \tag{15}$$

Even when considering the simplest case of leaving out static displacive and dynamical effects, and writing the replacive energy as a sum of nn pair interactions, this Ising model is impossible to "solve" in three dimensions: i.e., the summation in the partition function cannot be expressed in closed form. One must resort to Monte Carlo simulation or to variational solutions. To these several energies will correspond entropy contributions, particularly a configurational entropy associated with the replacive energy and a vibrational entropy associated with dynamical displacements.

Consider the configurational entropy. The idea of the CVM is to write the partition function not as a sum over all possible configurations but as a sum over selected cluster probabilities ($\mathbf{x}$) for small groups of atoms occupying the sites of 1-, 2-, 3-, 4-, many-point clusters; for example:

$$\{\mathbf{x}\} = \begin{cases} x_A, x_B & \text{points} \\ x_{AA}, x_{AB} & \text{pairs} \\ x_{AAA}, x_{AAB} & \text{triplets} \\ x_{AAAA} & \text{quads} \end{cases} \tag{16}$$

The transformation

$$Z = \sum_{\{\sigma\}} e^{-E(\sigma)/k_B T} \Rightarrow \sum_{\{\mathbf{x}\}} g(\mathbf{x}) e^{-E(\mathbf{x})/k_B T} = \sum_{\{\mathbf{x}\}} e^{-F(\mathbf{x})/k_B T} \tag{17}$$

introduces the "complexion" $g(\mathbf{x})$, or number of ways of creating configurations having the specified values of pair, triplet, ... probabilities $\mathbf{x}$. there corresponds a configurational entropy

$$S(\mathbf{x}) = k_B \log g(\mathbf{x}) \tag{18}$$

hence, a free energy $F = E - TS$, which is the one that appears in Equation 17. The optimal variational free energy is then obtained by minimizing $F(\mathbf{x})$ with respect to the selected set of cluster probabilities and subject to consistency constraints. The resulting $F(\mathbf{x}^*)$, with $\mathbf{x}^*$ being the values of the cluster probabilities at the minimum of $F(\mathbf{x})$, will be equal to or larger than the true free energy, hopefully close enough to the true free energy when an adequate set of clusters has been chosen.

Kikuchi (1951) gave the first derivation of the $g(\mathbf{x})$ factor for various cluster choices, which was followed by the more algebraic derivations of Barker (1953) and Hijmans and de Boer (1955). The simplest cluster approximation is of course that of the point, i.e., that of using only the lattice point as a cluster. The symbolic formula for the complexion in this approximation is

$$g(\mathbf{x}) = \frac{N!}{\{\bullet\}} \tag{19}$$

with the "Kikuchi symbol" (for binaries and for a single sublattice) defined by

$$\{\bullet\} = \prod_{I=A,B} (x_I N)! \tag{20}$$

The logarithm of $g(\mathbf{x})$ is required in the free energy, and it is seen, by making use of Stirling's approximation, that the entropy, calculated from Equations 19 and 20, is just the ideal entropy given in Equation 7. Thus the MF entropy is identical to the CVM point approximation.

To improve the approximation, it is necessary to go to higher clusters, pairs, triplets (triangles), quadruplets (tetrahedron, ...), and so on. Frequently used approximations for larger clusters are given in Figure 7, where the Kikuchi symbols now include pair, triplet, quadruplet, ..., cluster concentrations, as in Equation 16, instead of the simple x_I of Equation 20. The first panel (top left) shows the pair formula for the linear chain with nn interaction. This simplest of Ising models is treated exactly in this formulation. The extension of the entropy pair formula to higher-dimensional lattices is given in the top right panel, with ω being half the nn coordination number. The other formulas give better approximations; in fact, the "tetrahedron" formula is the lowest one that will treat ordering properly in fcc lattices. The octahedron-tetrahedron formula (Sanchez and de Fontaine, 1978) is more useful

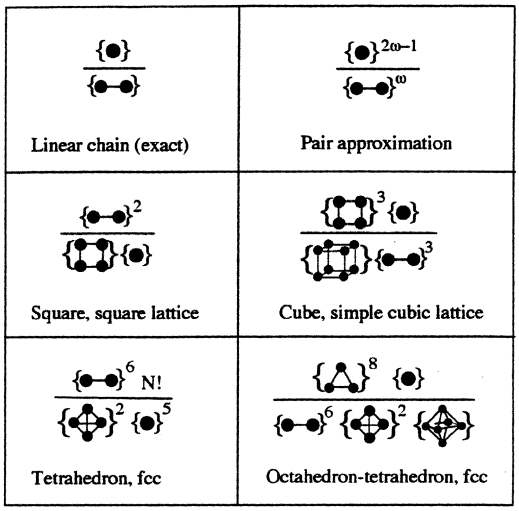


Figure 7. Various CVM approximations in the symbolic Kikuchi notation.

and more accurate for ordering on the fcc lattice because it contains next-nearest-neighbor (nnn) pairs, which are important when considering associated effective pair interactions (see below). Today, CVM entropy formulas can be derived "automatically" for arbitrary lattices and cluster schemes by computer codes based on group theory considerations. The choice of the proper cluster scheme to use is not a simple matter, although there now exist heuristic methods for selecting "good" clusters (Vul and de Fontaine, 1993; Finel, 1994).

The realization that configurational entropy in disordered systems was really a many-body thermodynamic expression led to the introduction of cluster probabilities in free energy functionals. In turn, cluster variables led to the development of a very useful cluster algebra, to be described very briefly here [see the original article (Sanchez et al., 1984) or various reviews (e.g., de Fontaine, 1994) for more details]. The basic idea was to develop complete orthonormal sets of cluster functions for expanding arbitrary functions of configurations, a sort of Fourier expansion for disordered systems. For simplicity, only binary systems will be considered here, without explicit consideration of sublattices. As will become apparent, the cluster algebra applied to the replacive (or configurational) energy leads quite naturally to a precise definition of effective cluster interactions, similar to the phenomenological w parameters of MF equations 8 and 9. With such a rigorous definition available, it will be in principle possible to calculate those parameters from "first principles," i.e., from the knowledge of only the atomic numbers of the constituents.

Any function of configuration $f(\boldsymbol{\sigma})$ can be expanded in a set of cluster functions as follows (Sanchez et al., 1984):

$$f(\sigma) = \sum_{\alpha} F_\alpha \Phi_\alpha(\sigma) \tag{21}$$

where the summation is over all clusters of lattice points (α designating a cluster of n_α points of a certain geometrical type: tetrahedron, square, ...), F_α are the expansion coefficients, and the Φ_α are suitably defined cluster functions. The cluster functions may be chosen in an infinity of ways but must form a complete set (Asta et al., 1991; Wolverton et al., 1991a; de Fontaine et al., 1994). In the binary case, the cluster functions may be chosen as products of σ variables ($= \pm 1$) over the cluster sites. The sets may even be chosen orthonormal, in which case the expansion (Equation 21) may be "inverted" to yield the cluster coefficients

$$F_\alpha = \rho_0 \sum_{\sigma} f(\sigma)\Phi_\alpha(\sigma) \equiv \langle \Phi_\alpha, f \rangle \tag{22}$$

where the summation is now over all possible 2^N configurations and ρ_0 is a suitable normalization factor. More general formulations have been given recently (Sanchez, 1993), but these simple formulas, Equations 21 and 22, will suffice to illustrate the method. Note that the formalism is exact and moreover computationally useful if series convergence is sufficiently rapid. In turn, this means that the burden and difficulty of treating disordered systems can now be carried by the determination of the cluster coefficients F_α. An analogy may be useful here: in solving linear partial differential equations, for example, one generally does not "solve" directly for the unknown function; one instead describes it by its representation in a complete set of functions, and the burden of the work is then carried by the determination of the coefficients of the expansion.

The most important application of the cluster expansion formalism is surely the calculation of the configurational energy of an alloy, though the technique can also be used to obtain such thermodynamic quantities as molar volume, vibrational entropy, and elastic moduli as a function of configuration. In particular, application of Equation 21 to the configurational energy E provides an exact expression for the expansion coefficients, called effective cluster interactions, or ECIs, V_α. In the case of pair interactions (for lattice sites p and q, say), the effective pair interaction is given by

$$V_{pq} = \tfrac{1}{4}(\bar{E}_{\mathrm{AA}} + \bar{E}_{\mathrm{BB}} - \bar{E}_{\mathrm{AB}} - \bar{E}_{\mathrm{BA}}) \tag{23}$$

where $\bar{E}_{IJ}$ ($I, J =$ A or B) is the average energy of all configurations having atom of type I at site p and of type J at q. Hence, it is seen that the V_α parameters, those required by the Ising model thermodynamics, are by no means "pair potentials," as they were often referred to in the past, but differences of energies, which can in fact be calculated. Generalizations of Equation 23 can be obtained for multiplet interactions such as V_{pqr} as a function of AAA, AAB, etc., average energies, and so on.

How, then, does one go about calculating the pair or cluster interactions in practice? One method makes use of the orthogonality property explicitly by calculating the average energies appearing in Equation 23. To be sure, not all possible configurations $\boldsymbol{\sigma}$ are summed over, as required by Equation 22, but a sampling is taken of, say, a few tens of configurations selected at random around a given IJ

pair, for instance, in a supercell of a few hundreds of atoms. Actually, there is a way to calculate directly the difference of average energies in Equation 23 without having to take small differences of large numbers, using what is called the method of direct configurational averaging, or DCA (Dreyssé et al., 1989; Wolverton et al., 1991b, 1993; Wolverton and Zunger, 1994). It is unfortunately very computer intensive and has been used thus far only in the tight binding approximation of the Hamiltonian appearing in the Schrödinger equation (see Electronic Structure Calculations section).

A convenient method of obtaining the ECIs is the structure inversion method (SIM), also known as the Connolly-Williams method after those who originally proposed the idea (Connolly and Williams, 1983). Consider a particular configuration $\boldsymbol{\sigma}_s$, for example, that of a small unit-cell ordered structure on a given lattice. Its energy can be expanded in a set of cluster functions as follows:

$$E(\sigma_s) = \sum_{r=1}^{n} m_r V_r \bar{\Phi}_r(\boldsymbol{\sigma}_s) \qquad (s = 1, 2, \ldots, q) \tag{24}$$

where $\bar{\Phi}_r(\boldsymbol{\sigma}_s)$ are cluster functions for that configuration averaged over the symmetry-equivalent positions of the cluster of type r, with m_r being a multiplicity factor equal to the number of r-clusters per lattice point. Similar equations are written down for a number (q) of other ordered structures on the same lattice. Usually, one takes $q > n$. The (rectangular) linear system (Equation 24) can then be solved by singular-value decomposition to yield the required ECIs V_r. Once these parameters have been calculated, the configurational energy of any other configuration, τ, on a supercell of arbitrary size can be calculated almost trivially:

$$E(\tau) = \sum_{s=1}^{q} C_s(\tau) E(\boldsymbol{\sigma}_s) \tag{25}$$

where the coefficients C_s may be obtained (symbolically) as

$$C_s(\tau) = \sum_{r=1}^{n} \bar{\Phi}_r^{-1}(\boldsymbol{\sigma}_s) \bar{\Phi}_r(\tau) \tag{26}$$

where $\bar{\Phi}_r^{-1}$ is the (generalized) inverse of the matrix of averaged cluster functions. The upshot of this little exercise is that, according to Equation 5, the energy of a large-unit-cell structure can be obtained as a linear combination of the energies of small-unit-cell structures, which are presumably easy to compute (e.g., by electronic structure calculation). So what about electronic structure calculations? The Electronic Structure Calculations section gives a quick overview of standard techniques in the "alloy theory" context, but before getting to that, it is worth considering the difference that the CVM already makes, compared to the MF approximation, in the calculation of a prototype "ordering" phase diagram.

Cluster Expansion Free Energy

Thermodynamic quantities are macroscopic ones, actually expectation values of microscopic ones. For the energy, for example, the expectation value $\langle E \rangle$ is needed in cluster-expanded form. This is achieved by taking the ensemble average of the general cluster expansion formula (Equation 22) to obtain, per lattice point,

$$\langle E(\boldsymbol{\sigma}) \rangle \sum_{\alpha} V_\alpha \xi_\alpha \tag{27}$$

with correlation variables defined by

$$\xi_\alpha = \langle \Phi_\alpha(\sigma) \rangle = \langle \sigma_1 \sigma_2 \ldots \sigma_{n_\alpha} \rangle \tag{28}$$

the latter equality being valid for binary systems. The entropy, already an averaged quantity, is expressed in the CVM by means of cluster probabilities $\mathbf{x}_\alpha$. These probabilities are related linearly to the correlation variables ξ_α by means of the so-called configuration matrix (Sanchez and de Fontaine, 1978), whose elements are most conveniently calculated by group theory computer codes. The required CVM free energy is obtained by combining energy (27) and entropy (18) contributions and using Stirling's approximation for the logarithm of the factorials appearing in the $g(\mathbf{x})$ expressions such as those illustrated in Figure 7, for example,

$$f \equiv f(\xi_1, \xi_2, \ldots) = \sum_r m_r V_r \xi_r - k_B T \sum_{k=1}^{K} m_k \gamma_k \times \sum_{\alpha_k} x_k(\sigma_k) \ln x_k(\sigma_k) \tag{29}$$

where the indices r and k denote sequential ordering of the clusters used in the expansion, K being the order of the largest cluster retained, and where the second summation is over all possible A/B configurations $\boldsymbol{\sigma}_k$ of cluster k. The correlations ξ_r are linearly independent variational parameters for the problem at hand, which means that the best choice for the correct free energy for the cluster approximation adopted is obtained by minimizing Equation 29 with respect to ξ_r. The γ_k coefficients in the entropy expression are the so-called Kikuchi-Barker coefficients, equal to the exponents of the Kikuchi symbols found in the symbolic formulas of Figure 7, for example. Minimization of the free energy (Equation 29) leads to systems of simultaneous algebraic equations (sometimes as many as a few hundred) in the ξ_r. This task greatly limits the size of clusters that can be handled in practice.

The CVM (Equation 29) and MF (Equation 9) free energy expressions look very different, raising the question of whether they are related to one another. They are, in the sense that the MF free energy must be the infinite-temperature limit of the CVM free energy. At high temperatures all correlations must disappear, which means that one can write the ξ_r as powers of the "point" correlations ξ_1. Since ξ_1 is equal to $2x - 1$, the energy term in Equation 29 is transformed, in the MF approximation, to a polynomial in x ($= x_B$, the average concentration of the B component). This is precisely the type of expression used by the CALPHAD group for the enthalpy, as mentioned above (see Mean-Field Approach). In the MF

limit, products of x_A and x_B must also be introduced in the cluster probabilities appearing in the configurational entropy. It is found that, for any cluster probability, the logarithmic terms in Equation 29 reduce to $n[x_A \log x_A + x_B \log x_B]$, precisely the MF (ideal) entropy term in Equation 7 multiplied by the number n of points in the cluster, thereby showing that the sum of the Kikuchi-Barker coefficients, weighted by n, must equal -1. This simple property may be used to verify the validity of a newly derived CVM entropy expression; it also shows that the type, or shape, of the clusters is irrelevant to the MF model, as only the number of points matters. This is another demonstration that the MF approximation ignores the true geometry of local atomic arrangements.

Note that, strictly speaking, the CVM is also a MF formalism, but it treats the correlations (almost) correctly for lattice distances included within the largest cluster considered in the approximation and includes SRO effects. Hence the CVM is a multisite MF model, whereas the "point" MF is a single-site model, the one previously referred to as the MF (ideal, regular solution, GBW, concentration wave, etc.) model.

It has been seen that the MF model is the same as the CVM point approximation, and the MF free energy can be derived by making the superposition ansatz, $x_{AB} \rightarrow x_A x_B$, $x_{AAB} \rightarrow x_A^2 x_B$, and so on. Of course, the point is very much simpler to handle than the "cluster" formalism. So how much is lost in terms of actual phase diagram results in going from MF to CVM? To answer that question, it is instructive to consider a very striking yet simple example, the case of ordering on the fcc lattice with only first-neighbor effective pair interactions (de Fontaine, 1979). The MF phase diagram of Figure 8A was constructed by the GBW MF approximation (Shockley, 1938), that of Figure 8B by the CVM in the tetrahedron approximation, whose symbolic formula is given in Figure 7 (de Fontaine, 1979, based on van Baal, 1973, and Kikuchi, 1974). It is seen that, even in a topological sense, the two diagrams are completely different: the GBW diagram (only half of the concentration axis is represented) shows a double second-order transition at the central composition ($x_B = 0.5$), whereas the CVM diagram shows three first-order transitions for the three ordered phases A_3B, AB_3 ($L1_2$-type ordered structure), and AB ($L1_0$). The short horizontal lines in Figure 8B are tie lines, as seen in Figure 1. The two very small three-phase regions are actually peritectic-like, topologically, as in Figure 1B.

There can be no doubt about which prototype diagram is correct, at least qualitatively: the CVM version, which in turn is very similar to the solid-state portion of the experimental Cu-Au phase diagram (Massalski, 1986; in this particular case, the first edition is to be preferred). The large discrepancy between the MF and CVM versions arises mainly from the basic geometrical figure of the fcc lattice being the nn equilateral triangle (and the associated nn tetrahedron). Such a figure leads to frustration where, as in this case, the first nn interaction favors unlike atomic pairs. This requirement can be satisfied for two of the nn pairs making up the triangle, but the third pair must necessarily be a "like" pair, hence the conflict. This frustration also lowers the transition temperature below

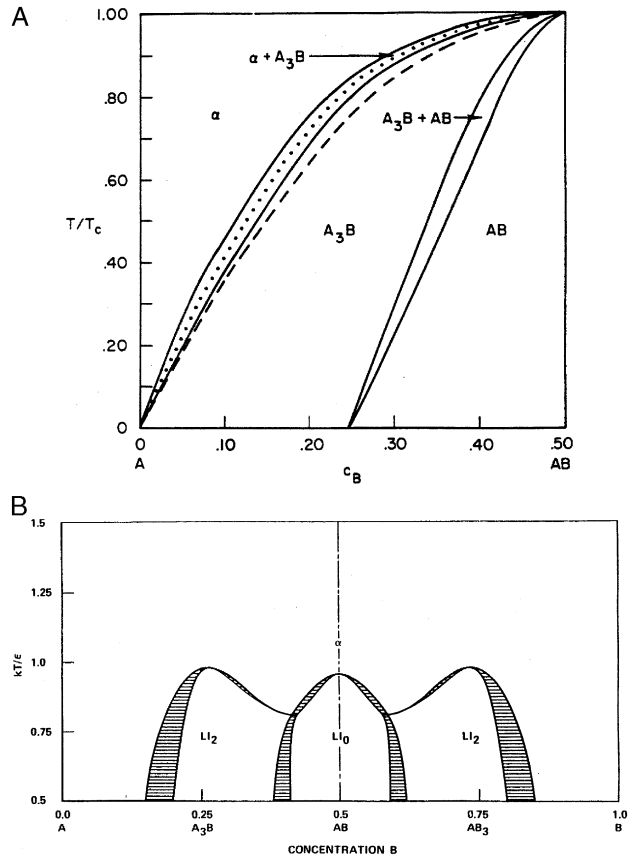


Figure 8. (**A**) Left half of fcc ordering phase diagram (full lines) with nn pair interactions in the GBW approximation; from de Fontaine (1979), according to W. Shockley (1938). Dotted and dashed lines are, respectively, loci of equality of free energies for disordered and ordered A_3B ($L1_2$) phases (so-called T_0 line) and ordering spinodal (not used in this unit). (**B**) The fcc ordering phase diagram with nn pair interactions in the CVM approximation; from de Fontaine (1979), according to C.M. van Baal (1973) and R. Kikuchi (1974).

what it would be in nonfrustrated systems. The cause of the problem with the MF approach is of course that it does not "know about" the triangle figure and the frustrated nature of its interactions. Numerically, also, the CVM configurational entropy is much more accurate than the GBW (MF), particularly if higher-cluster approximations are used.

These are not the only reasons for using the cluster approach: the cluster expansion provides a rigorous formulation for the ECI; see Equation 23 and extensions thereof. That means that the ECIs for the replacive (or strictly configurational) energy can now be considered not only as fitting parameters but also as quantities that can be rigorously calculated *ab initio*. Such atomistic computations are very difficult to perform, as they involve electronic structure calculations, but much progress has been realized in recent years, as briefly summarized in the next section.

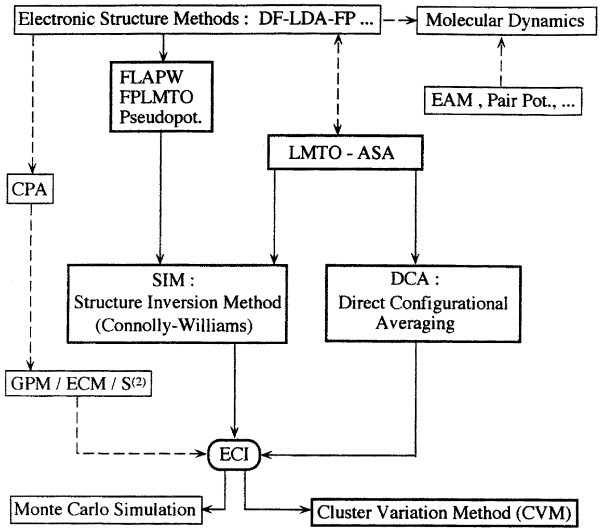


Figure 9. Flow chart for obtaining ECI from electronic structure calculations. Pot ≡ potentials.

Electronic Structure Calculations

One of most important breakthroughs in condensed-matter theory occurred in the mid-1960s with the development of the (exact) density functional theory (DFT) and its local density approximation (LDA), which provided feasible means for performing large-scale electronic structure calculations. It then took almost 20 years for practical and reliable computer codes to be developed and implemented on fast computers. Today many fully developed codes are readily accessible and run quite well on workstations, at least for moderate-size applications. Figure 9 introduces the uninitiated reader to some of the alphabet soup of electronic structure theory relevant to the subject at hand.

What the various methods have in common is solving the Schrödinger equation (sometimes the Dirac equation) in an electronically self-consistent manner in the local density approximation. The methods differ from one another in the choice of basis functions used in solving the relevant linear differential equation—augmented plane waves [(A)PW], muffin tin orbitals (MTOs), atomic orbitals in the tight binding (TB) method—generally implemented in a non-self-consistent manner. The methods also differ in the approximations used to represent the potential that an electron sees in the crystal or molecule, F or FP, designating "full potential." The symbol L designates a linearization of the eigenvalue problem, a characteristic not present in the "multiple-scattering" KKR (Korringa-Kohn-Rostocker) method. The atomic sphere approximation (ASA) is a very convenient simplification that makes the LMTO-ASA (linear muffin tin orbital in the ASA) a particularly efficient self-consistent method to use, although not as reliable in accuracy as the full-potential methods. Another distinction is that pseudopotential methods define separate "potentials" for $s, p, d, \ldots$, valence states, whereas the other self-consistent techniques are "all-electron" methods. Finally, note that plane-wave basis methods (pseudopotential, FLAPW) are the most convenient ones to use whenever forces on atoms need to be calculated. This is an important consideration as correct cohesive energies require that total energies be minimized with respect to lattice "constants" and possible local atomic displacements as well.

There is not universal agreement concerning the definition of "first-principles" calculations. To clarify the situation, let us define levels of *ab initio* character:

First level: Only the z values (atomic numbers) are needed to perform the electronic structure calculation, and electronic self-consistency is performed as part of the calculation via the density functional approach in the LDA. This is the most fundamental level; it is that of the APW, LMTO, and *ab initio* pseudopotentials. One can also use semiempirical pseudopotentials, whose parameters may be fitted partially to experimental data.

Second level: Electronic structure calculations are performed that do not feature electronic self-consistency. Then the Hamiltonian must be expressed in terms of parameters that must be introduced "from the outside." An example is the TB method, where indeed the Schrödinger equation is solved (the Hamiltonian is diagonalized, the eigenvalues summed up), and the TB parameters are calculated separately from a level 1 formulation, either from LMTO-ASA, which can give these parameters directly (hopping integrals, on-site energies) or by fitting to the electronic band structure calculated, for example, by a level 1 theory. The TB formulation is very convenient, can treat thousands of atoms, and often gives valuable band structure energies. Total energies, for which repulsive terms must be constructed by empirical means, are not straightforward to obtain. Thus, a typical level 2 approach is the combined LMTO-TB procedure.

Third level: Empirical potentials are obtained directly by fitting properties (cohesive energies, formation energies, elastic constants, lattice parameters) to experimentally measured ones or to those calculated by level 1 methods. At level 3, the "potentials" can be derived in principle from electronic structure but are in fact (partially) obtained by fitting. No attempt is made to solve the Schrödinger equation itself. Examples of such approaches are the embedded-atom method (EAM) and the TB method in the second-moment approximation (references given later).

Fourth level: Here the interatomic potentials are strictly empirical and the total energy, for example, is obtained by summing up pair energies. Molecular dynamics (MD) simulations operate with fourth- or third-level approaches, with the exception of the famous Car-Parrinello method, which performs *ab initio* molecular dynamics, combining MD with the level 1 approach.

In Figure 9, arrows emanate from electronic structure calculations, converging on the ECIs, and then lead to Monte Carlo simulation and/or the CVM, two forms of statistical thermodynamic calculations. This indicates the central role played by the ECIs in "first-principles thermodynamics." Three main paths lead to the ECIs: (a) DCA, based on the explicit application of Equations 22 and 23; (b) the SIM, based on solution of the linear system 24; and (c) methods based on the CPA (coherent potential approximation). These various techniques are described in some detail elsewhere (de Fontaine, 1994), where

original literature is also cited. Points that should be noted here are the following: (a) the DCA, requiring averaging over many configurations, has been implemented thus far only in the TB framework with TB parameters derived from the LMTO-ASA; (b) the SIM energies $E(\sigma_s)$ required in Equations 24 and 25 can be calculated in principle by any LDA method one chooses—FLAPW (Full-potential Linear APW), pseudopotentials, KKR, FPLMTO (Full-potential Linear MTO), or LMTO-ASA; and (c) the CPA, which creates an electronically self-consistent average medium possessing translational symmetry, must be "perturbed" in order to introduce short-range order through the general perturbation method (GPM) or embedded-cluster method (ECM) or $S^{(2)}$ method. The latter path (c), shown by dashed lines (meaning that it is not one trodden presently by most practitioners), must be implemented in electronic structure methods based on Green's function techniques, such as the KKR or TB methods. Another class of calculations is the very popular one of MD (also indicated in Fig. 9 by dashed arrow lines).

To summarize the procedure for doing first-principles thermodynamics, consider some paths in Figure 9. Start with an LDA electronic structure method to obtain cohesive or formation energies of a certain number of ordered structures on a given lattice; then set up a linear system as given by Equation 24 and solve it to obtain the relevant ECIs V_r. Or perform a DCA calculation, as required by Equation 22, by whatever electronic structure method appears feasible. Make sure that convergence is attained in the ECI computations; then insert these interaction parameters in an appropriate CVM free energy, minimize with respect to the correlation variables, and thereby obtain the best estimate of the free energy of the phase in question, ordered or disordered. A more complete description of the general procedure is shown in Figure 11 (see Ground-State Analysis section).

The DCA and SIM methods may appear to be quite different, but, in fact, in both cases, a certain number of configurations are being chosen: a set of random structures in the DCA, a set of small-unit-cell ordered structures in the SIM. The problem is to make sure that convergence has been attained by whatever method is used. It is clear that if all possible ordered structures within a given supercell are included, or if all possible "disordered" structures within a given cell are taken, and if the supercells are large enough, then the two methods will eventually sample the same set of structures. Which method is better in a particular case is mainly a question of convenience: the DCA generally uses small cells, the DCA large ones. The problem with the former is that certain low-symmetry configurations will be missed, and hence the description will be incomplete, which is particularly troublesome when displacements are to be taken into account. The trouble with the DCA is that it is very time consuming (to the point of being impractical) to perform first-principles calculations on large-unit-cell structures. There is no optimal panacea.

Ground-State Analysis

In the previous section, the computation of ECIs was discussed as if the only contribution to the energy was the strictly replacive one, as defined in Equations 13 and 15; yet it is clear from the latter equation that ECIs could in principle be calculated for the value of the combined function Ψ itself. For now, consider the ECIs limited to the replacive (or configurational, or Ising) energy, leaving for the next section a brief description of other contributions to the energy and free energy. In the present context, it is possible to break up the energy into different contributions. First of all it is convenient to subtract off the linear term in the total energy, according to the definition of Equation 4, applied here to E rather than to F, since only zero-temperature effects—i.e., ground-state energies—will be considered here. In Equation 4, the usual classical and MF notation was used; in the "cluster" world of the physics community it is customary to use the notation "formation energy" rather than "mixing energy," but the two are exactly the same. The symbol Δ will be used to emphasize the "difference" nature of the formation energy; it is by definition the total (replacive) energy minus the concentration-weighted average of the energies of the pure elements. This gives the following equation for binary systems:

$$F_{\text{mix}}(T=0) \equiv \Delta E_{\text{form}} = E - E_{\text{lin}} \equiv E - (x_{\text{A}}E_{\text{A}} + x_{\text{B}}E_{\text{B}}) \tag{30}$$

If the "linear" term in Equation 30 is taken as the reference state, then the formation energy can be plotted as a function of concentration for the hypothetical alloy A-B, as shown schematically in Figure 10 (de Fontaine, 1994). Three ordered structures (or compounds), of stoichiometry A_3B, AB, and AB_3, and only those, are assumed to be stable at very low temperatures, in addition to the pure elements A and B. The line linking those five structures forms the so-called convex hull (heavy polygonal line) for the system in question. The heavy dashed curve represents schematically the formation energy of the completely disordered state, i.e., the mixing energy of the hypothetical random state, the one that would have been obtained by a perfect

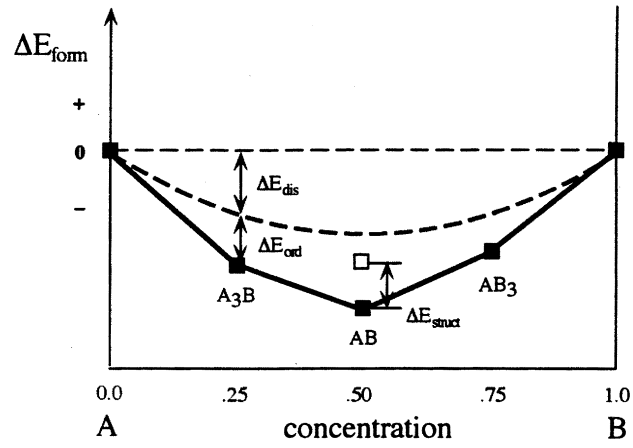


Figure 10. Schematic convex hull (full line) with equilibrium ground states (filled squares), energy of metastable state (open square), and formation energy of random state (dashed line); from de Fontaine (1994).

quench from a very high temperature state, assuming that no melting had taken place. The energy distance between the disordered energy of mixing and the convex hull is by definition the (zero-temperature) ordering energy. If other competing structures are identified, say the one whose formation energy is indicated by a square symbol in the figure, the energy difference between it and the one at the same stoichiometry (black square) is by definition the structural energy.

All the zero-temperature energies illustrated in Figure 10 can in principle be calculated by cluster expansion (CE) techniques, or, for the stoichiometric structures, by first-principles methods. The formation energy of the random state can be obtained by the cluster expansion

$$\Delta E_{\text{dis}} = \sum_{\alpha} V_{\alpha} \xi_1^{n_\alpha} \tag{31}$$

obtained from Equation 27 by replacing the correlation variable for cluster α by n_α times the point correlation ξ_1. Dissecting the energy into separate and physically meaningful contributions is useful when carrying out complex calculations.

Since the disordered-state energy in Figure 10 is a continuous curve, it necessarily follows that A and B must have the same structure and that the compounds shown can be considered as ordered superstructures, or "decorations," of the same parent structure, that of the elemental solids. But which decorations should one choose to construct the convex hull, and can one be sure that all the lowest-energy structures have been identified? Such questions are very difficult to answer in all generality; what is required is to "solve the ground-state problem". That expression can have various meanings: (1) among given structures of various stoichiometries, find the ones that lie on the true convex hull; (2) given a set of ECIs, predict the ordered structures of the parent lattice (or structure, such as hcp) that will lie on the convex hull; and (3) given a maximum range of interactions admissible, find all possible ordered ground states of the parent lattice and their domain of existence in ECI space. Problem (3) is by far the most difficult. It can be attacked by linear programming methods: the requirement that cluster concentration lie between 0 and 1 provides a set of linear constraints (via the configuration matrix, mentioned earlier) in the correlation variables, which are represented in ξ space by sets of hyperplanes. The convex polyhedral region that satisfies these constraints, i.e., the configuration polyhedron, has vertices that, in principle, have the correct ξ coordinates for the ground states sought. This method can actually predict totally new ground states, which one might never have guessed at. The problem is that sometimes the vertex determination produces ξ coordinates that do not correspond to any constructible structures—i.e., such coordinates are internally inconsistent with the requirement that the prediction will actually correspond to an actual decoration of the parent lattice. Another limitation of the method is that, in order to produce a nontrivial set of ground states, large clusters must often be used, and then the dimension of the ξ space gets too large to handle, even with the help of group theory–based computer codes (Ceder et al., 1994). Readers interested in this problem can do no better than to consult Ducastelle (1991), which also contains much useful information on the CVM, on electronic structure calculation (mostly TB based), and on the generalized perturbation method for calculating EPIs from electronic structure calculations. Another very readable coverage of ground-state problems is that given by Inden and Pitsch (1991), who also cover the derivation of multicomponent CVM equations. Accounts of the convex polyhedron method are also given in de Fontaine (1979, 1994).

Problem (2) is more tractable, since in this case the ECIs are assumed to be known, thereby determining the "objective function" of the linear programming problem. It is then possible to discover the corresponding vertex of the configurational polyhedron by use of the simplex algorithm. Problem (1) does not necessitate the knowledge or even use of the ECIs: the potentially competitive ordered structures are guessed at (in the words of Zunger, this amounts to "rounding up the usual suspects"), and direct calculations determine which of those structures will lie on the convex hull. In that case, however, the real convex hull structures may be missed altogether. Finally, it is also possible to set up a relatively large supercell and to populate it successively by all possible decoration of A and B atoms, then to calculate the energies of resulting structures by cluster expansion (Lu et al., 1991). There too, however, some optimal ordered structures may be missed, those whose unit cells will not fit exactly within the designated supercell.

Figure 11 summarizes the general method of calculating a phase diagram by cluster methods. The solution of the ground-state problem has presumably provided the correct lowest-energy ordered structures for a given parent lattice, which in turn determines the sublattices to consider. For each ordered structure, write down the CVM free energy, minimize it with respect to the configuration variables, then plot the resulting free energy curve. Now repeat these operations for another lattice, for which the required ordered states have been determined. Of course, when comparing structures on one parent lattice with another, it is necessary to know the structural energies involved, for example the difference between pure A and pure B in the fcc and bcc structures. Finally, as in the "classical" case, construct the lowest common tangents to all free energies present. Figure 11 shows schematically some ordered and disordered free energy curves at a given temperature, then at the lower panel the locus of common tangency points generating the required phase diagram, as illustrated for example in the MF approximation in Figure 5. The whole procedure is of course much more complicated than the classical or MF one, but consider what in principle is accomplished: with only the knowledge of the atomic numbers of the constituents, one will have calculated the (solid-state portion of the) phase diagram and also such useful thermodynamic quantities as the long-range, short-range order, equilibrium lattice parameters, formation and structural energies, bulk modulus (B), and metastable phase boundaries as a function of temperature and concentration. Note that calculations at a series of volumes must be performed in order to obtain quantities

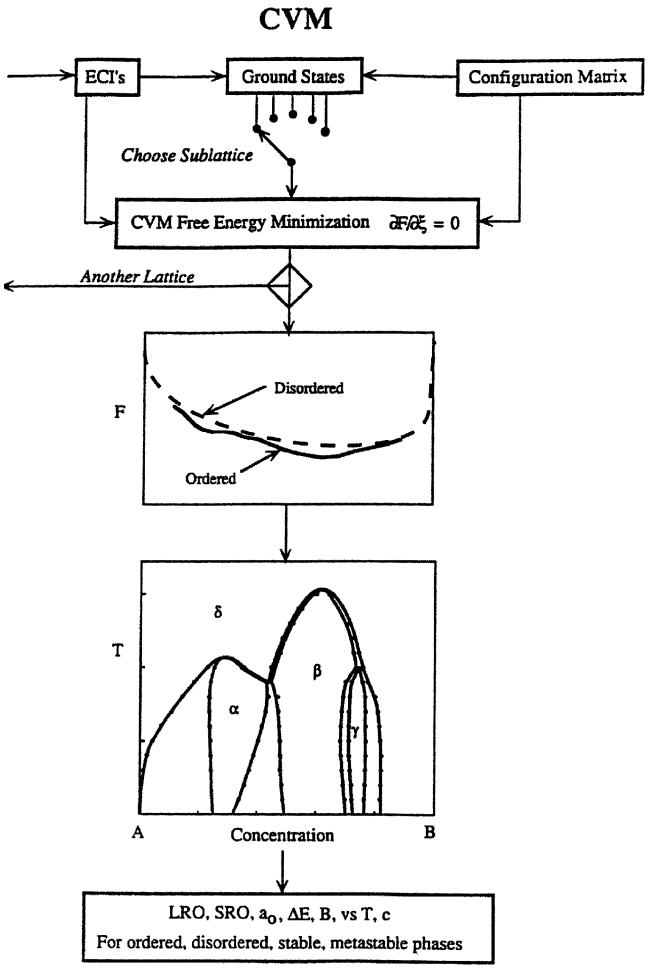


Figure 11. Flow chart for obtaining phase diagram from repeated calculations on different lattices and pertinent sublattices.

such as the equilibrium volume and the bulk modulus B. A summary of those binary and ternary systems for which the program had been at least partially completed up to 1994 was given in table form in de Fontaine (1994).

Up to now, only the strictly replacive (Ising) energy has been considered in calculating the ECIs. For simplicity, the other contributions, displacive and dynamic, that have been mentioned in previous sections have so far been left out of the picture, as indeed they often were in early calculations. It is now becoming increasingly clear that those contributions can be very large and, in case of large atomic misfit, perhaps dominant. The *ab initio* treatment of displacive interactions is a difficult topic, still being explored, and so will be summarized only briefly.

Static Displacive Interactions

In a review article, Zunger (1994) explained his views of displacive effects and stated that the only investigators who had correctly treated the whole set of phenomena, at least the static ones, were members of his group. At the time, this claim was certainly valid, indicating that far too little attention had been paid to this topic. Today, other groups have followed his lead, but the situation is still far from settled, as displacements, static and dynamic, are difficult to treat. It is useful (adopting Zunger's definitions) to examine in more detail the term E_{displ} which appears in Equations 13 and 15. From Equations 13, 15, and 30,

$$E = E_{\text{lin}} + E_{\text{stat}} = E_{\text{lin}} + \Delta E_{\text{VD}} + \Delta E_{\text{repl}} + \Delta E_{\text{def}} + \Delta E_{\text{relax}} \tag{32}$$

where E_{displ} of Equation 13 has been broken up into a volume deformation (VD) contribution, a "cell-external" deformation (ΔE_{def}), and a "cell-internal" deformation (relax). The various terms in Equation 32 are explained schematically on a simple square lattice in Figure 12 (Morgan et al., unpublished). From pure A and pure B on the same lattice but at their equilibrium volumes (the reference state E_{lin}), expand the smaller lattice parameter and contract that of the larger to a common intermediate parameter, assumed to be that of the homogeneous mixture, thus yielding the large contribution ΔE_{VD}. Now allow the atoms to rearrange at constant volume while remaining situated on the sites of the average lattice, thereby recovering energy ΔE_{repl}, discussed in the Basic Principles: Cluster Approach section under the assumption that only the "Ising energy" was contributing. Generally,

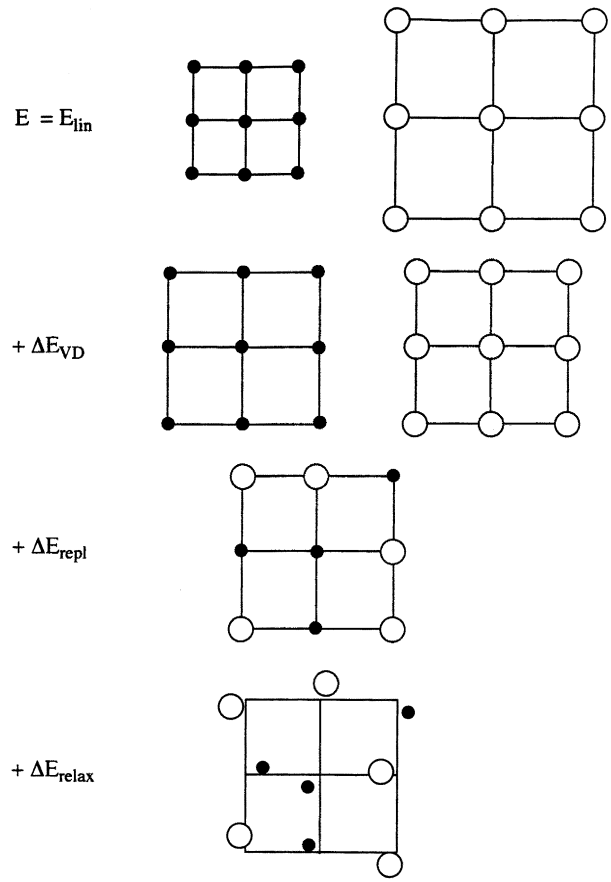


Figure 12. Schematic two-dimensional illustration of various energy contributions appearing in Equation 32; according to Morgan et al. (unpublished).

this atomic redistribution will lower the symmetry of the original states, so that the lattice parameters will extend or contract (such as c/a relaxation) and the cell angles vary at constant volume, a contribution denoted ΔE_{def} in Equation 32 but not indicated in Figure 12. Finally, the atoms may be displaced from their positions on the ideal lattice, producing a contribution denoted as ΔE_{relax}.

Equation 32 also indicates the order in which the calculations should be performed, given certain reasonable assumptions (Zunger, 1994; Morgan et al., unpublished). Actually, the cycle should be repeated until no significant atomic displacements take place. Such an iterative procedure was carried out in a strictly *ab initio* electronic structure calculation (FLMTO) by Amador et al. (1995) on the structures $L1_2$, $D0_{22}$, and $D0_{23}$ in Al_3Ti and Al_3Zr. In that case, the relaxation could be carried out without calculating forces on atoms simply by minimizing the energy, because the numbers of degrees of freedom available to atomic displacements were small. For more complicated unit cells of lower symmetry, minimizing the total energy for all possible types of displacements would be too time-consuming, so that actual atomic forces should be computed. For that, plane-wave electronic structure techniques are preferable, such as those given by FLAPW or pseudopotential codes.

Different techniques exist for calculating the various terms of Equation 32. The linear contribution E_{lin} requires only the total energies of the pure states and can generally be obtained in a straightforward way by electronic structure calculations. The volume deformation contribution is best calculated in a "classical" way, for instance by a method that Zunger has called the "ε-G" technique (Ferreira et al., 1988). If one makes the assumption that the equilibrium volume term depends essentially on the average concentration x, and not on the particular configuration $\boldsymbol{\sigma}$, then ΔE_{VD} can be generally approximated by the term $\Omega x(1-x)$, with Ω being a slowly varying function of x. The parameters of function Ω may also be obtained from electronic structure calculations (Morgan et al., unpublished). The "replacive" energy, which is the familiar "Ising" energy, can be calculated by cluster expansion techniques, as explained above, either in DCA or SIM mode. The cell-external deformation can also be obtained by electronic structure methods in which the lowest energy is sought by varying the cell parameters. Finally, various methods have been proposed to calculate the relaxation energy. One is to include the displacive and volume effects formally in the cluster expansion of the replacive term; this is done by using, as known energies $E(\sigma_s)$ in Equation 24 for the SIM, the energies of fully relaxed structures obtained from electronic structure calculations that can provide forces on atoms. This procedure can be very time consuming, as cluster expansions on relaxed structures generally converge more slowly than expansions dealing with strictly displacive effects. Also, if fairly large-unit-cell structures of low symmetry are not included in the SIM fit, the displacive effects may not get well represented. One may also define state functions (at zero kelvin) that are algebraic functions of atomic volume, c/a ratio (for hexagonal, tetragonal structures, ...), and so on, which, in a cluster expansion, will result in ECIs being functions of atomic volume, c/a ratio, etc. The resulting equilibrium energies may then be obtained by minimizing with respect to these displacive variables (Amador et al., 1995; Craievich et al., 1997). Not only energies but also free energies may thus be optimized, by minimizing the CVM free energy not only with respect to the ξ configuration variables but also with respect to volume, for instance, at various temperatures (Sluiter et al., 1989, 1990). This temperature dependence does not, however, take vibrational contributions into account (see Nonconfigurational Thermal Effects section).

Let us mention here the importance of calculating energies (or enthalpies) as a function of c/a, b/a ratios, ... at constant cell volume. It is possible, by a so-called Bain transformation, to go continuously from the fcc to the bcc structures, for example. Doing so is important not only for calculating structural energies, essential for complete phase diagram calculations, but also for ascertaining whether the higher-energy structures are metastable with respect to the ground state, or actually unstable. In the latter case, it is not permissible to use the standard CALPHAD extrapolation of phase boundaries (through extrapolation of empirical vibrational entropy functions) to obtain energy estimates of certain nonstable structures (Craievich et al., 1994). Other more general cell-external deformations have also been treated (Sob et al., 1997).

Elastic interactions, resulting from local relaxations, are characteristically long range. Hence, it is generally not feasible to use the type of cluster expansion described above (see Cluster Approach section), since very large clusters would be required in the expansion and the number of correlation variables required would be too large to handle. One successful approach is that of Zunger and collaborators (Zunger, 1994), who developed a k-space SIM with effective pair interactions calculated in Fourier space. One novelty of this approach is that the Fourier transformations feature a smoothing procedure, essential for obtaining reasonably rapid convergence; another is that of subtracting off the troublesome $k=0$ term by means of a concentration-dependent function. Previous Fourier space treatments were based on second-order expansions of the relaxation energy in terms of atomic displacements and so-called Kanzaki forces (de Fontaine, 1979; Khachaturyan, 1983)—the latter calculated by heuristic means, from the EAM (Asta and Foiles, 1996), or from an nn "spring model" (Morgan et al., unpublished).

Another option for taking static displacive effects into account is to use brute-force computer simulation techniques, such as MD or molecular statics. The difficulty here is that first-principles electronic structure methods are generally too slow to handle the millions of MD steps generally required to reach equilibrium over a sufficiently large region of space. Hence, empirical potentials, such as those provided by the EAM (Daw et al., 1993), are required. The vast literature that has grown up around these simulation techniques is too extensive to review here; furthermore, these methods are not particularly well suited to phase-diagram calculations, mainly because the average frequency for "hopping" (atomic interchange, or "replacement") is very much slower than that for atomic vibrations. Nonetheless, simulation techniques do have an

important role to play in dynamical effects, which are discussed below.

Nonconfigurational Thermal Effects

Mostly zero kelvin contributions described in the previous sections dealt only with configurational free energy. As was explained, the latter can be dealt with by the cluster variation method, which basically features a cluster expansion of the configurational entropy, Equation 29. But even there, there are difficulties: How does one handle long-range interactions, for example those coming from relaxation interactions? It has indeed been found that, in principle, all interactions needed in the energy expansion (Equation 27) must appear as clusters or subclusters of a maximal cluster appearing in the CVM configurational entropy. But, as mentioned before, the number of variables associated with large clusters increases exponentially with the number of points in the clusters: one faces a veritable "combinatorial explosion." One approximate and none-too-reliable solution is to treat the long-pair interactions by the GBW (superposition, MF) model.

A more fundamental question concerns the treatment of vibrational entropy, thermal expansion, and other excitations, such as electronic and magnetic ones. First, consider how the static and dynamic displacements can be entered in the CVM framework. Several methods have been proposed; in one of these (Kikuchi and Masuda-Jindo, 1997), the CVM free energy variables include continuous displacements, and summations are replaced by integrals. For computational reasons, though, the space is then discretized, leading to separate cluster variables at each point of the fine-mesh lattice thus introduced. Practically, this is equivalent to treating a multicomponent system with as many "components" as there are discrete points considered about each atom. Only model systems have been treated thus far, mainly containing one (real) component. In another approach, atomic positions are considered to be Gaussian distributed about their static equilibrium locations. Such is the "Gaussian CVM" of Finel (1994), which is simpler but somewhat less general than that of Kikuchi and Masuda-Jindo (1997). Again, mostly dynamic displacements have been considered thus far. One can also perform cluster expansions of the Debye temperature in a Debye-Grüneisen framework (Asta et al., 1993a) or of the average value of the logarithm of the vibrational frequencies (Garbulsky and Ceder, 1994) of selected ordered structures.

As for treating vibrational free energy along with the configurational one, it was originally considered sufficient to include a Debye-Grüneisen correction for each of the phases present (Sanchez et al., 1991). It has recently been suggested (Anthony et al., 1993; Fultz et al., 1995; Nagel et al., 1995), based on experimental evidence, that the vibrational entropy difference between ordered and disordered states could be quite large, in some cases comparable to the configurational one. This was a surprising result indeed, which perhaps could be explained away by the presence of numerous extended defects introduced by the method of preparation of the disordered phases, since in defective regions vibrational entropy would tend to be higher. Several quasiharmonic calculations were recently undertaken to test the experimental values obtained in the Ni_3Al system; one based on the Finnis-Sinclair (FS) potential (Ackland, 1994) and one based on EAM potentials (Althoff, 1977a,b). These studies gave results agreeing quite well with experimental values (Anthony et al., 1993; Fultz et al., 1995; Nagel et al., 1995) and with each other, namely, $\Delta S_{vib} = \sim 0.2k_B$ versus $0.2k_B$ to $0.3k_B$ reported experimentally. Integrated thermodynamic quantities obtained from EAM-based calculations, such as the isobaric specific heat (C_p), agree very well with available evidence for the $L1_2$ ordered phase (Kovalev et al., 1976). Figure 13 compares experimental (open circles) and calculated (curve) C_p values as a function of absolute temperature. The agreement is excellent at low temperature ranges, but the experimental points begin to deviate from the theoretical curve at ~900 K, when configurational entropy—which is not included in the present calculation—becomes important. At still higher temperatures, the quasiharmonic approximation may also introduce errors. Note that in the EAM and FS approaches, the disordered-state structures were fully relaxed (via molecular statics) and the temperature dependence of the atomic volume (by minimization of the quasiharmonic vibrational free energy) of the disordered state was fully taken into account. What is to be made of this? If the empirical potential results are valid, and so the experimental ones are as well, that means that vibrational entropy, usually neglected in the past, can play an important role in some phase diagram calculations and should be included. Still, more work is required to confirm both experimental and theoretical results.

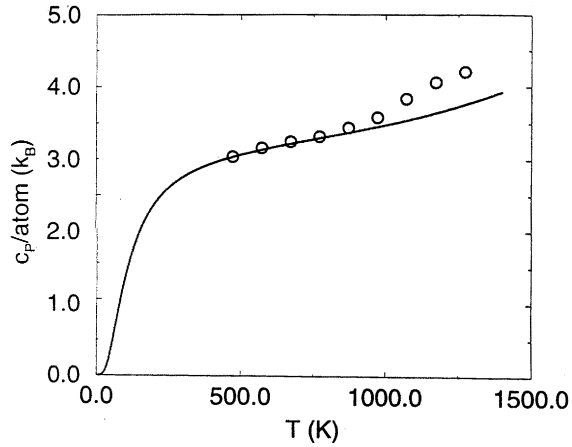


Figure 13. Isobaric specific heat as a function of temperature for ordered Ni_3Al as calculated in the quasiharmonic approximation (curve, Althoff et al., 1997a,b) and as experimentally determined (open circles, Kovalev et al., 1976).

Other contributing influences to be considered could be those of electronic and magnetic degrees of freedom. For the former, the reader should consult the paper by Wolverton and Zunger (1995). For the latter, an example of a "fitted" phase-diagram calculation with magnetic interactions will be presented in the Examples of Applications section.

EXAMPLES OF APPLICATIONS

The section on Cluster Expansion Free Energy showed a comparison between prototype fcc ordering in MF and CVM approaches. This section presents applications of various types of calculations to real systems in both fitted and first-principles schemes.

Aluminum-Nickel: Fitted, Mean-Field, and Hybrid Approaches

A very complete study of the Al-Ni phase diagram has been undertaken recently by Ansara and co-workers (1997). The authors used a CALPHAD method, in the so-called sublattice model, with parameters derived from an optimization procedure using all available experimental data, mostly from classical calorimetry but also from electromotive force (emf) measurements and mass spectrometry. Experimental phase boundary data and the assessed phase diagram are shown in Figure 14A, while the "modeled" diagram is shown in Figure 14B. The agreement is quite impressive, but of course it is meant to be so. There have also been some cluster/first-principles calculations for this system, in particular a full determination of the equilibrium phase diagram, liquid phase included, by Pasturel et al. (1992). In this work, the liquid free energy curve was obtained by CVM methods as if the liquid were an fcc crystal. This approach may sound incongruous but appears to work. In a more recent calculation for the Al-Nb alloy by these authors and co-workers (Colinet et al., 1997), the liquid-phase free energy was calculated by three different models—the MF subregular solution model, the pair CVM approximation (quasichemical), and the CVM tetrahedron—with very little difference in the results. Melting temperatures of the elements and entropies of fusion had to be taken from experimental data, as was true for the Al-Ni system. It would be unfair to compare the MF and "cluster" phase diagrams of Al-Ni because the latter was published before all the data used to construct the former were known and because the "first-principles" calculations were still more or less in their infancy at the time (1992).

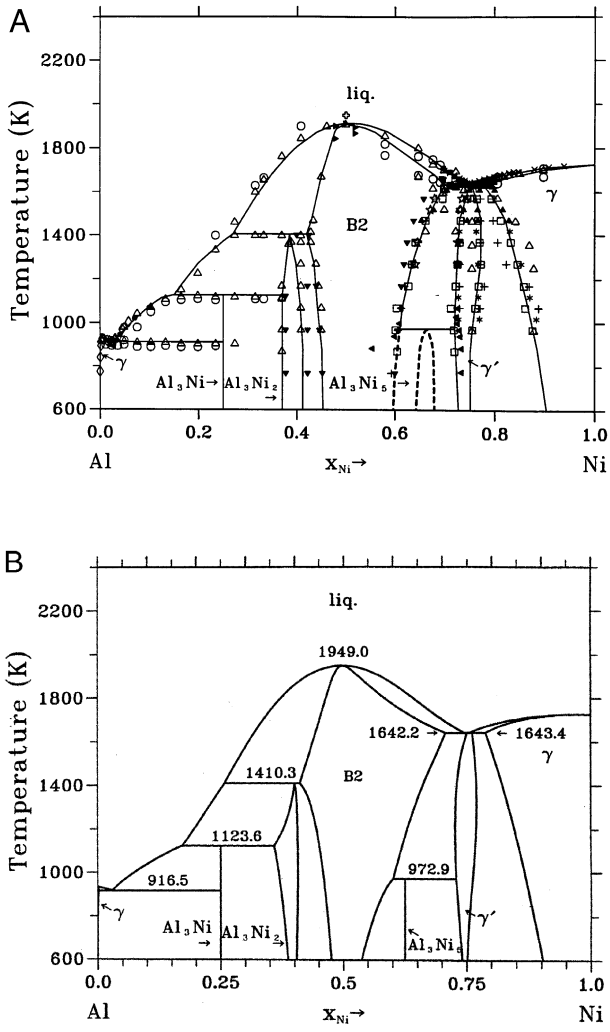


Figure 14. (**A**) Al-Ni assessed phase diagram with experimental data (various symbols; see Ansara et al., 1997, for references). (**B**) Al-Ni phase diagram modeled according to the CALPHAD method (Ansara et al., 1997).

Certainly, the MF method is by far the simpler one to use, but cannot make actual predictions from basic physical principles. Also, the physical parameters derived from an MF fit sometimes lead to unsatisfactory results. In the present Al-Ni case, the isobaric specific heat is not given accurately over a wide range of temperatures, whereas the brute-force quasiharmonic EAM calculation of the previous section gave excellent agreement with experimental data, as seen in Figure 13. Perhaps it would be more correct to refer to the CVM/*ab initio* calculations of Pasturel, Colinet, and co-workers as a hybrid method, combining electronic structure calculations, cluster expansions, and some CALPHAD-like fitting: a method that can provide theoretically valid and practically useful results.

Nickel-Platinum: Fitted, Cluster Approach

If one considers another Ni-based alloy, the Ni-Pt system, the phase diagram is quite different from those discussed in the previous section. A glance at the version of the experimental phase diagram (Fig. 15A) published in the American Society for Metals (ASM) handbook *Binary Alloy Phase Diagrams* (2nd ed.; Massalski, 1990) illustrates why it is often necessary to supplement experimental data with calculations: the Ni_3Pt and NiPt phase boundaries and the ferromagnetic transition are shown as disembodied lines, and the order/disorder Ni_3Pt (single dashed line) would suggest a second-order phase transition, which it surely is not. Consider now the cluster/fitted version—liquid phase not included: melting occurs above 1000 K in this system (Dahmani et al., 1985) of the phase diagram (Fig. 15B). The CVM tetrahedron approximation was used for the "chemical" interactions and also for magnetic interactions assumed to be nonzero only if at least two of the tetrahedron sites were occupied by Ni atoms. The chemical and magnetic interactions were obtained by fitting to recent (at the time, 1985) experimental data (Dahmani et al., 1985). It is seen that the CVM version (Fig. 15B) displays correctly three first-order transitions, and the disordering temperatures of Ni_3Pt and NiPt differ from those

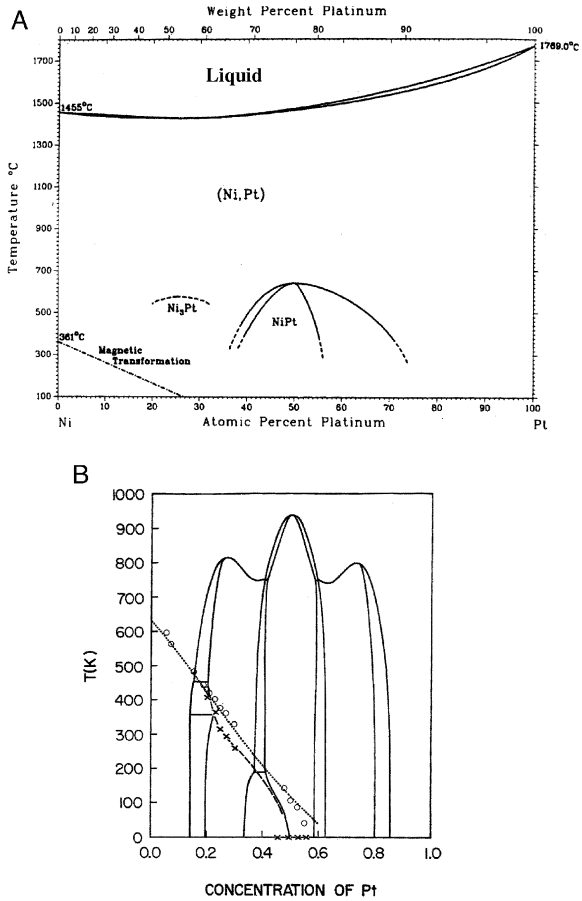


Figure 15. (**A**) Ni-Pt phase diagram according to the ASM handbook (Massalski, 1990). (**B**) Ni-Pt phase diagram calculated from a CVM free energy fitted to available experimental data (Dahmani et al., 1985).

indicated in the handbook version (Fig. 15A). It appears that the latter version, published several years after the work of Dahmani et al. (1985), is based on an older set of data; in fact the 1990 edition of the ASM handbook (Massalski, 1990) actually states that "no new data has been reported since the Hansen and Anderko phase-diagram handbook appeared in 1958!" Additionally, experimental values of the formation energies of the $L1_0$ NiPt phase were reported in 1970 by Walker and Darby (1970). The 1990 ASM handbook also fails to indicate the important perturbation of the "chemical" phase boundaries by the ferromagnetic transition, which the calculation clearly shows (Fig. 15A). There is still no unambiguous verification of the transition on the Pt-rich side of the phase diagram (platinum is expensive). Note also that a MF GBW model could not have produced the three first-order transitions expected; instead it would have given a triple second-order transition at 50% Pt, as in the MF prototype phase diagram of Figure 8A.

The solution to the three-transition problem in Ni-Pt might be to obtain the free energy parameters from electronic structure calculations, but surprises are in store there, as explained in some detail in the author's review (de Fontaine, 1994, Sect. 17). The KKR-CPA-$S^{(2)}$ calculations (Pinski et al., 1991) indicated that a 50:50 Ni-Pt solid solution would be unstable to a $\langle 100 \rangle$ concentration ordering wave, thereby presumably giving rise to an $L1_0$ ordered phase region, in agreement with experiment. However, that is true only for calculations based on what has been called the unrelaxed ΔE_{repl} energy of Equation 32. Actually, according to the Hamiltonian used in those $S^{(2)}$ calculations, the ordered $L1_0$ phase should be unstable to phase separation into Ni-rich and Pt-rich phases if relaxation were correctly taken into account, emphasizing once again that all the contributions to the energy in Equation 32 should be included. However, the $L1_0$ state is indeed the observed low-temperature state, so it might appear that the incomplete energy description was the correct one. The resolution of the paradox was provided by Lu, Wei, and Zunger (1993): The original $S^{(2)}$ Hamiltonian did not contain relativistic corrections, which are required since Pt is a heavy element. When this correction is included in a fully relaxed FLAPW calculation, the $L1_0$ phase indeed appears as the ground state in the central portion of the Ni-Pt phase diagram. Thus first-principles calculations can be helpful, but they must be performed with pertinent effects and corrections (including relativity!) taken into account—hence, the difficulty of the undertaking but also the considerable reward of deep understanding that may result.

Aluminum-Lithium: Fitted and Predicted

Aluminum-lithium alloys are strengthened by the precipitation of a metastable coherent Al_3Li phase of $L1_2$ structure (α'). Since this phase is less stable than the two-phase mixture of an fcc Al-rich solid solution (α) and a B32 bcc-based ordered structure AlLi (β), it does not appear on the experimentally determined equilibrium phase diagram (Massalski, 1986, 1990). It would therefore be useful to calculate the α/α' metastable phase boundaries and place them on the experimental phase diagram. This is what Sigli and Sanchez (1986) did using CVM free energies to fit the entire experimental phase diagram, as assessed by McAlister and reproduced here in Figure 16A from the first edition of the ASM handbook *Binary Alloy Phase Diagrams* (Massalski, 1986). Note that the newer version of the Al-Li phase diagram in the 2nd ed. of the handbook (Massalski, 1990) is incorrect (in all fairness, a correction appeared in an addendum, reproducing the older version). The CVM-fitted Al-Li phase diagram, including the metastable $L1_2$ phase region (dashed lines) and available experimental points, is shown in Figure 16B (Sigli and Sanchez, 1986). The Li-rich side of the α/α' two-phase region is shown to lie very close to the Al_3Li stoichiometry, along (as expected) with the highest disordering point, at ~825 K. A low-temperature metastable-metastable miscibility gap (because it is a metastable feature of a metastable equilibrium) is predicted to lie within the α/α' two-phase region. Note once again that the correct ordering behavior of the $L1_2$ structure by a first-order transition near $x_{\text{Li}} = 0.25$ could not have been obtained by a GBW MF model. How, then, could later authors (Khachaturyan et al., 1988) claim to have obtained a better fit to experimental points by using the method of concentration

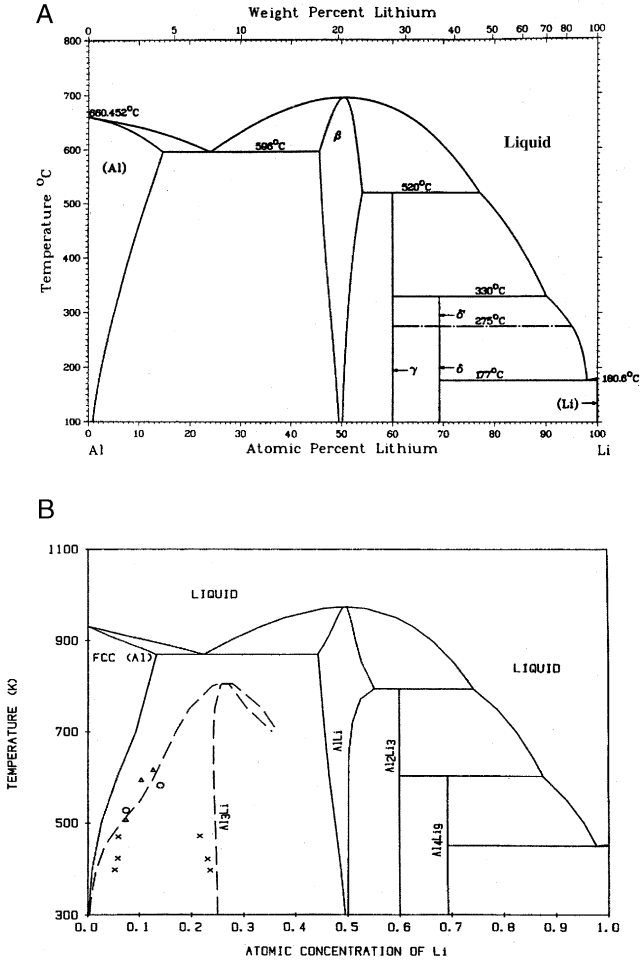


Figure 16. (**A**) Al-Li phase diagram from the first edition of the phase-diagram handbook (Massalski, 1986). (**B**) Al-Li phase diagram calculated from a CVM fit (full lines), $L1_2$ metastable phase regions (dashed lines), and corresponding experimental data (symbols; see Sigli and Sanchez, 1986).

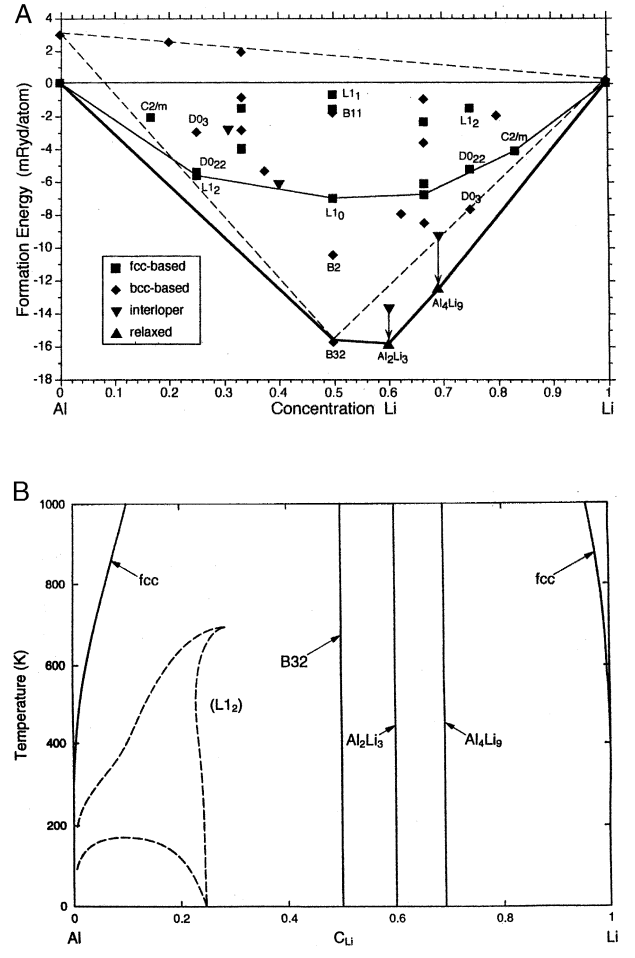


Figure 17. (**A**) Convex hull (heavy solid line) and formation energies (symbols) used in first-principles calculation of Al-Li phase diagram (from Sluiter et al., 1996). (**B**) Solid-state portion of the Al-Li phase diagram (full lines) with metastable $L1_2$ phase boundaries and metastable-metastable miscibility gap (dashed lines; Sluiter et al., 1996).

waves, which is completely equivalent to the GBW MF model? The answer is that these authors conveniently left out the high-temperature portion of their calculated phase boundaries so as not to show the inevitable outcome of their calculation, which would have incorrectly produced a triple second-order transition at concentration 0.5, as in Figure 8A.

A first-principles calculation of this system was attempted by Sluiter et al. (1989, 1990), with both fcc- and bcc-based equilibria present. The tetrahedron approximation of the CVM was used and the ECIs were obtained by structure inversion from FLAPW electronic structure calculations. The liquid free energy was taken to be that of a subregular solution model with parameters fitted to experimental data. Later (1996), the same Al-Li system was revisited by the same first author and other co-workers (Sluiter et al., 1996). This time the LMTO-ASA method was used to calculate the structures required for the SIM, and far more structures were used than previously. The calculated formation energies of structures used in the inversion are indicated in Figure 17A, an application to a real system of techniques illustrated schematically in Figure 10. All fcc-based structures are shown as filled squares, bcc based as diamonds, "interloper" structures (those that are not manifestly derived from fcc or bcc), unrelaxed and relaxed, as triangles pointing down and up, respectively. The solid thin line is the fcc-only, and the dashed thin line the bcc-only convex hull. The heavy solid line represents the overall (fcc, bcc, interloper) convex hull, its vertices being the predicted ground states for the Al-Li system. It is gratifying to see that the predicted ("rounding up the usual suspects") ground states are indeed the observed ones. The solid-state portion of the phase diagram, with free energies provided by the tetrahedron-octahedron CVM approximation (see Fig. 7), is shown in Figure 17B. The metastable α/α' ($L1_2$) phase boundaries, this time calculated completely from first principles, agree surprisingly well with those obtained by a tetrahedron CVM fit, as shown above in Figure 16B. Even the metastable MG is reproduced. Setting aside the absence of

the liquid phase, the first-principles phase diagram (no empirical parameters or fits employed) differs from the experimentally observed one (Fig. 16A) by the fact that the central B32 ordered phase has too narrow a domain of solubility. However, the calculations predict such things as the lattice parameters, bulk moduli, degree of long and short-range order as a function of temperature and concentration, and even the influence of pressure on the phase equilibria (Sluiter et al., 1996).

Other Applications

The case studies just presented are only a few among many, but should give the reader an idea of some of the difficulties encountered while attempting to gather, represent, fit, or calculate phase diagrams. Examples of other systems, up to 1994, are listed, for instance, in the author's review article (de Fontaine, 1994, Table V, Sect. 17). These have included noble metal alloys (Terakura et al., 1987; Mohri et al. 1988, 1991); Al alloys (Asta et al., 1993b,c, 1995; Asta and Johnson, 1997; Johnson and Asta, 1997); Cd-Mg (Asta et al., 1993a); Al-Ni (Sluiter et al., 1992; Turchi et al., 1991a); Cu-Zn (Turchi et al., 1991b); various ternary systems, such as Heusler-type alloys (McCormack and de Fontaine, 1996; McCormack et al., 1997); semiconductor materials (see the publications of the Zunger school); and oxides (e.g., Burton and Cohen, 1995; Ceder et al., 1995; Tepesch et al., 1995, 1996; Kohan and Ceder, 1996). In the oxide category, phase-diagram studies of the high-T_c superconductor YBCO (de Fontaine et al., 1987, 1989, 1990; Wille and de Fontaine, 1988; Ceder et al., 1991) are of special interest as the calculated oxygen-ordering phase diagram was derived by the CVM before experimental results were available and was found to agree reasonably well with available data. Material pertinent to the present unit can also be found in the proceedings of the joint NSF/CNRS Conference on Alloy Theory held in Mt. Ste. Odile, France, in October 1996 (de Fontaine and Dreyssé, 1997).

CONCLUSIONS

Knowledge of phase diagrams is absolutely essential in such fields as alloy development and processing. Yet, today, that knowledge is still incomplete: the phase diagram compilations and assessments that are available are fragmentary, often unreliable, and sometimes incorrect. It is therefore useful to have a range of theoretical models whose roles and degrees of sophistication differ widely depending on the aim of the calculation. The highest aim, of course, is that of predicting phase equilibria from the knowledge of the atomic number of the constituents, a true first-principles calculation. That goal is far from having been attained, but much progress is being realized currently. The difficulty is that so many different types of atomic/electronic interactions need to be taken into account, and the calculations are complicated and time consuming. Thus far, success along those lines belongs primarily to such quantities as heats of formation; configurational and vibrational entropies are more difficult to compute, and as a result, predictions of transition temperatures are often off by a considerable margin.

Fitting procedures can produce very accurate results, since with enough adjustable parameters, one can fit just about anything. Still, the exercise can be useful as such procedures may actually find inconsistencies in empirically derived phase diagrams, extrapolate into uncharted regions, or enable one to extract thermodynamic functions, such as free energy, from experiment using calorimetry, x-ray diffraction, electron microscopy, or other techniques. Still, one has to take care that the starting theoretical models make good physical sense. The foregoing discussion has demonstrated that GBW-type models, which produce acceptable results in the case of phase separation, cannot be used for order/disorder transformations, particularly in "frustrated" cases as encountered in the fcc lattice, without adding some rather unphysical correction terms. In these cases, the CVM, or Monte Carlo simulation, must be used. Perhaps a good compromise would be to try a combination of first-principles calculations for heats of formation supplemented by a CVM entropy plus "excess" terms fitted to available experimental data. Purists might cringe, but end users may appreciate the results. For a true understanding of physical phenomena, first-principles calculations, even if presently inaccurate regarding transition temperatures, must be carried out, but in a very complete fashion.

ACKNOWLEDGMENTS

The author thanks Jeffrey Althoff and Dane Morgan for helpful comments concerning an earlier version of the manuscript and also thanks I. Ansara, J. M. Sanchez, and M. H. F. Sluiter, who graciously provided figures. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences, under contract Nos. DE-AC04-94AL85000 (Sandia) and DE-AC03-76SF00098 (Berkeley).

LITERATURE CITED

Ackland, J. G. 1994. *In* Alloy Modeling and Design (G. Stocks and P. Turchi, eds.) p. 194. The Minerals, Metals and Materials Society, Warrendale, Pa.

Althoff, J. D., Morgan, D., de Fontaine, D., Asta, M. D., Foiles, S. M. and Johnson, D. D. 1997a. *Phys. Rev. B* 56:R5705.

Althoff, J. D., Morgan, D., de Fontaine, D., Asta, M. D., Foiles, S. M., and Johnson, D. D. 1997b. *Comput. Mater. Sci.* 9:1.

Amador, C., Hoyt, J. J., Chakoumakos, B. C., and de Fontaine, D. 1995. *Phys. Rev. Lett.* 74:4955.

Ansara, I., Dupin, N., Lukas, H. L., and Sundman, B. 1997. *J. Alloys Compounds* 247:20.

Anthony, L., Okamoto, J. K., and Fultz, B. 1993. *Phys. Rev. Lett.* 70:1128.

Asta, M. D. and Foiles, S. M. 1996. *Phys. Rev. B* 53:2389.

Asta, M. D., de Fontaine, D., and van Schilfgaarde, M. 1993b. *J. Mater. Res.* 8:2554.

Asta, M. D., de Fontaine, D., van Schilfgaarde, M., Sluiter, M., and Methfessel, M. 1993c. *Phys. Rev. B* 46:505.

Asta, M. D. and Johnson, D. D. 1997. *Comput. Mater. Sci.* 8:64.

Asta, M. D., McCormack, R., and de Fontaine, D. 1993a. *Phys. Rev. B* 48:748.

Asta, M. D., Ormeci, A., Wills, J. M., and Albers, R. C. 1995. *Mater. Res. Soc. Symp.* 1:157.

Asta, M. D., Wolverton, C., de Fontaine, D., and Dreyssé, H. 1991. *Phys. Rev. B* 44:4907.

Barker, J. A. 1953. *Proc. R. Soc. London, Ser. A* 216:45.

Burton, P. B. and Cohen, R. E. 1995. *Phys. Rev. B* 52:792.

Cayron, R. 1960. Etude Théorique des Diagrammes d'Equilibre dans les Systémes Quaternaires. Institut de Métallurgie, Louvain, Belgium.

Ceder, G. 1993. *Comput. Mater. Sci.* 1:144.

Ceder, G., Asta, M., and de Fontaine, D. 1991. *Physica C* 177:106.

Ceder, G., Garbulsky, G. D., Avis, D., and Fukuda, K. 1994. *Phys. Rev. B* 49:1.

Ceder, G., Garbulsky, G. D., and Tepesch, P. D. 1995. *Phys. Rev. B* 51:11257.

Colinet, C., Pasturel, A., Manh, D. Nguyen, Pettifor D. G., and Miodownik, P. 1997. *Phys. Rev. B* 56:552.

Connolly, J. W. D. and Williams, A. R. 1983. *Phys. Rev. B* 27:5169.

Craievich, P. J., Sanchez, J. M., Watson, R. E., and Weinert, M. 1997. *Phys. Rev. B* 55:787.

Craievitch, P. J., Weinert, M., Sanchez, J. M., and Watson, R. E. 1994. *Phys. Rev. Lett.* 72:3076.

Dahmani, C. E., Cadeville, M. C., Sanchez, J. M., and Moran-Lopez, J. L. 1985. *Phys. Rev. Lett.* 55:1208.

Daw, M. S., Foiles, S. M., and Baskes, M. I. 1993. *Mater. Sci. Rep.* 9:251.

de Fontaine, D. 1979. *Solid State Phys.* 34:73–274.

de Fontaine, D. 1994. *Solid State Phys.* 47:33–176.

de Fontaine, D. 1996. *MRS Bull.* 22:16.

de Fontaine, D., Ceder, G., and Asta, M. 1990. *Nature (London)* 343:544.

de Fontaine, D. and Dreyssé, H. (eds.) 1997. Proceedings of Joint NSF/CNRS Workshop on Alloy Theory. *Comput. Mater. Sci.* 8:1–218.

de Fontaine, D., Finel, A., Dreyssé, H., Asta, M. D., McCormack, R., and Wolverton, C. 1994. *In* Metallic Alloys: Experimental and Theoretical Perspectives, NATO ASI Series E, Vol. 256 (J. S. Faulkner and R. G. Jordan, eds.) p. 205. Kluwer Academic Publishers, Dordrecht, The Netherlands.

de Fontaine, D., Mann, M. E., and Ceder, G. 1989. *Phys. Rev. Lett.* 63:1300.

de Fontaine, D., Wille, L. T., and Moss, S. C. 1987. *Phys. Rev. B* 36:5709.

Dreyssé, H., Berera, A., Wille, L. T., and de Fontaine, D. 1989. *Phys. Rev. B* 39:2442.

Ducastelle, F. 1991. Order and Phase Stability in Alloys. North-Holland Publishing, New York.

Ferreira, L. G., Mbaye, A. A., and Zunger, A. 1988. *Phys. Rev. B* 37:10547.

Finel, A. 1994. *Prog. Theor. Phys. Suppl.* 115:59.

Fultz, B., Anthony, L., Nagel, J., Nicklow, R. M., and Spooner, S. 1995. *Phys. Rev. B* 52:3315.

Garbulsky, G. D. and Ceder, G. 1994. *Phys. Rev. B* 49:6327.

Gibbs, J. W. Oct. 1875–May 1876. On the equilibrium of heterogeneous substances. *Trans. Conn. Acad.* III:108.

Gibbs, J. W. May 1877–July 1878. *Trans. Conn. Acad.* III:343.

Hijmans, J. and de Boer, J. 1955. *Physica (Utrecht)* 21:471, 485, 499.

Inden, G. and Pitsch, W. 1991. *In* Phase Transformations in Materials (P. Haasen, ed.) pp. 497–552. VCH Publishers, New York.

Jansson, B., Schalin, M., Selleby, M., and Sundman, B. 1993. *In* Computer Software in Chemical and Extractive Metallurgy (C. W. Bale and G. A. Irons, eds.) pp. 57–71. The Metals Society of the Canadian Institute for Metallurgy, Quebec, Canada.

Johnson, D. D. and Asta, M. D. 1997. *Comput. Mater. Sci.* 8:54.

Khachaturyan, A. G. 1983. Theory of Structural Transformation in Solids. John Wiley & Sons, New York.

Khachaturyan, A. G., Linsey T. F., and Morris, J. W., Jr. 1988. *Met. Trans.* 19A:249.

Kikuchi, R. 1951. *Phys. Rev.* 81:988.

Kikuchi, R. 1974. *J. Chem. Phys.* 60:1071.

Kikuchi, R. and Masuda-Jindo, K. 1997. *Comput. Mater. Sci.* 8:1.

Kohan, A. F. and Ceder, G. 1996. *Phys. Rev. B* 53:8993.

Kovalev, A. I., Bronfin, M. B., Loshchinin, Yu V., and Vertogradskii, V. A. 1976. *High Temp. High Pressures* 8:581.

Lu, Z. W., Wei, S.-H., and Zunger, A. 1993. *Europhys. Lett.* 21:221.

Lu, Z. W., Wei, S.-H., Zunger, A., Frota-Pessoa, S., and Ferreira, L. G. 1991. *Phys. Rev. B* 44:512.

Massalski, T. B. (ed.) 1986. Binary Alloy Phase Diagrams, 1st ed. American Society for Metals, Materials Park, Ohio.

Massalski, T. B. (ed.) 1990. Binary Alloy Phase Diagrams, 2nd ed. American Society for Metals, Materials Park, Ohio.

McCormack, R., Asta, M. D., Hoyt, J. J., Chakoumakos, B. C., Mitsure, S. T., Althoff, J. D., and Johnson, D. D. 1997. *Comput. Mater. Sci.* 8:39.

McCormack, R. and de Fontaine, D. 1996. *Phys. Rev B* 54:9746.

Mohri, T., Terakura, K., Oguchi, T., and Watanabe, K. 1988. *Acta Metall.* 36:547.

Mohri, T., Terakura, K., Takizawa, S., and Sanchez, J. M., 1991. *Acta Metall. Mater.* 39:493.

Morgan, D., de Fontaine, D., and Asta, M. D. 1996. Unpublished.

Nagel, L. J., Anthony, L., and Fultz, B. 1995. *Philos. Mag. Lett.* 72:421.

Palatnik, L. S. and Landau, A. I. 1964. Phase Equilibria in Multicomponent Systems. Holt, Rinehart, and Winston, New York.

Pasturel, A., Colinet, C., Paxton, A. T., and van Schilfgaarde, M. 1992. *J. Phys. Condens. Matter* 4:945.

Pinski, F. J., Ginatempo, B., Johnson, D. D., Staunton, J. B., Stocks, G. M., and Györffy, B. L. 1991. *Phys. Rev. Lett.* 66:776.

Prince, A. 1966. Alloy Phase Equilibria. Elsevier, Amsterdam.

Sanchez, J. M. 1993. *Phys. Rev. B* 48:14013.

Sanchez, J. M., Ducastelle, F., and Gratias, D. 1984. *Physica* 128A:334.

Sanchez, J. M. and de Fontaine, D. 1978. *Phys. Rev. B* 17:2926.

Sanchez, J. M., Stark, J., and Moruzzi, V. L. 1991. *Phys. Rev. B* 44:5411.

Shockley, W. 1938. *J. Chem. Phys.* 6:130.

Sigli, C. and Sanchez, J. M. 1986. *Acta Metall.* 34:1021.

Sluiter, M. H. F., de Fontaine, D., Guo, X. Q., Podloucky, R., and Freeman, A. J. 1989. *Mater. Res. Soc. Symp. Proc.* 133:3.

Sluiter, M. H. F., de Fontaine, D., Guo, X. Q., Podloucky, R., and Freeman, A. J. 1990. *Phys. Rev. B* 42:10460.

Sluiter, M. H. F., Turchi, P. E. A., Pinski, F. J., and Johnson, D. D. 1992. *Mater. Sci. Eng. A* 152:1.

Sluiter, M. H. F., Watanabe, Y., de Fontaine, D., and Kawazoe, Y. 1996. *Phys. Rev. B* 53:6137.

Sob, M., Wang, L. G., and Vitek, V. 1997. *Comput. Mater. Sci.* 8:100.

Tepesch, P. D., Garbulsky, G. D., and Ceder, G. 1995. *Phys. Rev. Lett.* 74:2272.

Tepesch , P. D., Kohan, A. F., Garbulsky, G. D., and Cedex, G. 1996. *J. Am. Ceram. Soc.* 79:2033.

Terakura, K., Oguchi, T., Mohri, T., and Watanabe, K. 1987. *Phys. Rev. B* 35:2169.

Turchi, P. E. A., Sluiter, M. H. F., Pinski, F. J., and Johnson, D. D. 1991a. *Mater. Res. Soc. Symp.* 186:59.

Turchi, P. E. A., Sluiter, M. H. F., Pinski, F. J., Johnson, D. D., Nicholson, D. M. Stocks, G. M., and Staunton, J. B. 1991b. *Phys. Rev. Lett.* 67:1779.

van Baal, C. M., 1973. *Physica (Utrecht)* 64:571.

Vul, D. and de Fontaine, D. 1993. *Mater. Res. Soc. Symp. Proc.* 291:401.

Walker, R. A. and Darby, J. R. 1970. *Acta Metall.* 18:1761.

Wille, L. T. and de Fontaine, D. 1988. *Phys. Rev. B* 37:2227.

Wolverton, C., Asta, M. D., Dreyssé, H., and de Fontaine, D. 1991a. *Phys. Rev. B* 44:4914.

Wolverton, C., Ceder, G., de Fontaine, D., and Dreyssé, H. 1993. *Phys. Rev. B* 48:726.

Wolverton, C., Dreyssé, H., and de Fontaine, D. 1991b. *Mater. Res. Soc. Symp. Proc.* 193:183.

Wolverton, C. and Zunger, A. 1994. *Phys. Rev. B* 50:10548.

Wolverton, C. and Zunger, A. 1995. *Phys. Rev. B* 52:8813.

Zunger, A. 1994. *In* Statics and Dynamics of Phase Transformations (P. E. A. Turchi and A. Gonis, eds.). pp. 361–419. Plenum, New York.

KEY REFERENCES

Ceder, 1993. See above.

Original published text for the decoupling of configurational effects from other excitations; appeared initially in G. Ceder's Ph.D. Dissertation at U. C. Berkeley. The paper at first met with some (unjustified) hostility from traditional statistical mechanics specialists.

de Fontaine, 1979. See above.

A 200-page review of configurational thermodynamics in alloys, quite much and didactic in its approach. Written before the application of "first principles" methods to the alloy problem, and before generalized cluster techniques had been developed. Still, a useful review of earlier Bragg-Williams and concentration wave methods.

de Fontaine, 1994. See above.

Follows the 1979 review by the author; This time emphasis is placed on cluster expansion techniques and on the application of ab initio calculations. First sections may be overly general, thereby complicating the notation. Later sections are more readable, as they refer to simpler cases. Section 17 contains a useful table of published papers on the subject of CVM/ab initio calculations of binary and ternary phase diagrams, reasonably complete until 1993.

Ducastelle, 1991. See above.

"The" textbook for the field of statistical thermodynamics and electronic structure calculations (mostly tight binding and generalized perturbation method). Includes very complete description of ground-state analysis by the method of Kanamori et al. and also of the ANNNI model applied to long-period alloy superstructures.

Gibbs, 1875–1876. See above.

One of the great classics of 19th century physics, it launched a new field: that of chemical thermodynamics. Unparalleled for rigor and generality.

Inden and Pitsch, 1991. See above.

A very readable survey of the CVM method applied to phase diagram calculations. Owes much to the collaboration of A. Finel for the theoretical developments, though his name unfortunately does not appear as a co-author.

Kikuchi, 1951. See above.

The classical paper that ushered in the cluster variational method, affectionately known as "CVM" to its practitioners. Kikuchi employs a highly geometrical approach to the derivation of the CM equations. The more algebraic derivations found in later works by Ducastelle, Finel, and Sanchez (cited in this list) may be simpler to follow.

Lu et al., 1991. See above.

Fundamental paper from the "Zunger School" that recommends the Connolly-Williams method, also called the structure inversion method (SIM, in de Fontaine, 1994). That and later publications by the Zunger group correctly emphasize the major role played by elastic interactions in alloy theory calculations.

Palatnik and Landau, 1964. See above.

This little-known textbook by Russian authors (translated into English) is just about the only description of the mathematics (linear algebra, mostly) of multicomponent systems. Some of the main results given in this textbook are summarized in the book by Prince, listed below.

Prince, 1966. See above.

Excellent textbook on the classical thermodynamics of alloy phase diagram constructions, mainly for binary and ternary systems. This handsome edition contains a large number of phase diagram figures, in two colors (red and black lines), contributing greatly to clarity. Unfortunately, the book has been out of print for many years.

Sanchez et al., 1984. See above.

The original reference for the "cluster algebra" applied to alloy thermodynamics. An important paper that proves rigorously that the cluster expansion is orthogonal and complete.

DIDIER DE FONTAINE
University of California
Berkely, California

SIMULATION OF MICROSTRUCTURAL EVOLUTION USING THE FIELD METHOD

INTRODUCTION

Morphological pattern formation and its spatio-temporal evolution are among the most intriguing natural phenomena governed by nonlinear complex dynamics. These

phenomena form the major study subjects of many fields including biology, hydrodynamics, chemistry, optics, materials science, ecology, geology, geomorphology, and cosmology, to name a few. In materials science, the morphological pattern is referred to as microstructure, which is characterized by the size, shape, and spatial arrangement of phases, grains, orientation variants, ferroelastic/ferroelectric/ferromagnetic domains, and/or other structural defects. These structural features usually have an intermediate mesoscopic length scale in the range of nanometers to microns. Figure 1 shows several examples of microstructures observed in metals and ceramics.

Microstructure plays a critical role in determining the physical properties and mechanical behavior of materials. In fact, the use of materials as a science did not begin until human beings learned how to tailor the microstructure to modify the properties. Today the primary task of material design and manufacture is to optimize microstructures—by adjusting alloy composition and varying processing sequences—to obtain the desired properties.

Microstructural evolution is a kinetic process that involves the appearance and disappearance of various transient and metastable phases and morphologies. The optimal properties of materials are almost always associated with one of these nonequilibrium states, which is "frozen" at the application temperatures. Theoretical characterization of the transient and metastable states represents a typical nonequilibrium nonlinear and nonlocal multiparticle problem. This problem generally resists any analytical solution except in a few extremely simple cases.

With the advent of and easy access to high-speed digital computers, computer modeling is playing an increasingly important role in the fundamental understanding of the mechanisms underlying microstructural evolution. For many cases, it is now possible to simulate and predict critical microstructural features and their time evolution. In this unit, methods recently developed for simulating the formation and dynamic evolution of complex microstructural patterns will be reviewed. First, we will give a brief account of the basic features of each method, including the conventional front-tracking method (also referred to as the sharp-interface method) and the new techniques without front-tracking. Detailed description of the simulation technique and its practical application will focus on the field method since it is the one we are most familiar with and we feel it is a more flexible method with broader applications. The need to link the mesoscopic microstructural simulation with atomistic calculation as well as with macroscopic process modeling and property calculation will also be briefly discussed.

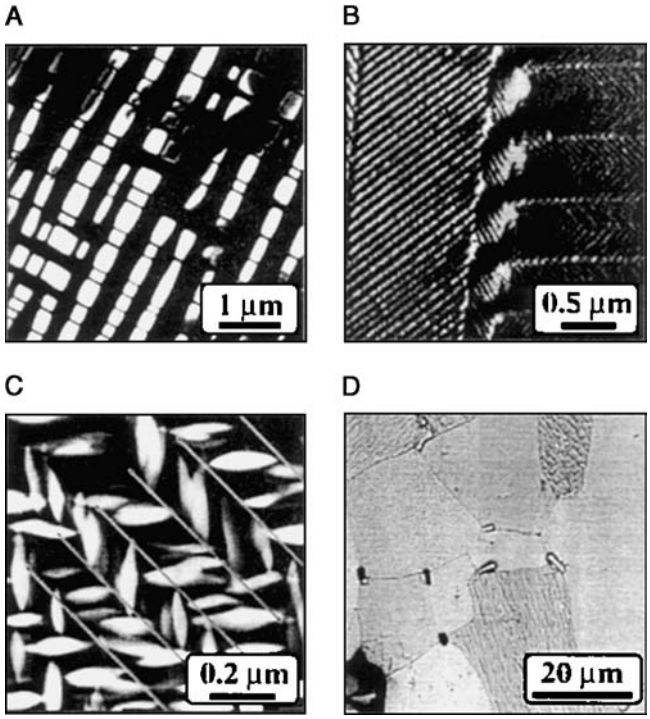


Figure 1. Examples of microstructures observed in metals and ceramics (**A**) Discontinuous rafting structure of γ' particles in Ni-Al-Mo (courtesy of M. Fahrmann). (**B**) Polytwin structure found in Cu-Au (Syutkina and Jakovleva, 1967). (**C**) Alternating band structure in Mg-PSZ (Bateman and Notis, 1992). (**D**) Polycrystalline microstructure in $SrTiO_3$-1 mol% Nb_2O_5 oxygen sensor sintered at 1550°C (Gouma et al., 1997).

Simulation Methods of Microstructural Evolution

Conventional Front-Tracking Methods. A microstructure is conventionally characterized by the geometry of interfacial boundaries between different structural components (phases, grains, etc.) that are assumed to be homogeneous up to their boundaries. The boundaries are treated as mathematical interfaces of zero thickness. The microstructural evolution is then obtained by solving the partial differential equations (PDEs) in each phase and/or domain with boundary conditions specified at the interfaces that are moving during a microstructural evolution. Such a moving-boundary or free-boundary problem is very problematic in numerical analysis and becomes extremely difficult to solve for complicated geometries. For example, the difficulties in dealing with topological changes during the microstructural evolution such as the appearance, coalescence, and disappearance of particles and domains during nucleation, growth, and coarsening processes have limited the application of the method mainly to two dimensions (2D). The extension of the calculation to 3D would introduce significant complications in both the algorithms and their implementation. In spite of these difficulties, important fundamental understanding has been developed concerning the kinetics and microstructural evolution during coarsening, strain-induced coarsening (Leo and Sekerka, 1989; Leo et al., 1990; Johnson et al., 1990; Johnson and Voorhees, 1992; Abinandanan and Johnson, 1993a,b; Thompson et al., 1994; Su and Voorhees, 1996a,b) and dendritic growth (Jou et al., 1994, 1997; Müller-Krumbhaar, 1996) by using, for example, the boundary integral method.

New Techniques Without Front Tracking. In order to take into account realistic microstructural features, novel simulation techniques at both mesoscopic and microscopic levels have been developed, which eliminate the need to track moving interfacial boundaries. Commonly used techniques include microscopic and continuum field methods, mesoscopic and atomistic Monte Carlo methods, the cellular automata method, microscopic master equations, the inhomogeneous path probability method, and molecular dynamics simulations.

Continuum Field Method (CFM). CFM (commonly referred to as the phase field method) is a phenomenological approach based on classical thermodynamic and kinetic theories (Hohenberg and Halperin, 1977; Gunton et al., 1983; Elder, 1993; Wang et al., 1996; Chen and Wang, 1996). Like conventional treatment, CFM describes microstructural evolution by PDEs. However, instead of using the geometry of interfacial boundaries, this method describes a complicated microstructure as a whole by using a set of mesoscopic field variables that are spatially continuous and time dependent. The most familiar examples of field variables are the concentration field used to characterize composition heterogeneity and the long-range order (lro) parameter fields used to characterize the structural heterogeneity (symmetry changes) in multiphase and/or polycrystalline materials. The spatio-temporal evolution of these fields, which provides all the information concerning a microstructural evolution at a mesoscopic level, can be obtained by solving the semiphenomenological dynamic equations of motion, for example, the nonlinear Cahn-Hilliard (CH) diffusion equation (Cahn, 1961, 1962) for the concentration field and the time-dependent Ginzburg-Landau (TDGL) equation (see Gunton et al., 1983 for a review) for the lro parameter fields.

CFM offers several unique features. First, it can easily describe any arbitrary morphology of a complex microstructure by using the field variables, including such details as shapes of individual particles, grains, or domains and their spatial arrangement, local curvatures of surfaces (e.g., surface groove and dihedral angle), and internal boundaries. Second, CFM can take into account various thermodynamic driving forces associated with both short- and long-range interactions, and hence allows one to explore the effects of internal and applied fields (e.g., strain, electrical, and magnetic fields) on the microstructure development. Third, CFM can simulate different phenomena such as nucleation, growth, coarsening, and applied field-induced domain switching within the same physical and mathematical formalism. Fourth, the CH diffusion equation allows a straightforward characterization of the long-range diffusion, which is the dominant process taking place (e.g., during sintering, precipitation of second-phase particles, solute segregation, etc. Fifth, the time, length, and temperature scales in the CFM are determined by the semiphenomenological constants used in the CH and TDGL equations, which in principle can be related to experimentally measured or *ab initio* calculated quantities of a particular system. Therefore, this method can be directly applied to specific material systems. Finally, it is an easy technique and its implementation in both 2D and 3D is rather straightforward. These unique features have made CFM a very attractive method in dealing with microstructural evolution in a wide range of advanced material systems. Table 1 summarizes its major recent applications (Chen and Wang, 1996).

Mesoscopic Monte Carlo (MC) and Cellular Automata (CA) Methods. MC and CA are discrete methods developed for the study of the collective behavior of a large number of elementary events. These methods are not limited to any particular physical process, and have a rich variety of applications in very different fields. The general principles of these two methods can be found in the monographs by Allen and Tildesley (1987) and Wolfram (1986). Their applications to problems related to materials science can be found in recent reviews by Satio (1997) and by Ozawa et al. (1996) for both mesoscopic and microscopic MC and in Wolfram (1984) and Lepinoux (1996) for CA.

The mesoscopic MC technique was developed by Srolovitz et al. (1983), who coarse grained the atomistic Potts model—a generalized version of the Ising model that allows more than two degenerate states—to a subgrain (particle) level for application to mesoscopic microstructural evolution, in particular grain growth in polycrystalline materials. The CA method was originally developed for simulating the formation of complicated morphological patterns during biological evolution (von Neumann, 1966; Young and Corey, 1990). Since most of the time the microscopic mechanisms of growth are unknown, CA was adopted as a trial-and-error computer experiment (i.e., by matching the simulated morphological patterns to observations) to develop a fundamental understanding of the elementary rules dominating cell duplication. The applications of CA in materials science started from work by Lepinoux and Kubin (1987) on dynamic evolution of dislocation structures and work by Hesselbarth and Göbel (1991) on grain growth during primary recrystallization.

The MC and CA methods share several common characteristics. For example, they describe mesoscopic microstructure and its evolution in a similar way. Instead of using PDEs, they map an initial microstructure on to a discrete lattice (square or triangle in 2D and cubic in 3D) with each lattice site or unit cell assigned a microstructural state (e.g., the crystallographic orientation of grains in polycrystalline materials). The interfacial boundaries are described automatically wherever the change of microstructural state takes place. The lattices are so-called coarse-grained lattices, which have a length scale much larger than the atomic scale but substantially smaller than the typical particle or domain size. Therefore, these methods can easily describe complicated morphological features defined at a mesoscopic level using the current generation of computers. The dynamic evolution of the microstructure is described by updating the microstructural state at each lattice site following either certain probabilistic procedures (in MC) or some simple deterministic transition rules (in CA). The implementation of these probabilistic procedures or deterministic rules provides certain

Table 1. Examples of CFM Applications

Type of Processes	Field variables[a]	References
Spinodal decomposition:		
Binary systems	c	Hohenberg and Halperin (1977); Rogers et al. (1988); Nishimori and Onuki (1990); Wang et al. (1993)
Ternary systems	c_1, c_2	Eyre (1993); Chen (1993a, 1994)
Ordering and antiphase domain coarsening	η	Allen and Cahn (1979); Oono and Puri (1987)
Solidification in:		
Single-component systems	η	Langer (1986); Caginalp (1986); Wheeler et al. (1992); Kobayashi (1993)
Alloys	c, η	Wheeler et al. (1993); Elder et al. (1994); Karma (1994); Warren and Boettinger (1995); Karma and Rappel (1996, 1998); Steinbach and Schmitz (1998)
Ferroelectric domain formation		
180° domain	P	Yang and Chen (1995)
90° domain	P_1, P_2, P_3	Nambu and Sagala (1994); Hu and Chen (1997); Semenovskaya and Khachaturyan (1998)
Precipitation of ordered inter-metallics with:		
Two kinds of ordered domains	c, η	Eguchi et al. (1984); Wang et al. (1994); Wang and Khachaturyan (1995a, b)
Four kinds of ordered domains	c, η_1, η_2, η_3	Wang et al. (1998); Li and Chen (1997a)
Applied stress	c, η_1, η_2, η_3	Li and Chen (1997b, c, d, 1998a, b)
Tetragonal precipitates in a cubic matrix	c, η_1, η_2, η_3	Wang et al. (1995); Fan and Chen (1995a)
Cubic → tetragonal displacive transformation	η_1, η_2, η_3	Fan and Chen (1995b)
Three-dimensional martensitic transformation	η_1, η_2, η_3	Wang and Khachaturyan (1997); Semenovskaya and Khachaturyan (1997)
Grain growth in:		
Single-phase material	η_1, η_2, . . . , η_Q	Chen and Yang (1994); Fan and Chen (1997a, b)
Two-phase material	c, η_1, η_2, . . . , η_Q	Chen and Fan (1996)
Sintering	ρ, η_1, η_2, . . . , η_Q	Cannon (1995); Liu et al. (1997)

[a] c, composition; η, lro parameter; P, polarization; ρ, relative density; Q, total number of lro parameters required.

ways of sampling the coarse-grained phase space for the free energy minimum. Because both time and space are discrete in these methods, they are direct computational methods and have been demonstrated to be very efficient for simulating the fundamental phenomena taking place during grain growth (Srolovitz et al., 1983, 1984; Anderson et al., 1984; Gao and Thompson, 1996; Liu et al., 1996; Tikare and Cawley, 1997), recrystallization (Hesselbarth and Göbel, 1991; Holm et al., 1996), and solidification (Rappaz and Gandin, 1993; Shelton and Dunand, 1996).

However, both approaches suffer from similar limitations. First, it becomes increasingly difficult for them to take into account long-range diffusion and long-range interactions, which are the two major factors dominating microstructural evolution in many advanced material systems. Second, the choice of the coarse-grained lattice type may have a strong effect on the interfacial energy anisotropy and boundary mobility, which determine the morphology and kinetics of the microstructural evolution. Therefore, careful consideration must be given to each type of application (Holm et al., 1996). Third, the time and length scales in MC/CA simulations are measured by the number of MC/CA time steps and space increments, respectively. They do not correspond, in general, to the real time and space scales. Extra justifications have to be made in order to relate these quantities to real values (Mareschal and Lemarchand, 1996; Lepinoux, 1996; Holm et al., 1996). Some recent simulations have tried to incorporate real time and length scales by matching the simulation results to either experimental observations or constitutive relations (Rappaz and Gandin, 1993; Gao and Thompson, 1996; Saito, 1997).

Microscopic Field Model (MFM). MFM is the microscopic counterpart of CFM. Instead of using a set of mesoscopic field variables, it describes an arbitrary microstructure by a microscopic single-site occupation probability function, which describes the probability that a given lattice

site is occupied by a solute atom of a given type at a given time. The spatio-temporal evolution of the occupation probability function following the microscopic lattice-site diffusion equation (Khachaturyan, 1968, 1983) yields all the information concerning the kinetics of atomic ordering and decomposition as well as the corresponding microstructural evolution within the same physical and mathematical model (Chen and Khachaturyan, 1991a; Semenovskaya and Khachaturyan, 1991; Wang et al., 1993; Koyama et al., 1995; Poduri and Chen, 1997, 1998).

The CH and TDGL equations in CFM can be obtained from the microscopic diffusion equation in MFM by a limit transition to the continuum (Chen, 1993b; Wang et al., 1996). Both the macroscopic CH/TDGL equations and the microscopic diffusion equation are actually different forms of the phenomenological dynamic equation of motion, the Onsager equation, which simply assumes that the rate of relaxation of any nonequilibrium field toward equilibrium is proportional to the thermodynamic driving force, which is the first variational derivative of the total free energy. Therefore, both equations are valid within the same linear approximation with respect to the thermodynamic driving force.

The total free energy in MFM is usually approximated by a mean-field free energy model, which is a functional of the occupation probability function. The only input parameters in the free energy model are the interatomic interaction energies. In principle, these engeries can be determined from experiments or theoretical calculations. Therefore, MFM does not require any *a priori* assumptions on the structure of equilibrium and metastable phases that might appear along the evolution path of a nonequilibrium system. In CFM, however, one must specify the form of the free energy functional in terms of composition and long-range order parameters, and thus assume that the only permitted ordered phases during a phase transformation are those whose long-range order parameters are included in the continuum free energy model. Therefore, there is always a possibility of missing important transient or intermediate ordered phases not described by the *a priori* chosen free energy functional. Indeed, MFM has been used to predict transient or metastable phases along a phase-transformation path (Chen and Khachaturyan, 1991b, 1992).

By solving the microscopic diffusion equations in the Fourier space, MFM can also deal with interatomic interactions with very long interation ranges, such as elastic interactions (Chen et al., 1991; Semenovskaya and Khachaturyan, 1991; Wang et al., 1991a, b, 1993), Coulomb interactions (Chen and Khachaturyan, 1993; Semenovskaya et al., 1993), and magnetic and electric dipole-dipole interactions.

However, CFM is more generic and can describe larger systems. If all the equilibrium and metastable phases that may exist along a transformation path are known in advance from experiments or calculations [in fact, the free energy functional in CFM can always be fitted to the mean-field free energy model or the more accurate cluster variation method (CVM) calculations if the interatomic interaction energies for a particular system are known], CFM can be very powerful in describing morphological evolution, as we will see in more detail in the two examples given in Practical Aspects of the Method. CFM can be applied essentially to every situation in which the free energy can be written as a functional of conserved and nonconserved order parameters that do not have to be the lro parameters, whereas MFM can only be applied to describe diffusional transformations such as atomic ordering and decomposition on a fixed lattice.

Microscopic Master Equations (MEs) and Inhomogeneous Path Probability Method (PPM). A common feature of CFM and MFM is that they are both based on a free energy functional, which depends only on a local order parameter/local atomic density (in CFM) or point probabilities (in MFM). Two-point or higher-order correlations are ignored. Both CFM and MFM describe the changes of order parameters or occupation probabilities with respect to time as linearly proportional to the thermodynamic driving force. In principle, the description of kinetics by linear models is only valid when a system is not too far from equilibrium, i.e., the driving force for a given diffusional process is small. Furthermore, the proportionality constants in these models are usually assumed to be independent of the values of local order parameters or the occupation probabilities.

One way to overcome the limitations of CFM and MFM is to use inhomogeneous microscopic ME (Van Baal, 1985; Martin, 1994; Chen and Simmons, 1994; Dobretsov et al., 1995, 1996; Geng and Chen, 1994, 1996), or the inhomogeneous PPM (R. Kikuchi, pers. comm., 1996). In these methods, the kinetic equations are written with respect to one-, two-, or n-particle cluster probability distribution functions (where n is the number of particles in a given cluster), depending on the level of approximation. Rates of changes of those cluster correlation functions are proportional to the exponential of an activation energy for atomic diffusion jumps. With the input of interaction parameters, the activation energy for diffusion, and the initial microstructure, the temporal evolution of these nonequilibrium distribution functions describes the kinetics of diffusion processes such as ordering, decomposition, and coarsening, as well as the accompanying microstructural evolution. In both ME and PPM, the free energy functional does not explicitly enter into the kinetic equations of motion. High-order atomic correlations such as pair and tetrahedral correlations can be taken into account. The rate of change of an order parameter is highly nonlinear with respect to the thermodynamic driving force. The dependence of atomic diffusion or exchange on the local atomic configuration is automatically considered. Therefore, it is a substantial improvement over the continuum and microscopic field equations derived from a free energy functional in terms of describing rates of transformations. At equilibrium, both ME and PPM produce the same equilibrium states as one would calculate from the CVM at the same level of approximation.

However, if one's primary concern is in the sequence of a microstructural evolution (e.g., the appearance and disappearance of various transient and metastable structural states, rather than the absolute rate of the transformation), CFM and MFM are simpler and more powerful methods. Alternatively, as in MFM, ME and PPM can

only be applied to diffusional processes on a fixed lattice and the system sizes are limited by the total number of lattice sites that can be considered. Furthermore, it is very difficult to incorporate long-range interactions in ME and PPM. Finally, the formulation becomes very tedious and complicated as high-order probability distribution functions are included.

Atomistic Monte Carlo Method. In an atomistic MC simulation, a given morphology or microstructure is described by occupation numbers on an Ising lattice, representing the instantaneous configuration of an atomic assembly. The evolution of the occupation numbers as a function of time is determined by Metropolis types of algorithms (Allen and Tildesley, 1987). Similar to MFM and the microscopic ME, the atomistic MC may be applied to microstructural evolution during diffusional processes such as ordering transformation, phase separation, and coarsening. This method has been extensively employed in the study of phase transitions involving simple atomic exchanges on a rigid lattice in binary alloys (Bortz et al., 1974; Lebowitz et al., 1982; Binder, 1992; Fratzl and Penrose, 1995). However, there are several differences between the atomistic MC method and the kinetic equation approaches. First of all, the probability distribution functions generated by the kinetic equations are averages over a time-dependent nonequilibrium ensemble, whereas in MC, a series of snapshots of instantaneous atomic configurations along the simulated Markov chain are produced (Allen and Tildesley, 1987). Therefore, a certain averaging procedure has to be designed in the MC technique in order to obtain information about local composition or local order. Second, while the time scale is clearly defined in CFM/MFM and microscopic ME, it is rather difficult to relate the MC time steps to real time. Third, in many cases, simulations based on kinetic equations are computationally more efficient than MC, which is also the main reason why most of the equilibrium phase diagram calculations in alloys were performed using CVM instead of MC. The main disadvantage of the microscopic kinetic equations, as mentioned in the previous section, is the fact that the equations become increasingly tedious when increasingly higher-order correlations are included, whereas in MC, essentially all correlations are automatically included. Furthermore, MC is easier to implement than ME and PPM. Therefore, atomistic MC simulation remains a very popular method for modeling the diffusional transformations in alloy systems.

Molecular Dynamics Simulations. Since microstructure evolution is realized through the movement of individual atoms, atomistic simulation techniques such as the atomistic molecular dynamics (MD) approach should, in principle, provide more direct and accurate descriptions by solving Newton's equations of motion for each atom in the system. Unfortunately, such a brute force approach is still not feasible computationally when applied to studying overall mesoscale microstructural evolution. For example, even with the use of state-of-the-art parallel computers, MD simulations cannot describe dynamic processes exceeding nanoseconds in systems containing several million atoms. In microstructural evolution, however, many morphological patterns are measured in submicrons and beyond, and the transport phenomena that produce them may occur in minutes and hours. Therefore, it is unlikely that the advancements in computer technology will make such length and time scales accessible by atomistic simulations in the near future.

However, the insight into microscopic mechanisms obtained from MD simulations provides critical information needed for coarser scale simulations. For example, the semiphenomenological thermodynamic and kinetic coefficients in CFM and the fundamental transition rules in MC and CA simulations can be obtained from the atomistic simulations. On the other hand, for some localized problems of microstructural evolution (such as the atomic rearrangement near a crack tip and within a grain boundary, deposition of thin films, sintering of nanoparticles, and so on), atomistic simulations have been very successful (Abraham, 1996; Gumbsch, 1996; Sutton, 1996; Zeng et al., 1998; Zhu and Averback, 1996; Vashista et al., 1997). The readers should refer to the corresponding units (see, e.g., PREDICTION OF PHASE DIAGRAMS) of this volume for details.

In concluding this section, we must emphasize that each method has its own advantages and limitations. One should make the best use of these methods for the problems that he is solving. For complicated problems, one may use more than one method to utilize the complementary characteristics of each of them.

PRINCIPLES OF THE METHOD

CFM is a mesoscopic technique that solves systems of coupled nonlinear partial differential equations of a set of coarse-grained field variables. The solution provides spatio-temporal evolution of the field variables from which detailed information can be obtained on the sizes and shapes of individual particles/domains and their spatial arrangements at each time moment during the microstructural evolution. By subsequent statistical analysis of the simulation results, more quantitative microstructural information such as average particle size and size distribution as well as constitutive relations characterizing the time evolution of these properties can be obtained. These results may serve as valuable input for macroscopic process modeling and property calculations. However, no structural information at an atomic level (e.g., the atomic arrangements of each constituent phase/domain and their boundaries) can be obtained by this method. On the other hand, it is still too computationally intensive at present to use CFM to directly simulate microstructural evolution during an entire material process such as metal forming or casting. Also, the question of what kinds of microstructure provide the optimal properties cannot be answered by the microstructure simulation itself. To answer this question we need to integrate the microstructure simulation with property calculations. Both process modeling and property calculation belong to the domain of macroscopic simulation, which is characterized by the use of constitutive equations. Typical macroscopic simulation techniques include the finite element method (FEM) for

solids and the finite difference method (FDM) for fluids. The generic relationships between simulation techniques at different length and time scales will be discussed in Future Trends, below.

Theoretical Basis of CFM

The principles of CFM are based on the classical thermodynamic and kinetic theories. For example, total free energy reduction is the driving force for the microstructural evolution, and the atomic and interface mobility determines the rate of the evolution. The total free energy reduction during microstructural evolution usually consists of one or several of the following: reduction in bulk chemical free energy; decrease of surface and interfacial energies; relaxation of elastic strain energy; reduction in electrostatic/magnetostatic energies; and relaxation of energies associated with an external field such as applied stress, electrical, or magnetic field. Under the action of these forces, different structural components (phases/domains) will rearrange themselves, through either diffusion or interface controlled kinetics, into new configurations with lower energies. Typical events include nucleation (or continuous separation) of new phases and/or domains and subsequent growth and coarsening of the resulting multiphase/multidomain heterogeneous microstructure.

Coarse-Grained Approximation. According to statistical mechanics, the free energy of a system is determined by all the possible microstates characterized by the electronic, vibrational, and occupational degrees of freedom of the constituent atoms. Because it is still impossible at present to sample all these microscopic degrees of freedom, a coarse-grained approximation is usually employed in the existing simulation techniques where a free energy functional is required. The basic idea of coarse graining is to reduce the $\sim 10^{23}$ microscopic degrees of freedom to a level that can be handled by the current generation of computers. In this approach, the microscopic degrees of freedom are separated into "fast" and "slow" variables and the computational models are built on the slow ones. The remaining fast degrees of freedom are then incorporated into a so-called coarse-grained free energy. Familiar examples of the fast degrees of freedom are (a) the occupation number at a given lattice site in a solid solution that fluctuates rapidly and (b) the magnetic spin at a given site in ferromagnetics that flip rapidly. The corresponding slow variables in the two systems will be the concentration and magnetization, respectively. These variables are actually the statistical averages over the fast degrees of freedom. Therefore, the slow variables are a small set of mesoscopic variables whose dynamic evolution is slow compared to the microscopic degrees of freedom, and whose spatial variation is over a length scale much larger than the interatomic distance. After such a coarse graining, the $\sim 10^{23}$ microscopic degrees of freedom are replaced by a much smaller number of mesoscopic variables whose spatio-temporal evolution can be described by the semiphenomenological dynamic equations of motion. On the other hand, these slow variables are the quantities that are routinely measured experimentally, and hence the calculation results can be directly compared with experimental observations.

Under the coarse-grained approximation, computer simulation of microstructural evolution using the field method is reduced to the following four major steps:

1. Find the proper slow variables for the particular microstructural features under consideration.
2. Formulate the coarse-grained free energy, F_{cg}, as a function of the chosen slow variables following the symmetry and basic thermodynamic behavior of the system.
3. Fit the phenomenological coefficients in F_{cg} to experimental data or more fundamental calculations.
4. Set up appropriate initial and/or boundary conditions and numerically solve the field kinetic equations (PDEs).

The solution will yield the spatio-temporal evolution of the field variables that describes every detail of the microstructure evolution at a mesoscopic level. The computer programming in both 2D and 3D is rather straightforward; the following diagram summarizes the main steps involved in a computer simulation (Fig. 2).

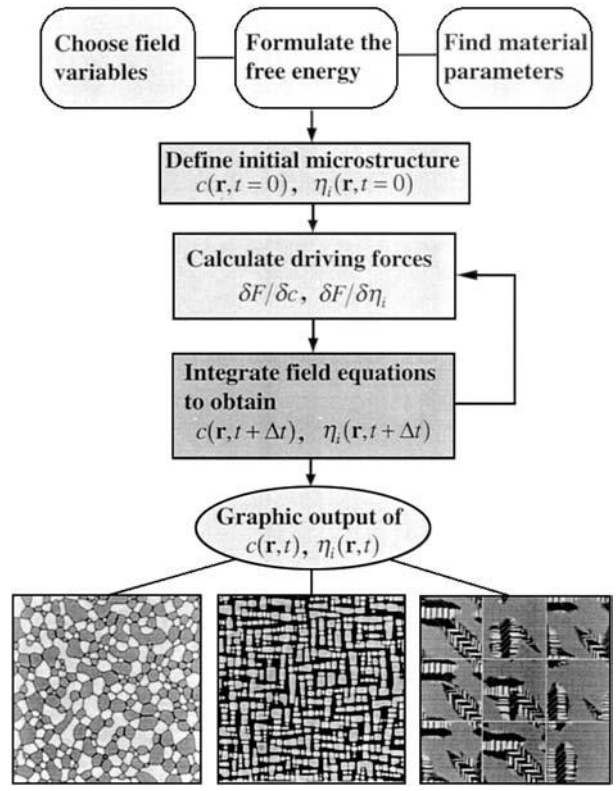


Figure 2. The basic procedures of CFM.

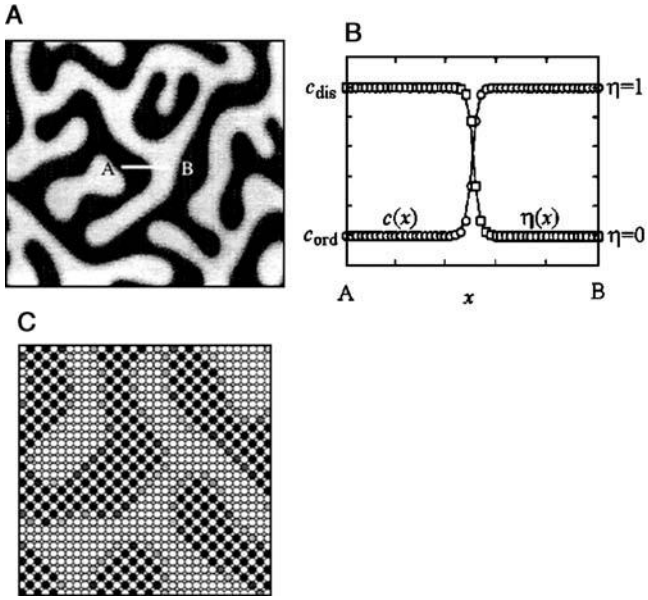


Figure 3. Typical 2D contour plot of concentration field (**A**) 1D plot of concentration and lro parameter profiles across an interface (**B**), and a plot of the occupation probability function (**C**) for a two-phase mixture of ordered and discarded phases.

Selection of Slow Variables. The mesoscopic slow variables used in CFM are defined as spatially continuous and time-dependent field variables. As already mentioned, the most commonly used field variables are the concentration field, $c(\mathbf{r}, t)$ (where $\mathbf{r}$ is the spatial coordinate), which describes compositional inhomogeneity and the lro parameter field, $\eta(\mathbf{r}, t)$, which describes structural inhomogeneity. Figure 3 shows the typical 2D contour plots (Fig. 3A) and one-dimensional (1D) profiles (Fig. 3B) of these fields for a two-phase alloy where the coexisting phases have not only different compositions but different crystal symmetry as well. To illustrate the coarse graining scheme, the atomic structure described by a microscopic field variable, $n(\mathbf{r}, t)$, the occupation probability function of a solute atom at a given lattice site, is also presented (Fig. 3C).

The selection of field variables for a particular system is a fundamental question that has to be answered first. In principle, the number of variables selected should be just enough to describe the major characteristics of an evolving microstructure (e.g., grains and grain boundaries in a polycrystalline material, second-phase particles of two-phase alloys, orientation variants in a martensitic crystal, and electric/magnetic domains in ferroelectrics and ferromagnetics). Worked examples can be found in Table 1. Detailed procedures for the selection of field variables for a particular application can be found in the examples presented in Practical Aspects of the Method.

Diffuse-Interface Nature. From Figure 3B, one may notice that the field variables $c(\mathbf{r}, t)$ and $\eta(\mathbf{r}, t)$ are continuous across the interfaces between different phases or structural domains, even though they have very large gradients in these regions. For this reason, CFM is also referred to as a diffuse interface approach. However, mesoscopic simulation techniques [e.g., the conventional sharp interface approach, the discrete lattice approach (MC and CA), or the diffuse interface CFM] ignore atomic details, and hence give no direct information on the atomic arrangement of different phases and interfacial boundaries. Therefore, they cannot give an accurate description of the thickness of interfacial boundaries. In the front tracking method, for example, the thickness of the boundary is zero. In MC, the boundary is defined as the regions between adjacent lattice sites of different microstructural states. Therefore, the thickness of the boundary is roughly equal to the lattice parameter of the coarse-grained lattice. In CFM, the boundary thickness is also commensurate with the mesh size of the numerical algorithm used for solving the PDEs (~3 to 5 mesh grids).

Formulation of the Coarse-Grained Free Energy. Formulation of the nonequilibrium coarse-grained free energy as a functional of the chosen field variables is central to CFM because the free energy defines the thermodynamic behavior of a system. The minimization of the free energy along a transformation path yields the transient, metastable, and equilibrium states of the field variables, which in turn define the corresponding microstructural states. All the energies that are microstructure dependent, such as those mentioned earlier, have to be properly incorporated. These energies fall roughly into two distinctive categories: the ones associated with short-range interatomic interactions (e.g., the bulk chemical free energy and interfacial energy) and the ones resulting from long-range interactions (e.g., the elastic energy and electrostatic energy). These energies play quite different roles in driving the microstructural evolution.

Bulk Chemical Free Energy. Associated with the short-range interatomic interactions (e.g., bonding between atoms), the bulk chemical free energy determines the number of phases present and their relative amounts at equilibrium. Its minimization gives the information that is found in an equilibrium phase diagram. For a mixture of equilibrium phases, the total chemical free energy is morphology independent, i.e., it does not depend on the size, shape, and spatial arrangement of coexisting phases and/or domains. It simply follows the Gibbs additive principle and depends only on the volume fraction of each phase. In the field approach, the bulk chemical free energy is described by a local free energy density function, f, which can be approximated by a Landau-type of polynomial expansion with respect to the field variables. The specific form of the polynomial is usually obtained from general symmetry and stability considerations that do not require reference to the atomic details (Izyumov and Syromyatnikov, 1990). For an isostructural spinodal decomposition, for example, the simplest polynomial providing a double-well structure of the f-c diagram for a miscibility gap (Fig. 4) can be formulated as

$$f(c) = -\frac{1}{2}A_1(c - c_\alpha)^2 + \frac{1}{4}A_2(c - c_\beta)^4 \qquad (1)$$

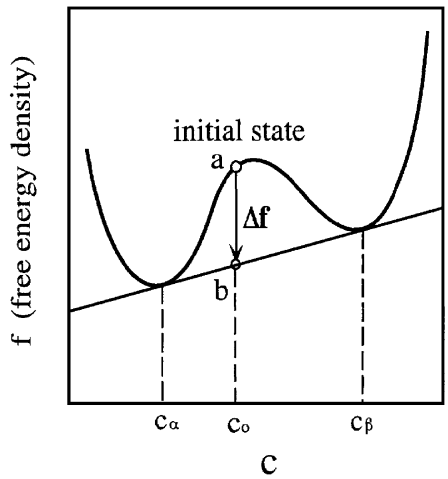


Figure 4. A schematic free energy vs. composition plot for spinodal decomposition.

where A_1, A_2, c_α, and c_β are positive constants. In general, the phenomenological expansion coefficients A_1 and A_2 are functions of temperature and their temperature dependence determines the specific thermodynamic behavior of the system. For example, A_1 should be negative when the temperatures are above the miscibility gap.

If an alloy is quenched from its homogenization temperature within a single-phase field in the phase diagram into a two-phase field, the free energy of the initial homogeneous single-phase, which is characterized by $c(\mathbf{r}, t) = c_0$, is given by point a in the free energy curve in Figure 4. If the alloy decomposes into a two-phase mixture of compositions c_α and c_β, the free energy is given by point b, which corresponds to the minimum free energy at the given composition. The bulk chemical free energy reduction (per unit volume) is the free energy difference at the two points a and b, i.e., Δf, which is the driving force for the decomposition.

Interfacial Energy. The interfacial energy is an excess free energy associated with the compositional and/or structural inhomogeneities occurring at interfaces. It is part of the chemical free energy and can be introduced into the coarse-grained free energy by adding a gradient term (or gradient terms for complex systems where more than one field variable is required) to the bulk free energy (Rowlinson, 1979; Ornstein and Zernicke, 1918; Landau, 1937a, b; Cahn and Hilliard, 1958), for example,

$$F_{\text{chem}} = \int_V \left[f(c) + \frac{1}{2}\kappa_C (\nabla c)^2 \right] dV \qquad (2)$$

where F_{chem} is the total chemical free energy, κ_C is the gradient energy coefficient, which, in principle, depends on both temperature and composition and can be expressed in terms of interatomic interaction energies (Cahn and Hilliard, 1958; Khachaturyan, 1983), (∇) is the gradient operator, and V is the system volume. The second term in the bracket of the integrand in Equation (2) is referred to as the gradient energy term. Please note that the gradient energy, (volume integral of the gradient) is not the interfacial energy. The interfacial energy by definition is the total excess free energy associated with inhomogeneities at interfaces and is given by (Cahn and Hilliard, 1958)

$$\sigma_{\text{int}} = \int_{c_\alpha}^{C_\beta} \sqrt{2\kappa_C [f(c) - f_0(c)]}\, dc \qquad (3)$$

where $f_0(c)$ is the equilibrium free energy density of a two-phase mixture. It is represented by the common tangent line through c_α and c_β in Figure 4. Therefore, the gradient energy is only part of the interfacial energy.

Elastic Energy. Microstructural evolution in solids usually involves crystal lattice rearrangement, which leads to a lattice mismatch between the coexisting phases/domains. If the interphase boundaries are coherent, i.e., the lattice planes are continuous across the boundaries, elastic strain fields will be generated in the vicinity of the boundaries to eliminate the lattice discontinuity. The elastic interaction associated with the strain fields has an infinite long-range asymptote decaying as $1/r^3$, where r is the separation distance between two interacting finite elements of the coexisting phases/domains. Accordingly, the elastic energy arising from such a long-range interaction is very different from the bulk chemical free energy and the interfacial energy, which are associated with the short-range interactions. For example, the bulk chemical free energy depends only on the volume fraction of each phase, and the interfacial energy depends only on the total interfacial area, while the elastic energy depends on both the volume and the morphology of the coexisting phases. The contribution of elastic energy to the total free energy results in a very special situation where the total bulk free energy becomes morphology dependent. In this case, the shape, size, orientation, and spatial arrangement of phases/domains become internal thermodynamic parameters similar to concentration and lro parameter profiles. Their equilibrium values are determined by minimizing the total free energy. Therefore, elastic energy relaxation is a key driving force dominating the microstructural evolution in coherent systems. It is responsible for the formation of various collective mesoscopic domain structures.

The elasticity theory for calculating the elastic energy of an arbitrary microstructure was proposed about 30 years ago by Khachaturyan (1967, 1983) and Khachaturyan and Shatalov (1969) using a sharp-interface description and a shape function to characterize a microstructure. It has been reformulated for the diffuse-interface description using a composition field, $c(\mathbf{r}, t)$, if the strain is predominantly caused by the concentration heterogeneity (Wang et al., 1991a,b, 1993; Chen et al., 1991; Onuki, 1989a,b), or lro parameter fields, $\eta_p^2(\mathbf{r}, t)$ (where p represents the pth orientation variant of the product phase), if the strain is mainly caused by the structural heterogeneity (Chen et al., 1992; Wang et al., 1993; Fan and Chen, 1995a; Wang et al., 1995). The elastic energy

has a closed form in Fourier space in terms of concentration or lro parameter fields,

$$E_{\text{el}} = \frac{1}{2}\int \frac{d^3k}{(2\pi)^3} B(\mathbf{e})\,|\tilde{c}(\mathbf{k})|^2 \tag{4}$$

$$E_{\text{el}} = \frac{1}{2}\sum_{pq}\int \frac{d^3k}{(2\pi)^3} B_{pq}(\mathbf{e})\{\eta_p^2(\mathbf{r},t)\}_{\mathbf{k}}\{\eta_q^2(\mathbf{r},t)\}_{\mathbf{k}}^{*} \tag{5}$$

where $\mathbf{e} = \mathbf{k}/k$ is a unit vector in reciprocal space, $\tilde{c}(\mathbf{k})$ and $\{\eta_p^2(\mathbf{r},t)\}_{\mathbf{k}}$ are the Fourier transforms of $c(\mathbf{r}, t)$ and $\eta_p^2(\mathbf{r},t)$, respectively, and $\{\eta_q^2(\mathbf{r},t)\}_{\mathbf{k}}^{*}$ is the complex conjugate of $\{\eta_q^2(\mathbf{r},t)\}_{\mathbf{k}}$. The sign $\int$ in Equations 4 and 5 means that a volume of $(2\pi)^3/V$ about $\mathbf{k} = 0$ is excluded from the integration.

$$B(\mathbf{e}) = c_{ijkl}\varepsilon_{ij}^0\varepsilon_{kl}^0 - e_i\sigma_{ij}^0\Omega_{jk}(\mathbf{e})\sigma_{kl}^0 e_l \tag{6}$$

$$B_{pq}(\mathbf{e}) = c_{ijkl}\varepsilon_{ij}^0(p)\varepsilon_{kl}^0(q) - e_i\sigma_{ij}^0(p)\Omega_{jk}(\mathbf{e})\sigma_{kl}^0(q)e_l \tag{7}$$

where e_i is the ith component of vector $\mathbf{e}$, c_{ijkl} is the elastic modulus tensor, ε_{ij}^0 and $\varepsilon_{ij}^0(p)$ are the stress-free transformation strain transforming the parent phase into the product phase, and the pth orientation variant of the product phase, respectively, $\sigma_{ij}^0(p) = c_{ijkl}\varepsilon_{kl}^0(p)$ and $\Omega_{ij}(\mathbf{e})$ is a Green function tensor, which is inverse to the tensor $\Omega(\mathbf{e})_{ij}^{-1} = c_{ijkl}e_k e_l$. In Equations 6 and 7, repeated indices imply summation.

The effect of elastic strain on coherent microstructural evolution can be simply included into the field formalism by adding the elastic energy (Equation 4 or 5) into the total coarse-grained free energy, F_{cg}. The function $B(\mathbf{e})$ in Equation 4 or $B_{pq}(\mathbf{e})$ in Equation 5 characterizes the elastic properties and crystallography of the phase transformation of a system through the Green function $\Omega_{ij}(\mathbf{e})$ and the coefficients σ_{ij}^0 and $\sigma_{ij}^0(p)$, respectively, while all the information concerning the morphology of the mesoscopic microstructure enters in the field variables $c(\mathbf{r}, t)$ and $\eta_p(\mathbf{r},t)$. It should be noted that Equations 4 and 5 are derived for homogeneous modulus cases where the coexisting phases are assumed to have the same elastic constants. Modulus misfit may also have an important effect on the microstructural evolution of coherent precipitates (Onuki and Nishimori, 1991; Lee, 1996a; Jou et al., 1997), particularly when an external stress field is applied (Lee, 1996b; Li and Chen 1997a). This effect can also be incorporated in the field approach (Li and Chen, 1997a) by utilizing the elastic energy equation of an inhomogeneous solid formulated by Khachaturyan et al. (1995).

Energies Associated with External Fields. Microstructures will, in general, respond to external fields. For example, ferroelastic, ferroelectric, or ferromagnetic domain walls will move under externally applied stress, electric, or magnetic fields, respectively. Therefore, one can tailor a microstructure by applying external fields. Practical examples include stress aging of superalloys to produce rafting structures (e.g., Tien and Copley, 1971) and thermomagnetic aging to produce highly anisotropic microstructures for permanent magnetic materials (e.g., Cowley et al., 1986). The driving force due to an external field can be incorporated into the field model by including a coupling term between an internal field and an applied field in the coarse-grained free energy: i.e.,

$$F_{\text{coup}} = -\int_V xY\,dV \tag{8}$$

where x represents an internal field such as a local strain field, a local electrical polarization field, or a local magnetic moment field, and Y represents the corresponding applied stress, electric, and magnetic field. The internal field can either be a field variable itself, chosen to represent a microstructure in the field model, or it can be represented by a field variable (Li and Chen, 1997a, b, 1998a, b). If the applied field is homogeneous, the coupling potential energy becomes

$$F_{\text{coup}} = -V\bar{x}Y \tag{9}$$

where $\bar{x}$ is the average value of the internal field over the entire system volume, V.

If a system is homogeneous in terms of its elastic properties and its dielectric and magnetic susceptibilities, this coupling term is the only contribution to the total driving force produced by the applied fields. However, the problem becomes much more complicated if the properties of the material are inhomogeneous. For example, for an elastically inhomogeneous material, such as a two-phase solid of which each phase has a different elastic modulus, an applied field will produce a different elastic deformation within each phase. Consequently, an applied stress field will cause a lattice mismatch in addition to the lattice mismatch due to their stress-free lattice parameter differences. In general, the contribution from external fields to the total driving force for an inhomogeneous material has to be computed numerically for a given microstructure. However, for systems with very small inhomogeneity, as shown by Khachaturyan et al. (1995) for the case of elastic inhomogeneity, the problem can be solved analytically using the effective medium approximation, and the contribution from externally applied stress fields can be directly incorporated into the field model, as will be shown later in one of the examples given in Practical Aspects of the Method.

The Field Kinetic Equations. In most systems, the microstructure evolution involves both compositional and structural changes. Therefore, we generally need two types of field variables, conserved and nonconserved, to fully define a microstructure. The most familiar example of a conserved field variable is the concentration field. It is conserved in a sense that its volume integral is a constant equal to the total number of solute atoms. An example of a nonconserved field is the lro parameter field, which describes the distinction between the symmetry of different phases or the crystallographic orientation between different structural domains of the same phase. Although the lro parameter does not enter explicitly into the equilibrium thermodynamics that are determined solely by the dependence of the free energy on compositions, it may not be

ignored in microstructure modeling. The reason is that it is the evolution of the lro parameter toward equilibrium that describes the crystal lattice rearrangement, atomic ordering, and, in particular, the metastable and transient structural states.

The temporal evolution of the concentration field is governed by a generalized diffusion equation that is usually referred to as the Cahn-Hillaird equation (Cahn, 1961, 1962):

$$\frac{\partial c(\mathbf{r},t)}{\partial t} = -\nabla \bullet \mathbf{J} \tag{10}$$

where $\mathbf{J} = -M\nabla\mu$ is the flux of solute atoms, M is the diffusion mobility, and

$$\mu = \frac{\delta F_{\mathrm{cg}}}{\delta c} = \frac{\delta F_{\mathrm{chem}}}{\delta c} + \frac{\delta F_{\mathrm{elast}}}{\delta c} + \cdots \tag{11}$$

is the chemical potential. The temporal evolution of the lro parameter field is described by a relaxation equation, often called the time-dependent Ginzburg-Landau (TDGL) equation or the Allen-Cahn equation,

$$\frac{\partial \eta(\mathbf{r},t)}{\partial t} = -L\frac{\delta F_{\mathrm{cg}}}{\partial \eta(\mathbf{r},t)} \tag{12}$$

where L is the relaxation constant. Various forms of the field kinetic equations can be found in Gunton et al. (1983). Equations 10 and 12 are coupled through the total coarse-grained free energy, F_{cg}. With random thermal noise terms added, both types of equations become stochastic and their applications to studying critical dynamics have been extensively discussed (Hohenberg and Halperin, 1977; Gunton et al., 1983).

PRACTICAL ASPECTS OF THE METHOD

In the above formulation, simulating microstructural evolution using CFM has been reduced to finding solutions of the field kinetic equations Equations 10 and/or 12 under conditions corresponding to either a prototype model system or a real system. The major input parameters include the kinetic coefficients in the kinetic equations and the expansion coefficients and phenomenological constants in the free energy equations. These parameters are listed in Table 2.

Because of its simplicity, generality, and flexibility, CFM has found a rich variety of applications in very different material systems in the past two decades. For example, by choosing different field variables, the same formalism has been adapted to study solidification in both single- and two-phase materials; grain growth in single-phase materials; coupled grain growth and Ostward ripening in two-phase composites; solid- and liquid-state sintering; and various solid-state phase transformations in metals and ceramics including superalloys, martensitic alloys, transformation toughened ceramics, ferroelectrics, and superconductors (see Table 1 for a summary and references). Later in this section we present a few typical examples to illustrate how to formulate a realistic CFM for a particular application.

Table 2. Major Input Parameters of CFM

Input Coefficients in CFM	Fitting Properties
M, L—The kinetic coefficients in Equations 11 and 13	In general, they can be fitted to atomic diffusion and grain/interfacial boundary mobility data
$A_i (i = 1, 2, \ldots)$—The polynomial expansion coefficient in free energy Equation 1	These parameters can be fitted to the typical free energy vs. concentration curve (and also free energy vs. lro parameter curves for ordering systems), which can be obtained from either the CALPHAD[a] program or CVM[b,c] calculations
κ_C—The gradient coefficients in Equation 2	It can be fitted to available interfacial energy data
c_{ijkl} and ε_0—Elastic constants and lattice misfit	Experimental data are available for many systems

[a]Sundman, B., Jansson, and Anderson, J.-O. 1985. CALPHAD 9:153.
[b]Kikuchi, R. 1951. *Phys. Rev.* 81:988–1003.
[c]Sanchez, J. M., Ducastelle, F., and Gratias, D. 1984. *Physica* 128A:334–350.

Microstructural Evolution of Coherent Ordered Precipitates

Precipitation hardening by small dispersed particles of an ordered intermetallic phase is the major strengthening mechanism for many advanced engineering materials including high-temperature and ultralightweight alloys that are critical to the aerospace and automobile industries. The Ni-based superalloys are typical prototype examples. In these alloys, the ordered intermetallic γ'-Ni_3X $L1_2$ phase precipitates out from the compositionally disordered γ face-centered cubic (fcc) parent phase. Experimental studies have revealed a rich variety of interesting morphological patterns of γ', which play a critical role in determining the mechanical behavior of the material. Well-known examples include rafting structures and split patterns.

A typical example of the discontinuous rafting structure (chains of cuboid particles) formed in a Ni-based superalloy during isothermal aging can be found in Figure 1A, where the ordered particles (bright shades) formed via homogeneous nucleation are coherently embedded in the disordered parent phase matrix (dark shades). They have cuboidal shapes and are very well aligned along the elastically soft $\langle 100 \rangle$ directions. These ordered particles distinguish themselves from the disordered matrix by differences in composition, crystal symmetry, and lattice parameters. In addition, they also distinguish themselves from each other by the relative positions of the origin of their sublattices with respect to the origin of their parent phase lattice, which divides them into four distinctive types of ordered domains called antiphase domains (APDs). When these APDs impinge on each other during

growth and coarsening, a particular structural defect called an antiphase domain boundary (APB) is produced. The energy associated with APBs is usually much higher than the interfacial energy.

The complication associated with the combined effect of ordering and coherency strain in this system has made the microstructure simulation a difficult challenge. For example, all of these structural features have to be considered in a realistic model because they all contribute to the microstructural evolution.

The composition and structure changes associated with the $L1_2$ ordering can be described by the concentration field, $c(\mathbf{r}, t)$, and the lro parameter field, $\eta(\mathbf{r}, t)$, respectively. According to the concentration wave representation of atomic ordering (Khachaturyan, 1978), the lro parameters represent the amplitudes of the concentration waves that generate the atomic ordering. Since the $L1_2$ ordering is generated by superposition of three concentration waves along the three cubic directions in reciprocal space, we need three lro parameters to define the $L1_2$ structure. The combination of the three lro parameters will automatically characterize the four types of APDs and three types of APBs associated with the $L1_2$ ordering. The concentration field plus the three lro parameter fields will fully define the system at a mesoscopic level.

After selecting the field variables, the next step is to formulate the coarse-grained free energy as a functional of these fields. A general form of the polynomial approximation of the bulk chemical free energy can be written as a Taylor expansion series

$$\begin{aligned} f(c,\eta_1,\eta_2,\eta_3) = f_0(c,T) &+ \sum_{p=1}^{3} A_p(c,T)\eta_p \\ &+ \frac{1}{2}\sum_{p,q=1}^{3} A_{pq}(c,T)\eta_p\eta_q \\ &+ \frac{1}{3}\sum_{p,q,r=1}^{3} A_{pqr}(c,T)\eta_p\eta_q\eta_r \\ &+ \frac{1}{4}\sum_{p,q,r,s=1} A_{pqrs}(c,T)\eta_p\eta_q\eta_r\eta_s + \cdots \end{aligned} \tag{13}$$

where $A_p(c,T)$ to $A_{pqrs}(c,T)$ are the expansion coefficients. The free energy $f(c,\eta_1,\eta_2,\eta_3)$, must be invariant with respect to the symmetry operations of an fcc lattice. This requirement drastically simplifies the polynomial. For example, it reduces Equation 13 to the following (Wang et al., 1998):

$$\begin{aligned} f(c,\eta_1,\eta_2,\eta_3) = f_0(c,T) &+ \frac{1}{2}A_2(c,T)(\eta_1^2+\eta_2^2+\eta_3^2) \\ &+ \frac{1}{3}A_3(c,T)\eta_1\eta_2\eta_3 + \frac{1}{4}A_4(c,T)(\eta_1^4+\eta_2^4+\eta_3^4) \\ &+ \frac{1}{4}A_4'(c,T)(\eta_1^2+\eta_2^2+\eta_3^2)^2 + \cdots \end{aligned} \tag{14}$$

According to Equation 14, the minimum energy state is always 4-fold degenerate with respect to the lro parameters. For example, there are four sets of lro parameters that correspond to the free energy extremum, i.e.,

$$\begin{aligned} (\eta_1,\eta_2,\eta_3) = (\eta_0,\eta_0,\eta_0)(\eta_0,-\eta_0,-\eta_0),(-\eta_0,\eta_0,-\eta_0), \\ (-\eta_0,-\eta_0,\eta_0) \end{aligned} \tag{15}$$

where η_0 is the equilibrium lro parameter. These parameters satisfy the following extremum condition

$$\frac{\partial f(c,\eta_1,\eta_2,\eta_3)}{\partial \eta_p} = 0 \tag{16}$$

where $p = 1, 2, 3$. The four sets of lro parameters given in Equation 16 describe the four energetically and structurally equivalent APDs of the $L1_2$ ordered phase. Such a free energy polynomial for the $L1_2$ ordering was first obtained by Lifshitz (1941, 1944) using the theory of irreducible representation of space groups. Similar formulations were also used by Lai (1990), Braun et al. (1996), and Li and Chen (1997a).

The total chemical free energy with the contributions from both concentration and lro parameter inhomogeneities can be formulated by adding the gradient terms. A general form can be written as

$$\begin{aligned} F = \int_V \Big\{ &\frac{1}{2}\kappa_C[\nabla_C(\mathbf{r},t)]^2 + \frac{1}{2}\sum_{p=1}^{3}\kappa_{ij}(p)\nabla_i\eta_p(\mathbf{r},t)\nabla_j\eta_p(\mathbf{r},t) \\ &+ f(c,\eta_1,\eta_2,\eta_3)\Big\} dV \end{aligned} \tag{17}$$

where κ_C and $\kappa_{ij}(p)$ are the gradient energy coefficients and V is the total volume of the system. Different from Equation 2, additional gradient terms of lro parameters are included in Equation 18, which contribute to both interfacial and APB energies as a result of the structural inhomogeneity at these boundaries. These gradient terms introduce correlation between the compositions at neighboring points and also between the lro parameters at these points. In general, the gradient coefficients κ_C and $\kappa_{ij}(p)$ can be determined by fitting the calculated interfacial and APB energies to experimental data for a given system. Next, we present the simplest procedures to fit the polynomial approximation of Equation 14 to experimental data of a particular system.

In principle, with a sufficient number of the polynomial expansion coefficients as fitting parameters, the free energy can always be described with the desired accuracy. For the sake of simplicity, we approximate the free energy density $f(c,\eta_1,\eta_2,\eta_3)$ by a fourth-order polynomial. In this case, the expression of the free energy (Equation 14) for a single-ordered domain with $\eta_1 = \eta_2 = \eta_3 = \eta$ becomes

$$\begin{aligned} f(c,\eta) = &\frac{1}{2}B_1(c-c_1)^2 + \frac{1}{2}B_2(c_2-c)\eta^2 \\ &- \frac{1}{3}B_3\eta^3 + \frac{1}{4}B_4\eta^4 \end{aligned} \tag{18}$$

where

$$\frac{1}{2}B_1(c-c_1)^2 = f_0(c,T_0) \tag{19}$$

$$\frac{1}{2}B_2(c_2-c) = \frac{3}{2}A_2(c,T_0) \tag{20}$$

$$B_3 = -A_3(c,T_0) \tag{21}$$

$$\frac{1}{4}B_4 = \frac{3}{4}A_4(c,T_0) + \frac{9}{4}A_4'(c,T_0) \tag{22}$$

T_0 is the isothermal aging temperature and B_1, B_2, B_3, B_4, c_1, and c_2 are positive constants. The first requirement for the polynomial approximation in Equation 18 is that the signs of the coefficients should be determined in such a way that they provide a global minimum of $f(c,\eta_1,\eta_2,\eta_3)$ at the four sets of lro parameters given in Equation 15. In this case, minimization of the free energy will produce all four possible antiphase domains.

Minimizing $f(c,\eta)$ with respect to η gives the equilibrium value of the lro parameter $\eta = \eta_0(c)$. Substituting $\eta = \eta_0(c)$ into Equation 18 yields $f[c,\eta_0(c)]$, which becomes a function of c only. A generic plot of the *f-c* curve, which is typical for an fcc + $L1_2$ two-phase region in systems like Ni-Al superalloys, is shown in Figure 5. This plot differs from the plot for isostructural decomposition shown in Figure 4 in having two branches. One branch describes the ordered phase and the other describes the disordered phase. Their common tangent determines the equilibrium compositions of the ordered and disordered phases. The simplest way to find the coefficients in Equation 18 is to fit the polynomial to the *f-c* curves shown in Figure 5. If we assume that the equilibrium compositions of γ and γ' phases at $T_0 = 1270$ K are $c_\gamma = 0.125$ and $c_{\gamma'} = 0.24$ (Okamoto, 1993), the lro parameter for the ordered phase is $\eta_{\gamma'} = 0.99$, and the typical driving force of the transformation $|\Delta f| = f(c^*)$ [where c^* is the composition at which the γ and γ' phase have the same free energy (Fig. 5)] is unity, then the coefficients in Equation 19 will have the values $B_1 = 905.387$, $B_2 = 211.611$, $B_3 = 165.031$, $B_4 = 135.637$, $c_1 = 0.125$, and $c_2 = 0.383$, where B_1 to B_4 are measured in units of $|\Delta f|$. The *f-c* diagram produced by this set of parameters is identical to the one shown in Figure 5.

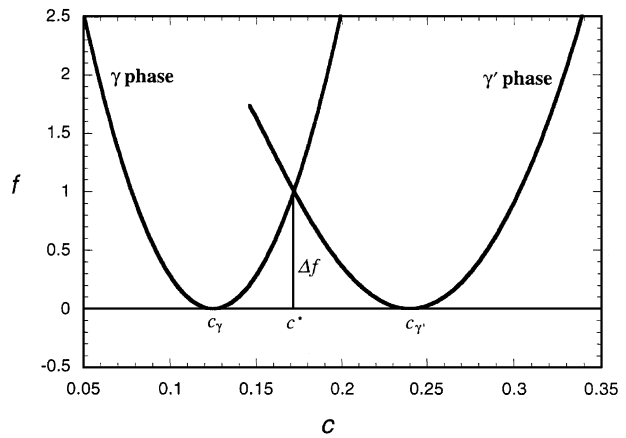


Figure 5. A typical free energy vs. composition plot for a γ/γ' two-phase Ni-Al alloy. The parameter f is presented in a reduced unit.

Note that the above fitting procedure is only semiquantitative because it fits only three characteristic points: the equilibrium compositions c_γ and $c_{\gamma'}$ and the typical driving force $|\Delta f|$ (whose absolute value will be determined below). More accurate fitting can be obtained if more data are available from either calculations or experiments.

In Ni-Al, the lattice parameters of the ordered and disordered phases depend on their compositions only. The strain energy of such a coherent system is then given by Equation 4. It is a functional of the concentration profile $c(\mathbf{r})$ only. The following experimental data have been used in the elastic energy calculation: lattice misfit between γ'/γ phases: $\varepsilon_0 = (a_{\gamma'} - a_\gamma)/a_\gamma \approx 0.0056$ (Miyazaki et al., 1982); elastic constants: $c_{11} = 2.31$, $c_{12} = 1.49$, $c_{44} = 1.17 \times 10^{12}$ erg/cm^3 (Pottenbohm et al., 1983). The typical elastic energy to bulk chemical free energy ratio, $\mu = (c_{11}+c_{12})^2\varepsilon_0^2/|\Delta f|c_{11}$, has been chosen to be 1600 in the simulation, which yields a typical driving force $|\Delta f| = 1.85 \times 10^7$ erg/cm^3.

As with any numerical simulation, it is more convenient to work with dimensionless quantities. If we divide both sides of Equations 10 and 12 by the product $L|\Delta f|$ and introduce a length unit l to the increment of the simulation grid to identify the spatial scale of the microstructure described by the solutions of these equations, we can present all the dimensional quantities entering the equations in the following dimensionless forms:

$$\tau = L|\Delta f|t, \quad \rho = x/l, \quad \varphi_1 = \frac{\kappa_C}{|\Delta f|l^2}, \quad \varphi_2 = \frac{\kappa}{|\Delta f|l^2}, \quad M^* = \frac{M}{Ll^2} \tag{23}$$

where $\tau, \rho, \varphi_1, \varphi_2$ and M^* are the dimensionless time, and length, and gradient energy coefficients and the diffusion mobility, respectively. For simplicity, the tensor coefficient $\kappa_{ij}(p)$ in Equation 17 has been assumed to be $\kappa_{ij}(p) = \kappa\delta_{ij}$ (where δ_{ij} is the Kronecker delta symbol), which is equivalent to assuming an isotropic APB energy. The following numerical values have been used in the simulation: $\varphi_1 = -50.0, \varphi_2 = 5.0, M^* = 0.4$. The value of φ_1 could be positive or negative for systems undergoing ordering, depending on the atomic interaction energies. Here φ_1 is assigned a negative value. By fitting the calculated interfacial energy between γ and γ', $\sigma_{\text{siml}} = \sigma^*_{\text{siml}}|\Delta f|l$ ($\sigma^*_{\text{siml}} = 4.608$ is the calculated dimensionless interfacial energy), to the experimental value, $\sigma_{\text{exp}} = 14.2$ erg/cm^3 (Ardell, 1968), we can find the length unit l of our numerical grid: $l = 16.7$ Å. The simulation result that we will show below is obtained for a 2D system with 512×512 grid points. Therefore, the size of the system is $512l = 8550$ Å. The time scale in the simulation can also be determined if we know the experimental data or calculated values of the kinetic coefficient L or M for the system considered.

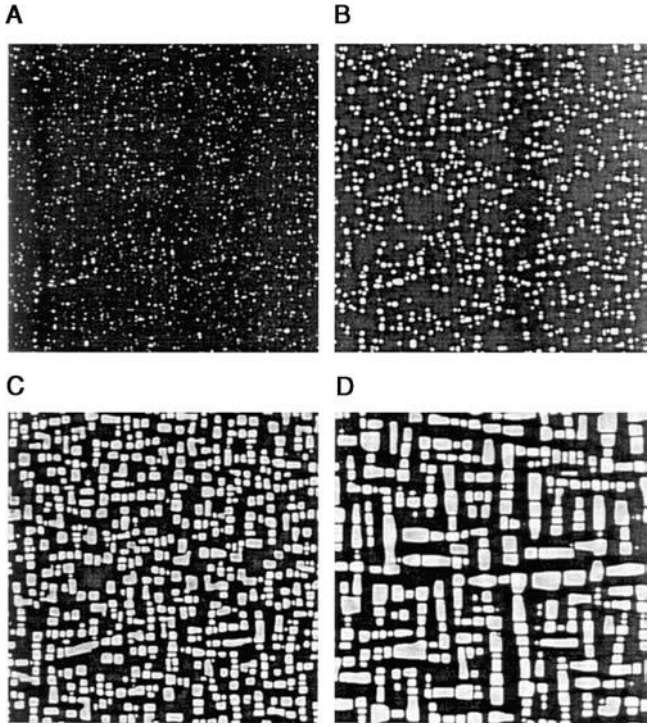


Figure 6. Microstructural evolution of γ' precipitates in Ni-Al obtained by CFM for a 2D system. (**A**) $\tau = 0.5$, (**B**) $\tau = 2$, (**C**) $\tau = 10$, (**D**) $\tau = 100$, where τ is a reduced time.

Since the model is able to describe both ordering and long-range elastic interaction and the material parameters have been fitted to experimental data, numerical solutions of the coupled field kinetic equations, Equations 10 and 12 should be able to reproduce the major characteristics of the microstructural development. Figure 6 shows a typical example of the simulated microstructural evolution for an alloy with 40% (volume fraction) of the ordered phase, where the microstructure is presented through shades of gray in accordance with the concentration field, $c(\mathbf{r}, t)$. The higher the value of $c(\mathbf{r}, t)$, the brighter the shade. The simulation was started from an initial state corresponding to a homogeneous supersaturated solid solution characterized by $c(\mathbf{r},t,0) = \bar{c}, \eta_1(\mathbf{r},t,0) = \eta_2(\mathbf{r},t,0) = \eta_3(\mathbf{r},t,0) = 0$, where $\bar{c}$ is the average composition of the alloy. The nucleation process of the transformation was simulated by the random noise terms added to the field kinetic equations, Equations 10 and 12. Because of the stochastic nature of the noise terms, the four types of ordered domains generated by the $L1_2$ ordering are randomly distributed and have comparable populations.

The microstructure predicted in Figure 6D shows a remarkable agreement with experimental observations (e.g., Fig. 1A), even though the simulation results were obtained for a 2D system. The major morphological features of γ' precipitates observed in the experiments—for example, both the cuboidal shapes of the particles and the discontinuous rafting structures (discrete cuboid particles aligned along the elastically soft directions)—have been reproduced. However, if any of the structural features of the system mentioned earlier (such as the crystal lattice misfit or the ordered nature of the precipitates) is not included, none of these features appear (see Wang et al., 1998, for details).

This example shows that CFM is able to capture the main characteristics of the microstructural evolution of misfitting ordered precipitates involving four types of ordered domains during all three stages of precipitation (e.g., nucleation, growth, and coarsening). Furthermore, it is rather straightforward to include the effect of applied stress in the homogeneous modulus approximation (Li and Chen, 1997a, b) and in cases where the modulus inhomogeneity is small (Khachaturyan et al., 1995; Li and Chen, 1997a). Roughly speaking, an applied stress produces two effects. One is an additional lattice mismatch due to the difference in the elastic constants of the precipitates and the matrix (e.g., the elastic inhomogeneity),

$$\Delta\varepsilon \sim \frac{\Delta\lambda}{\bar{\lambda}^2}\sigma^a \tag{24}$$

where $\Delta\varepsilon$ is the additional mismatch strain caused by the applied stress, $\Delta\lambda$ is the modulus difference between precipitates and matrix, $\bar{\lambda}$ is the average modulus of the two-phase solid, and σ^a is the magnitude of the applied stress. The other is a coupling potential energy,

$$\begin{aligned} F_{\text{coup}} &= -\int_V \bar{\varepsilon}\sigma^a \, d^3r \\ &\sim -V\sum_p \omega_p \varepsilon(p)\sigma^a \end{aligned} \tag{25}$$

where $\bar{\varepsilon}$ is the homogeneous strain and ω_p is the volume fraction of the precipitates.

For an elastically homogeneous system, the coupling potential energy term is the only contribution that can affect the microstructure of a two-phase material. For the γ' precipitates in the Ni-based superalloys discussed above, both γ and γ' are cubic, and thus the lattice misfit is dilatational. The γ' morphology can be affected by applied stresses only when the elastic modulus is inhomogeneous. According to Khachaturyan et al. (1995), if the difference between elastic constants of the precipitates and the matrix is small, the elastic energy of a coherent microstructure can be calculated by (1) replacing the elastic constant in the Green's function tenor, $\Omega_{ij}(e)$, with the average value

$$\bar{c}_{ijkl} = c^*_{ijkl}\omega + c_{ijkl}(1-\omega) \tag{26}$$

where c^*_{ijkl} and c_{ijkl} are the elastic constants of precipitates and the matrix, respectively; and by (2) replacing σ^0_{ij} in Equation 6 with

$$\sigma^*_{ij} = \bar{c}_{ijkl}\varepsilon^0_{kl} - \Delta c_{ijkl}\bar{\varepsilon}_{kl} \tag{27}$$

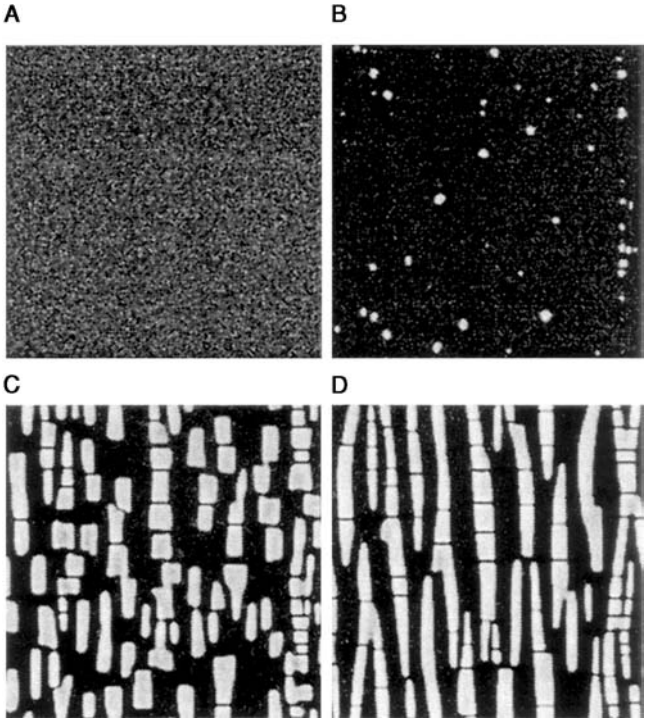


Figure 7. Microstructural evolution during precipitation of γ' particles in γ matrix under a vertically applied tensile strain. (**A**) $\tau = 100$, (**B**) $\tau = 300$, (**C**) $\tau = 1500$, (**D**) $\tau = 5000$, where τ is a reduced time.

where $\bar{\varepsilon}_{ij}$ is the homogeneous strain and $\Delta c_{ijkl} = c^*_{ijkl} - c_{ijkl}$. When the system is under a constant strain constraint, the homogeneous strain, $\bar{\epsilon}_{ij}$, is the applied strain. In this case, $bar\epsilon_{ij} = \bar{\epsilon}ij^a = \bar{S}_{ijkl}\bar{\sigma}^a_{kl}$, where $\bar{\epsilon}^a_{ij}$ is the constraint strain caused by initially applied stress $\bar{\sigma}^a_{kl}$. The parameter $\bar{S}_{ijkl}$ is the average fourth-rank compliance tensor. Figure 7 shows an example of the simulated microstructural evolution during precipitation of the γ' phase from a γ matrix under a constant tensile strain along the vertical direction (Li and Chen, 1997a). In the simulation, the kinetic equations were discretized using 256×256 square grids, and essentially the same thermodynamic model was employed as described above for the precipitation of γ' ordered particles from a γ matrix without an externally applied stress. The elastic constants of γ' and γ employed are (in units of 10^{10} ergs/cm^3) $C^*_{11} = 16.66$, $C^*_{12} = 10.65$, $C^*_{44} = 9.92$, and $C_{11} = 11.24$, $C_{12} = 6.27$, and $C_{44} = 5.69$, respectively. The magnitude of the initially applied stress is 4×10^8 ergs/cm^3 ($= 40$ MPa). It can been seen in Figure 7 that γ' precipitates were elongated along the applied tensile strain direction and formed a raft-like microstructure. Comparing this finding to the results obtained in the previous case (Fig. 6D), where the rafts are equally oriented along the elastically soft directions ([10] and [01] in the 2D system considered), the rafts in the presence of an external field are oriented uniaxially. In principle, one could use the same model to predict the effect of composition, phase volume fraction, stress state, and magnitude on the directional coarsening (or rafting) behavior.

Microstructural Evolution During Diffusion-Controlled Grain Growth

Controlling grain growth, and hence the grain size, is a critical issue for the processing and application of polycrystalline materials. One of the effective methods of controlling grain growth is the development of multiphase composites or duplex microstructures. An important example is the Al_2O_3-ZrO_2 two-phase composite, which has been extensively studied because of its toughening and superplastic behaviors (for a review, see Harmer et al., 1992). Coarsening of such a two-phase microstructure is driven by the reduction in the total grain and interphase boundary energy. It involves two quite different but simultaneous processes: grain growth through grain boundary migration and Ostwald ripening via long-range diffusion.

To model such a process, one can describe an arbitrary two-phase polycrystalline microstructure using the following set of field variables (Chen and Fan, 1996; Fan and Chen, 1997c, d, e),

$$\eta^\alpha_1(\mathbf{r},t), \eta^\alpha_2(\mathbf{r},t), \ldots, \eta^\alpha_p(\mathbf{r},t), \eta^\beta_1(\mathbf{r},t), \eta^\beta_2(\mathbf{r},t), \ldots, \eta^\beta_q(\mathbf{r},t), c(\mathbf{r},t) \tag{28}$$

where $\eta^\alpha_i (i = 1, \ldots, p)$ and $\eta^\beta_j (j = 1, \ldots, q)$ are called orientation field variables, with each representing grains of a given crystallographic orientation of a given phase. Those variables change continuously in space and assume continuous values ranging from -1.0 to 1.0. For example, a value of 1.0 for $\eta^\alpha_1(\mathbf{r},t)$, with values for all the other orientation variables 0.0 at $\mathbf{r}$, means that the material at position ($\mathbf{r}$, t) belongs to a α grain with the crystallographic orientation labeled as 1. Within the grain boundary region between two α grains with orientation 1 and 2, $\eta^\alpha_1(\mathbf{r},t)$ and $\eta^\alpha_2(\mathbf{r},t)$ will have absolute values intermediate between 0.0 and 1.0. The composition field is $c(\mathbf{r},t)$, which takes the value of c_α within an α grain and c_β within a β grain. The parameter $c(\mathbf{r},t)$ has intermediate values between c_α and c_β across an α/β interphase boundary.

The total free energy of such a two-phase system, F_{cg}, can be written as

$$F_{cg} = \int \Bigg[f_0(c(\mathbf{r},t), \eta^\alpha_1(\mathbf{r},t), \eta^\alpha_2(\mathbf{r},t), \ldots, \eta^\alpha_p(\mathbf{r},t), \eta^\beta_1(\mathbf{r},t), \eta^\beta_2(\mathbf{r},t), \ldots, \eta^\beta_q(\mathbf{r},t) + \frac{\kappa_c}{2}\left(\nabla c(\mathbf{r},t)\right)^2 + \sum_{i=1}^{p} \frac{\kappa^\alpha_i}{2}\left(\nabla \eta^\alpha_i(\mathbf{r},t)\right)^2 + \sum_{i=1}^{q} \frac{\kappa^\beta_i}{2}\left(\nabla \eta^\beta_i(\mathbf{r},t)\right)^2 \Bigg] d^3r \tag{29}$$

where f_0 is the local free energy density, κ_C, κ^α_i, and κ^β_i are the gradient energy coefficients for the composition field and orientation fields, and p and q represent the number of orientation field variables for α and β. The cross-gradient energy terms have been ignored for simplicity. In order to simulate the coupled grain growth and Ostwald ripening for a given system, the free energy density

function f_0 should have the following characteristics: (1) if the values of all the orientation field variables are zero, it describes the dependence of the free energy of the liquid phase on composition; (2) the free energy density as a function of composition in a given α-phase grain is obtained by minimizing f_0 with respect to the orientation field variable corresponding to that grain under the condition that all other orientation field variables are zero; (3) the free energy density as a function of composition of a given β phase grain may be obtained in a similar way. Therefore, all the phenomenological parameters in the free energy model, in principle, may be fixed using the information about the free energies of the liquid, solid α phase, and solid β phase.

Other main requirements for f_0 is that it has p degenerate minima with equal depth located at $(\eta_1^\alpha, \eta_2^\alpha, \ldots, \eta_p^\alpha) = (1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,0,\ldots,1)$in p-dimension orientation space at the equilibrium concentration c_α, and that it have q degenerate minima located at $(\eta_1^\beta, \eta_2^\beta, \ldots, \eta_q^\beta) = (1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,0,\ldots,1)$ at c_β. These requirements ensure that each point in space can only belong to one crystallographic orientation of a given phase.

Once the free energy density is obtained, the gradient energy coefficients can be fitted to the grain boundary energies of α and β as well as the α/β boundary energy by numerically solving Equations 10 and 12. The kinetic coefficients, in principle, can be fitted to grain boundary mobility and atomic diffusion data.

It should be emphasized that unless one is interested in phase transformations between α and β, the exact form of the free energy density function may not be very important in modeling microstructural evolution during coarsening of a two-phase solid. The reason is that the driving force for grain growth is the reduction in total grain and interphase boundary energy. Other important parameters are the diffusion coefficients and boundary mobilities. In other words, we assume that the values of the grain and interphase boundary energies together with the kinetic coefficients completely control the kinetics of microstructural evolution, irrespective of the form of the free energy density function.

Below we show an example of investigating the micostructural evolution and the grain growth kinetics in the Al_2O_3-ZrO_2 system using the continuum field model. It was reported (Chen and Xue, 1990) that the ratio of grain boundary energy in Al_2O_3 (denoted as α phase) to the interphase boundary energy between Al_2O_3 and ZrO_2 is $R_\alpha = \sigma^\alpha_{\text{alu}}/\sigma^{\alpha\beta}_{\text{int}} = 1.4$, and the ratio of grain boundary energy in ZrO_2 (denoted as β phase) to the interphase boundary energy is $R_\beta = \sigma^\beta_{\text{zir}}/\sigma^{\alpha\beta}_{\text{int}} = 0.97$. The gradient coefficients and phenomenological parameters in the free energy density function are fitted to reproduce these experimentally determined ratios. The following assumptions are made in the simulation: the grain boundary energies and the interphase boundary energy are isotropic, the grain boundary mobility is isotropic, and the chemical diffusion coefficient is the same in both phases. The systems of CH and TDGL equations are solved using the finite difference in space and explicit Euler method in time (Press et al., 1992).

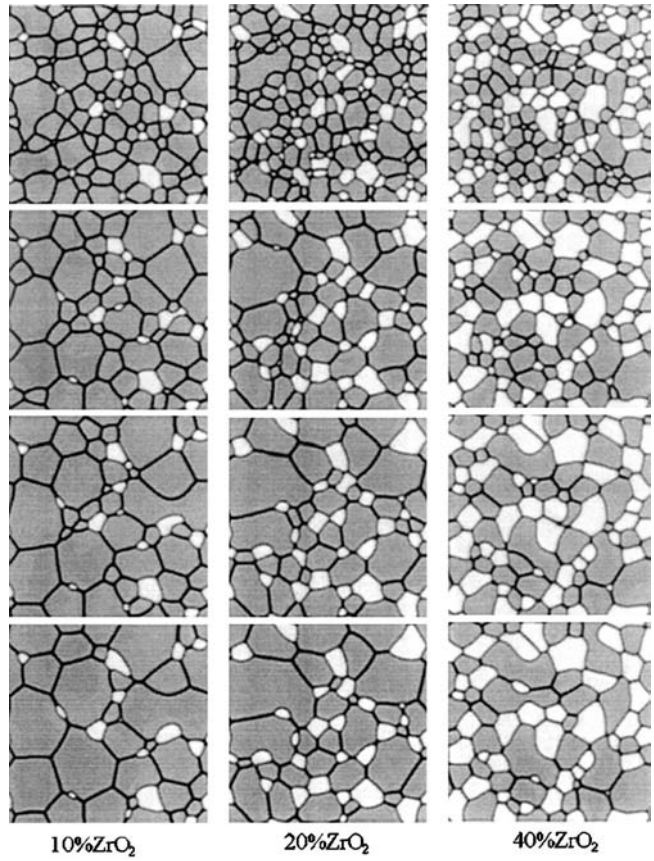


Figure 8. The temporal microstructural evolution in ZrO_2-Al_2O_3 two-phase solids with volume fraction of 10%, 20%, and 40% ZrO_2.

Examples of microstructures with 10%, 20%, and 40% volume fraction of ZrO_2 are shown in Figure 8. The computer simulations were carried out in two dimensions with 256×256 points and with periodic boundary conditions applied along both directions. Bright regions will be β grains (ZrO_2), gray regions are α grains (Al_2O_3), and the dark lines are grain or interphase boundaries. The total number of orientation field variables $(p + q)$ is 30. The initial microstructures were generated from fine grain structures produced by a normal grain growth simulation, and then by randomly assigning all the grains to either α or β according to the desired volume fractions. The simulated microstructures agree very well with those observed experimentally (Lange and Hirlinger, 1987; French et al., 1990; Alexander et al., 1994). More importantly, the main features of coupled grain growth and Ostwald ripening, as observed experimentally, are predicted by the computer simulations (for details, see Fan and Chen, 1997e).

One can obtain all the kinetic data and size distributions with the temporal microstructural evolution. For example, we showed that for both phases, the average size $(R)_t$ as a function of time (t) follows the growth-power law $R_t^m - R_0^m = kt$, with $m = 3$, which is independent of the volume fraction of the second phase, indicating that the coarsening is always controlled by the long-range diffusion process in two-phase solids (Fan and Chen, 1997e). The predicted volume fraction dependence of the kinetic

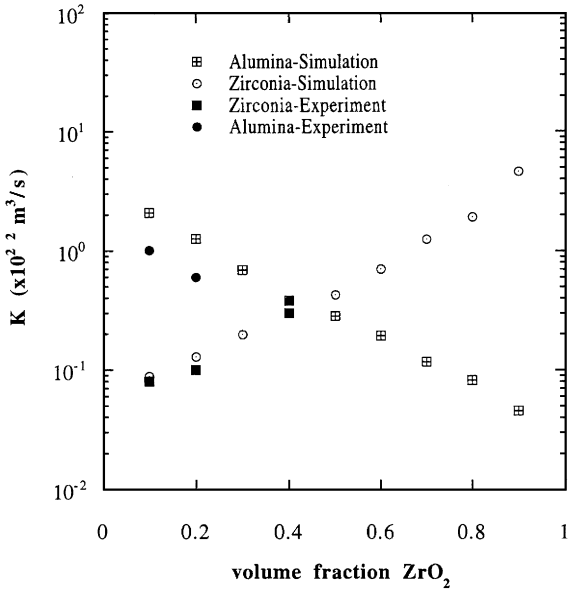


Figure 9. Comparison of simulation and experiments on the dependence of kinetic coefficient k in each phase on volume fraction of ZrO_2 in (Al_2O_3-ZrO_2). (Adopted from Alexander et al., 1994.)

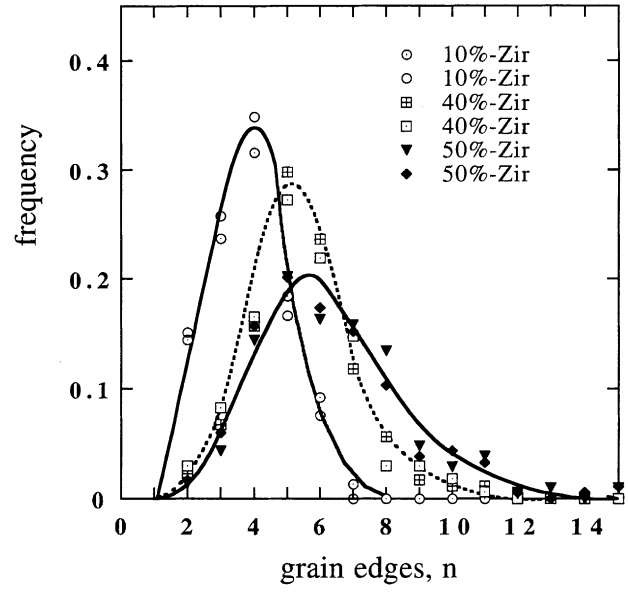


Figure 11. The effect of volume fraction on the topological distributions of ZrO_2 (β) Phase in Al_2O_3-ZrO_2 systems.

coefficients is compared with experimental results in the Ce-doped Al_2O_3-ZrO_2 (CeZTA) system (Alexander et al., 1994) in Figure 9. In Figure 9, the kinetic coefficients were normalized with the experimental values for CeZTA at 40% ZrO_2 (Alexander et al., 1994).

One type of distribution that we can obtain from the simulated microstructures is the topological distribution that is a plot of the frequency of grains with a certain number of sides. The topological distributions for the α and β phases in the 10% ZrO_2 system are compared in Figure 10. It can be seen that the second phase (10% ZrO_2) has much narrower distributions and the peaks have shifted to the four-sided grain while the distributions for the matrix phase (90% Al_2O_3) are much wider with the peaks still at the six-sided grains. The effect of volume fractions on topological distributions of the ZrO_2 phase is summarized in Figure 11. It shows that, as the volume fraction of ZrO_2 increases, topological distributions become wider and peak frequencies lower; this is accompanied by a shift of the peak position from four- to five-sided grains. Similar dependence of topological distributions on volume fractions were observed experimentally on 2D cross-sections of 3D microstructures in the Al_2O_3-ZrO_2 two-phase composites (Alexander, 1995).

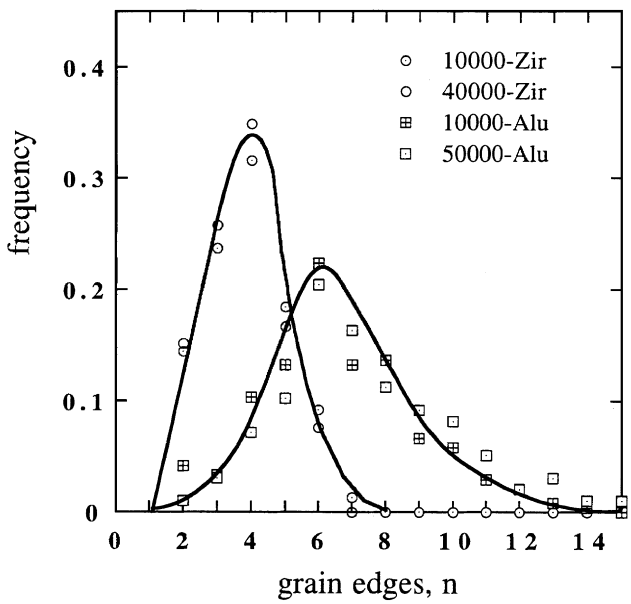


Figure 10. The comparison of topological distributions between Al_2O_3 (α) phase and ZrO_2 (β) Phase in 10% ZrO_2 + 90% Al_2O_3 system. The numbers in the legend represent the time steps.

FUTURE TRENDS

Though process modeling has enjoyed great success in the past years and many commercial software packages are currently available, simulating microstructural evolution is still at its infant stage. It significantly lags behind process modeling. To our knowledge, no commercial software programs are available to process design engineers. Most of the current effort focuses on model development, algorithm improvement, and analyses of fundamental problems. Much more effort is needed toward the development of commercial computational tools that are able to provide optimized processing conditions for desired microstructures. This task is very challenging and the following aspects should be properly addressed.

Links Between Microstructure Modeling and Atomistic Simulations, Continuum Process Modeling, and Property Calculation

One of the greatest challenges associated with materials research is that the fundamental phenomena that

determine the properties and performance of materials take place over a wide range of length and time scales, from microscopic through mesoscopic to macroscopic. As already mentioned, microstructure is defined by mesoscopic features and the CFM discussed in this unit is accordingly a mesoscopic technique. At present, no single simulation method can cover the full range of the three scales. Therefore, integrating simulation techniques across different length and time scales is the key to success in computational characterization of the generic processing/structure/property/performance relationship in material design. The mesoscopic simulation technique of microstructural evolution plays a critical role. It forms a bridge between the atomic-level fundamental calculation and the macroscopic semiempirical process and performance modeling. First of all, fundamental material properties obtained from the atomistic calculations, such as thermodynamic and transport properties, interfacial and grain boundary energies, elastic constants, and crystal structures and lattice parameters of equilibrium and metastable phases, provide valuable input for the mesoscopic microstructural simulations. The output of the mesoscopic simulations (e.g., the detailed microstructural evolution in the form of constitutive relations) can then serve as important input for the macroscopic modeling. Integrating simulations across different length scales and especially time scales is critical to the future virtual prototyping of industrial manufacturing processes, and it will remain the greatest challenge in the next decade for computational materials science (see, e.g., Kirchner et al., 1996; Tadmore et al., 1996; Cleri et al., 1998; Broughton and Rudd, 1998).

More Efficient Numerical Algorithms

From the examples presented above, one can see that simulating the microstructural evolution in various material systems using the field method has been reduced to finding solutions of the corresponding field kinetic equations with all the relevant thermodynamic driving forces incorporated and with the phenomenological parameters fitted to the observed or calculated values. Since the field equations are, in general, nonlinear coupled integral-differential equations, their solution is quite computationally intensive. For example, the typical microstructure evolution shown in Figure 6, obtained with 512×512 uniform grid points using the simple forward Euler technique for numerical integration, takes about 2 h of CPU time and 40 MB of memory of the Cray-C90 supercomputer at the Pittsburg Supercomputing Center. A typical 3D simulation of martensitic transformation using $64 \times 64 \times 64$ uniform grid points (Wang and Khachaturyan, 1997) requires about 2 h of CPU time and 48 MB of memory. Therefore, development of faster and less memory-demanding algorithms and techniques plays a critical role in practical applications of the method.

The most frequently used numerical algorithm in CFM is the finite difference method. For periodic boundary conditions, the fast Fourier transform algorithm (also called spectral method) has also been used, particularly in systems with long-range interactions. In this case, the Fourier transform converts the integral-deferential equations into algebraic equations. In reciprocal space, the simple forward Euler differencing technique can be employed for the numerical solution of the equations. It is a single-step explicit scheme allowing explicit calculation of quantities at time step $t + \Delta t$ in terms of only quantities known at time step t. The major advantages of this single-step explicit algorithm is that it takes little storage, requires a minimal number of Fourier transforms, and executes quickly because it is fully vectorizable (vectorizing a loop will speed up execution by roughly a factor of 10). The disadvantage is that it is only first-order accurate in Δt. In particular, when the field equations are solved in real space the stability with respect to mesh size, numerical noise, and time step could be a major problem. Therefore, more accurate methods are strongly recommended. The time steps allowed in the single-step forward Euler method in reciprocal space fall basically in the range of 0.1 to 0.001 (in reduced units).

Recently, Chen and Shen (1998) proposed a semi-implicit Fourier-spectral method for solving the field equations. For a single TDGL equation, it is demonstrated that for a prescribed accuracy of 0.5% in both the equilibrium profile of an order parameter and the interface velocity, the semi-implicit Fourier-spectral method is $\sim$200 and 900 times more efficient than the explicit Euler finite-difference scheme in 2D and 3D, respectively. For a single CH equation describing the strain-dominated microstructural evolution, the time step that one can use in the first-order semi-implicit Fourier-spectral method is about 400 to 500 times larger than the explicit Fourier-spectral method for the same accuracy.

In principle, the semi-implicit schemes can also be efficiently applied to the TDGL and CH equations with Dirichlet, Neumann, or mixed boundary conditions by using the fast Legendre- or Chebyshev-spectral methods developed (Shen, 1994, 1995) for second- and fourth-order equations. However, the semi-implicit scheme also has its limitations. It is most efficient when applied to problems whose principal elliptic operators have constant coefficients, although problems with variable coefficients can be treated with slightly less efficiency by an iterative procedure (Shen, 1994) or a collocation approach (see, e.g., Canuto et al., 1987). Also, since the proposed method uses a uniform grid for the spatial variables, it may be difficult to resolve extremely sharp interfaces with a moderate number of grid points. In this case, an adaptive spectral method may become more appropriate (see, e.g., Bayliss et al., 1990).

There are two generic features of microstructure evolution that can be utilized to improve the efficiency of CFM simulations. First, the linear dimensions of typical microstructure features such as average particle or domain size keep increasing in time. Second, the compositional and structural heterogeneities of an evolving microstructure usually assume significant values only at interphase or grain boundaries, particularly for coarsening, solidification, and grain growth. Therefore, an ideal numerical algorithm for solving the field kinetic equations should have the potential of generating (1) uniform-mesh grids that change scales in time and (2) adaptive nonuniform grids

that change scales in space, with the finest grids on the interfacial boundaries. The former will make it possible to capture structures and patterns developing at increasingly coarser length scales and the latter will allow us to simulate larger systems. Such adaptive algorithms are under active development. For example, Provatas et al. (1998) recently developed an adaptive grid algorithm to solve the field equations applied to solidification by using adaptive refinement of a finite element grid. They showed that, in two dimensions, solution time will scale linearly with the system size, rather than quadratically as one would expect in a uniform mesh. This finding allows one to solve the field model in much larger systems and for longer simulation times. However, such an adaptive method is generally much more complicated to implement than the uniform grid. In addition, if the spatial scale is only a few times larger than the interfacial width, such an adaptive scheme may not be advantageous over the uniform grid. It seems that the parallel processing techniques developed in massive MD simulations (see, e.g., Vashista et al., 1997) are more attractive in the development of more efficient CFM codes.

PROBLEMS

The accuracy of time and spatial discretization of the kinetic field equations may have a significant effect on the interfacial energy anisotropy and boundary mobility. A common error that could result from a rough discretization of space in the field method is the so-called dynamic frozen result (the microstructure stops to evolve with time artificially). Therefore, careful consideration must be given to a particular type of application, particularly when quantitative information on the absolute growth and coarsening rate of a microstructure is desired. Validation of the growth and coarsening kinetics of CFM simulations against known analytical solutions is strongly recommended. Recently, Bösch et al. (1995) proposed a new algorithm to overcome these problems by randomly rotating and shifting the coarse-grained lattice.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the National Science Foundation under Career Award DMR9703044 (YW) and under grant number DMR 96-33719 (LQC) as well as from the Office of Naval Research under grant number N00014-95-1-0577 (LQC). We would also like to thank Bruce Patton for his comments on the manuscript and Depanwita Banerjee and Yuhui Liu for their help in preparing some of the micrographes.

LITERATURE CITED

Abinandanan, T. A. and Johnson, W. C. 1993a. Coarsening of elastically interacting coherent particles—I. Theoretical formulations. *Acta Metall. Mater*. 41:17–25.

Abinandanan, T. A. and Johnson, W. C. 1993b. II. Simulations of preferential coarsening and particle migrations. *Acta Metall. Mater*. 41:27–39.

Abraham, F. F. 1996. A parallel molecular dynamics investigation of fracture. *In* Computer Simulation in Materials Science—Nano/Meso/Macroscopic Space and Time Scales (H. O. Kirschner, K. P. Kubin, and V. Pontikis, eds.) pp. 211–226, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Alexander, K. B. 1995. Grain growth and microstructural evolution in two-phase systems: alumina/zirconia composites, short course on "Sintering of Ceramics" at the American Ceramic Society Annual Meeting, Cincinnati, Ohio.

Alexander, K. B., Becher, P. F., Waters, S. B., and Bleier, A. 1994. Grain growth kinetics in alumina-zirconia (CeZTA) composites. *J. Am. Ceram. Soc.* 77:939–946.

Allen, M. P. and Tildesley, D. J. 1987. Computer Simulation of Liquids. Oxford University Press, New York.

Allen, S. M. and Cahn, J. W. 1979. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metall.* 27:1085–1095.

Anderson, M. P., Srolovitz, D. J., Grest, G. S., and Sahni, P. S. 1984. Computer simulation of grain growth—I. Kinetics. *Acta Metall*. 32:783–791.

Ardell, A. J. 1968. An application of the theory of particle coarsening: the γ' precipitates in Ni-Al alloys. *Acta Metall*. 16:511–516.

Bateman, C. A. and Notis, M. R. 1992. Coherency effect during precipitate coarsening in partially stabilized zirconia. *Acta Metall. Mater*. 40:2413.

Bayliss, A., Kuske, R., and Matkowsky, B. J. 1990. A two dimensional adaptive pseudo-spectral method. *J. Comput. Phys*. 91:174–196.

Binder, K. (ed). 1992. Monte Carlo Simulation in Condensed Matter Physics. Springer-Verlag, Berlin.

Bortz, A. B., Kalos, M. H., Lebowitz, J. L., and Zendejas, M. A. 1974. Time evolution of a quenched binary alloy: Computer simulation of a two-dimensional model system. *Phys. Rev. B* 10:535–541.

Bösch. A., Müller-Krumbhaar, H., and Schochet, O. 1995. Phase-field models for moving boundary problems: controlling metastability and anisotrophy. *Z. Phys. B* 97:367–377.

Braun, R. J., Cahn, J. W., Hagedorn, J., McFadden, G. B., and Wheeler, A. A. 1996. Anisotropic interfaces and ordering in fcc alloys: A multiple-order-parameter continuum theory. *In* Mathematics of Microstructure Evolution (L. Q. Chen, B. Fultz, J. W. Cahn, J. R. Manning, J. E. Morral, and J. A. Simmons, eds.) pp. 225–244. The Minerals, Metals & Materials Society, Warrendale, Pa.

Broughton, J. Q. and Rudd, R. E. 1998. Seamless integration of molecular dynamics and finite elements. *Bull. Am. Phys. Soc.* 43:532.

Caginalp, G. 1986. An analysis of a phase field model of a free boundary. *Arch. Ration. Mech. Anal*. 92:205–245.

Cahn, J. W. 1961. On spinodal decomposition. *Acta Metall*. 9:795.

Cahn, J. W. 1962. On spinodal decomposition in cubic crystals. *Acta Metall.* 10:179.

Cahn, J. W. and Hilliard, J. E. 1958. Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.* 28:258–267.

Cannon, C. D. 1995. Computer Simulation of Sintering Kinetics. Bachelor Thesis, The Pennsylvania State University.

Canuto, C., Hussaini, M. Y., Quarteroni, A., and Zang, T. A. 1987. Spectral Methods In Fluid Dynamics. Springer-Verlag, New York.

Chen, I.-W. and Xue, L. A. 1990. Development of Superplastic Structural Ceramics. *J. Am. Ceram. Soc.* 73:2585–2609.

Chen, L. Q. 1993a. A computer simulation technique for spinodal decomposition and ordering in ternary systems, *Scrip. Metall. Mater.* 29:683–688.

Chen, L. Q. 1993b. Non-equilibrium pattern formation involving both conserved and nonconserved order parameters. *Mod. Phys. Lett. B* 7:1857–1881.

Chen, L. Q. 1994. Computer simulation of spinodal decomposition in ternary systems. *Acta Metall. Mater.* 42:3503–3513.

Chen, L. Q. and Fan, D. N. 1996. Computer Simulation of Coupled Grain-Growth in Two-Phase Systems, *J. Am. Ceram. Soc.* 79:1163–1168.

Chen, L. Q. and Khachaturyan, A. G. 1991a. Computer simulation of structural transformations during precipitation of an ordered intermetallic phase. *Acta Metall. Mater.* 39:2533–2551.

Chen, L. Q. and Khachaturyan, A. G. 1991b. On formation of virtual ordered phase along a phase decomposition path. *Phys. Rev. B* 44:4681–4684.

Chen, L. Q. and Khachaturyan, A. G. 1992. Kinetics of virtual phase formation during precipitation of ordered intermetallics. *Phys. Rev. B* 46:5899–5905.

Chen, L. Q. and Khachaturyan, A. G. 1993. Dynamics of simultaneous ordering and phase separation and effect of long-range coulomb interactions. *Phys. Rev. Lett.* 70:1477–1480.

Chen, L. Q. and Shen, J. 1998. Applications of semi-implicit Fourier-spectral method to phase field equations. *Comp. Phys. Commun.* 108:147–158.

Chen, L. Q. and Simmons, J. A. 1994. Master equation approach to inhomogeneous ordering and clustering in alloys—direct exchange mechanism and single-siccccccccccte approximation. *Acta Metall. Mater.* 42:2943–2954.

Chen, L. Q. and Wang, Y. 1996. The continuum field approach to modeling microstructural evolution. *JOM* 48:13–18.

Chen, L. Q., Wang, Y., and Khachaturyan, A. G. 1991. Transformation-induced elastic strain effect on precipitation kinetics of ordered intermetallics. *Philos. Mag. Lett.*, 64:241–251.

Chen, L. Q., Wang, Y., and Khachaturyan, A. G. 1992. Kinetics of tweed and twin formation in an ordering transition. *Philos. Mag. Lett.*, 65:15–23.

Chen, L. Q. and Yang, W. 1994. Computer simulation of the dynamics of a quenched system with large number of nonconserved order parameters—the grain growth kinetics. *Phys. Rev. B* 50:15752–15756.

Cleri, F., Phillpot, S. R., Wolf, D., and Yip, S. 1998. Atomistic simulations of materials fracture and the link between atomic and continuum length scales. *J. Am. Ceram. Soc.* 81:501–516.

Cowley, S. A., Hetherington, M. G., Jakubovics, J. P., and Smith, G. D. W. 1986. An FIM-atom probe and TEM study of ALNICO permanent magnetic material. *J. Phys.* 47: C7-211–216.

Dobretsov, V. Y., Martin, G., Soisson, F., and Vaks, V. G. 1995. Effects of the interaction between order parameter and concentration on the kinetics of antiphase boundary motion. *Europhys. Lett.* 31:417–422.

Dobretsov, V. Yu and Vaks, V. G. 1996. Kinetic features of phase separation under alloy ordering, *Phys. Rev. B* 54:3227.

Eguchi, T., Oki, K. and Matsumura, S. 1984. Kinetics of ordering with phase separation. *In* MRS Symposia Proceedings, Vol. 21, pp. 589–594. North Holland, New York.

Elder, K. 1993. Langevin simulations of nonequilibrium phenomena. *Comp. Phys.* 7: 27–33.

Elder, K. R., Drolet, F., Kosterlitz, J. M., and Grant, M. 1994. Stochastic eutectic growth. *Phys. Rev. Lett.* 72 677–680.

Eyre, D. 1993. Systems of Cahn-Hilliard equations. *SIAM J. Appl. Math.* 53:1686–1712.

Fan, D. N. and Chen, L. Q. 1995a. Computer simulation of twin formation in Y_2O_3-ZrO_2. *J. Am. Ceram. Soc.* 78:1680–1686.

Fan, D. N. and Chen, L. Q. 1995b. Possibility of spinodal decomposition in yttria-partially stabilized zirconia (ZrO_2-Y_2O_3) system—A theoretical investigation. *J. Am. Ceram. Soc.* 78: 769–773.

Fan, D. N. and Chen, L. Q. 1997a. Computer simulation of grain growth using a continuum field model. *Acta Mater.* 45:611–622.

Fan, D. N. and Chen, L. Q. 1997b. Computer simulation of grain growth using a diffuse-interface model: Local kinetics and topology. *Acta Mater.* 45:1115–1126.

Fan, D. N. and Chen, L. Q. 1997c. Diffusion-controlled microstructural evolution in volume-conserved two-phase polycrystals. *Acta Mater.* 45:3297–3310.

Fan, D. N. and Chen, L. Q. 1997d. Topological evolution of coupled grain growth in 2-D volume-conserved two-phase materials. *Acta Mater.* 45:4145–4154.

Fan, D. N. and Chen, L. Q. 1997e. Kinetics of simultaneous grain growth and ostwald ripening in alumina-zirconia two-phase composites—a computer simulation. *J. Am. Ceram. Soc.* 89:1773–1780.

Fratzl, P. and Penrose, O. 1995. Ising model for phase separation in alloys with anisotropic elastic interaction—I. Theory. *Acta Metall. Mater.* 43:2921–2930.

French, J. D., Harwmer, M. P., Chan, H. M., and Miller G. A. 1990. Coarsening-resistant dual-phase interpenetrating microstructures. *J. Am. Ceram. Soc.* 73:2508.

Gao, J. and Thompson, R. G. 1996. Real time-temperature models for Monte Carlo simulations of normal grain growth. *Acta Mater.* 44:4565–4570.

Geng, C. W. and Chen, L. Q. 1994. Computer simulation of vacancy segregation during spinodal decomposition and Ostwald ripening. *Scr. Metall. Mater.* 31:1507–1512.

Geng, C. W. and Chen, L. Q. 1996. Kinetics of spinodal decomposition and pattern formation near a surface. *Surf. Sci.* 355:229–240.

Gouma, P. I., Mills, M. J., and Kohli, A. 1997. Internal report of CISM. The Ohio State University, Columbus, Ohio.

Gumbsch, P. 1996. Atomistic modeling of failure mechanisms. *In* Computer Simulation in Materials Science—Nano/Meso/Macroscopic Space and Time Scales (H. O. Kirschner, K. P. Kubin, and V. Pontikis, eds.) pp. 227–244. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Gunton, J. D., Miguel, M. S., and Sahni, P. S. 1983. The dynamics of first-order phase transitions. *In* Phase Transitions and Critical Phenomena, Vol. 8 (C. Domb and J. L. Lebowitz, eds.) pp. 267–466. Academic Press, New York.

Harmer, M. P., Chan, H. M., and Miller, G. A. 1992. Unique opportunities for microstructural engineering with duplex and laminar ceramic composites. *J. Am. Ceram. Soc.* 75:1715–1728.

Hesselbarth, H. W. and Göbel, I. R. 1991. Simulation of recrystallization by cellular automata. *Acta Metall. Mater.* 39:2135–2143.

Hohenberg, P. C. and Halperin, B. I. 1977. Theory of dynamic critical phenomena. *Rev. Mod. Phys.* 49:435–479.

Holm, E. A., Rollett, A. D., and Srolovitz, D. J. 1996. Mesoscopic simulations of recrystallization. *In* Computer Simulation in Materials Science, Nano/Meso/Macroscopic Space and Time Scales, NATO ASI Series E: Applied Sciences, Vol. 308 (H. O. Kirchner, L. P. Kubin, and V. Pontikis, eds.) pp. 273–389. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Hu, H. L. and Chen, L. Q. 1997. Computer simulation of 90° ferroelectric domain formatio in two dimensions. *Mater. Sci. Eng. A* 238:182–189.

Izyumov, Y. A. and Syromoyatnikov, V. N. 1990. Phase Transitions and Crystal Symmetry. Kluwer Academic Publishers, Boston.

Johnson, W. C., Abinandanan, T. A., and Voorhees, P. W. 1990. The coarsening kinetics of two misfitting particles in an anisotropic crystal. *Acta Metall. Mater.* 38:1349–1367.

Johnson, W. C. and Voorhees, P. W. 1992. Elastically-induced precipitate shape transitions in coherent solids. *Solid State Phenom.* 23 and 24:87–103.

Jou, H. J., Leo, P. H., and Lowengrub, J. S. 1994. Numerical calculations of precipitate shape evolution in elastic media. *In* Proceedings of an International Conference on Solid to Solid Phase Transformations (Johnson, W. C., Howe, J. M., Laughlin, D. E., and Soffa, W. A., eds.) pp. 635–640. The Minerals, Metals & Materials Society.

Jou, H. J., Leo, P. H., and Lowengrub, J. S. 1997. Microstructural evolution in inhomogeneous elastic solids. *J. Comput. Phys.* 131:109–148.

Karma, A. 1994. Phase-field model of eutectic growth. *Phys. Rev. E* 49:2245–2250.

Karma, A. and Rappel, W.-J. 1996. Phase-field modeling of crystal growth: New asymptotics and computational limits. *In* Mathematics of Microstructures, Joint SIAM-TMS proceedings (L. Q. Chen, B. Fultz, J. W. Cahn, J. E. Morral, J. A. Simmons, and J. A. Manning, eds.) pp. 635–640.

Karma, A. and Rappel, W.-J. 1998. Quantitative phase-field modeling of dendritic growth in two and three dimensions. *Phys. Rev. E* 57:4323–4349.

Khachaturyan, A. G. 1967. Some questions concerning the theory of phase transformations in solids. *Sov. Phys. Solid State,* 8:2163–2168.

Khachaturyan, A. G. 1968. Microscopic theory of diffusion in crystalline solid solutions and the time evolution of the diffuse scatering of x-ray and thermal neutrons. *Soviet Phys. Solid State.* 9:2040–2046.

Khachaturyan, A. G. 1978. Ordering in substitutional and interstitial solid solutions. *In* Progress in Materials Science, Vol. 22 (B. Chalmers, J. W. Christian, and T. B. Massalsky, eds.) pp. 1–150. Pergamon Press, Oxford.

Khachaturyan, A. G. 1983. Theory of Structural Transformations in Solid. John Wiley & Sons, New York.

Khachaturyan, A. G., Semenovskaya, S., and Tsakalakos, T. 1995. Elastic strain energy of inhomogeneous solids. *Phys. Rev. B* 52:1–11.

Khachaturyan, A. G. and Shatalov, G. A. 1969. Elastic-interaction potential of defects in a crystal. *Sov. Phys. Solid State* 11:118–123.

Kiruchi, R. 1951. *Phys. Rev.* 81:988–1003.

Kirchner, H. O., Kubin, L. P., and Pontikis, V. (eds.), 1996. Computer Simulation in Materials Science, Nano/Meso/Macroscopic Space and Time Scales. Kluwer Academic Publishers Dordrecht, The Netherlands.

Kobayashi, R. 1993. Modeling and numerical simulations of dendritic crystal growth. *Physica D* 63:410–423.

Koyama, T., Miyazaki, T., and Mebed, A. E. 1995. Computer simulation of phase decomposition in real alloy systems based on the Khachaturyan diffusion equation. *Metall. Mater. Trans. A* 26:2617–2623.

Lai, Z. W. 1990. Theory of ordering dynamics for Cu_3Au. *Phys. Rev. B* 41:9239–9256.

Landau, L. D. 1937a. *JETP* 7:19.

Landau, L. D. 1937b. *JETP* 7:627.

Lange, F. F. and Hirlinger, M. M. 1987. Grain growth in two-phase ceramics: Al_2O_3 inclusions in ZrO_2. J. Am. Ceram. Soc. 70:827–830.

Langer, J. S. 1986. Models of pattern formation in first-order phase transitions. *In* Directions in Condensed Matter Physics (G. Grinstein and G. Mazenko, eds.) pp. 165–186. World Scientific, Singapore.

Lebowitz, J. L., Marro, J., and Kalos, M. H. 1982. Dynamical scaling of structure function in quenched binary alloys. *Acta Metall.* 30:297–310.

Lee, J. K. 1996a. A study on coherency strain and precipitate morphology via a discrete atom method. *Metall. Mater. Trans. A* 27:1449–1459.

Lee, J. K. 1996b. Effects of applied stress on coherent precipitates via a discrete atom method. *Metals Mater.* 2:183–193.

Leo, P. H., Mullins, W. W., Sekerka, R. F., and Vinals, J. 1990. Effect of elasticity on late stage coarsening. *Acta Metall. Mater.* 38:1573–1580.

Leo, P. H. and Sekerka, R. F. 1989. The effect of elastic fields of the marphological stability of a precipitate growth from solid solution. *Acta Metall.* 37:3139–3149.

Lepinoux, J. 1996. Application of cellular automata in materials science. *In* Computer Simulation in Materials Science, Nano/Meso/Macroscopic Space and Time Scales, NATO ASI Series E: Applied Sciences, Vol. 308 (H. O. Kirchner, L. P. Kubin and V. Pontikis, eds.) pp. 247–258. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Lepinoux, J. and Kubin, L. P. 1987. The dynamic organization of dislocation structures: a simulation. *Scr. Metall.* 21:833–838.

Li, D. Y. and Chen, L. Q. 1997a. Computer simulation of morphological evolution and rafting of particles in Ni-based superalloys under applied stress. *Scr. Mater.* 37:1271–1277.

Li, D. Y. and Chen, L. Q. 1997b. Shape of a rhombohedral coherent $Ti_{11}Ni_{14}$ precipitate in a cubic matrix and its growth and dissolution during constraint aging. *Acta Mater.* 45:2435–2442.

Li, D. Y. and Chen, L. Q. 1998a. Morphological evolution of coherent multi-variant $Ti_{11}Ni_{14}$ precipitates in Ti-Ni alloys under an applied stress—a computer simulation study. *Acta Mater.* 46:639–649.

Li, D. Y. and Chen, L. Q. 1998b. Computer simulation of stress-oriented nucleation and growth of precipitates in Al-Cu alloys. *Acta Mater.* 46:2573–2585.

Lifshitz, E. M. 1941. *JETP* 11:269.

Lifshitz, E. M. 1944. *JETP* 14:353.

Liu, Y., Baudin, T., and Penelle, R. 1996. Simulation of normal grain growth by cellular automata, *Scr. Mater.* 34:1679–1683.

Liu, Y., Ciobanu. C., Wang, Y., and Patton, B. 1997. Computer simulation of microstructural evolution and electrical conductivity calculation during sintering of electroceramics. Presentation in the 99th annual meeting of the American Ceramic Society, Cincinnati, Ohio.

Mareschal, M. and Lemarchand, A. 1996. Cellular automata and mesoscale simulations. *In* Computer Simulation in Materials Science, Nano/Meso/Macroscopic Space and Time Scales, NATO ASI Series E: Applied Sciences, Vol. 308 (H. O. Kirchner, L. P. Kubin, and V. Pontikis, eds.) pp. 85–102. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Martin, G. 1994. Relaxation rate of conserved and nonconserved order parameters in replacive transformations. *Phys. Rev. B* 50:12362–12366.

Miyazaki, T., Imamura, M., and Kokzaki, T. 1982. The formation of "γ' precipitate doublets" in Ni-Al alloys and their energetic stability. *Mater. Sci. Eng*. 54:9–15.

Müller-Krumbhaar, H. 1996. Length and time scales in materials science: Interfacial pattern formation. *In* Computer Simulation in Materials Science, Nano/Meso/Macroscopic Space and Time Scales, NATO ASI Series E: Applied Sciences, Vol. 308 (H. O. Kirchner, L. P. Kubin, and V. Pontikis, eds.) pp. 43–59. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Nambu, S. and Sagala, D. A. 1994. Domain formation and elastic long-range interaction in ferroelectric perovskites. *Phys. Rev. B* 50:5838–5847.

Nishimori, H. and Onuki, A. 1990. Pattern formation in phase-separating alloys with cubic symmetry. *Phys. Rev. B* 42:980–983.

Okamoto, H. 1993. Phase diagram update: Ni-Al. J. *Phase Equilibria*. 14:257–259.

Onuki, A. 1989a. Ginzburg-Landau approach to elastic effects in the phase separation of solids. *J. Phy. Soc. Jpn*. 58:3065–3068.

Onuki, A. 1989b. Long-range interaction through elastic fields in phase-separating solids. *J. Phy. Soc. Jpn*. 58:3069–3072.

Onuki, A. and Nishimori, H. 1991. On Eshelby's elastic interaction in two-phase solids. *J. Phy. Soc. Jpn*. 60:1–4.

Oono, Y. and Puri, S. 1987. Computationally efficient modeling of ordering of quenched phases. *Phys. Rev. Lett*. 58:836–839.

Ornstein, L. S. and Zernicke, F. 1918. *Phys. Z*. 19:134.

Ozawa, S., Sasajima, Y., and Heermann, D. W. 1996. Monte Carlo Simulations of film growth. *Thin Solid Films* 272:172–183.

Poduri, R. and Chen, L. Q. 1997. Computer simulation of the kinetics of order-disorder and phase separation during precipitation of δ' (Al_3Li) in Al-Li alloys. *Acta Mater*. 45:245–256.

Poduri, R. and Chen, L. Q. 1998. Computer simulation of morphological evolution and coarsening kinetics of δ' (AlLi) precipitates in Al-Li alloys. *Acta Mater*. 46:3915–3928.

Pottenbohm, H., Neitze, G., and Nembach, E. 1983. Elastic properties (the stiffness constants, the shear modulus a nd the dislocation line energy and tension) of Ni-Al solid solutions and of the Nimonic alloy PE16. *Mater. Sci. Eng*. 60:189–194.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. 1992. Numerical recipes in Fortran 77, second edition, Cambridge University Press.

Provatas, N., Goldenfeld, N., and Dantzig, J. 1998, Efficient computation of dendritic microstructures using adaptive mesh refinement. *Phys. Rev. Lett*. 80:3308–3311.

Rappaz, M. and Gandin, C.-A. 1993. Probabilistic modelling of microstructure formation in solidification process. *Acta Metall. Mater*, 41:345–360.

Rogers, T. M., Elder, K. R., and Desai, R. C. 1988. Numerical study of the late stages of spinodal decomposition. *Phys. Rev. B* 37:9638–9649.

Rowlinson, J. S. 1979. Translation of J. D. van der Waals' thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. *J. Stat. Phys*. 20:197. [Originally published in Dutch *Verhandel. Konink. Akad. Weten. Amsterdam* Vol. 1, No. 8, 1893.]

Saito, Y. 1997. The Monte Carlo simulation of microstructural evolution in metals. *Mater. Sci. Eng*. A223:114–124.

Sanchez, J. M., Ducastelle, F., and Gratias, D. 1984. *Physica* 128A:334–350.

Semenovskaya, S. and Khachaturyan, A. G. 1991. Kinetics of strain-reduced transformation in $YBa_2Cu_3O_{7-}$. *Phys. Rev. Lett*. 67:2223–2226.

Semenovskaya, S. and Khachaturyan, A. G. 1997. Coherent structural transformations in random crystalline systems. *Acta Mater*. 45:4367–4384.

Semenovskaya, S. and Khachaturyan, A. G. 1998. Development of ferroelectric mixed state in random field of static defects. *J. App. Phys*. 83:5125–5136.

Semenovskaya, S., Zhu, Y., Suenaga, M., and Khachaturyan, A. G. 1993. Twin and tweed microstructures in $YBa_2Cu_3O_{7-\delta}$. doped by trivalent cations. *Phy. Rev. B* 47:12182–12189.

Shelton, R. K. and Dunand, D. C. 1996. Computer modeling of particle pushing and clustering during matrix crystallization. *Acta Mater*. 44:4571–4585.

Shen, J. 1994. Efficient spectral-Galerkin method I. Direct solvers for second- and fourth-order equations by using Legendre polynomials. *SIAM J. Sci. Comput*. 15:1489–1505.

Shen, J. 1995. Efficient spectral-Galerkin method II. Direct solvers for second- and fourth-order equations by using Chebyshev polynomials. *SIAM J. Sci. Comput*. 16:74–87.

Srolovitz, D. J., Anderson, M. P., Grest, G. S., and Sahni, P. S. 1983. Grain growth in two-dimensions. *Scr. Metall*. 17:241–246.

Srolovitz, D. J., Anderson, M. P., Sahni P. S., and Grest, G. S. 1984. Computer simulation of grain growth—II. Grain size distribution, topology, and local dynamics. *Acta Metall*. 32: 793–802.

Steinbach, I. and Schmitz, G. J. 1998. Direct numerical simulation of solidification structure using the phase field method. *In* Modeling of Casting, Welding and Advanced Solidification Processes VIII (B. G. Thomas and C. Beckermann, eds.) pp. 521–532, The Minerals, Metals & Materials Society, Warrendale, Pa.

Su, C. H. and Voorhees, P. W. 1996a. The dynamics of precipitate evolution in elastically stressed solids—I. Inverse coarsening. *Acta Metall*. 44:1987–1999.

Su, C. H. and Voorhees, P. W. 1996b. The dynamics of precipitate evolution in elastically stressed solids—II. Particle alignment. *Acta Metall*. 44:2001–2016.

Sundman, B., Jansson, B., and Anderson, J.-O. 1985. CALPHAD 9:153.

Sutton, A. P. 1996. Simulation of interfaces at the atomic scale. *In* Computer Simulation in Materials Science—Nano/Meso/Macroscopic Space and Time Scales (H. O. Kirchner, K. P. Kubin, and V. Pontikis, eds.) pp. 163–187, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Syutkina, V. I. and Jakovleva, E. S. 1967. *Phys. Status Solids* 21:465.

Tadmore, E. B., Ortiz, M., and Philips, R. 1996. Quasicontinuum analysis of defects in solids. *Philos. Mag. A* 73:1529–1523.

Tien, J. K. and Coply, S. M. 1971. *Metall. Trans*. 2:215; 2:543.

Tikare, V. and Cawley, J. D. 1998. Numerical Simulation of Grain Growth in Liquid Phase Sintered Materials—II. Study of Isotropic Grain Growth. *Acta Mater*. 46:1343.

Thompson, M. E., Su, C. S., and Voorhees, P. W. 1994. The equilibrium shape of a misfitting precipitate. *Acta Metall. Mater*. 42:2107–2122.

Van Baal, C. M. 1985. Kinetics of crystalline configurations. III. An alloy that is inhomogeneous in one crystal direction. *Physica A* 129:601–625.

Vashishta, R., Kalia, R. K., Nakano, A., Li, W., and Ebbsjö, I. 1997. Molecular dynamics methods and large-scale simulations of amorphous materials. *In* Amorphous Insulators and Semiconductors (M. F. Thorpe and M. I. Mitkova, eds.) pp. 151–213. Kluwer Academic Publishers, Dordrecht, The Netherlands.

von Neumann, J. 1966. Theory of Self-Reproducing Automata (edited and completed by A. W. Burks). University of Illinois, Urban, Ill.

Wang, Y., Chen, L.-Q., and Khachaturyan, A. G. 1991a. Shape evolution of a precipitate during strain-induced coarsening—a computer simulation. *Script. Metall. Mater.* 25:1387–1392.

Wang, Y., Chen, L.-Q., and Khachaturyan, A. G. 1991b. Strain-induced modulated structures in two-phase cubic alloys. *Script. Metall. Mater.* 25:1969–1974.

Wang, Y., Chen, L.-Q., and Khachaturyan, A. G. 1993. Kinetics of strain-induced morphological transformation in cubic alloys with a miscibility gap, *Acta Metall. Mater.* 41:279–296.

Wang, Y., Chen, L.-Q., and Khachaturyan, A. G. 1994. Computer simulation of microstructure evolution in coherent solids. *In* Solid → Solid phase Transformations (W. C. Johnson, J. M. Howe, D. E. Laughlin and W. A. Soffa, eds.) The Minerals, Metals & Materials Society, pp. 245–265, Warrendale, Pa.

Wang, Y., Chen, L.-Q., and Khachaturyan, A. G. 1996. Modeling of dynamical evolution of micro/mesoscopic morphological patterns in coherent phase transformations. *In* Computer Simulation in Materials Science—Nano/Meso/Macroscopic Space and Time Scales (H. O. Kirchner, K. P. Kubin and V. Pontikis, eds.) pp. 325–371, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Wang, Y. and Khachaturyan, A. G. 1995a. Effect of antiphase domains on shape and spatial arrangement of coherent ordered intermetallics. *Scripta Metall. Mater.* 31:1425–1430.

Wang, Y. and Khachaturyan, A. G. 1995b. Microstructural evolution during precipitation of ordered intermetallics in multi-particle coherent systems. *Philos. Mag. A* 72:1161–1171.

Wang, Y. and Khachaturyan, A. G. 1997. Three-dimensional field model and computer modeling of martensitic transformations. *Acta Mater.* 45:759–773.

Wang, Y., Banerjee, D., Su, C. C., and Khachaturyan, A. G. 1998. Field kinetic model and computer simulation of precipitation of $L1_2$ ordered intermetallics from fcc solid solution. *Acta Mater* 46:2983–3001.

Wang, Y., Wang, H. Y., Chen, L. Q., and Khachaturyan, A. G. 1995. Microstructural development of coherent tetragonal precipitates in Mg-partially stabilized zirconia: A computer simulation. *J. Am. Ceram. Soc.* 78:657–661.

Warren, J. A. and Boettinger, W. J. 1995. Prediction of dendritic growth and microsegregation patterns in a binary alloy using the phase-field method. *Acta Metall. Mater.* 43:689–703.

Wheeler, A. A., Boettinger, W. J., and McFadden, G. B. 1992. Phase-field model for isothermal phase transitions in binary alloys. *Phys. Rev. A* 45:7424–7439.

Wheeler, A. A., Boettinger, W. J., and McFadden, G. B. 1993. Phase-field model of solute trapping during solidification. *Phys. Rev. A* 47:1893.

Wolfram, S. 1984. Cellular automata as models of complexity. *Nature* 311:419–424.

Wolfram, S. 1986. Theory and Applications of Cellular Automata, World Scientific, Singapore.

Yang, W. and Chen, L.-Q. 1995. Computer simulation of the kinetics of 180° ferroelectric domain formation. *J. Am. Ceram. Soc.* 78:2554–2556.

Young, D. A. and Corey, E. M. 1990. Lattice model of biological growth. *Phys. Rev. A* 41, 7024–7032.

Zeng, P., Zajac, S., Clapp, P. C., and Rifkin, J. A. 1998. Nanoparticle sintering simulations. *Mater. Sci. Eng. A* 252:301–306.

Zhu, H. and Averback, R. S. 1996. Sintering processes of metallic nano-particles: A study by molecular-dynamics simulations. *In* Sintering Technology (R. M. German, G. L. Messing, and R. G. Cornwall, eds.) pp. 85–92. Marcel Dekker, New York.

KEY REFERENCES

Wang et al., 1996; Chen and Wang, 1996. See above.

These two references provide overall reviews of the practical applications of both the microscopic and continuum field method in the context of microstructural evolution in different materials systems. The generic relationship between the microscopic and continuum field methods are also discussed.

Gunton et al., 1983. See above.

Provides a detailed theoretical picture of the continuum field method and various forms of the field kinetic equation. Also contains detailed discussion of applications in critical phenomena.

INTERNET RESOURCES

http://www.procast.com
http://www.abaqus.com
http://www.deform.com
http://marc.com

Y. WANG
Ohio State University
Columbus, Ohio

L.-Q. CHEN
Pennsylvania State University
University Park, Pennsylvania

BONDING IN METALS

INTRODUCTION

The conditions controlling alloy phase formation have been of concern for a long time, even in the period prior to the advent of the first computer-based electronic structure calculations in the 1950s. In those days, the role of such factors as atomic size and valence were considered, and significant insight was gained, associated with names such as Hume-Rothery. In recent times, electronic structure calculations have yielded the total energies of systems, and hence the heats of formation of competing real and hypothetical phases. Doing this accurately has become of increasing importance as the groups capable of thermodynamic measurements have dwindled in number.

This unit will concentrate on the alloys of the transition metals and will consider both the old and the new (for some early views of non-transition metal alloying, see Hume-Rothery, 1936). In the first part, we will inspect some of the ideas concerning metallic bonding that derive from the early days. Many of these ideas are still applicable and offer insights into alloy phase behavior that are computationally cheaper, and in some cases rather different, from those obtainable from detailed computation. In the second part, we will turn to the present-day total-energy calculations. Comparisons between theoretical and experimental heats of formation will not be made, because the spreads between experimental heats and their computed

counterparts are not altogether satisfying for metallic systems. Instead, the factors controlling the accuracy of the computational methods and the precision with which they are carried out will be inspected. All too often, consideration of computational costs leads to results that are less precisely defined than is needed. The discussion in the two sections necessarily reflects the biases and preoccupations of the authors.

FACTORS AFFECTING PHASE FORMATION

Bonding-Antibonding and Friedel's *d*-Band Energetics

If two electronic states in an atom, molecule, or solid are allowed by symmetry to mix, they will. This hybridization will lead to a "bonding" level lying lower in energy than the original unhybridized pair plus an "antibonding" level lying above. If only the bonding level is occupied, the system gains by the energy lowering. If, on the other hand, both are occupied, there is no gain to be had since the center of gravity of the energy of the pair is the same before and after hybridization. Friedel (1969) applied this observation to the d-band bonding of the transition metals by considering the ten d-electron states of a transition metal atom broadening into a d band due to interactions with neighboring atoms in an elemental metal. He presumed that the result was a rectangular density of states with its center of gravity at the original atomic level. With one d state per atom, the lowest tenth of the density of states would be occupied, and adding another d electron would necessarily result in less gain in band-broadening energy since it would have to lie higher in the density of states. Once the d bands are half-filled, there are only antibonding levels to be occupied and the band-broadening energy must drop, that energy being

$$E_{d\,\mathrm{band}} \propto n(10-n)W \tag{1}$$

where n is the d-electron count and W the bandwidth. This expression does quite well—i.e., the maximum cohesion (and the maximum melting temperatures) does occur in the middle of the transition metal rows. This trend has consequences for alloying because the heat of formation of an A_xB_{1-x} alloy involves the competition between the energy of the alloy and those of the elemental solids, i.e.,

$$\Delta H = E[A_xB_{1-x}] - xE[A] - (1-x)E[B] \tag{2}$$

If either A or B resides in the middle of a transition metal row, those atoms have the most to lose, and hence the alloy the least to gain, upon alloy formation.

Crystal Structures of the Transition Metals

Sorting out the structures of either the transition metals or the main group elements is best done by recourse to electronic structure calculations. In a classic paper, Pettifor (1970) did tight-binding calculations that correctly yielded the hcp→ bcc→hcp→fcc sequence that occurs as one traverses the transition metal rows, if one neglects the magnetism of the 3d elements (where hcp signifies hexagonal close-packed, bcc is body-centered cubic, and fcc is face-centered cubic). The situation is perhaps better described by Figure 1.

The Topologically Close-Packed Phases

While it is common to think of the metallic systems as forming in the fcc, bcc, and hcp structures, many other crystalline structures occur. An important class is the abovementioned Frank-Kasper phases. Consider packing atoms of equal size starting with a pair lying on a bond line and then adding other atoms at equal bond lengths lying on its equator. Adding two such atoms leads to a tetrahedron formed by the four atoms; adding others leads to additional tetrahedra sharing the bond line as a common edge. The number of regular tetrahedra that can be packed with a common edge (q_{tetra}) is nonintegral.

$$q_{\mathrm{tetra}} = \frac{2\pi}{\cos^{-1}(1/3)} = 5.104299\ldots \tag{3}$$

Thus, geometric factors frustrate the topological close packing associated with having an integral number of such neighbors common to a bond line. Frank and Kasper (1958, 1959) introduced a construct for crystalline structures that adjusts to this frustration. The construct involves four species of atomic environments and, having some relation to the above ideal tetrahedral packing, the resulting structures are known as tcp phases. The first environment is the 12-fold icosahedral environment where the atom at the center shares $q = 5$ common nearest

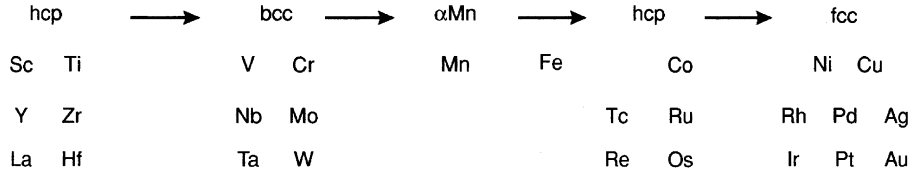


Figure 1. Crystalline structures of the transition metals as a function of d-band filling. αMn has a structure closely related to the topologically close-packed (tcp)– or Frank-Kasper (1958, 1959)–type phases and forms because of its magnetism; otherwise it would have the hcp structure. Similarly, Co would be fcc if it were not for its magnetism. Ferromagnetic iron is bcc and as a nonmagnetic metal would be hcp.

neighbors with each of the 12 neighbors. The other three environments are 14-, 15-, and 16-fold coordinated, each with twelve $q = 5$ neighbor bond lines plus two, three, and four neighbors, respectively, sharing $q = 6$ common nearest neighbors. Nelson (1983) has pointed out that the $q = 6$ bond lines may be viewed as $-72°$ rotations or disclinations leading to an average common nearest neighbor count that is close in value to q_{tetra}. Hence, the introduction of disclinations has caused a relaxation of the geometric frustration leading to a class of tcp phases that have, on average, tetrahedral packing. This involves nonplanar arrays of atomic sites of differing size. The nonplanarity encourages the systems to be nonductile. If compound formation is not too strong, transition metal alloys tend to form in structures determined by their average d-electron count—i.e., the same sequence, hcp → bcc → tcp → hcp → fcc, is found for the elements. For example, in the Mo-Ir system the four classes of structures, bcc → fcc, are traversed as a function of Ir concentration. One consequence of this is the occurrence of half-filled d-band alloys having high melting temperatures that are brittle tcp phases—i.e., high-temperature, mechanically unfavorable phases. The tcp phases that enter this sequence tend to have higher coordinated sites in the majority and include the Cr_3Si, A15 (the structure associated with the highest of the low-T_c superconductors), and the σCrFe structure. Other tcp phases where the larger atoms are in the minority tend to be line compounds and do not fall in the above sequence. These include the Laves phases and a number of structures in which the rare earth 3d hard magnets form. There are also closely related phases, such as αMn (Watson and Bennett, 1985), that involve other atomic environments in addition to the four Frank-Kasper building blocks. In fact, αMn itself may be viewed as an alloy, not between elements of differing atomic number, but instead between atoms of differing magnetic moment and hence of differing atomic size. There are also some hard magnet structures, closely related to the Laves phase, that involve building blocks other than the four of Frank and Kasper.

Bonding-Antibonding Effects

When considering stronger compound formation between transition metals, such as that involving one element at the beginning and another at the end of the transition metal rows, there are strong deviations from having the maximum heats of formation at 50:50 alloy composition. This skewing can be seen in the phase diagrams by drawing a line between the melting temperatures of the elemental metals and inspecting the deviations from that line for the compounds in between. This can be understood (Watson et al., 1989) in terms of the filling of "bonding" levels, while leaving the "antibonding" unfilled, but with unequal numbers of the two. Essential to this is that the centers of gravity of the d bands intrinsic to the two elements be separated. In general, the d levels of the elements to the left in the transition metal rows lie higher in energy, so that this condition is met. The consequences of this are seen in Figure 2, where the calculated total densities of states for observed phases of Pd_3Ti and PdTi are plotted

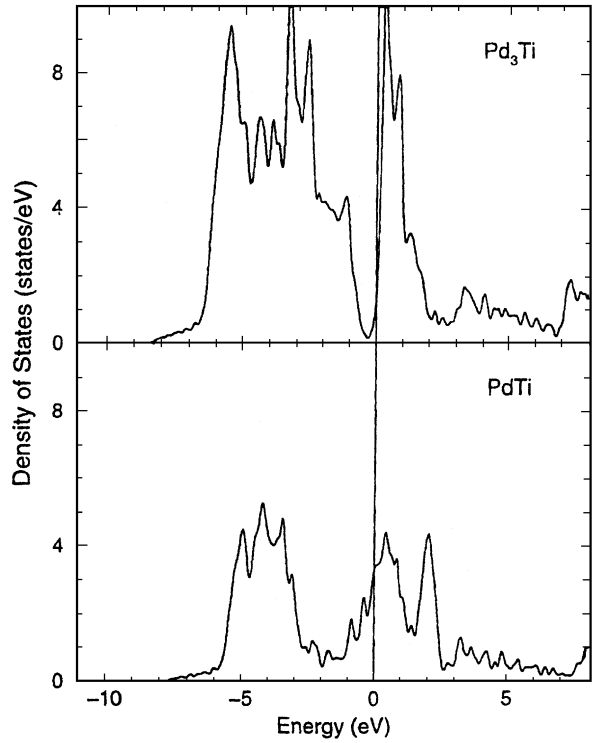


Figure 2. Calculated densities of states for Pd_3Ti (Cu_3Au structure) and PdTi (CsCl structure), which are the observed phases of the Pd-Ti system as obtained by the authors of this unit. The zeroes of the energy coordinate are the Fermi energies of the system. Some of the bumpiness of the plots is due to the finite sets of **k**-points (165 and 405 special points for the two structures), which suffice for determining the total energies of the systems.

(the densities of states in the upper plot are higher because there are twice the number of atoms in the crystal unit cell). The lower energy peak in each plot is Pd 4d in origin, while the higher energy peak is Ti 3d. However, there is substantial hybridization of the other atom's d and non-d character into the peaks. In the case of 1:1 PdTi, accommodation of the Pd-plus-Ti charge requires filling the low-lying "bonding" peak plus part of the antibonding. Going on to 3:1, all but a few states are accommodated by the bonding peak alone, resulting in a stabler alloy; the heat of formation (measured per atom or per mole-atom) is calculated to be some 50% greater than that for the 1:1. Hence, the phase diagram is significantly skewed. Because of the hybridization, very little site-to-site charge transfer is associated with this band filling—i.e., emptying the Ti peaks does not deplete the Ti sites of charge by much, if at all. Note that the alloy concentration at which the bonding peak is filled and the antibonding peak remains empty depends on the d counts of the two constituents. Thus, Ti alloyed with Rh would have this bonding extreme closer to 1:1, while for Ru, which averaged with Ti has half-filled d bands, the bonding peak filling occurs at 1:1.

When alloying a transition metal with a main group element such as Al there is only one set of d bands and the situation is different. Elemental Al has free electronlike bands, but this is not the case when alloyed with a transition metal where several factors are at play, including

hybridization with the transition metal and a separate narrowing of the *s*- and *p*-band character intrinsic to Al, since the Al atoms are more dilute and further apart. One consequence of this is that the *s*-like wave function character at the Al sites is concentrated in a peak deep in the conduction bands and contributes to the bonding to the extent that transition metal hybridization has contributed to this isolation. In contrast, the Al *p* character hybridizes with the transition metal wave function character throughout the bands, although it occurs most strongly deep in the transition metal *d* bands, thus again introducing a contribution to the bonding energy of the alloy. Also, if the transition metal *d* bands have a hollow in their density of states, hybridization, particularly with the Al *p*, tends to deepen and more sharply define the hollow. Occasionally this even leads to compounds that are insulators, such as Al_2Ru.

Size Effects

Another factor important to compound formation is the relative sizes of the constituent elements. Substitutional alloying is increasingly inhibited as the difference in size of the constituents is increased, since differences in size introduce local strains. Size is equally important to ordered phases. Consider the CsCl structure, common to many 1:1 intermetallic phases, which consists of a bcc lattice with one constituent occupying the cube centers and the other the cube corners. All of the fourteen first and second nearest neighbors are chemically important in the bcc lattice, which for the CsCl structure involves eight closest unlike and six slightly more removed like neighbors. Having eight closest unlike neighbors favors phase formation if the two constituents bond favorably. However, if the constituents differ markedly in size, and if the second-neighbor distance is favorable for bonding between like atoms of one constituent, it is unfavorable for the other. As a result, the large 4*d* and 5*d* members of the Sc and Ti columns do not form in the CsCl structure with other transition metals; instead the alloys are often found in structures such as the CrB structure, which has seven, rather than eight, unlike near neighbors and again six somewhat more distant like near neighbors among the large constituent atoms. In the case of the smaller atoms, the situation is different; there are but four like neighbors, two of which are at closer distances than the unlike neighbors, and thus this structure better accommodates atoms of measurably different size than does that of CsCl. Another bcc-based structure is the antiphase Heusler BiF_3 structure. Many ternaries such as N_2MA (where N and M are transition metals and A is Al, Ga, or In) form this structure, in which N atoms are placed at the cube corners, while the A and M atoms are placed in the cube centers but alternate along the *x*, *y*, and *z* directions. Important to the occurrence of these phases is that the A and the M not be too different in size.

Size is also important in the tcp phases. Those appearing in the abovementioned transition metal alloy sequences normally involve the 12-, 14-, and 15-fold Frank-Kasper sites. Their occurrence and some significant range of stoichiometries are favored (Watson and Bennett, 1984) if the larger atoms are in a range of 20% to 40% larger in elemental volume than the smaller constituent. In contrast, the Laves structures involve 12- and 16-fold sites, and these only occur if the large atoms have elemental volumes that are at least 30% greater than the small; in some cases the difference is a factor of 3. These size mismatches imply little or no deviation from strict stoichiometry in the Laves phases.

The role of size as a factor in controlling phase formation has been recognized for many years and still other examples could be mentioned. Size clearly has impact on the energetics of present-day electronic structure calculations, but its impact is often not transparent in the calculated results.

Charge Transfer and Electronegativities

In addition to size and valence (*d*-band filling) effects, a third factor long considered essential to compound formation is charge transfer. Although presumably heavily screened in metallic systems, charge transfer is normally attributed to the difference in electronegativities (i.e., the propensity of one element to attract valence electron charge from another element) of the constituent elements. Roughly speaking, there are as many electronegativity scales as there are workers who have considered the issue. One classic definition is that of Mulliken (Mulliken, 1935; Watson et al., 1983), where the electronegativity is taken to be the average ionization energy and electron affinity of the free ion. The question arises whether the behavior of the free atom is characteristic of the atom in a solid. One alternative choice for solids is to measure the heights of the elemental chemical potentials—i.e., their Fermi levels—with respect to each other inside a common crystalline environment or, barring this, to assign the elemental work functions to be their electronegativities (e.g., Gordy and Thomas, 1956). Miedema has employed such a metallurgist's scale, somewhat adjusted, in his effective Hamiltonian for the heats of formation of compounds (de Boer et al., 1988), where the square of the difference in electronegativities is the dominant binding term in the resultant heat. The results are at best semiquantitative and, as often as not, they get the direction of the abovementioned distortion of heats with concentration in the wrong direction. Nevertheless, these types of models represent a "best buy" when one compares quality of results to computational effort. Present-day electronic structure calculations provide much better estimates of heats of formation, but with vastly greater effort.

The difference in the electronegativities of two constituents in a compound may or may not provide some measure of the "transfer" of charge from one site to another. Different reasonably defined electronegativity scales will not even agree as to the direction of the transfer, but more serious are questions associated with defining such transfer in a crystal when given experimental data or the results of an electronic structure calculation. Some measurements, such as those of hyperfine effects or of chemically induced core or valence level shifts in photoemission, are complicated by more than one contribution to a shift. Even given a detailed charge density map from calculation

or from x-ray diffraction, there remain problems of attribution: one must first choose what part of the crystal is to be attributed to which atomic site and then the conclusions concerning any net transfer of charge will be a figment of that choice (e.g., assigning equal site volumes to atoms of patently different size does not make sense). Even if atomic cells can be satisfactorily chosen for some system, problems remain. As quantum chemists have known for years, the charge at a site is best understood as arising from three contributions: the charge intrinsic to the site, the tails of charge associated with neighboring atoms (the "medium" in which the atom finds itself), and an overlap charge (Watson et al., 1991). The overlap charge is often on the order of one electron's worth. Granted that the net charge transfer in a metallic system is small, the attribution of this overlap charge determines any statement concerning the charge at, and intrinsic to, a site. In the case of gold-alkali systems, computations indicate (Watson and Weinert, 1994) that while it is reasonable to assign the alkalis as being positively charged, their sites actually are negatively charged due to Au and overlap charge lapping into the site.

There are yet other complications to any consideration of "charge transfer." As noted earlier, hybridization in and out of d bands intrinsic to some particular transition metal involves the wave function character of other constituents of the alloy in question. This affects the on-site d-electron count. In alloys of Ni, Pd, or Pt with other transition metals or main group elements, for example, depletion by hybridization of the d count in their already filled d-band levels is accompanied by a filling of the unoccupied levels, resulting in the metals becoming noble metal-like with little net change in the charge intrinsic to the site. In the case of the noble metals, and particularly Cu and Au, whose d levels lie closer to the Fermi level, the filled d bands are actively involved in hybridization with other alloy constituents, resulting in a depletion of d-electron count that is screened by a charge of non-d character (Watson et al., 1971). One can argue as to how much this has to do with the simple view of "electronegativity." However, subtle shifts in charge occur on alloying, and depending on how one takes their measure, one can have them agree or disagree in their direction with some electronegativity scale.

Wave Function Character of Varying *l*

Usually, valence electrons of more than one l (orbital angular momentum quantum number) are involved in the chemistry of an atom's compound formation. For the transition and noble metals, it is the d and the non-d s-p electrons that are active, while for the light actinides there are f-band electrons at play as well. For the main group elements, there are s and p valence states and while the p may dominate in semiconductor compound formation, the s do play a role (Simons and Bloch, 1973; St. John and Bloch, 1974). In the case of the Au metal alloying cited above, Mössbauer isomer shifts, which sample changes in s-like contact charge density, detected charge flow onto Au sites, whereas photoelectron core-level shifts indicated deeper potentials and hence apparent charge transfer off Au sites. The Au $5d$ are more compact than the $6s$-$6p$, and any change in d count has approximately twice the effect on the potential as changes in the non-d counts (this is the same for the transition metals as well). As a result, the Au alloy photoemission shifts detected the d depletion due to hybridization into the d bands, while the isomer shift detected an increase in non-d charge. Neither experiment alone indicated the direction of any net charge transfer. Often the sign of transition metal core-level shifts upon alloying is indicative (Weinert and Watson, 1995) of the sign of the change in d-electron count, although this is less likely for atoms at the surface of a solid (Weinert and Watson, 1995). Because of the presence of charge of differing l, if the d count of one transition metal decreases upon alloying with another, this does not imply that the d count on the other site increases; in fact, as often as not, the d count increases or decreases at the two sites simultaneously. There can be subtle effects in simple metals as well. For example, it has long been recognized that Ca has unoccupied $3d$ levels not far above its Fermi level that hybridize during alloying. Clearly, having valence electrons of more than one l is important to alloy formation.

Loose Ends

A number of other matters will not be attended to in this unit. There are, e.g., other physical parameters that might be considered. When considering size for metallic systems, the atomic volume or radius appears to be appropriate. However, for semiconductor compounds, another measure of size appears to be relevant (Simons and Bloch, 1973; St. John and Bloch, 1974), namely, the size of the atomic cores in a pseudopotential description. Of course, the various parameters are not unrelated to one another, and the more such sets that are accumulated, the greater may be the linearities in any scheme employing them. Given such sets of parameters, there has been a substantial literature of so-called structural maps that employ the averages (as in d-band filling) and differences (as in sizes) as coordinates and in which the occurrence of compounds in some set of structures is plotted. This is a cheap way to search for possible new phases and to weed out suspect ones. As databases increase in size and availability, such mappings, employing multivariant analyses, will likely continue. These matters will not be discussed here.

WARNINGS AND PITFALLS

Electronic structure calculations have been surprisingly successful in reproducing and predicting physical properties of alloys. Because of this success, it is easy to be lulled into a false sense of security regarding the capabilities of these types of calculations. A critical examination of reported calculations must consider issues of both precision and accuracy. While these two terms are often used interchangeably, we draw a distinction between them.

We use the word "accuracy" to describe the ability to calculate, on an absolute scale, the physical properties of a system. The accuracy is limited by the physical effects,

which are included in the model used to describe the actual physical situation. "Precision," on the other hand, relates to how well one does solve a given model or set of equations. Thus, a very precise solution to a poor model will still have poor accuracy. Conversely, it is possible to get an answer in excellent agreement with experiment by inaccurately solving a poor model, but the predictive power of such a model would be questionable at best.

Issues Relating to Accuracy

The electronic structure problem for a real solid is a many-body problem with on the order of 10^{23} particles, all interacting via the Coulomb potential. The main difficulty in solving the many-body problem is not simply the large number of particles, but rather that they are interacting. Even solving the few-electron problem is nontrivial, although recently Monte Carlo techniques have been used to generate accurate wave functions for some atoms and small molecules where on the order of 10 electrons have been involved.

The H_2 molecule already demonstrates one of the most important conceptual problems in electronic structure theory, namely, the distinction between the band theory (molecular orbital) and the correlated-electron (Heitler-London) points of view. In the molecular orbital theory, the wave function is constructed in terms of Slater determinants of one-particle wave functions appropriate to the systems; for H_2, the single-particle wave functions are $\{\phi_a + \phi_b\}/\sqrt{2}$, where ϕ_a is a hydrogen-like function centered around the nucleus at R_a. This choice of single-particle wave functions means that the many-body wave function $\Psi(1,2)$ will have terms such as $\phi_a(1)\phi_a(2)$, i.e., there is a significant probability that both electrons will be on the same site. When there is significant overlap between the atomic wave functions, this behavior is what is expected; however, as the nuclei are separated, these terms remain and hence the H_2 molecule does not dissociate correctly. The Heitler-London approximation corrects this problem by simply neglecting those terms that would put two electrons on a single site. This additional Coulomb correlation is equivalent to including an additional large repulsive potential U between electrons on the same site within a molecular orbital picture, and is important in "localized" systems.

The simple band picture is roughly valid when the band width W (a measure of the overlap) is much larger than U; likewise the Heitler-London highly correlated picture is valid when U is much greater than W. It is important to realize that few, if any, systems are completely in one or the other limit and that many of the systems of current interest, such as the high-T_c superconductors, have $W \sim U$. Which limit is a better starting point in a practical calculation depends on the system, but it is naive, as well as incorrect, to claim that band theory methods are inapplicable to correlated systems. Conversely, the limitations in the treatment of correlations should always be kept in mind.

Most modern electronic structure calculations are based on density functional theory (Kohn and Sham, 1965). The basic idea, which is discussed in more detail elsewhere in this unit (SUMMARY OF ELECTRONIC STRUCTURE METHODS), is that the total energy (and related properties) of a many-electron system is a unique functional of the electron density, $n(r)$, and the external potential. In the standard Kohn-Sham construct, the many-body problem is reduced to a set of coupled one-particle equations. This procedure generates a mean-field approach, in which the effective one-particle potentials include not only the classical Hartree (electrostatic) and electron-ion potentials but also the so-called "exchange-correlation" potential. It is this exchange-correlation functional that contains the information about the many-body problem, including all the correlations discussed above.

While density functional theory is in principle correct, the true exchange-correlation functional is not known. Thus, to perform any calculations, approximations must be made. The most common approximation is the local density approximation (LDA), in which the exchange-correlation energy is given by

$$E_{xc} = \int dr\, n(r)\varepsilon_{xc}(r) \tag{4}$$

where $\varepsilon_{xc}(r)$ is the exchange-correlation energy of the homogeneous electron gas of density n. While this approximation seems rather radical, it has been surprisingly successful. This success can be traced in part to several properties: (1) the Coulomb potential between electrons is included correctly in a mean-field way and (2) the regions where the density is rapidly changing are either near the nucleus, where the electron-ion Coulomb potential is dominant, or in tail regions of atoms, where the density is small and will give only small exchange-correlation contributions to the total energy.

Although the LDA works rather well, there are several notable defects, including generally overestimated cohesive energies and underestimated lattice constants relative to experiment. A number of modifications to the exchange-correlation potential have been suggested. The various generalized gradient approximations (GGAs; Perdew et al., 1996, and references cited therein) include the effect of gradients of the density to the exchange-correlation potential, improving the cohesive energies in particular. The GGA treatment of magnetism also seems to be somewhat superior: the GGA correctly predicts the ferromagnetic bcc state to be most stable (Bagano et al., 1989; Singh et al., 1991; Asada and Terakura, 1992). This difference is apparently due in part to the larger magnetic site energy (Bagano et al., 1989) in the GGA than in the LDA; this increased magnetic site energy may also help resolve a discrepancy (R.E. Watson and M. Weinert, pers. comm.) in the heats of formation of nonmagnetic Fe compounds compared to experiment. Although the GGA seems promising, there are also some difficulties. The lattice constants for many systems, including a number of semiconductors, seem to be overestimated in the GGA, and consequently, magnetic moments are also often overestimated. Furthermore, proper introduction of gradient terms generally requires a substantial computational increase in the fast Fourier transform meshes in order to properly describe the potential terms arising from the density gradients.

Other correlations can be included by way of additional effective potentials. For example, the correlations associated with the on-site Coulomb repulsion characterized by U in the Heitler-London picture can be modeled by a local effective potential. Whether or not these various additional potentials can be derived rigorously is still an open question, and presently they are introduced mainly in an *ad hoc* fashion.

Given a density functional approximation to the electronic structure, there are still a number of other important issues that will limit the accuracy of the results. First, the quantum nature of the nuclei is generally ignored and the ions simply provide an external electrostatic potential as far as the electrons are concerned. The justification of the Born-Oppenheimer approximation is that the masses of the nuclei are several orders of magnitude larger than those of the electrons. For low-Z atoms such as hydrogen, zero-point effects may be important in some cases. These effects are especially of concern in so-called "first-principles molecular dynamics" calculations, since the semiclassical approximation used for the ions limits the range of validity to temperatures such that k_bT is significantly higher than the zero-point energies.

In making comparisons between calculations and experiments, the idealized nature of the calculational models must be kept in mind. Generally, for bulk systems, an infinite periodic system is assumed. Real systems are finite in extent and have surfaces. For many bulk properties, the effect of surfaces may be neglected, even though the electronic structure at surfaces is strongly modified from the bulk. Additionally, real materials are not perfectly ordered but have defects, both intrinsic and thermally activated. The theoretical study of these effects is an active area of research in its own right.

The results of electronic structure calculations are often compared to spectroscopic measurements, such as photoemission experiments. First, the eigenvalues obtained from a density functional calculation are not directly related to real (quasi) particles but are formally the Lagrange multipliers that ensure the orthogonality of the auxiliary one-particle functions. Thus, the comparison of calculated eigenvalues and experimental data is not formally sanctioned. This view justifies any errors, but is not particularly useful if one wants to make comparisons with experiments. In fact, experience has shown that such comparisons are often quite useful. Even if one were to assume that the eigenvalues correspond to real particles, it is important to realize that the calculations and the experiments are not describing the same situation. The band theory calculations are generally for the ground-state situation, whereas the experiments inherently measure excited-state properties. Thus, a standard band calculation will not describe satellites or other spectral features that result from strong final-state effects. The calculations, however, can provide a starting point for more detailed calculations that include such effects explicitly.

While the issues that have been discussed above are general effects that may affect the accuracy of the predictions, another class of issues is how particular effects—spin-orbit, spin-polarization, orbital effects, and multiplets—are described. These effects are often ignored or included in some approximation. Here, we will briefly discuss some aspects that should be kept in mind.

As is well known from elementary quantum mechanics, the spin-orbit interaction in a single-particle system can be related to dV/dr, where V is the (effective) potential. Since the density functional equations are effectively single-particle equations, this is also the form used in most calculations. However, for a many-electron atom, it is known (Blume and Watson, 1962, 1963) that the spin-orbit operator includes a two-electron piece and hence cannot be reduced to simply dV/dr. While the usual manner of including spin-orbit is not ideal, neglecting it altogether for heavy elements may be too drastic an approximation, especially if there are states near the Fermi level that will be split.

Magnetic effects are important in a large class of materials. A long-standing discussion has revolved around the question of whether itinerant models of magnetism can describe systems that are normally treated in a localized model. The exchange-correlation potential appropriate for magnetic systems is even less well known than that for nonmagnetic systems. Nevertheless, there have been notable successes in describing the magnetism of Ni and Fe, including the prediction of enhanced moments at surfaces.

Within band theory, these magnetic systems are most often treated by means of spin-polarized calculations, where different potentials and different orbitals are obtained for electrons of differing spin. A fundamental question arises in that the total spin of the system S is not a good quantum number in spin-polarized band theory calculations. Instead, S_z is the quantity that is determined. Furthermore, in most formulations the spin direction is uncoupled from the lattice; only by including the spin-orbit interaction can anisotropies be determined. Because of these various factors, results obtained for magnetic systems can be problematic. Of particular concern are those systems in which different types of magnetic interactions are competing, such as antiferromagnetic systems with several types of atoms. As yet, no general and practical formulation applicable to all systems exists, but a number of different approaches for specific cases have been considered.

While in most solids orbital effects are quenched, for some systems (such as Co and the actinides), the orbital moment may be substantial. The treatment of orbital effects until now has been based in an *ad hoc* manner on an approximation to the Racah formalism for atomic multiplet effects. The form generally used is strictly only valid (e.g., Watson et al., 1994) for the configuration with maximal spin, a condition that is not met for the metallic systems of interest here. However, even with these limitations, many computed orbital effects appear to be qualitatively in agreement with experiment. The extent to which this is accidental remains to be seen.

Finally, in systems that are localized in some sense, atomic multipletlike structure may be seen. Clearly, multiplets are closely related to both spin polarization and orbital effects. The standard LDA or GGA potentials cannot describe these effects correctly; they are simply beyond the physics that is built in, since multiplets have both total

spin and orbital momenta as good quantum numbers. In atomic Hartree-Fock theory, where the symmetries of multiplets are correctly obeyed, there is a subtle interplay between aspherical direct Coulomb and aspherical exchange terms. Different members of a multiplet have different balances of these contributions, leading to the same total energy despite different charge and spin densities. Moreover, many of the states are multideterminantal, whereas the Kohn-Sham states are effectively single-determinantal wave functions. An interesting point to note (e.g., Watson et al., 1994) is that for a given atomic configuration, there may be states belonging to different multiplets, of differing energy (with different S and L) that have identical charge and spin densities. An example of this in the d^4 configuration is the ^{5}D, $M_L = 1$, $M_S = 0$, which is constructed from six determinants—the ^{3}H, $M_L = 5$, $M_S = 0$, which is constructed from two determinants; and the ^{1}I, $M_L = 5, M_S = 0$, which is also constructed from two determinants. All three have identical charge and spin densities (and share a common M_S value), and thus both LDA- and GGA-like approximations will incorrectly yield the same total energies for these states. The existence of these states shows that the true exchange-correlation potential must depend on more than simply the charge and magnetization densities, perhaps also the current density. Until these issues are properly resolved, an *ad hoc* patch, which would be an improvement on the above-mentioned approximation for orbital effects, might be to replace the aspherical LDA (GGA) exchange-correlation contributions to the potential at a magnetic atomic site by an explicit estimate of the Hartree-Fock aspherical exchange terms (screened to crudely account for correlation), but the cost would be the necessity to carry electron wave function as well as density information during the course of a calculation.

Issues Relating to Precision

The majority of first-principles electronic structure calculations are done within the density functional framework, in particular either the LDA or the GGA. Once one has chosen such a model to describe the electronic structure, the ultimate accuracy has been established, but the precision of the solution will depend on a number of technical issues. In this section, we will address some of them.

There are literally orders-of-magnitude differences in computational cost depending on the approximations made. For understanding some properties, the simplest calculations may be completely adequate, but it should be clear that "precise" calculations will require detailed attention to the various approximations. One of the difficulties is that often the only way to show that a given approximation is adequate is to show that it agrees with a more rigorous calculation without that approximation.

First Principles Versus Tight Binding. The first distinction is between "first-principles" calculations and various parameterized models, such as tight binding. In a tight-binding scheme, the important interactions and matrix elements are parameterized based on some model systems, either experimental or theoretical, and these are then used to calculate the electronic structure of other systems. The advantage is that these methods are extremely fast, and for well-tested systems such as silicon and carbon, the results appear to be quite reliable. The main difficulties are related to those cases where the local environments are so different that the parameterizations are no longer adequate, including the case in which several different atomic species are present. In this case, the interactions among the different atoms may be strongly modified from the reference systems. Generating good tight-binding parameterizations is a nontrivial undertaking and involves a certain amount of art as well as science.

First-principles methods, on the other hand, start with the basic set of density functional equations and then must decide how to best solve them. In principle, once the nuclear charges and positions (and other external potentials) are given, the problem is well defined and there should be a density functional "answer." In practice, there are many different techniques, the choice of which is a matter of personal taste.

Self-Consistency. In density functional theory, the basic idea is to determine the minimum of the total energy with respect to the density. In the Kohn-Sham procedure, this minimization is cast into the form of a self-consistency condition on the density and potential. In both the Kohn-Sham procedure and a direct minimization, the precision of the result is related to the quality of the density. Depending on the properties of interest, the requirements can vary. The total energy is one of the least sensitive, since errors are second order in δn. However, when comparing the total energies of different systems, these second-order differences may be of the same scale as the desired answer. If one needs accurate single-particle wave functions in order to predict the quantity of interest, then there are more stringent requirements on self-consistency since the errors are first order. Decreasing δn requires more computer time, hence there is always some tradeoff between precision and practical considerations. In making these decisions, it is crucial to know what level of uncertainty one is willing to accept beforehand and to be prepared to monitor it.

All-Electron Versus Pseudopotential Methods. While self-consistency in some form is common to all methods, a major division exists between all-electron and pseudopotential methods. In all-electron methods, as the name implies, all the valence and core electrons are treated explicitly, although the core electrons may be treated in a slightly different fashion than the valence electrons. In this case, the electrostatic and exchange-correlation effects of all the electrons are explicitly included, even though the core electrons are not expected to take part in the chemical bonding. Because the core electrons are so strongly bound, the total energy is dominated by the contributions from the core. In the pseudopotential methods, these core electrons are replaced by an additional potential that attempts to mimic the effect of the core electrons on the valence state in the important bonding region of the solid. Here the calculated total energy is

significantly smaller in magnitude since the core electrons are not treated explicitly. In addition, the valence wave functions have been replaced by pseudo–wave functions that are significantly smoother than the all-electron wave functions in the core region. An important feature of pseudopotentials is the arbitrariness/freedom in the definition of the pseudopotentials that can be used to optimize them for different problems. While there are a number of advantages to pseudopotentials, including simplicity of computer codes, it must be emphasized that they are still an approximation to the all-electron result; furthermore, the large magnitude of the all-electron total energy does not cause any inherent problems.

Generating a pseudopotential that is transferable, i.e., that can be used in many different situations, is both nontrivial and an art; in an all-electron method, this question of transferability simply does not exist. The details of the pseudopotential generation scheme can affect both their convergence and the applicability. For example, near the beginning of the transition metal rows, the core cutoff radii are of necessity rather large, but strong compound formation may cause the atoms to have particularly close approaches. In such cases, it may be difficult to internally monitor the transferability of the pseudopotential, and one is left with making comparisons to all-electron calculations as the only reliable check. These comparisons between all-electron and pseudopotential results need to be done on the same systems. In fact, we have found in the Zr-Al system (Alatalo et al., 1998) that seemingly small changes in the construction of the pseudopotentials can cause qualitative differences in the relative stability of different phases. While this case may be rather extreme, it does provide a cautionary note.

Basis Sets. One of the major factors determining the precision of a calculation is the basis set used, both its type and its convergence properties. The physically most appealing is the linear combination of atomic orbitals (LCAOs). In this method, the basis functions are atomic orbitals centered at the atoms. From a simple chemical point of view, one would expect to need only nine functions ($1s$, $3p$, $5d$) per transition metal atom. While such calculations give an intuitive picture of the bonding, the general experience has been that significantly more basis functions per atom are needed, including higher l orbitals, in order to describe the changes in the wave functions due to bonding. One further related difficulty is that straightforwardly including more and more excited atomic orbitals is generally not the most efficient approach: it does not guarantee that the basis set is converging at a reasonable rate.

For systemic improvement of the basis set, plane waves are wonderful because they are a complete set with well-known convergence properties. It is equally important that plane waves have analytical properties that make the formulas generally quite simple. The disadvantage is that, in order to describe the sharp structure in the atomic core wave functions and in the nuclear-electron Coulomb interaction, exceedingly large numbers of plane waves are required. Pseudopotentials, because there are no core electrons and the very small wavelength structures in the valence wave functions have been removed, are ideally suited for use with plane-wave basis sets. However, even with pseudopotentials, the convergence of the total energy and wave functions with respect to the plane-wave cutoff must be monitored closely to ensure that the basis sets used are adequate, especially if different systems are to be compared. In addition, the convergence of the plane-wave energy cutoff may be rather slow, depending on both the particular atoms and the form of the pseudopotential used.

To treat the all-electron problem, another class of methods besides the LCAOs is used. In the augmented methods, the basis functions around a nucleus are given in terms of numerical solutions to the radial Schrödinger equation using the actual potential. Various methods exist that differ in the form of the functions to be matched to in the interstitial region: plane waves [e.g., various augmented plane wave (APW) and linear APW (LAPW) methods], r^{-l} (LMTOs), Slater-type orbitals (LASTOs), and Bessel/Hankel functions (KKR-type). The advantage of these methods is that in the core region where the wave functions are changing rapidly, the correct functions are used, ensuring that, e.g., cusp conditions are satisfied. These methods have several different cutoffs that can affect the precision of the methods. First, there is the spherical partial-wave cutoff l and the number of different functions inside the sphere that are used for each l. Second, the choice and number of tail functions are important. The tail functions should be flexible enough to describe the shape of the actual wave functions in that region. Again, plane waves (tails) have an advantage in that they span the interstitial well and the basis can be systematically improved. Typically, the number of augmented functions needed to achieve convergence is smaller than in pseudopotential methods, since the small-wavelength structure in the wave functions is described by numerical solutions to the corresponding Schrödinger equation.

The authors of this unit have employed a number of different calculational schemes, including the full-potential LAPW (FLAPW) (Wimmer et al., 1981; Weinert et al., 1982) and LASTO (Fernando et al., 1989) methods. The LASTO method, which uses Slater-type orbitals (STOs) as tail functions in the interstitial region, has been used to estimate heats of formation of transition metal alloys in a diverse set of well-packed and ill-packed compound crystal structures. Using a single s-, p-, and d-like set of STOs is inadequate for any serious estimate of compound heats of formation. Generally, $2s$, $2p$, $2d$, and an f-like STO were employed for a transition metal element when searching for the compound's optimum lattice volume, c/a, b/a, and any internal atomic coordinates not controlled by symmetry. Once this was done, a final calculation was performed with an additional d plus a g-like STO (clearly the same basis must be employed when determining the reference energy of the elemental metal). Usually, the effect on the calculated heat of formation was measurably < 0.01 eV/atom, but occasionally this difference was larger. These basis functions were only employed in the interstitial region, but nevertheless any claim of a computational precision of, say, 0.01 eV/atom required a substantial basis for such metallic transition metal compounds.

There are lessons to be learned from this experience for other schemes relying on localized basis functions and/or small basis sets such as the LMTO and LCAO methods.

Full Potentials. Over the years a number of different approximations have been made to the shape of the potential. In one simple picture, a solid can be thought of as being composed of spherical atoms. To a first approximation, this view is correct. However, the bonding will cause deviations and the charge will build up between the atoms. The charge density and the potential thus have rather complicated shapes in general. Schemes that account for these shapes are normally termed "full potential." Depending on the form of the basis functions and other details of the computational technique, accounting for the detailed shape of the potential throughout space may be costly and difficult.

A common approximation is the muffin-tin approximation, in which the potential in the spheres around the atoms is spherical and in the interstitial region it is flat. This form is a direct realization of the simple picture of a solid given above, and is not unreasonable for metallic bonding of close-packed systems. A variant of this approximation is the atomic-sphere approximation (ASA), in which the complications of dealing with the interstitial region are avoided by eliminating the interstices completely; in order to conserve the crystal volume, the spheres are increased in size. In the ASA, the spheres are thus of necessity overlapping. While these approximations are not too unreasonable for some systems, care must be exercised when comparing heats of formation of different structures, especially ones that are less well packed.

In a pure plane-wave method, the density and potential will again be given in terms of a plane-wave expansion. Even the Coulomb potential is easy to describe in this basis; for this reason, plane-wave (pseudopotential) methods are generally full potential to begin with. For the other methods, the solution of the Poisson equation becomes more problematic since either there are different representations in different regions of space or the site-centered representation is not well adapted to the long-range nonlocal nature of the Coulomb potential. Although this aspect of the problem can be dealt with, the largest computational cost of including the full potential is in including the nonspherical contributions consistently in the Hamiltonian.

Many of the present-day augmented methods, starting with the FLAPW method, include a full-potential description. However, even within methods that claim full potentials, there are differences. For example, the FLAPW method solves the Poisson equation correctly for the complete density described by different representations in different regions of space. On the other hand, certain "full-potential" LMTO methods solve for the nonspherical terms in the spheres but then extrapolate these results in order to obtain the potential in the interstitial. Even with this approximation, the results are significantly better than LMTO-ASA.

At this time, it can be argued that a calculation of a heat of formation or almost any other observable should not be deemed "precise" if it does not involve a full-potential treatment.

Brillouin Zone Sampling. In solving the electronic structure of an ordered solid, one is effectively considering a crystal of N unit cells with periodic boundary conditions. By making use of translational symmetry, one can reduce the problem to finding the electronic structure at N **k**-points in the Brillouin zone. Thus, there is an exact 1:1 correspondence between the effective crystal size and the number of **k**-points. If there are additional rotational (or time-reversal) symmetries, then the number of **k**-points that must be calculated is reduced. Note that using **k**-points to increase the effective crystal size always has linear scaling. Because the figure of merit is the effective crystal size, using large unit cells (but small numbers of **k**-points) to model real systems may be both inefficient and unreliable in describing the large-crystal limit. As a simple example, the electronic structure of a monoatomic fcc metal calculated using only 60 special **k**-points in the irreducible wedge is identical to that calculated using a supercell with 2048 atoms and the single Γ **k**-point. While this effective crystal size (number of **k**-points) is marginal for many purposes, a supercell of 2048 atoms is still not routinely done.

The number of **k**-points needed for a given system will depend on a number of factors. For filled-band systems such as semiconductors and insulators, fewer points will be needed. For metallic systems, enough **k**-points to adequately describe the Fermi surface and any structure in the density of states nearby are needed. Although tricks such as using very large temperature broadenings can be useful in reaching self-consistency, they cannot mimic structures in the density of states that are not included because of a poor sampling of the Brillouin zone. When comparisons are to be made between different structures, it is necessary that tests regarding the convergence in the number of **k**-points be made for each of the structures.

Besides the number of **k**-points that are used, it is important to use properly symmetrized **k**-points. A surprisingly common error in published papers is that only the **k**-points in the irreducible wedge appropriate for a high-symmetry (typically cubic) system is used, even when the irreducible wedge that should be used is larger because the overall symmetry has been reduced. Depending on the system, the effects of this type of error may not be so large that the results are blatantly wrong but may still be of significance.

Structural Relaxations. For simple systems, reliable experimental data exist for the structural parameters. For more complicated structures, however, there is often little, if any, data. Furthermore, LDA (or GGA) calculations do not exactly reproduce the experimental lattice constants or volumes, although they are reasonably good. The energies associated with these differences in lattice constants may be relatively small, on the order of a few hundredths of an electron volt per atom, but these energies may be significant when comparing the heats of formation of competing phases, both at the same and at different concentrations. For complicated structures, the internal structural parameters are even less well known experimentally, but they may have a relatively large effect on the heats of formation (the energies related to variations

of the internal parameters are on the scale of optical phonon modes).

To calculate consistent numbers for properties such as heats of formation, the total energy should be minimized with respect to both external (e.g., lattice constants) and internal structural parameters. Often the various parameters are coupled, making a search for the minimum rather time-consuming, although the use of forces and stresses are helpful. Furthermore, while the changes in the heats may be small, it is necessary to do the minimization in order to find out how close the initial guess was to the minimum.

Final Cautions

There is by now a large body of work suggesting that modern electronic structure theory is able to reliably predict many properties of materials, often to within the accuracy that can be obtained experimentally. In spite of this success, the underlying assumptions, both physical and technical, of the models used should always be kept in mind. In assessing the quality and applicability of a calculation, both accuracy and precision must be considered. Not all physical systems can be described using band theory, but the results of a band theory may still provide a good starting point or input to other theories. Even when the applicability of band theory is in question, a band calculation may give insights into the additional physics that needs to be included.

Technical issues, i.e., precision, are easier to quantify. In practice there are trade-offs between precision and computational cost. For many purposes, simpler methods with various approximations may be perfectly adequate. Unfortunately, often it is only possible to assess the effect of an approximation by (partially) removing it. Technical issues that we have found to be important in studies of heats of formation of metallic systems, regardless of the specific method used, include the flexibility of the basis set, adequate Brillouin zone sampling, and a full-potential description. Although the computational costs are relatively high, especially when the convergence checks are included, this effort is justified by the increased confidence in both the numerical results and the physics that is extracted from the numbers.

Over the years there has been a concern with the accuracy of the approximations employed. Too often, arguments concerning relative accuracies of, say, LDA versus GGA, for some particular purpose, have been mired in the inadequate precision of some of the computed results being compared. Verifying the precision of a result can easily increase the computational cost by one or more orders of magnitude but, on occasion, is essential. When the increased precision is not needed, simple economics dictates that it be avoided.

CONCLUSIONS

In this unit, we have attempted to address two issues associated with our understanding of alloy phase behavior. The first is the simple metallurgical modeling of the kind associated with Hume-Rothery's name (although many workers have been involved over the years). This modeling has involved parameters such as valence (d-band filling), size, and electronegativity. As we have attempted to indicate, partial contact can be made with present-day electronic structure theory. The ideas of Hume-Rothery, Friedel, and others concerning the role of d-band filling in alloy phase formation has an immediate association with the densities of states, such as those shown for Pd-Ti in Figure 2. In contrast, the relative sizes of alloy constituents are often essential to what phases are formed, but the role of size (except as manifested in a calculated total energy) is generally not transparent in band theory results. The third parameter, electronegativity, attempts to summarize complex bonding trends in a single parameter. Modern-day computation and experiment have shed light on these complexities. For the alloys at hand, "charge transfer" involves band hybridization of charge components of various l, of band filling, and of charge intrinsic to neighboring atoms overlapping onto the atom in question. All of these effects cannot be incorporated easily into a single numbered scale. Despite this, scanning for trends in phase behavior in terms of such parameters is more economical than doing vast arrays of electronic structure calculations in search of hitherto unobserved stable and metastable phases, and has a role in metallurgy to this day.

The second matter, which overlaps other units in this chapter, deals with the issues of accuracy and precision in band theory calculations and, in particular, their relevance to estimates of heats of formation. Too often, published results have not been calculated to the precision that a reader might reasonably presume or need.

ACKNOWLEDGMENTS

The work at Brookhaven was supported by the Division of Materials Sciences, U.S. Department of Energy, under Contract No. DE-AC02-76CH00016, and by a grant of computer time at the National Energy Research Scientific Computing Center, Berkeley, California.

LITERATURE CITED

Alatalo, M., Weinert, M., and Watson, R. E. 1998. *Phys. Rev. B* 57:R2009-R2012.

Asada, T. and Terakura, K. 1992. *Phys. Rev. B* 46:13599.

Bagano, P., Jepsen, O., and Gunnarson, O. 1989. *Phys. Rev. B* 40:1997.

Blume, M. and Watson, R. E. 1962. *Proc. R. Soc. A* 270:127.

Blume, M. and Watson, R. E. 1963. *Proc. R. Soc. A* 271:565.

de Boer, F. R., Boom, R., Mattens, W. C. M., Miedema, A. R., and Niessen, A. K. 1988. Cohesion in Metals. North-Holland, Amsterdam, The Netherlands.

Fernando, G. W., Davenport, J. W., Watson, R. E., and Weinert, M. 1989. *Phys. Rev. B* 40:2757.

Frank, F. C. and Kasper, J. S. 1958. *Acta Crystallogr.* 11:184.

Frank, F. C. and Kasper, J. S. 1959. *Acta Crystallogr.* 12:483.

Friedel, J. 1969. *In* The Physics of Metals (J. M. Ziman, ed.). Cambridge University Press, Cambridge.

Gordy, W. and Thomas, W. J. O. 1956. *J. Chem. Phys.* 24:439.

Hume-Rothery, W. 1936. The Structure of Metals and Alloys. Institute of Metals, London.

Kohn, W. and Sham, L. J. 1965. *Phys. Rev.* 140:A1133.

Mulliken, R. S. 1934. *J. Chem. Phys.* 2:782.

Mulliken, R. S. 1935. *J. Chem. Phys.* 3:573.

Nelson, D. R. 1983. *Phys. Rev. B* 28:5155.

Perdew, J. P., Burke, K., and Ernzerhof, M. 1996. *Phys. Rev. Lett.* 77:3865.

Pettifor, D. G. 1970. *J. Phys. C* 3:367.

Simons, G. and Bloch, A. N. 1973. *Phys. Rev. B* 7:2754.

Singh, D. J., Pickett, W. E., and Krakauer, H. 1991. *Phys. Rev. B* 43:11628.

St. John, J. and Bloch, A. N. 1974. *Phys. Rev. Lett.* 33:1095.

Watson, R. E. and Bennett, L. H. 1984. *Acta Metall.* 32:477.

Watson, R. E. and Bennett, L. H. 1985. *Scripta Metall.* 19:535.

Watson, R. E., Bennett, L. H., and Davenport, J. W. 1983. *Phys. Rev. B* 27:6429.

Watson, R. E., Hudis, J., and Perlman, M. L. 1971. *Phys. Rev. B* 4:4139.

Watson, R. E. and Weinert, M. 1994. *Phys. Rev. B* 49:7148.

Watson, R. E., Weinert, M., Davenport, J. W., and Fernando, G. W. 1989. *Phys. Rev. B* 39:10761.

Watson, R. E., Weinert, M., Davenport, J. W., Fernando, G. W., and Bennett, L. H. 1994. *In* Statics and Dynamics of Alloy Phase Transitions (P. E. A. Turchi and A. Gonis, eds.) pp. 242–246. Plenum, New York.

Watson, R. E., Weinert, M., and Fernando, G. W. 1991. *Phys. Rev. B* 43:1446.

Weinert, M. and Watson, R. E. 1995. *Phys. Rev. B* 51:17168.

Weinert, M., Wimmer, E., and Freeman, A. J. 1982. *Phys. Rev. B* 26:4571.

Wimmer, E., Krakauer, H., Weinert, M., and Freeman, A. J. 1981. *Phys. Rev. B* 24:864.

KEY REFERENCES

Freidel, 1969. See above.

Provides the simple bonding description of d bands that yields the fact that the maximum cohesion of transition metals occurs in the middle of the row.

Harrison, W. A. 1980. Electronic Structure and the Properties of Solids. W. H. Freeman, San Francisco.

Discusses the determination of the properties of materials from simplified electronic structure calculations, mainly based on a linear combination of atomic orbital (LCAO) picture. Gives a good introduction to tight-binding (and empirical) pseudopotential approaches. For transition metals, Chapter 20 is of particluar interest.

Moruzzi, V. L., Janak, J. F., and Williams, A. R. 1978. Calculated Electronic Properties of Metals. Pergamon Press, New York.

Provides a good review of the trends in the calculated electronic properties of metals for Z″ 49 (In). Provides density of states and other calculated properties for the elemental metals in either bcc or fcc structure, depending on the observed ground state. (The hcp elements are treated as fcc.) While newer calculations are undoubtedly better, this book provides a consistent and useful set of results.

Pearson, W. B. 1982. *Philos. Mag. A* 46:379.

Considers the role of the relative sizes of an ordered compound's constituents in determining the crystal structure that occurs.

Pettifor, D. J. 1977. *J. Phys.* F7:613, 1009.

Provides a tight-binding description of the transition metal that predicts the crystal structures across the row.

Waber, J. T., Gschneidner, K., Larson, A. C., and Price, M. Y. 1963. *Trans. Metall. Soc.* 227:717.

Employs "Darken Gurry" plots in which the tendency for substitutional alloy formation is correlated with small differences in atomic size and electronegativities of the constituents: the correlation works well with size but less so for electronegativity (where a large difference would imply strong bonding and, in turn, the formation of ordered compounds).

R. E. WATSON
M. WEINERT
Brookhaven National Laboratory
Upton, New York

BINARY AND MULTICOMPONENT DIFFUSION

INTRODUCTION

Diffusion is intimately related to the irreversible process of mixing. Almost two centuries ago it was observed for the first time that gases tend to mix, no matter how carefully stirring and convection are avoided. A similar tendency was later also observed in liquids. Already in 1820 Faraday and Stodart made alloys by mixing of metal powders and subsequent annealing. The scientific study of diffusion processes in solid metals started with the measurements of Au diffusion in Pb by Roberts-Austen (1896). The early development was discussed by, for example, Mehl (1936) during a seminar on atom movements (ASM, 1951) in Chicago. The more recent theoretical development concentrated on multicomponent effects, and extensive discussions can be found in the textbooks by Adda and Philibert (1966) and Kirkaldy and Young (1987).

Diffusion processes play an important role during all processing and usage of materials at elevated temperatures. Most materials are based on a major component and usually several alloy elements that may all be important for the performance of the materials, i.e., the materials are multicomponent systems. In this unit, we shall present the theoretical basis for analyzing diffusion processes in multicomponent solid metals. However, despite their limited practical applicability, for pedagogical reasons we shall also discuss binary systems.

We shall present the famous Fick's law and mention its relation to other linear laws such as Fourier's and Ohm's laws. We shall formulate a general linear law and introduce the concept of mobility. A special section is devoted to the choice of frame of reference and how this leads to the various diffusion coefficients, e.g., individual diffusion coefficients, chemical diffusion, tracer diffusion, and self-diffusion.

We shall then turn to the general multicomponent for mulation, i.e., the Fick-Onsager law, and extend the discussion on the frames of reference to the multicomponent systems.

In a multicomponent system with n components, $n - 1$ concentration variables may be varied independently, and

the concentration of one of the components has to be chosen as a dependent variable. This choice is arbitrary, but it is often convenient to choose one of the major components as the dependent component and sometimes it is even necessary to change the choice of dependent component. Then, a new set of diffusion coefficients have to be calculated from the old set. This is discussed in the section Change of Dependent Concentration Variable.

The diffusion coefficients, often called diffusivities, express the linear relation between diffusive flux and concentration gradients. However, from a strict thermodynamic point of view, concentration gradients are not forces because real forces are gradients of some potential, e.g., electrical potential or chemical potential, and concentration is not a true potential. The kinetic coefficients relating flux to gradients in chemical potential are called mobilities. In a system with n components, $n(n-1)$ diffusion coefficients are needed to give a full representation of the diffusion behavior whereas there are only n independent mobilities, one for each diffusing component. This means that the diffusivities are not independent, and experimental data should thus be stored as mobilities rather than as diffusivities. For binary systems it does not matter whether diffusivities or mobilities are stored, but for ternary systems there are six diffusivities and only three mobilities. For higher-order systems the difference becomes larger, and in practice it is impossible to construct a database based on experimentally measured diffusivities. A database should thus be based upon mobilities and a method to calculate the diffusivities whenever needed. A method to represent experimental information in terms of mobilities for diffusion in multicomponent interstitial and substitutional metallic systems is presented, and the temperature and concentration dependences of the mobilities are discussed. The influence of magnetic and chemical order will be discussed.

In this unit we shall not discuss the mathematical solution of diffusion problems. There is a rich literature on analytical solutions of heat flow and diffusion problems; see, e.g., the well-known texts by Carslaw and Jaeger (1946/1959), Jost (1952), and Crank (1956). It must be emphasized that the analytical solutions are usually less applicable to practical calculations because they cannot take into account realistic conditions, e.g., the concentration dependence of the diffusivities and more complex boundary conditions. In binary systems it may sometimes be reasonable to approximate the diffusion coefficient as constant and obtain satisfactory results with analytical solutions. In multicomponent systems the so-called off-diagonal diffusivities are strong functions of concentration and may be approximated as constant only when the concentration differences are small. For multicomponent systems, it is thus necessary to use numerical methods based on finite differences (FDM) or finite elements (FEM).

THE LINEAR LAWS

Fick's Law

The formal basis for the diffusion theory was laid by the German physiologist Adolf Fick, who presented his theoretical analysis in 1855. For an isothermal, isobaric, one-phase binary system with diffusion of a species B in one direction z, Fick's famous law reads

$$J_{\mathrm{B}} = -D_{\mathrm{B}} \frac{\partial c_{\mathrm{B}}}{\partial z} \tag{1}$$

where J_{B} is the flux of B, i.e., the amount of B that passes per unit time and unit area through a plane perpendicular to the z axis, and c_{B} is the concentration of B, i.e., the amount of B per unit volume. Fick's law thus states that the flux is proportional to the concentration gradient, and D_{B} is the proportionality constant called the diffusivity or the diffusion coefficient of B and has the dimension length squared over time. Sometimes it is called the diffusion constant, but this term should be avoided since D is not a constant at all. For instance, its variation with temperature is very important. It is usually represented by the following mathematical expression, the so-called Arrhenius relation:

$$D = D_0 \exp\left(\frac{-Q}{RT}\right) \tag{2}$$

where Q is the activation energy for diffusion, D_0 is the frequency factor, R is the gas constant, and T is the absolute temperature.

The flux J must be measured in a way analogous to the way used for the concentration c. If J is measured in moles per unit area and time, c must be in moles per volume. If J is measured in mass per unit area and time, c must be measured in mass per volume. Notice that the concentration in Fick's law must never be expressed as mass percent or mole fraction. When the primary data are in percent or fraction, they must be transformed. We thus obtain, e.g.,

$$c_{\mathrm{B}}\left(\frac{\mathrm{mol\,B}}{m^3}\right) = \frac{x_{\mathrm{B}}}{V_m}\left(\frac{\mathrm{mol\,B/mol}}{m^3/\mathrm{mol}}\right) \tag{3}$$

and

$$dc_{\mathrm{B}} = \left(1 - \frac{d\ln V_m}{d\ln x_{\mathrm{B}}}\right)\frac{dx_{\mathrm{B}}}{V_m} \tag{4}$$

where V_m is the molar volume and x_{B} the mole fraction. We can thus write Fick's law in the form

$$J_{\mathrm{B}} = -\frac{D^x_{\mathrm{B}}}{V_m}\frac{\partial x_{\mathrm{B}}}{\partial z} \tag{5}$$

where $D^x_{\mathrm{B}} = (1 - d\ln V_m/d\ln x_{\mathrm{B}})D_{\mathrm{B}}$. It may often be justified to approximate D^x_{B} with D_{B}, and in the following we shall adopt this approximation and drop the superscript x.

Experimental data usually indicate that the diffusion coefficient varies with concentration. With D evaluated experimentally as a function of concentration, Fick's law may be applied for practical calculations and yields precise results. Although such a concentration dependence certainly makes it more difficult to find analytical solutions to diffusion, it is easily handled by numerical methods on

the computer. However, it must be emphasized that the diffusivity is not allowed to vary with the concentration gradient.

There are many physical phenomena that obey the same type of linear law, e.g., Fourier's law for heat conduction and Ohm's law for electrical conduction.

General Formulation

Fick's law for diffusion can be brought into a more fundamental form by introducing the appropriate thermodynamic potential, i.e., the chemical potential μ_B, instead of the concentration. Assuming that $\mu_B = \mu_B(c_B)$,

$$J_B = -D_B \frac{\partial c_B}{\partial z} = -L_B \frac{\partial \mu_B}{\partial z} = -L_B \frac{d\mu_B}{dc_B} \frac{\partial c_B}{\partial z} \tag{6}$$

i.e.,

$$D_B = L_B \frac{d\mu_B}{dc_B} \tag{7}$$

where L_B is a kinetic quantity playing the same role as the heat conductivity in Fourier's law and the electric conductivity in Ohm's law. Noting that a potential gradient is a force, we may thus state all such laws in the same form: Flux is proportional to force. This is quite different from the situation of Newtonian mechanics, which states that the action of a force on a body causes the body to accelerate with an acceleration proportional to the force. However, if the body moves through a media with friction, a linear relation between velocity and force is observed after a transient time.

Note that for ideal or dilute solutions we have

$$\mu_B = \mu_B^0 + RT \ln x_B = \mu_B^0 + RT \ln(c_B V_m) \tag{8}$$

and we find for the case when V_m may be approximated as constant that

$$D_B = \frac{RT}{x_B} V_m L_B \tag{9}$$

Mobility

From a more fundamental point of view we may understand that Fick's law should be based on $\partial\mu_B/\partial z$, which is the force that acts on the B atoms in the z direction. One should thus expect a "current density" that is proportional to this force and also to the number of B atoms that feel this force, i.e., the number of B atoms per volume, $c_B = x_B/V_m$. The constant of proportionality that results from such a consideration may be regarded as the mobility of B. This quantity is usually denoted by M_B and is defined by

$$J_B = -M_B \frac{x_B}{V_m} \frac{\partial \mu_B}{\partial z} \tag{10}$$

By comparison, we find $M_B x_B/V_m = L_B$, and thus

$$D_B = M_B \frac{x_B}{V_m} \frac{d\mu_B}{dc_B} \tag{11}$$

For dilute or ideal solutions we find

$$D_B = M_B RT \tag{12}$$

This relation was initially derived by Einstein. At this stage we may notice that Equation 10 has important consequences if we want to predict how the diffusion of B is affected when one more element is added to the system. When element C is added to a binary A-B alloy, we have $\mu_B = \mu_B(c_B, c_C)$, and consequently

$$\begin{aligned} J_B &= -M_B c_B \frac{\partial \mu_B}{\partial z} = -M_B c_B \left(\frac{\partial \mu_B}{\partial c_B} \frac{\partial c_B}{\partial z} + \frac{\partial \mu_B}{\partial c_C} \frac{\partial c_C}{\partial z} \right) \\ &= -D_{BB} \frac{\partial c_B}{\partial z} - D_{BC} \frac{\partial c_C}{\partial z} \end{aligned} \tag{13}$$

i.e., we find that the diffusion of B not only depends on the concentration gradient of B itself but also on the concentration gradient of C. Thus the diffusion of B is described by two diffusion coefficients:

$$D_{BB} = M_B c_B \frac{\partial \mu_B}{\partial c_B} \tag{14}$$

$$D_{BC} = M_B c_B \frac{\partial \mu_B}{\partial c_C} \tag{15}$$

To describe the diffusion of both B and C, we would thus need at least four diffusion coefficients.

DIFFERENT FRAMES OF REFERENCE AND THEIR DIFFUSION COEFFICIENTS

Number-Fixed Frame of Reference

The form of Fick's law depends to some degree upon how one defines the position of a certain point, i.e., the frame of reference. One method is to count the number of moles on each side of the point of interest and to require that these numbers stay constant. For a binary system this method will automatically lead to

$$J_A + J_B = 0 \tag{16}$$

if J_A and J_B are expressed in moles. This particular frame of reference is usually called the number-fixed frame of reference. If we now apply Fick's law, we shall find

$$-\frac{D_A}{V_m} \frac{\partial x_A}{\partial z} = J_A = -J_B = \frac{D_B}{V_m} \frac{\partial x_B}{\partial z} \tag{17}$$

Because $\partial x_A/\partial z = -\partial x_B/\partial z$, we must have $D_A = D_B$. The above method to define the frame of reference is thus equivalent to assuming that $D_A = D_B$. In a binary system we can thus only evaluate one diffusion coefficient if the number-fixed frame of reference is used. This diffusion

coefficient describes how rapidly concentration gradients are leveled out, i.e., the mixing of A and B, and is called the "chemical diffusion coefficient" or the "interdiffusion coefficient."

Volume-Fixed Frame of Reference

Rather than count the number of moles flowing in each direction, one could look at the volume and define a frame of reference from the condition that there would be no net flow of volume, i.e.,

$$J_A V_A + J_B V_B = 0 \tag{18}$$

where V_A and V_B are the partial molar volumes of A and B, respectively. This frame of reference is called the volume-fixed frame of reference. Also in this frame of reference it is possible to evaluate only one diffusion coefficient in a binary system.

Lattice-Fixed Frame of Reference

Another method of defining the position of a point is to place inert markers in the system at the points of interest. One could then measure the flux of A and B atoms relative to the markers. It is then possible that more A atoms will diffuse in one direction than will B atoms in the opposite direction. In that case one must work with different diffusion coefficients D_A and D_B. Using this method, we may thus evaluate individual diffusion coefficients for A and B. Sometimes the term "intrinsic diffusion coefficients" is used. Since the inert markers are regarded as fixed to the crystalline lattice, this frame of reference is called the lattice-fixed frame of reference.

Suppose that it is possible to evaluate the fluxes for A and B individually in a binary system, i.e., the lattice-fixed frame of reference. In this case, we may apply Fick's law for each of the two components:

$$J_A = -\frac{D_A}{V_m}\frac{\partial x_A}{\partial z} \tag{19}$$

$$J_B = -\frac{D_B}{V_m}\frac{\partial x_B}{\partial z} \tag{20}$$

The individual, or intrinsic, diffusion coefficients D_A and D_B usually have different values. It is convenient in a binary system to choose one independent concentration and eliminate the other one. From $\partial x_A/\partial z = -\partial x_B/\partial z$, we obtain

$$J_A = D_A \frac{1}{V_m}\frac{\partial x_B}{\partial z} = -D_{AB}^{A}\frac{1}{V_m}\frac{\partial x_B}{\partial z} \tag{21}$$

$$J_B = -D_B \frac{1}{V_m}\frac{\partial x_B}{\partial z} = -D_{BB}^{A}\frac{1}{V_m}\frac{\partial x_B}{\partial z} \tag{22}$$

where we have introduced the notation D_{AB}^{A} and D_{BB}^{A} as the diffusion coefficients for A and B, respectively, with respect to the concentration gradient of B when the concentration of A has been chosen as the dependent variable. Observe that

$$D_{AB}^{A} = -D_A \tag{23}$$

$$D_{BB}^{A} = D_B \tag{24}$$

Transformation between Different Frames of Reference

We may transform between two arbitrary frames of reference by means of the relations

$$J'_A = J_A - \frac{\vartheta}{V_m}x_A \tag{25}$$

$$J'_B = J_B - \frac{\vartheta}{V_m}x_B \tag{26}$$

where the primed frame of reference moves with the migration rate ϑ relative the unprimed frame of reference. Summation of Equations 25 and 26 yields

$$J'_A + J'_B = J_A + J_B - \frac{\vartheta}{V_m} \tag{27}$$

We shall now apply Equation 27 to derive a relation between the individual diffusion coefficients and the interdiffusion coefficient. Evidently, if the unprimed frame of reference corresponds to the case where J_A and J_B are evaluated individually and the primed to $J'_A + J'_B = 0$, we have

$$\frac{\vartheta}{V_m} = J_A + J_B \tag{28}$$

Here, ϑ is called the Kirkendall shift velocity because it was first observed by Kirkendall and his student Smigelskas that the markers appeared to move in a specimen as a result of the diffusion (Smigelskas and Kirkendall, 1947).

Introducing the mole fraction of A, x_A, as the dependent concentration variable and combining Equations 21 through 28, we obtain

$$J'_B = -\left[(1-x_B)D_{BB}^{A} - x_B D_{AB}^{A}\right]\frac{1}{V_m}\frac{\partial x_B}{\partial z} \tag{29}$$

We may thus define

$$D_{BB}^{A'} = \left[(1-x_B)D_{BB}^{A} - x_B D_{AB}^{A}\right] \tag{30}$$

which is essentially a relation first derived by Darken (1948). The interdiffusion coefficient is often denoted $\tilde{D}_{AB}$. However, we shall not use this notation because it is not suitable for extension to multicomponent systems.

We may introduce different frames of reference by the generalized relation

$$a_A J_A + a_B J_B = 0 \tag{31}$$

Table 1. Diffusion Coefficients and Frames of Reference in Binary Systems

1. Fick's law

x_A dependent variable	x_B dependent variable	relations (Eq. 59)
$J_B = -\frac{D^A_{BB}}{V_m}\frac{\partial x_B}{\partial z}$	$J_B = -\frac{D^B_{BA}}{V_m}\frac{\partial x_A}{\partial z}$	$D^A_{BB} = -D^B_{BA}$
$J_A = -\frac{D^A_{AB}}{V_m}\frac{\partial x_B}{\partial z}$	$J_A = -\frac{D^B_{AA}}{V_m}\frac{\partial x_A}{\partial z}$	$D^A_{AB} = -D^B_{AA}$

2. Frames of reference
   (a) Lattice fixed: J_A and J_B independent $\rightarrow$ D^A_{BB} and D^A_{AB} independent, two individual or intrinsic diffusion coefficients.
   Kirkendall shift velocity: $\vartheta = -(D^A_{BB} + D^A_{AB})\frac{\partial x_B}{\partial z}$
   (b) Number fixed:
   $J_{A'} + J_{B'} = 0 \rightarrow D^{A'}_{BB} = -D^{A'}_{AB} \qquad D^{A'}_{BB} = [(1 - x_B)D^A_{BB} - x_B D^A_{AB}]$
   One chemical diffusion coefficient or interdiffusion coefficient

For example, if A is a substitutional atom and B is an interstitial solute, it is convenient to choose a frame of reference where $a_A = 1$, $a_B = 0$. In this frame of reference there is no flow of substitutional atoms. They merely constitute a background for the interstitial atoms. Table 1 provides a summary of some relations in differing frames of reference.

The Kirkendall Effect and Vacancy Wind

We may observe a Kirkendall effect whenever the atoms move with different velocities relative to the lattice, i.e., when their individual diffusion coefficients differ. However, such a behavior is incompatible with the so-called place interchange theory, which was predominant before the experiments of Smigelskas and Kirkendall (1947). This theory was based on the idea that diffusion takes place by direct interchange between the atoms. In fact, the observation that different elements may diffuse with different rates is strong evidence for the vacancy mechanism, i.e., diffusion takes place by atoms jumping to neighboring vacant sites. The lattice-fixed frame of reference would then be regarded as a number-fixed frame of reference if the vacancies are considered as an extra component. We may then introduce a vacancy flux J_{Va} by the relation

$$J_A + J_B + J_{Va} = 0 \tag{32}$$

and we find that

$$J_{Va} = -\frac{\vartheta}{V_m} \tag{33}$$

The above discussion not only is an algebraic exercise but also has some important physical implications. The flux of vacancies is a real physical phenomenon manifested in the Kirkendall effect and is frequently referred to as the vacancy wind. Vacancies would be consumed or produced at internal sources such as grain boundaries. However, under certain conditions very high supersaturations of vacancies will be built up and pores may form, resulting in so-called Kirkendall porosity. This may cause severe problems when heat treating joints of dissimilar materials because the mechanical properties will deteriorate.

It is thus clear that by the chemical diffusion coefficient we may evaluate how rapidly the components mix but will have no information about the Kirkendall effect or the risk for pore formation. To predict the Kirkendall effect, we would need to know the individual diffusion coefficients.

Tracer Diffusion and Self-diffusion

Consider a binary alloy A-B. One could measure the diffusion of B by adding some radioactive B, a so-called tracer of B. The radioactive atoms would then diffuse into the inactive alloy and mix with the inactive B atoms there. The rate of diffusion depends upon the concentration gradient of the radioactive B atoms, $\partial c^*_B/\partial z$. We denote the corresponding diffusion constant by D^*_B and the mobility by M^*_B. Since the tracer is added in very low amounts, it may be considered as a dilute solution and Equation 12 holds; i.e., we have

$$D^*_B = M^*_B RT \tag{34}$$

From a chemical point of view, the radioactive B atoms are almost identical to the inactive B atoms, and their mobilities should be very similar:

$$M^*_B = M_B \tag{35}$$

We thus obtain with high accuracy

$$D_B = \frac{D^*_B}{RT}\frac{x_B}{V_m}\frac{d\mu_B}{dc_B} \tag{36}$$

It must be noted that these two diffusion constants have different character. Here, the diffusion coefficient D^*_B describes how the radioactive and inactive B atoms mix with each other and is called the tracer diffusion coefficient. If the experiment is performed by adding the tracer to pure B, D^*_B will describe the diffusion of the inactive B atoms as well as the radioactive ones. It is justified to

assume that the same amount of inactive B atoms will flow in the opposite direction to the flow of the radioactive B atoms. Thus, in pure B, D_B^* concerns self-diffusion and is called the "self-diffusion coefficient." More generally we may denote the intrinsic diffusion of B in pure B as the self-diffusion of B.

DIFFUSION IN MULTICOMPONENT ALLOYS: FICK-ONSAGER LAW

For diffusion in multicomponent solutions, a first attempt of representing the experimental data may be based on the assumption that Fick's law is valid for each diffusing substance k, i.e.,

$$J_k = -D_k \frac{\partial c_k}{\partial z} \tag{37}$$

The diffusivities D_k may then be evaluated by taking the ratio between the flux and concentration gradient of k. However, when analyzing experimental data, it is found, even in rather simple systems, that this procedure results in diffusivities that also depend on the concentration gradients. This would make Fick's law useless for all practical purposes, and a different multicomponent formulation must be found. This problem was first solved by Onsager (1945), who suggested the following multicomponent extension of Fick's law:

$$J_k = -\sum_{j}^{n-1} D_{kj} \frac{\partial c_j}{\partial z} \tag{38}$$

We can rewrite this equation if instead we introduce the chemical potential of the various species μ_k and assume that the μ_k are unique functions of the composition. Introducing a set of parameters L_{kj}, we may write

$$J_k = -\sum_i L_{ki} \frac{\partial \mu_i}{\partial z} = -\sum_i \sum_j L_{ki} \frac{\partial \mu_i}{\partial c_j} \frac{\partial c_j}{\partial z} \tag{39}$$

and may thus identify

$$D_{kj} = \sum_i L_{ki} \frac{\partial \mu_i}{\partial c_j} \tag{40}$$

Onsager's extension of Fick's law includes the possibility that the concentration gradient of one species may cause another species to diffuse. A well-known experimental demonstration of this effect was given by Darken (1949), who studied a welded joint between two steels with initially similar carbon contents but quite different silicon contents. The welded joint was heat treated at a high temperature for some time and then examined. Silicon is substitutionally dissolved and diffuses very slowly despite the strong concentration gradient. Carbon, which was initially homogeneously distributed, dissolves interstitially and diffuses much faster from the silicon-rich to the silicon-poor side. In a thin region, the carbon diffusion is actually opposite to what is expected from Fick's law in its original formulation, so-called up-hill diffusion.

Equation 10 may be regarded as a special case of Equation 40 if we assume that it applies also in multicomponent systems:

$$J_k = -M_k \frac{x_k}{V_m} \frac{\partial \mu_k}{\partial z} \tag{41}$$

Onsager went one step further and suggested the general formulation

$$J_k = -\sum L_{ki} X_i \tag{42}$$

where J_k stands for any type of flux, e.g., diffusion of a species, heat, or electric charge, and X_i is a force, e.g., a gradient of chemical potential, temperature, or electric potential. The off-diagonal elements of the matrix of phenomenological coefficients L_{ki} represent coupling effects, e.g., diffusion caused by a temperature gradient, the "Soret effect," or heat flow caused by a concentration gradient, the "Dufour effect." Moreover, Onsager (1931) showed that the matrix L_{ki} is symmetric if certain requirements are fulfilled, i.e.,

$$L_{ki} = L_{ik} \tag{43}$$

In 1968 Onsager was given the Nobel Prize for the above relations, usually called the Onsager reciprocity relations.

FRAMES OF REFERENCE IN MULTICOMPONENT DIFFUSION

The previous discussion on transformation between different frames of reference is easily extended to systems with n components. In general, a frame of reference is defined by the relation

$$\sum_{k=1}^{n} a_k J'_k = 0 \tag{44}$$

and the transformation between two frames of reference is given by

$$J'_k = J_k - \frac{\vartheta}{V_m} x_k \qquad (k = 1, \ldots, n) \tag{45}$$

where the primed frame of reference moves with the migration rate ϑ relative to the unprimed frame of reference. Multiplying each equation with a_k and summing $k = 1, \ldots, n$ yield

$$\sum_{k=1}^{n} a_k J'_k = \sum_{k=1}^{n} a_k J_k - \frac{\vartheta}{V_m} \sum_{k=1}^{n} a_k x_k \tag{46}$$

Combination of Equations 44 and 46 yields

$$\frac{\vartheta}{V_m} = \sum_{k=1}^{n} \frac{a_k J_k}{a_m} \tag{47}$$

where we have introduced $a_m = \sum_{k=1}^{n} a_k x_k$. Equation 45 then becomes

$$J'_k = J_k - \frac{x_k}{a_m}\sum_{i=1}^{n} a_i J_i$$
$$= \sum_{i=1}^{n}\left(\delta_{ik} - \frac{x_k a_i}{a_m}\right) J_i \qquad (k = 1, \ldots, n) \qquad (48)$$

where the Kronecker delta $\delta_{ik} = 1$ when $i = k$ and $\delta_{ik} = 0$ otherwise.

If the diffusion coefficients are known in the unprimed frame of reference and we have arbitrarily chosen the concentration of component n as the dependent concentration, i.e.,

$$J_k = -\frac{1}{V_m}\sum_{j=1}^{n-1} D_{kj}^{n} \frac{\partial x_j}{\partial z} \qquad (49)$$

we may directly express the diffusivities in the primed framed of reference by the use of Equation 47 and obtain

$$D_{kj}^{n'} = \sum_{i=1}^{n}\left(\delta_{ik} - \frac{x_k a_i}{a_m}\right) D_{ij}^{n} \qquad (50)$$

It should be noticed that Equations 48 and 50 are valid regardless of the initial frame of reference. In other words, the final result does not depend on whether one transforms directly from, for example, the lattice-fixed frame of reference to the number-fixed frame of reference or if one first transforms to the volume-fixed frame of reference and subsequently from the volume-fixed frame of reference to the number-fixed frame of reference.

By inserting Equation 42 in Equation 48, we find a similar expression for the phenomenological coefficients of Equation 42:

$$L'_{kj} = \sum_{i=1}^{n}\left(\delta_{ik} - \frac{x_k a_i}{a_m}\right) L_{ij} \qquad (51)$$

As a demonstration, the transformations from the lattice-fixed frame of reference to the number-fixed frame of reference in a ternary system A-B-C are given as

$$D_{\mathrm{BB}}^{\mathrm{A}'} = \left[(1 - x_{\mathrm{B}}) D_{\mathrm{BB}}^{\mathrm{A}} - x_{\mathrm{B}} D_{\mathrm{AB}}^{\mathrm{A}} - x_{\mathrm{B}} D_{\mathrm{CB}}^{\mathrm{A}}\right]$$
$$D_{\mathrm{BC}}^{\mathrm{A}'} = \left[x_{\mathrm{B}}(1 - x_{\mathrm{B}}) D_{\mathrm{BC}}^{\mathrm{A}} - x_{\mathrm{B}} D_{\mathrm{AC}}^{\mathrm{A}} - x_{\mathrm{B}} D_{\mathrm{CC}}^{\mathrm{A}}\right]$$
$$D_{\mathrm{CB}}^{\mathrm{A}'} = \left[x_{\mathrm{C}}(1 - x_{\mathrm{C}}) D_{\mathrm{CC}}^{\mathrm{A}} - x_{\mathrm{C}} D_{\mathrm{AB}}^{\mathrm{A}} - x_{\mathrm{C}} D_{\mathrm{BB}}^{\mathrm{A}}\right]$$
$$D_{\mathrm{CC}}^{\mathrm{A}'} = \left[(1 - x_{\mathrm{C}}) D_{\mathrm{CC}}^{\mathrm{A}} - x_{\mathrm{C}} D_{\mathrm{AC}}^{\mathrm{A}} - x_{\mathrm{C}} D_{\mathrm{CC}}^{\mathrm{A}}\right] \qquad (52)$$

For a system with n components, we thus need $(n-1) \times (n-1)$ diffusion coefficients to obtain a general description of how the concentration differences evolve in time. Moreover, since the diffusion coefficients usually are functions not only of temperature but also of composition, it is almost impossible to obtain a complete description of a real system by experimental investigations alone, and a powerful method to interpolate and extrapolate the information is needed. This will be discussed under Modeling Diffusion in Substitutional and Interstitial Metallic Systems. However, both the $(n-1)n$ individual diffusion coefficients and the $(n-1)(n-1)$ interdiffusion coefficients can be calculated from n mobilities (see Equation 41) provided that the thermodynamic properties of the system are known.

CHANGE OF DEPENDENT CONCENTRATION VARIABLE

In practical calculations it is often convenient to change the dependent concentration variable. If the diffusion coefficients are known with one choice of dependent concentration variable, it then becomes necessary to recalculate them all. In a binary system the problem is trivial because $\partial x_{\mathrm{A}}/\partial z = -\partial x_{\mathrm{B}}/\partial z$ and a change in dependent concentration variable simply means a change in sign of the diffusion coefficients, i.e.,

$$D_{\mathrm{AB}}^{\mathrm{A}} = -D_{\mathrm{AA}}^{\mathrm{B}} \qquad (53)$$

$$D_{\mathrm{BB}}^{\mathrm{A}} = -D_{\mathrm{BA}}^{\mathrm{B}} \qquad (54)$$

In the general case we may derive a series of transformations as

$$J_k = -\frac{1}{V_m}\sum_{j=1}^{n} D_{kj}^{l} \frac{\partial x_j}{\partial z} \qquad (55)$$

Observe that we may perform the summation over all components if we postulate that

$$D_{kl}^{l} = 0 \qquad (56)$$

Since $\partial x_l/\partial z = -\sum_{j \neq l} \partial x_j/\partial z$, we may change from x_l to x_i as dependent concentration variable by rewriting Equation 55:

$$J_k = -\frac{1}{V_m}\sum_{j=1}^{n} (D_{kj}^{l} - D_{ki}^{l}) \frac{\partial x_j}{\partial z} \qquad (57)$$

i.e., we have

$$J_k = -\frac{1}{V_m}\sum_{j=1}^{n} D_{kj}^{i} \frac{\partial x_j}{\partial z} \qquad (58)$$

and we may identify the new set of diffusion coefficients as

$$D_{kj}^{i} = D_{kj}^{l} - D_{ki}^{l} \qquad (59)$$

It is then immediately clear that $D_{ki}^{i} = 0$, i.e., Equation 56 that was postulated must hold. We may further conclude that the binary transformation rules are obtained as special cases. Note that Equation 89 holds irrespective of the frame of reference used because it is a trivial mathematical consequence on the interdependence of the concentration variables.

MODELING DIFFUSION IN SUBSTITUTIONAL AND INTERSTITIAL METALLIC SYSTEMS

Frame of Reference and Concentration Variables

In cases where we have both interstitial and substitutional solutes, it is convenient to define a number-fixed frame of reference with respect to the substitutional atoms. In a system (A, B)(C, Va), where A and B are substitutionally dissolved atoms, C is interstitially dissolved, and Va denotes vacant interstitial sites, we define a number-fixed frame of reference from $J_A + J_B = 0$. To be consistent, we should then not use the mole fraction x as the concentration variable but rather the fraction u defined by

$$\begin{aligned} u_B &= \frac{x_B}{x_A + x_B} = \frac{x_B}{1 - x_C} \qquad u_A = 1 - u_B \\ u_C &= \frac{x_C}{x_A + x_B} = \frac{x_C}{1 - x_C} \end{aligned} \tag{60}$$

The fraction u is related to the concentration c by

$$c_k = \frac{x_k}{V_m} = \frac{u_k(1 - x_C)}{V_m} = \frac{u_k}{V_S} \tag{61}$$

where $V_S = V_m/(1 - x_C)$ is the molar volume counted per mole of substitutional atom and is usually more constant than the ordinary molar volume. If we approximate it as constant, we obtain

$$\frac{\partial c_k}{\partial z} = \frac{1}{V_S}\frac{\partial u_k}{\partial z} \tag{62}$$

Mobilities and Diffusivities

We shall now discuss how experimental data on diffusion may be represented in terms of simple models. As already mentioned, the experimental evidence suggests that diffusion in solid metals occurs mainly by thermally activated atoms jumping to nearby vacant lattice sites. In a random mixture of the various kinds of atoms $k = 1, 2, \ldots, n$ and vacant sites, the probability that an atom k is nearest neighbor with a vacant site is given by the product $y_k y_{Va}$, where y_k is the fraction of lattice sites occupied by k and y_{Va} the fraction of vacant lattice sites.

From absolute reaction rate theory, we may derive that the fluxes in the lattice-fixed frame of reference are given by

$$J_k = -y_k y_{Va}\frac{\Omega_{kVa}}{V_S}\left[\frac{\partial \mu_k}{\partial z} - \frac{\partial \mu_{Va}}{\partial z}\right] \tag{63}$$

where Ω_{kVa} represents how rapidly k jumps to a nearby vacant site. Assuming that we everywhere have thermodynamic equilibrium, at least locally, $\mu_{Va} = 0$, and thus Equation 63 becomes

$$J_k = -y_k y_{Va}\frac{\Omega_{kVa}}{V_S}\frac{\partial \mu_k}{\partial z} \tag{64}$$

The fraction of vacant substitutional lattice sites is usually quite small, and it is convenient to introduce the mobility $M_k = y_{Va}\Omega_{kVa}$. However, the fraction of vacant interstitial site is usually quite high, and if k is an interstitial solute, it is more convenient to set $M_k = \Omega_{kVA}$. Introducing $c_k = u_k/V_S$ and $u_k \cong y_k$, we have

$$J_k = -c_k M_k \frac{\partial \mu_k}{\partial z} = -L_{kk}\frac{\partial \mu_k}{\partial z} \tag{65}$$

if k is substitutional and

$$J_k = -c_k y_{Va} M_k \frac{\partial \mu_k}{\partial z} = -L_{kk}\frac{\partial \mu_k}{\partial z} \tag{66}$$

if k is interstitial. (The double index kk does not mean summation in Equations 65 and 66.)

When the thermodynamic properties are established, i.e., when the chemical potentials μ_i are known as functions of composition, Equation 40 may be used to derive a relation between mobilities and the diffusivities in the lattice-fixed frame of reference. Choosing component n as the dependent component and using u fractions as concentration variables, i.e., the functions $\mu_i(u_1, u_2, \ldots, u_{n-1})$, we obtain, for the intrinsic diffusivities,

$$D^n_{kj} = u_k M_k \frac{\partial \mu_k}{\partial u_j} \tag{67}$$

if k is substitutional and

$$D^n_{kj} = u_k y_{Va} M_k \frac{\partial \mu_k}{\partial u_j} \tag{68}$$

if k is interstitial.

One may change to a number-fixed frame of reference with respect to the substitutional atoms defined by

$$\sum_{k \in s} J'_k = 0 \tag{69}$$

where $k \in s$ indicates that the summation is performed over the substitutional fluxes only. Equation 50 yields

$$D^{n'}_{kj} = u_k M_k \frac{\partial \mu_k}{\partial u_j} - u_k \sum_{i \in s}^{n} u_i M_i \frac{\partial \mu_i}{\partial u_j} \tag{70}$$

when k is substitutional and

$$D^{n'}_{kj} = u_k y_{Va} M_k \frac{\partial \mu_k}{\partial u_j} - u_k \sum_{i \in s}^{n} u_i M_i \frac{\partial \mu_i}{\partial u_j} \tag{71}$$

when k is interstitial.

If we need to change the dependent concentration variable by means of Equation 59, it should be observed that we can never choose an interstitial component as the dependent one. From the definition of the u fractions, it is evident that the u fractions of the interstitial components are independent, whereas for the substitutional components we have

$$\sum_{k \in s} u_k = 1 \tag{72}$$

We thus obtain

$$D^i_{kj} = D^l_{kj} - D^l_{ki} \tag{73}$$

when j is substitutional and

$$D^i_{kj} = D^l_{kj} \tag{74}$$

when j is interstitial.

If the thermodynamic properties, the functions $\mu_i(u_1, u_2, \ldots, u_{n-1})$, are known, we may use Equations 67 to 74 to extract mobilities from experimental data on all the various diffusion coefficients. For example, if the tracer diffusion coefficient D^*_k is known, we have

$$M_k = \frac{D^*_k}{RT} \tag{75}$$

Temperature and Concentration Dependence of Mobilities

From absolute reaction rate arguments, we expect that the mobility obeys the following type of temperature dependence:

$$M_k = \frac{1}{RT}\nu\delta^2 \exp\left(-\frac{\Delta G^*_k}{RT}\right) \tag{76}$$

where ν is the atomic vibrational frequency of the order of magnitude 10^{13} s^{-1}; δ is the jump distance and of the same order of magnitude as the interatomic spacing, i.e., a few angstroms; and ΔG^*_k is the Gibbs energy activation barrier and may be divided into an enthalpy and an entropy part, i.e., $\Delta G^*_k = \Delta H^*_k - \Delta S^*_k T$. We may thus write Equation 76 as

$$M_k = \frac{1}{RT}\nu\delta^2 \exp\left(\frac{\Delta S^*_k}{R}\right)\exp\left(\frac{-\Delta H^*_k}{RT}\right) \tag{77}$$

If ΔH^*_k and ΔS^*_k are temperature independent, one obtains the Arrhenius expression

$$M_k = \frac{1}{RT} M^0_k \exp\left(-\frac{Q_k}{RT}\right) \tag{78}$$

where Q_k is the activation energy and $M^0_k = \nu\delta^2 \exp(\Delta S^*_k/R)$ is a preexponential factor. The experimental information may thus be represented by an activation energy and a preexponential factor for each component. In general, the experimental information suggests that both these quantities depend on composition. For example, the mobility of a Ni atom in pure Ni will generally differ from the mobility of a Ni atom in pure Al.

It seems reasonable that the entropy part of ΔG^*_k should obey the same type of concentration dependence as the enthalpy part. However, the division of ΔG^*_k into enthalpy and entropy parts is never trivial and cannot be made from the experimental temperature dependence of the mobility unless the quantity $\nu\delta^2$ is known, which is seldom the case. To avoid any difficulty arising from the division into enthalpy and entropy parts, it is suggested that the quantity $RT\ \ln(RTM_k)$ rather than M^0_k and Q_k should be expanded as a function of composition, i.e., the function

$$\begin{aligned} RT\ln(RTM_k) &= RT\ln(\nu\delta^2) - (\Delta H^*_k - \Delta S^*_k T) \\ &= RT\ln(M^0_k) - Q_k \end{aligned} \tag{79}$$

The above function may then be expanded by means of polynomials of the composition, i.e., for a substitutional solution we write

$$RT\ln(RTM_k) = \Phi_k = \sum_i x_i {}^0\Phi_{ki} + \sum_i \sum_{j>i} x_i x_j \Phi_{kij} \tag{80}$$

where ${}^0\Phi_{ki}$ is $RT\ln(RTM_k)$ in pure i in the structure under consideration. The function Φ_{kij} represents the deviation from a linear interpolation between the endpoint values; ${}^0\Phi_{ki}$ and Φ_{kij} may be arbitrary functions of temperature but linear functions correspond to the Arrhenius relation. For example, in a binary A-B alloy we write

$$RT\ln(RTM_A) = x_A {}^0\Phi_{AA} + x_B {}^0\Phi_{AB} + x_A x_B \Phi_{AAB} \tag{81}$$

$$RT\ln(RTM_B) = x_A {}^0\Phi_{BA} + x_B {}^0\Phi_{BB} + x_A x_B \Phi_{BAB} \tag{82}$$

It should be observed that the four parameters of the type ${}^0\Phi_{AA}$ may be directly evaluated if the self-diffusivities and the diffusivities in the dilute solution limits are known because then Equation 75 applies and we may write ${}^0\Phi_{AA} = RT\ln(D_A^{\text{in pure A}})$ and $\Phi_{AB} = RT\ \ln(D_A^{\text{in pure B}})$ and similar equations for the mobility of B.

In the general case, the experimental information shows that the parameters Φ_{kij} will vary with composition and may be represented by the so-called Redlich-Kister polynomial

$$\Phi_{kij} = \sum_{r=0}^{m} {}^r\Phi_{kij}(x_i - x_j)^r \tag{83}$$

For substitutional and interstitial alloys, the u fraction is introduced and Equation 80 is modified into

$$RT\ln(RTM_k) = \Phi_k = \sum_i \sum_j u_i u_j \left(\frac{1}{c}\right) {}^0\Phi_{kij} + \cdots \tag{84}$$

where i stands for substitutional elements and j for interstitials and c denotes the number of interstitial sites per substitutional site. The parameter ${}^0\Phi_{kij}$ stands for the value of $RT\ln(RTM_k)$ in the hypothetical case when there is only i on the substitutional sublattice and all the interstitial sites are occupied by j. Thus we have ${}^0\Phi_{ki} = {}^0\Phi_{ki\text{Va}}$.

Effect of Magnetic Order

The experimental information shows that ferromagnetic ordering below the Curie temperature lowers the mobility M_k by a factor $\Gamma^{\text{mag}}_k < 1$. This yields a strong deviation from the Arrhenius behavior. One may include the possibility of a temperature-dependent activation energy by the general definition

$$Q_k = -R\frac{\partial \ln(RTM_k)}{\partial(1/T)} \tag{85}$$

It is then found that there is a drastic increase in the activation energy in the temperature range around the Curie temperature T_C and the ferromagnetic state below T_C has a higher activation energy than the paramagnetic state at high temperatures. It is suggested (Jönsson, 1992) that this effect is taken into account by an extra term $RT \ln \Gamma_k^{mag}$ in Equation 79, i.e.,

$$RT \ln(RTM_k) = RT \ln(M_k^{0para}) - Q_k^{para} + RT \ln \Gamma_k^{mag} \quad (86)$$

The superscript para means that the preexponential factor and the activation energy are taken from the paramagnetic state. The magnetic contribution is then given by

$$RT \ln \Gamma_k^{mag} = (6RT - Q_k^{para})\alpha\xi \quad (87)$$

For body-centered cubic (bcc) alloys $\alpha = 0.3$, but for face-centered cubic (fcc) alloys the effect of the transition is small and $\alpha = 0$. The state of magnetic order at the temperature T under consideration is represented by ξ, which is unity in the fully magnetic state at low temperatures and zero in the paramagnetic state. It is defined as

$$\xi = \frac{\Delta H_T^{mag}}{\Delta H_0^{mag}} \quad (88)$$

where $\Delta H_T^{mag} = \int_T^\infty c_P^{mag} dT$ and c_P^{mag} is the magnetic contribution to the heat capacity.

Diffusion in B2 Intermetallics and Effect of Chemical Order

Tracer Diffusion and Mobility. Chemical order is defined as any deviation from a disordered mixture of atoms on a crystalline lattice. Like the state of magnetic order, the state of chemical order has a strong effect on diffusion. Most experimental studies are on intermetallics with the B2 structure, which may be regarded as two interwoven simple-cubic sublattices. At sufficiently high temperatures there is no long-range order and the two sublattices are equivalent; this structure is called A2, i.e., the normal bcc structure. Upon cooling, the long-range order onsets at a critical temperature, i.e., there is a second-order transition to the B2 state. Most of the experimental observations concern tracer diffusion in binary A2-B2 alloys, and the effect of chemical ordering is then very similar to that of magnetic ordering; see, e.g., the first studies on β-brass by Kuper et al. (1956). Close to the critical temperature there is a drastic increase in the activation energy defined as $Q_k = -R\,\partial(\ln D_k^*)/\partial(1/T)$, and in the ordered state it is substantially higher than in the disordered state. In practice, this means that the tracer diffusion and the corresponding mobility are drastically decreased by ordering.

In the theoretical analysis of diffusion in partially ordered systems, it is common, as in the case of disordered systems, to adopt the local equilibrium hypothesis. This means that both the vacancy concentration and the state of order are governed by the thermodynamic equilibrium for the composition and temperature under consideration. For a binary A-B system with A2-B2 ordering, Girifalco (1964) found that the ordering effect may be represented by the expression

$$Q_k = Q_k^{dis}(1 + \alpha_k s^2) \quad (89)$$

where $k =$ A or B, and Q_k^{dis} is the activation energy for tracer diffusion in the disordered state at high temperatures, α_k may be regarded as an adjustable parameter, and s is the long-range order parameter. For A2-B2, ordering s is defined as $y'_B - y''_B$, where y'_B and y''_B are the B occupancy on the first and second sublattices, respectively. For the stoichiometric composition $x_B = \frac{1}{2}$, it may be written as $s = 2y'_B - 1$.

The experimental verification of Equation 89 indicates that there is essentially no change in diffusion mechanism as the ordering onsets but a vacancy mechanism would also prevail in the partially ordered state. However, the details on the atomic level may be more complex. It has been argued that after a B atom has completed its jump to a neighboring vacant site, it is on the wrong sublattice; an antisite defect has been created. There is thus a deviation from the local equilibrium that must be restored in some way. A number of mechanisms involving several consecutive jumps have been suggested; see the review by Mehrer (1996). However, it should be noted that thermodynamic equilibrium is a statistical concept and does not necessarily apply to each individual atomic jump. For each B atom jumping to the wrong sublattice there may be, on average, a B atom jumping to the right sublattice, and as a whole the equilibrium is not affected.

In multicomponent alloys, several order parameters are needed, and Equation 89 does not apply. Using the site fractions y'_A, y'_B, y''_A, and y''_B and noting that they are interrelated by means of $y'_A + y'_B = 1$, $y''_A + y''_B = 1$, and $y'_B + y''_B = 2x_B$, i.e., there is only one independent variable, we have

$$\left[y'_A y''_B + y'_B y''_A - 2x_A x_B\right] = \frac{1}{2}s^2 \quad (90)$$

and Equation 89 may be recast into

$$Q_k = Q_k^{dis} + \Delta Q_k^{order}[y'_A y''_B + y'_B y''_A - 2x_A x_B] \quad (91)$$

where $\Delta Q_k^{order} + 2Q_k^{dis}\alpha_k$. It then seems natural to apply the following multicomponent generalization of Equations 89 and 91:

$$Q_k = Q_k^{dis} + \sum_i \sum_{i \neq j} \Delta Q_{kij}^{order}[y'_i y''_j - x_i x_j] \quad (92)$$

As already mentioned, there is no indication of an essential change in diffusion mechanism due to ordering, and Equation 64 is expected to hold also in the partially ordered case. We may thus directly add the effect of ordering to Equation 80:

$$RT \ln(RTM_k) = \Phi_k = \sum_i x_i{}^0\Phi_{ki} + \sum_i \sum_{j>i} x_i x_j \Phi_{kij} + \sum_i \sum_{i \neq j} \Delta Q_{kij}^{order}[y'_i y'_j - x_i x_j] \quad (93)$$

When Equation 93 is used, the site fractions, i.e., the state of order, must be obtained in some way, e.g., from experimental data or from a thermodynamic calculation.

Interdiffusion. As already mentioned, the experimental observations show that ordering decreases the mobility, i.e., diffusion becomes slower. This effect is also seen in the experimental data on interdiffusion, which shows that the interdiffusion coefficient decreases as ordering increases. In addition, there is a particularly strong effect at the critical temperature or composition where order onsets. This drastic drop in diffusivity may be understood from the thermodynamic factors $\partial\mu_k/\partial\mu_j$ in Equation 70. For the case of a binary substitutional system A-B, we may apply the Gibbs-Duhem equation, and Equation 70 may be recast into

$$D_{BB}^{A'} = [x_A M_B + x_B M_A] x_B \frac{\partial \mu_B}{\partial x_B} \qquad (94)$$

The derivative $\partial\mu_B/\partial x_B$ reflects the curvature of the Gibbs energy as a function of composition and drops to much lower values as the critical temperature or composition is passed and ordering onsets. In principle, the change is discontinuous but short-range order tends to smooth the behavior. The quantity $x_B\,\partial\mu_B/\partial x_B$ exhibits a maximum at the equiatomic composition and thus opposes the effect of the mobility, which has a minimum when there is a maximum degree of order. It should be noted that although M_A and M_B usually have their minima close to the equiatomic composition, the quantity $[x_A M_B + x_B M_A]$ generally does not except in the special case when the two mobilities are equal. This is in agreement with experimental observations showing that the minimum in the interdiffusion coefficient is displaced from the equiatomic composition in NiAl (Shankar and Seigle, 1978).

CONCLUSIONS: EXAMPLES OF APPLICATIONS

In this unit, we have discussed some general features of the theory of multicomponent diffusion. In particular, we have emphasized equations that are useful when applying diffusion data in practical calculations. In simple binary solutions, it may be appropriate to evaluate the diffusion coefficients needed as functions of temperature and composition and represent them by any type of mathematical functions, e.g., polynomials. When analyzing experimental data in higher order systems, where the number of diffusion coefficients increases drastically, this approach becomes very complicated and should be avoided. Instead, a method based on individual mobilities and the thermodynamic properties of the system has been suggested. A suitable code would then contain the following parts: (1) numerical solution of the set of coupled diffusion equations, (2) thermodynamic calculation of Gibbs energy and all derivatives needed, (3) calculation of diffusion coefficients from mobilities and thermodynamics, (4) a thermodynamic database, and (5) a mobility database.

This method has been developed by the present author and his group and has been implemented in the code DICTRA (Ågren, 1992). This code has been applied to many practical problems [see, e.g., the work on composite steels by Helander and Ågren (1997)] and is now commercially available from Thermo-Calc AB (www.thermo-calc.se).

LITERATURE CITED

Adda, Y. and Philibert, J. 1966. La Diffusion dans les Solides. Presses Universitaires de France, Paris.

Ågren, J. 1992. Computer simulations of diffusional reactions in complex steels. *ISIJ Int*. 32:291–296.

American Society of Metals (ASM). 1951. Atom Movements. ASM, Cleveland.

Carslaw, S. H. and Jaeger, C. J. 1946/1959. Conduction of Heat in Solids. Clarendon Press, Oxford.

Crank, J. 1956. The Mathematics of Diffusion. Clarendon Press, Oxford.

Darken, L. S. 1948. Diffusion, mobility and their interrelation through free energy in binary metallic systems. *Trans. AIME* 175:184–201.

Darken, L. S. 1949. Diffusion of carbon in austenite with a discontinuity in composition. *Trans. AIME* 180:430–438.

Girifalco, L. A. 1964. Vacancy concentration and diffusion in order-disorder alloys. *J. Phys. Chem. Solids* 24:323–333.

Helander, T. and Ågren, J. 1997. Computer simulation of multicomponent diffusion in joints of dissimilar steels. *Metall. Mater. Trans. A* 28A:303–308.

Jönsson, B. 1992. On ferromagnetic ordering and lattice diffusion—a simple model. *Z. Metallkd.* 83:349–355.

Jost, W. 1952. Diffusion in Solids, Liquids, Gases. Academic Press, New York.

Kirkaldy, J. S. and Young, D. J. 1987. Diffusion in the Condensed State. Institute of Metals, London.

Kuper, A. B., Lazarus, D., Manning, J. R., and Tomiuzuka, C. T. 1956. Diffusion in ordered and disordered copper-zinc. *Phys. Rev*. 104:1436–1541.

Mehl, R. F. 1936. Diffusion in solid metals. *Trans. AIME* 122:11–56.

Mehrer, H. 1996. Diffusion in intermetallics. *Mater. Trans. JIM* 37:1259–1280.

Onsager, L. 1931. Reciprocal relations in irreversible processes. II. *Phys. Rev*. 38:2265–2279.

Onsager, L. 1945/46. Theories and problems of liquid diffusion. *N.Y. Acad. Sci. Ann*. 46:241–265.

Roberts-Austen, W. C. 1896. *Philos. Trans. R. Soc. Ser. A* 187:383–415.

Shankar, S. and Seigle, L. L. 1978. Interdiffusion and intrinsic diffusion in the NiAl (δ) phase of the Al-Ni system. *Metall. Trans. A* 9A:1467–1476.

Smigelskas, A. C. and Kirkendall, E. O. 1947. Zinc diffusion in alpha brass. *Trans. AIME* 171:130–142.

KEY REFERENCES

Darken, 1948. See above.

Derives the well-known relation between individual coefficients and the chemical diffusion coefficient.

Darken, 1949. See above.

Experimentally demonstrates for the first time the thermodynamic coupling between diffusional fluxes.

A comprehensive textbook on multicomponent diffusion theory.

Onsager, 1931. See above.

Discusses recirprocal relations and derives them by means of statistical considerations.

JOHN ÅGREN
Royal Institute of Technology
Stockholm, Sweden

MOLECULAR-DYNAMICS SIMULATION OF SURFACE PHENOMENA

INTRODUCTION

Molecular-dynamics (MD) simulation is a well-developed numerical technique that involves the use of a suitable algorithm to solve the classical equations of motion for atoms interacting with a known interatomic potential. This method has been used for several decades now to illustrate and understand the temperature and pressure dependencies of dynamical phenomena in liquids, solids, and liquid-solid interfaces. Extensive details of this technique and its applications can be found in Allen and Tildesley (1987) and references therein.

MD simulation techniques are also well suited for studying surface phenomena, as they provide a qualitative understanding of surface structure and dynamics. This unit considers the use of MD techniques to better understand surface disorder and premelting. Specifically, it examines the temperature dependence of structure and vibrational dynamics at surfaces of face-centered cubic (fcc) metals—mainly Ag, Cu, and Ni. It also makes contact with results from other theoretical and experimental methods.

While the emphasis in this chapter is on metal surfaces, the MD technique has been applied over the years to a wide variety of surfaces including those of semiconductors, insulators, alloys, glasses, and simple or binary liquids. A full review of the pros and cons of the method as applied to these very interesting systems is beyond the scope of this chapter. However, it is worth pointing out that the success of the classical MD simulation depends to a large extent on the exactness with which the forces acting on the ion cores can be determined. On semiconductor and insulator surfaces, the success of *ab initio* molecular dynamics simulations has made them more suitable for such calulations, rather than classical MD simulations.

Understanding Surface Behavior

The interface with vacuum makes the surface of a solid very different from the material in the bulk. In the bulk, periodic arrangement of atoms in all directions is the norm, but at the surface such symmetry is absent in the direction normal to the surface plane. This broken symmetry is responsible for a range of striking features. Quantized modes, such as phonons, magnons, plasmons, and polaritons, have surface counterparts that are localized in the surface layers—i.e., the maximum amplitude of these excitations lies in the top few layers.

A knowledge of the distinguishing features of surface excitations provides information about the nature of the bonds between atoms at the surface and about how these bonds differ from those in the bulk. The presence of the surface may lead to a rearrangement or reassignment of bonds between atoms, to changes in the local electronic density of states, to readjustments of the electronic structure, and to characteristic hybridizations between atomic orbitals that are quite distinct from those in the bulk solid.

The surface-induced properties in turn may produce changes in the atomic arrangement, the reactivity, and the dynamics of the surface atoms. Two types of changes in the atomic geometry are possible: an inward or outward shift of the ionic positions in the surface layers, called surface relaxation, or a rearrangement of ionic positions in the surface layer, called surface reconstruction. Certain metallic and semiconductor surfaces show striking reconstruction and appreciable relaxation, while others may not exhibit a pronounced effect. Well-known examples of surface reconstruction include the 2×1 reconstruction of Si(100), consisting of buckled dimers (Yang et al., 1983); the "missing row" reconstruction of the (110) surface of metals, such as Pt, Ir, and Au (Chan et al., 1980; Binnig et al., 1983; Kellogg, 1985); and the "herringbone" reconstruction of Au(111) (Perdereau et al., 1974; Barth et al., 1990).

The rule of thumb is that the more open the surface, the more the propensity to reconstruct or relax; however, not all open surfaces reconstruct. The elemental composition of the surface may be a determining factor in such structural transitions. On some solid surfaces, structural changes may be initiated by an absorbed overlayer or other external agent (Onuferko et al., 1979; Coulman et al. 1990; Sandy et al., 1992).

Whether a surface relaxes or reconstructs, the observed features of surface vibrational modes demonstrate changes in the force fields in the surface region (Lehwald et al., 1983). The changes in surface force fields are sensitive to the details of surface geometry, atomic coordination, and elemental composition. They also reflect the modification of the electronic structure at surfaces, and subsequently of the chemical reactivity of surfaces. A good deal of research on understanding the structural electronic properties and the dynamics at solid surfaces is linked to the need to comprehend technologically important processes, such as catalysis and corrosion.

Also, in material science and in nanotechnology, a comprehensive understanding of the characteristics of solid surfaces at the microscopic level is critical for purposes such as the production of good-quality thin films that are grown epitaxially on a substrate. Although technological developments in automation and robotics have made the industrial production of thin films and wafers routine, the quality of films at the nanometer level is still not guaranteed.

To control film defects, such as the presence of voids, cracks, and roughness, on the nanometer scale, it is neces-

sary to understand atomic processes including sticking, attachment, detachment, and diffusion of atoms, vacancies, and clusters. Although these processes relate directly to the mobility and the dynamics of atoms impinging on the solid surface (the substrate), the substrate itself is an active participant in these phenomena. For example, in epitaxial growth, whether the film is smooth or rough depends on whether the growth process is layer-by-layer or in the form of three-dimensional clusters. This, in turn, is controlled by the relative magnitudes of the diffusion coefficients for the motion of adatoms, dimers, and atomic clusters and vacancies on flat terraces and along steps, kinks, and other defects generally found on surfaces. Thus, to understand the atomistic details of thin film growth, we need to also develop an understanding of the temperature-dependent behavior of solid surfaces.

Surface Behavior Varies with Temperature

The diffusion coefficients, as well as the structure, dynamics, and stability of the surface, vary with surface temperature. At low temperatures the surface is relatively stable and the mobility of atoms, clusters, and vacancies is limited. At high temperatures, increased mobility may allow better control over growth processes, but it may also cause the surface to disorder. The distinction between low and high temperatures depends on the material, the surface orientation, and the presence of defects such as steps and kinks.

Systematic studies of the structure and dynamics of flat, stepped, and kinked surfaces have shown that the onset of enhanced anharmonic vibrations leads not just to surface disordering (Lapujoulade, 1994) but to several other striking phenomena, whose characteristics depend on the metal, the geometry, and the surface atomic coordination. As disordering continues with increasing temperature, there comes a stage at which certain surfaces undergo a structural phase transition called roughening (van Biejeren and Nolden, 1987). With further increase in temperature, some surfaces premelt before the bulk melting temperature is reached (Frenken and van der Veen, 1985).

Existence of surface roughening has been confirmed by experimental studies of the temperature variation of the surface structure of several elements: Cu(110) (Zeppenfeld et al., 1989; Bracco, 1994), Ag(110) (Held et al., 1987; Bracco et al., 1993), and Pb(110) (Heyraud and Metois, 1987; Yang et al., 1989), as well as several stepped metallic surfaces (Conrad and Engel, 1994). Surface premelting has so far been confirmed only on two metallic surfaces: Pb(110) (Frenken and van der Veen, 1985) and Al(110) (Dosch et al., 1991). Additionally, anomalous thermal behavior has been observed in diffraction measurements of Ni(100) (Cao and Conrad, 1990a), Ni(110) (Cao and Conrad, 1990b) and Cu(100) (Armand et al., 1987).

Computer-simulation studies using molecular-dynamics (MD) techniques provide a qualitative understanding of the processes by which these surfaces disorder and premelt (Stoltze, et al., 1989; Barnett and Landman, 1991; Loisel et al., 1991; Yang and Rahman, 1991; Yang et al., 1991; Hakkinen and Manninen, 1992; Beaudet et al., 1993; Toh et al., 1994). One MD study demonstrated the appearance of the roughening transition on Ag(110) (Rahman et al., 1997). Interestingly, certain other surfaces, such as Pb(111), have been reported to superheat, i.e., maintain their structural order even beyond the bulk melting temperature (Herman and Elsayed-Ali, 1992).

Experimental and theoretical studies of metal surfaces have revealed trends in the temperature dependence of structural order and how this relationship varies with the bulk melting temperature of the solid. Geometric and orientational effects may also play an important role in determining the temperature at which disorder sets in. Less close-packed surfaces, or surfaces with regularly spaced steps and terraces, called vicinal surfaces, may undergo structural transformations at temperatures lower than those observed for their flat counterparts. There are, of course, exceptions to this trend, and surface geometry alone is not a good indicator of temperature-driven transformations at solid surfaces. Regardless, the propensity of a surface to disorder, roughen, or premelt is ultimately related to the strength and extent of the anharmonicity of the interatomic bonds at the surface.

Experimental and theoretical investigations in the past two decades have revealed a variety of intriguing phenomena at metal surfaces. An understanding of these phenomena is important for both fundamental and technological reasons: it provides insight about the temperature dependence of the chemical bond between atoms in regions of reduced symmetry, and it aids in the pursuit of novel materials with controlled thermal behavior.

Molecular Dynamics and Surface Phenomena

Simply speaking, the studies of surface phenomena may be classified into studies of surface structure and studies of surface dynamics. Studies of surface structure imply a knowledge of the equilibrium positions of atoms corresponding to a specific surface temperature. As mentioned above, one concern is surface relaxation, in which the spacings between atoms in the topmost layers at their 0 K equilibrium configuration may be different from that in the bulk-terminated positions. To explore thermal expansion at the surface, this interlayer separation is examined as a function of the surface temperature. Studies of surface structure also include examination of quantities like surface stress and surface energetics: surface energy (the energy needed to create the surface) and activation barriers (the energy needed to move an atom from one position to another).

In surface dynamics, the focus is on atomic vibrations about equilibrium positions. This involves investigation of surface phonon frequencies and their dispersion, mean-square atomic vibrational amplitudes, and vibrational densities of states. At temperatures for which the harmonic approximation is valid and phonons are eigenstates of the system, knowledge of the vibrational density of states provides information on the thermodynamic properties of surfaces: for example, free energy, entropy, and heat capacity. At higher temperatures, anharmonic contributions to the interatomic potential become important; they manifest themselves in thermal expansion, in the

softening of phonon frequencies, in the broadening of phonon line-widths, and in a nonlinear variation of atomic vibrational amplitudes and mean-square displacements with respect to temperature.

In the above description we have omitted mention of some other intriguing structural and dynamical properties of solid surfaces, e.g., that of a roughening transition (van Biejeren and Nolden, 1987) on metal surfaces, or a deroughening transition on amorphous metal surfaces (Ballone and Rubini, 1996), which have also been examined by the MD technique. Amorphous metal surfaces are striking, as they provide a cross-over between liquidlike and solidlike behavior in structural, dynamical, and thermodynamical properties near the glass transition. Another surface system of potential interest here is that of alloys that exhibit peculiarities in surface segregation. These surfaces offer challenges to the application of MD simulations because of the delicate balances in the detailed description of the system necessary to mimic experimental observations.

The discussion below summarizes a few other simulation and theoretical methods that are used to explore structure and dynamics at metal surfaces, and provides details of some molecular-dynamics studies.

Related Theoretical Techniques

The accessibility of high-powered computers, together with the development of simulation techniques that mimic realistic systems, have led in recent years to a flurry of studies of the structure and dynamics at metal surfaces. Theoretical techniques range from those based on first-principles electronic-structure calculations to those utilizing empirical or semiempirical model potentials.

A variety of first-principles or *ab initio* calculations are familiar to physicists and chemists. Studies of the structure and dynamics of surfaces are generally based on density functional theory in the local density approximation (Kohn and Sham, 1965). In this method, quantum mechanical equations are solved for electrons in the presence of ion cores. Calculations are performed to determine the forces that would act on the ion cores to relax them to their minimum energy equilibrium position, corresponding to 0 K. Quantities such as surface relaxation, stress, surface energy, and vibrational dynamics are then obtained with the ion cores at the minimum energy equilibrium configuration. These calculations are very reliable and provide insight into electronic and structural changes at surfaces. Accompanied by lattice-dynamics methods, these *ab initio* techniques have been successful in accurately predicting the frequencies of surface vibrational modes and their displacement patterns. However, this approach is computationally very demanding and the extension of the method to calculate properties of solids at finite temperatures is still prohibitive. Readers are referred to a review article (Bohnen and Ho, 1993) for further details and a summary of results.

First-principles electronic-structure calculations have been enhanced in recent years (Payne et al., 1992) by the introduction of new schemes (Car and Parrinello, 1985) for obtaining solutions to the Kohn-Sham equations for the total energy of the system. In an extension of the *ab initio* techniques, classical equations of motion are solved for the ion cores with forces obtained from the self-consistent solutions to the electronic-structure calculations. This is the so-called first-principles molecular-dynamics method. Ideally, the forces responsible for the motion of the ion cores in a solid are evaluated at each step in the calculation from solutions to the Hamiltonian for the valence electrons. Since there are no adjustable parameters in the theory, this approach has good predictive power and is very useful for exploring the structure and dynamics at any solid surface. However, several technical obstacles have generally limited its applicability to metallic systems. Until improved *ab initio* methods become feasible for use in studies of the dynamics and structure of systems with length and time scales in the nano regime, finite temperature studies of metal surfaces will have to rely on model interaction potentials.

Studies of structure and dynamics using model, many-body interaction potentials are performed either through a combination of energy minimization and lattice-dynamics techniques, as in the case of first-principles calculations, or through classical molecular-dynamics simulations. The lattice-dynamics method has the advantage that once the dynamical force constant matrix is extracted from the interaction potential, calculation of the vibrational properties is straightforward. The method's main drawback is in the use of the harmonic or the quasi-harmonic approximation, which restricts its applicability to temperatures of roughly one-half the bulk melting temperature. Nevertheless, the method provides a reasonable way to analyze experimental data on surface phonons (Nelson et al., 1989; Yang et al., 1991; Kara et al., 1997a). Moreover, it provides a theoretical framework for calculating the temperature dependence of the diffusion coefficient for adatom and vacancy motion along selected paths on surfaces (Kürpick et al., 1997). These calculations agree qualitatively with experimental results. Also, this theoretical approach allows a linkage to be established between the calculated thermodynamic properties of vicinal surfaces (Durukanoglu et al., 1997) and the local atomic coordination and environment.

The lattice dynamical approach, being quasi-analytic, provides a good grasp of the physics of systems and insights into the fundamental processes underlying the observed surface phenomena. It also provides measures for testing results of molecular-dynamics simulations for surface dynamics at low temperatures.

PRINCIPLES AND PRACTICE OF MD TECHNIQUES

Unprecedented developments in computer and computational technology in the past two decades have provided a new avenue for examining the characteristics of systems. Computer simulations now routinely supplement purely theoretical and experimental modes of investigation by providing access to information in regimes inaccessible otherwise. The main idea behind MD simulations is to mimic real-life situations on the computer. After preparing a sample system and determining the fundamental forces

that are responsible for the microscopic interactions in the system, the system is first equilibrated to the temperature of interest. Next, the system is allowed to evolve in time through time steps of magnitude relevant to the characteristics of interest. The goal is to collect substantial statistics of the system over a reasonable time interval, so as to provide reliable average values of the quantities of interest. Just as in experiments, the larger the gathered statistics, the smaller is the effect of random noise. The statistics that emerge at the end of an MD simulation consist of the phase space (positions and velocities) description of the system at each time step. Suitable correlation functions and other averaged quantities are then calculated to provide information on the structural and dynamical properties of the system.

For the simulation of surface phenomena, MD cells consisting of a few thousand atoms arranged in several (ten to twenty) layers are sufficient. Some other analytical applications—for example, the study of the propagation of cracks in solids—may use classical MD simulations with millions of atoms in the cell (Zhou et al., 1997). The minimum size of the cell depends very much on the nature of the dynamical property of interest. In studying surface phenomena, periodic boundary conditions are applied in the two directions parallel to the surface while no such constraint is imposed in the direction normal to it. For the results presented here, Nordsieck's algorithm with a time step of 10^{-15} s is used to solve Newton's equations for all the atoms in the MD cell.

For any desired temperature, a preliminary simulation for the bulk system (in this case, 256 particles arranged in an fcc lattice with periodic boundary conditions applied in all three directions) is carried out under conditions of constant temperature and constant pressure (NPT) to obtain the lattice constant at that temperature. A system with a particular surface crystallographic orientation is then generated using the bulk-terminated positions. Under conditions of constant volume and constant temperature (NVT), this surface system is equilibrated to the desired temperature, usually obtained in simulations of ~10 to 20 ps, although longer runs are necessary at temperatures close to a transition. Next, the system is isolated from its surroundings and allowed to evolve in a much longer run, anywhere from hundreds of picoseconds to tens of nanoseconds, while its total energy remains constant (a microcanonical or NVE ensemble). The output of the MD simulation is the phase space information of the system—i.e., the positions and velocities of all atoms at every time step. Statistics on the positions and velocities of the atoms are recorded. All information about the structural properties and the dynamics of the system is obtained from appropriate correlation functions, calculated from recorded atomic positions and velocities.

An essential ingredient of the above technique is the interatomic interaction potentials that provide the forces with which atoms move. Semiempirical potentials based on the embedded-atom method (EAM; Foiles et al., 1986) have been used. These are many-body potentials, and therefore do not suffer from the unrealistic constraints that pair-potentials would impose on the elastic constants and energetics of the system.

For the six fcc metals (Ni, Pd, Pt, Cu, Ag, and Au) and their intermetallics, the simulations using many-body potentials reproduce many of the observed characteristics of the bulk metal (Daw et al., 1993). More importantly, the sensitivity of these interaction potentials to the local atomic coordination makes them suitable for use in examining the structural properties and dynamics of the surfaces of these metals. The EAM is one of a genre of realistic many-body potentials methods that have become available in the past 15 years (Finnis and Sinclair, 1984; Jacobsen et al., 1987; Ercolessi et al, 1988).

The choice of the EAM to derive interatomic potentials introduces some ambiguity into the calculated values of physical properties, but this is to be expected as these are all model potentials with a set of adjustable parameters. Whether a particular choice of interatomic potentials is capable of characterizing the temperature-dependent properties of the metallic surface under consideration can only be ascertained by performing a variety of tests and comparing the results to what is already known experimentally and theoretically. The discussion below summarizes some of the essentials of these potentials; the original papers provide further details.

The EAM potentials are based on the concept that the solid is assembled atom by atom and the energy required to embed an atom in a homogeneous electron background depends only on the electron density at that particular site. The total internal energy of a solid can thus be written as in Equation 1:

$$E_{\text{tot}} = \sum_i F_i(\rho_{h,i}) + \frac{1}{2}\sum_{i,j} \phi_{ij}(R_{ij}) \tag{1}$$

$$\rho_{h,i} = \sum_{j \neq i} f_j(R_{ij}) \tag{2}$$

where $\rho_{h,i}$ is the electron density at atom i due to all other (host) atoms, f_j is the electron density of atom j as a function of distance from its center, R_{ij} is the distance between atoms i and j, F_i is the energy required to embed atom i into the homogeneous electron density due to all other atoms, and ϕ_{ij} is a short-range pair potential representing an electrostatic core-core repulsion.

In Equation 2 a simplifying assumption connects the host electron density to a superposition of atomic charge densities. It makes the local charge densities dependent on the local atomic coordination. It is not clear *a priori* how realistic this assumption is, particularly in regions near a surface or interface. However, with the assumption, the calculations turn out to be about as computer intensive as those based on simple pair potentials. Another assumption implicit in the above equations is that the background electron density is a constant. Even with these two approximations, it is not yet feasible to obtain the embedding energy from first principles. These functions are obtained from a fit to a large number of experimentally observed properties, such as the lattice constants, the elastic constants, heats of sublimation, vacancy formation energies, and bcc-fcc energy differences.

DATA ANALYSIS AND INTERPRETATION

Structure and Dynamics at Room Temperature

To demonstrate some results of MD simulations, the three low-Miller-index surfaces (100), (110), and (111) of fcc solids are considered. The top layer geometry and the associated two-dimensional Brillouin zones and crystallographic axes are displayed in Figure 1. The z axis is in each case normal to the surface plane. This section presents MD results for the structure and dynamics of these surfaces for Ag, Cu, and Ni at room temperature.

As anharmonic effects are small at room temperature, the MD results can be compared with available results from other calculations, most of which are based on the harmonic approximation in solids. Room temperature also provides a good opportunity to compare MD results with a variety of experimental data that are commonly taken at room temperature. In addition, at room temperature quantum effects are not expected to be important. The discussion below begins by exploring anisotropy in the mean-square vibrational amplitudes at the (100), (110), and (111) surfaces. This is followed by an examination of the interlayer relaxations at these surfaces.

Mean-Square Vibrational Amplitudes. The mean-square vibrational amplitudes of the surface and the bulk atoms are calculated using Equation 3:

$$\langle u_{j\alpha}^2 \rangle = \frac{1}{N_j} \sum_{i=1}^{N_j} \langle [r_{i\alpha}(t) - \langle r_{i\alpha}(t-\tau) \rangle \tau]^2 \rangle \qquad (3)$$

where the left-hand side represents the Cartesian component α of the mean-square vibrational amplitudes of atoms in layer j, and $r_{i\alpha}$ the instantaneous position of atom i along α. The terms in brackets, $\langle \ldots \rangle$, are time averages with τ as a time interval. These calculated vibrational amplitudes provide useful information for the analysis of structural data on surfaces. Without the direct information provided by MD simulations, values for surface vibrational amplitudes are estimated from an assigned surface Debye temperature—an *ad hoc* notion in itself.

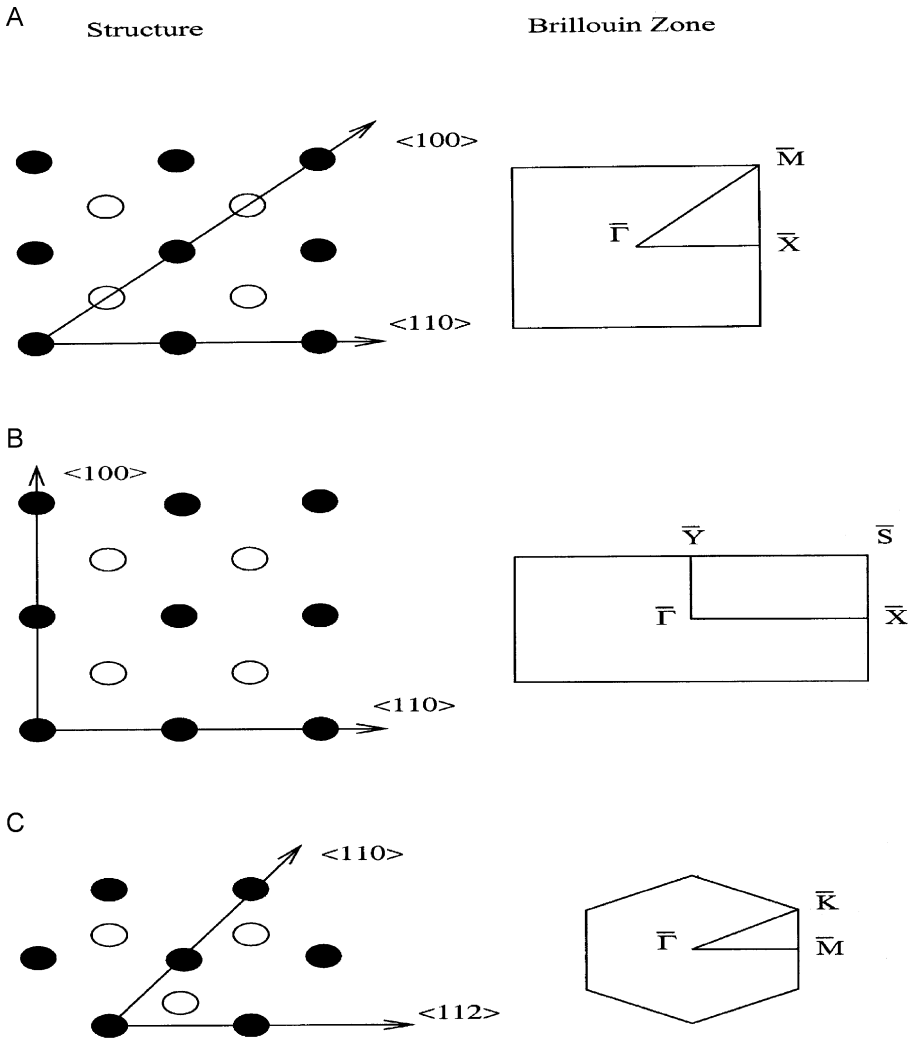


Figure 1. The structure, crystallographics directions, and surface Brillouin zone for the (**A**) (100), (**B**) (110), and (**C**) (111), surfaces of fcc crystals.

Table 1. Simulated Mean-square Displacements in Units of 10^{-2} A^2 at 300 K

	$\langle u_x^2 \rangle$	$\langle u_y^2 \rangle$	$\langle u_z^2 \rangle$
Ag(100)	1.38	1.38	1.52
Cu(100)	1.3	1.3	1.28
Ni(100)	0.7	0.7	0.81
Ag(110)	1.43	1.95	1.86
Cu(110)	1.13	1.95	1.31
Ni(110)	0.73	1.13	0.86
Ag(111)	1.10	1.17	1.59
Cu(111)	0.95	0.95	1.27
Ni(111)	0.55	0.58	0.94
Ag(bulk)	0.8	0.8	0.8
Cu(bulk)	0.7	0.7	0.7
Ni(bulk)	0.39	0.39	0.39

A few comments about mean-square vibrational amplitudes provide some background to the significance of this measurement. In the bulk metal, the mean-square vibrational amplitudes, $\langle u_{j\alpha}^2 \rangle$, have to be isotropic. However, at the surface, there is no reason to expect this to be the case. In fact, there are indications that the mean-square vibrational amplitudes of the surface atoms are anisotropic. Although experimental data for the in-plane vibrational amplitudes are difficult to extract, anisotropy has been observed in some cases. Moreover, as the force constants between atoms at the surface are generally different from those in the bulk, and as this deviation does not follow a simple pattern, it is intriguing to see if there is a universal relationship between surface vibrational amplitudes and those of an atom in the bulk. In view of these questions, it is also interesting to examine the values obtained from MD simulations for the three surfaces of Ag, Cu, and Ni at 300 K. These are summarized in Table 1.

The main message from Table 1 is that the vibrational amplitude normal to the surface (out of plane) in most cases is not larger than the in-plane vibrational amplitudes. This finding contradicts expectations from earlier calculations that did not take into account the deviations in surface force constants from bulk values (Clark et al., 1965). For the more open (110) surface, the amplitude in the $\langle 001 \rangle$ direction is larger than that in the other two directions. On the (100) surface, the amplitudes are almost isotropic, but on the closed-packed (111) surface, the out-of-plane amplitude is the largest. In each case, the surface amplitude is larger than that in the bulk, but the ratio depends on the metal and the crystallographic orientation of the surface.

It has not been possible to compare quantitatively the MD results with experimental data, mainly because of the uncertainties in the extraction of these numbers. Qualitatively, there is agreement with the results on Ag(100) (Moraboit et al., 1969) and Ni(100) (Cao and Conrad, 1990a), where the vibrational amplitudes are found to be isotropic. For Cu(100), though, the in-plane amplitude is reported to be larger than the out-of-plane amplitude by 30% (Jiang et al., 1991), while the MD results indicate that they are isotropic. On Ag(111), the MD results imply that the out-of-plane amplitude is larger, but experimental data suggest that it is isotropic (Jones et al., 1966). As the techniques are further refined and more experimental data become available for these surfaces, a direct comparison with the values in Table 1 could provide further insight into vibrational amplitudes on metal surfaces.

Interlayer Relaxation. Simulations of interlayer relaxation for the (100), (111), and (110) surfaces of Ag, Cu, and Ni using EAM potentials are found to be in reasonable agreement with experimental data, except for Ni(110) and Cu(110), for which the EAM underestimates the values. A problem with these comparisons, particularly for the (110) surface, is the wide range of experimental values obtained by various techniques. Different types of EAM potentials and other semiempirical potentials also yield a range of values for the interlayer relaxation. The reader is referred to the article by Bohnen and Ho (1993) for a comparison of the values obtained by various experimental and theoretical techniques.

Table 2 displays the values of the interlayer relaxations obtained in MD simulations at 300 K. In general, these values are reasonable and in line with those extracted from experimental data. For Cu(110), results available from first principles calculations show an inward relaxation of 9.2% (Rodach et al., 1993), compared to 4.73% in the MD simulations. This discrepancy is not surprising given the rudimentary nature of the EAM potentials. However, despite this discrepancy with interlayer relaxation, the MD-calculated values of the surface phonon frequencies are in very good agreement with experimental data and also with available first-principles calculations (Yang and Rahman 1991; Yang et al., 1991).

Structure and Dynamics at Higher Temperatures

This section discusses the role of anharmonic effects as reflected in the following: surface phonon frequency shifts and line-width broadening; mean-square vibrational amplitudes of the atoms; thermal expansion; and the onset of surface disordering. It also considers the appearance of premelting as the surface is heated to the bulk melting temperature.

Surface Phonons of Metals. When the MD/EAM approach is used, the surface phonon dispersion is in reasonable agreement with experimental data for most modes on

Table 2. Simulated Interlayer Relaxation at 300 K in Å[a]

Surface	d_{bulk}	d_{12}	$(d_{12} - d_{\text{bulk}})/d_{\text{bulk}}$
Ag(100)	2.057	2.021	−1.75%
Cu(100)	1.817	1.795	−1.21%
Ni(100)	1.76	1.755	−0.30%
Ag(110)	1.448	1.391	−3.94%
Cu(110)	1.285	1.224	−4.73%
Ni(110)	1.245	1.211	−2.70%
Ag(111)	2.376	2.345	−1.27%
Cu(111)	2.095	2.071	−1.11%
Ni(111)	2.04	2.038	−0.15%

[a] Here, d_{bulk} and d_{12} are the interlayer spacing in the bulk and between the top two layers, respectively.

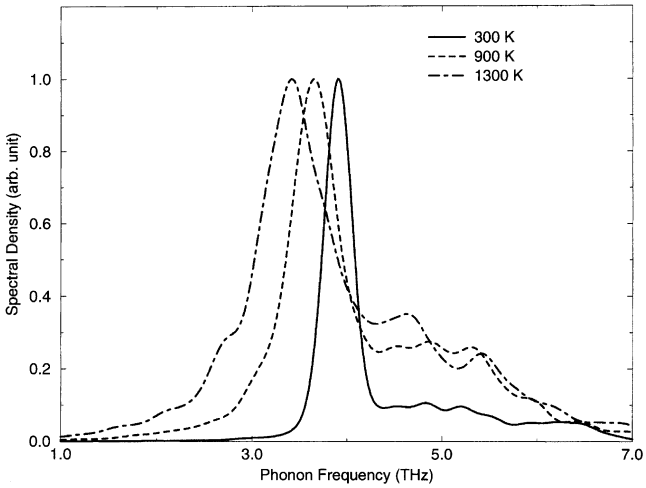


Figure 2. Phonon spectral densities representing the shear vertical mode at $\bar{M}$ on Ni(111) at three different temperatures (from Al-Rawi and Rahman, in press).

the (100), (110), and (111) surfaces of Ag, Cu, and Ni. In MD simulations, the phonon spectral densities in the one-phonon approximation can be obtained from Equation 4 (Hansen and Klein, 1976; Wang et al., 1988):

$$\rho_{j\alpha\alpha}(\vec{Q}_{||\tau}) = \int e^{i\omega t}\left[\sum_{i=1}^{N_j} e^{i\vec{Q}_{||}\vec{R}_i^{\alpha}}\langle v_i^{\alpha}(t)v_0^{\alpha}(0)\rangle\right]dt \qquad (4)$$

with

$$\langle v_2^{\alpha}(t)v_i^{\alpha}(0)\rangle = \frac{1}{MN_j}\sum_{m=1}^{M}\sum_{j=1}^{N_j} v_{i+j}^{\alpha}(t+z_m)v_i^{\alpha}(\tau_m)$$

where $\rho_{j\alpha\alpha}$ is the spectral density for the displacements along the direction $\alpha(=x,y,z)$ of the atoms in layer j, N_j is the number of atoms in layer j, $\vec{Q}_{||}$ is the two dimensional wave-vector parallel to the surface, and R_i^0 is the equilibrium position of atom i whose velocity is $\vec{v}_i$ and M is the number of MD steps.

Figure 2 shows the phonon spectral densities arising from the displacement of the atoms in the first layer with wave-vectors corresponding to $\bar{M}$ point in the Brillouin zone (see Fig. 1) for Ni(111). The spectral density for the shear vertical mode (the Rayleigh wave) is plotted for three different surface temperatures. The softening of the phonon frequency and the broadening of the line width with increasing temperatures attest to the presence of anharmonic effects (Maradudin and Fein, 1962). Such shifts in the frequency of a vibrational mode and the broadening of its line width also have been found for Cu(110) and Ag(110), both experimentally (Baddorf and Plummer, 1991; Bracco et al., 1996) and in MD simulations (Yang and Rahman, 1991; Rahman and Tian, 1993).

Mean-Square Atomic Displacements. The temperature dependence of the mean-square displacements of the top three layers of Ni(110), Ag(110), and Cu(110) shows very

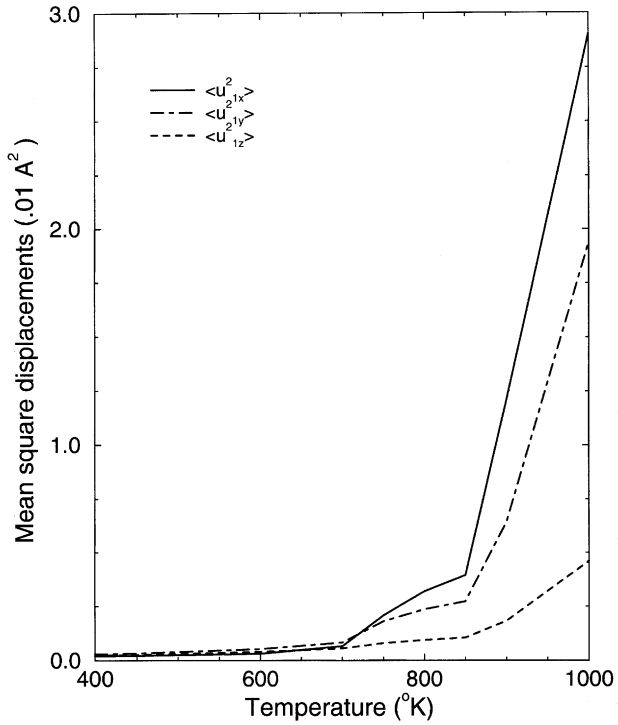


Figure 3. Mean-square displacement of top layer Ag(110) atoms (taken from Rahman et al., 1997).

similar behavior. In general, for the top layer atoms of the (110) surface, the amplitude along the y axis (the $\langle 001\rangle$ direction, perpendicular to the close-packed rows) is the largest at low temperatures. The in-plane amplitudes continue to increase with rising temperature; $\langle u_{1y}^2\rangle$ dominates until adatoms and vacancies are created. Beyond this point, $\langle u_{1x}^2\rangle$ becomes larger than $\langle u_{1y}^2\rangle$, indicating a preference for the adatoms to diffuse along the close-packed rows (Rahman et al., 1997). Figure 3 shows this behavior for the topmost atoms of Ag(110). Above 700 K, large anharmonic effects come into play, and adatoms and vacancies appear on the surface, causing the in-plane mean-square displacements to increase abruptly. Figure 3 also shows that the out-of-plane vibrational amplitude does not increase as rapidly as the in-plane amplitudes, indicating that interlayer diffusion is not as dominant as intralayer diffusion at medium temperatures.

The temperature variation of the mean-square displacements is a good indicator of the strength of anharmonic effects. For a harmonic system, $\langle u^2\rangle$ varies linearly with temperature. Hence, the slope of $\langle u^2\rangle/T$ can serve as a measure of anharmonicity. Figure 4 shows $\langle u^2\rangle/T$. This is plotted so that the relative anharmonicity in the three directions for the topmost atoms of Ag(110) can be examined over the range of 300 to 700 K. Once again, on the (110) surface, the $\langle 001\rangle$ direction is the most anharmonic and the vertical direction is the least so. The slight dip for $\langle u_{1z}^2\rangle$ at 450 K is due to statistical errors in the simulation results. The anharmonicity is largest along the y axis, at lower temperatures.

In experimental studies, such as those with low-energy electron diffraction (LOW-ENERGY ELECTRON DIFFRACTION), the

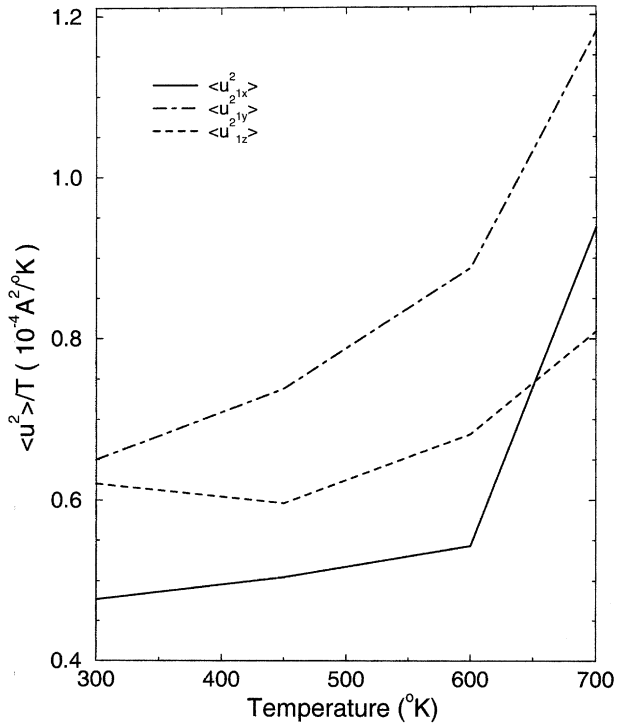


Figure 4. Anharmonicity in mean-square vibrational amplitudes for top layer atoms of Ag(110) (taken from Rahman et al., 1997).

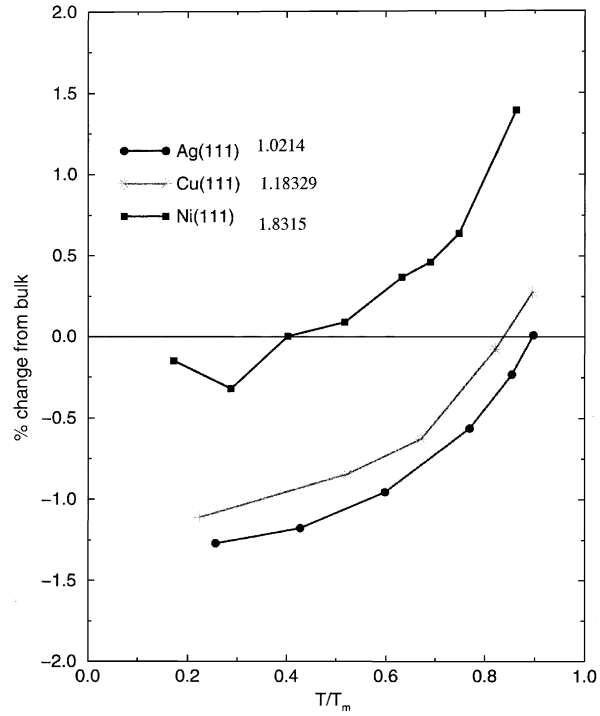


Figure 5. Thermal expansion of Ag(111), Cu(111), and Ni(111). From Al-Rawi and Rahman (in press). Here, T_m is the bulk melting temperature.

intensities of the diffracted beams are most sensitive to the vibrational amplitudes perpendicular to the surface (through the Debye-Waller factor). For several surfaces, the intensity of the beam appears to attenuate dramatically beginning at a specific temperature (Cao and Conrad, 1990a,b).

On both Cu(110) and Cu(100) surfaces, previous studies indicate the onset of enhanced anharmonic behavior at 600 K (Yang and Rahman, 1991; Yang et al., 1991). For Ag(110) the increment is more gradual, and there is already an onset of anharmonic effects at ~450 K. More detailed examination of the MD results show that adatoms and vacancies begin to appear at ~750 K on Ag(110) (Rahman et al., 1997) and ~850 K on Cu(110) (Yang and Rahman, 1991). However, the (100) and (111) surfaces of these metals do not display the appearance of a significant number of adatom/vacancy pairs. Instead, long-range order is maintained up to almost the bulk melting temperature (Yang et al., 1991; Hakkinen and Manninen, 1992; Al-Rawi et al., 2000).

A comparative experimental study of the onset of surface disordering on Cu(100) and Cu(110) using helium-atomic-beam diffraction (Gorse and Lapujoulade, 1985) adds credibility to the conclusions from MD simulations. This experimental work shows that although Cu(100) maintains its surface order until 1273 K (the bulk melting temperature), Cu(110) begins to disorder at ~500 K.

Surface Thermal Expansion. The molecular-dynamics technique is ideally suited to the examination of thermal expansion at surfaces, as it takes into account the full extent of anharmonicity of the interaction potentials. In the direction normal to the surface, the atoms are free to move and establish new equilibrium interlayer separation, although within each layer atoms occupy equilibrium positions as dictated by the bulk lattice constant for that temperature. Here again, the more open (110) surfaces of fcc metals show a propensity for appreciable surface thermal expansion over a given temperature range. For example, temperature-dependent interlayer relaxation on Ag(110) shows a dramatic change, from −4% at 300 K to + 4% at 900 K (Rahman and Tian, 1993). A similar increment (10% over the temperature range 300 to 1700 K) was found for Ni(110) in MD simulations (Beaudet et al., 1993) and also experimentally (Cao and Conrad, 1990b).

Results from MD simulations for the thermal expansion of the (111) surface indicate behavior very similar to that in the bulk. Figure 5 plots surface thermal expansion for Ag(111), Cu(111), and Ni(111) over a broad range of temperatures. Although all interlayer spacings expand with increasing temperature, that between the surface layers of Ag(111) and Cu(111) hardly reaches the bulk value even at very high temperatures. This result is different from those for the (110) surfaces of the same metals. Some experimental data also differ from the results presented in Figure 5 for Ag(111) and Cu(111). This is a point of ongoing discussion: ion-scattering experiments report a 10% increment over bulk thermal expansion at between 700 K and 1150 K (Statiris et al., 1994) for Ag(111), while very recent x-ray scattering data (Botez et al., in press)

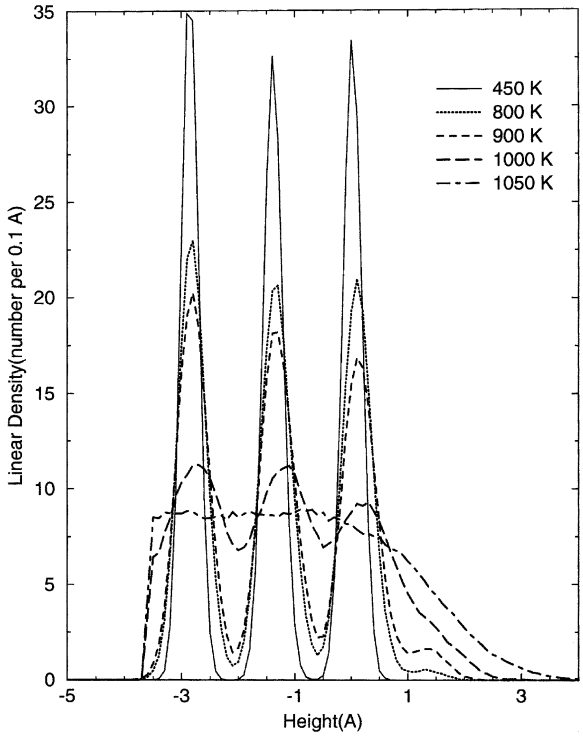


Figure 6. Time-averaged linear density of atoms initially in the top three layers of Ag(110).

and MD simulations do not display this sudden expansion at elevated temperature (Lewis, 1994; Kara et al., 1997b).

Temperature Variation of Layer Density. As the surface temperature rises, the number of adatoms and vacancies on the (110) surface increases and the surface begins to disorder. This tendency continues until there is no order maintained in the surface layers. In the case of Ag(110), the quasi-liquid-like phase occurs at ~1000 K. One measure of this premelting, the layer density for the atoms in the top three layers, is plotted in Figure 6.

At 450 K, the atoms are located in the first, second, and third layers. At ~800 K the situation is much the same, except that some of the atoms have moved to the top to form adatoms. The effect is more pronounced at 900 K, and is dramatic at 1000 K: the three peaks are very broad, with a number of atoms floating on top. At 1050 K, the three layers are submerged into one broad distribution. At this temperature, the atoms in the top three layers are diffusing freely over the entire extent of the MD cell. This is called premelting. The bulk melting temperature of Ag calculated with an EAM potential is 1170 K (Foiles et al., 1986).

The calculated structure factor for the system also indicates a sudden decrease in the layer-by-layer order. Surface melting is also defined by a sharp decline of the in-plane orientational order, at the characteristic temperature (Rahman et al., 1997; Toh et al., 1994). In contrast to the (110) surface, the (111) surface of the same metals do not appear to premelt (Al-Rawi and Rahman, in press), and vacancies and adatom pairs do not appear until close to the melting temperature.

PROBLEMS

MD simulations of surface phenomena provide valuable information on the temperature dependence of surface structure and dynamics. Such data may be difficult to obtain with other theoretical techniques. However, the MD method has limitations that may make its application questionable in some circumstances.

First, the MD technique is based on classical equations of motion, and therefore, it cannot be applied in cases where quantum effects are bound to play a significant role. Secondly, the method is not self-consistent. Model interaction potentials need to be provided as input to any calculation. The predictive power of the method is thus tied to the accuracy with which the chosen interaction potentials describe reality.

Thirdly, in order to depict phenomena at the ionic level, the time steps in MD simulations have to be quite a bit smaller than atomic vibrational frequencies ($\sim 10^{-13}$ s). Thus, an enormously large number of MD time steps are needed to obtain a one-to-one correspondence with events observed in the laboratory. Typically, MD simulations are performed in the nanosecond time scale and these results are extrapolated to provide information about events occurring in the microsecond and millisecond regime in the laboratory. Related to the limitation on time scale is the limitation on length scale. At best, MD simulations can be extended to millions of atoms in the MD cell. This is still a very small number compared to that in a real sample. Efforts are underway to extend the time scales in MD simulations by applying methods like the accelerated MD (Voter, 1997).

There is also a price to be paid for periodic boundary conditions that are imposed to remove edge effects from MD simulations. (These arise from the finite size of the MD cell.) Phenomena that change the geometry and shape of the MD cell, such as surface reconstruction, need extra effort and become more time consuming and elaborate.

Finally, to properly simulate any event, enough statistics have to be collected. This requires running the program a large number of times with new initial conditions. Even with these precautions it may not be possible to record all events that take place in a real system. Rare events are, naturally, the most problematic for MD simulations.

ACKNOWLEDGMENTS

This is work was supported in part by the National Science Foundation and the Kansas Center for Advanced Scientific Computing. Computations were performed on the Convex Exemplar multiprocessor supported in part by the NSF under grants No. DMR-9413513, and CDA-9724289. Many thanks to my co-workers A. Al-Rawi, A. Kara, Z. Tian, and L. Yang for their hard work, whose results form the basis of this unit.

LITERATURE CITED

Al-Rawi, A. and Rahman, T. S. Comparative Study of Anharmonic Effects on Ag(111), Cu(111) and Ni(111). In press.

Al-Rawi, A., Kara, A., and Rahman, T. S. 2000. Anharmonic effects on Ag(111): A molecular dynamics study. *Surf. Sci.* 446:17–30.

Allen, M. P. and Tildesley, D. J. 1987. Computer Simulation of Liquids. Oxford Science Publications.

Armand, G., Gorse, D., Lapujoulade, J., and Manson, J. R. 1987. The Dynamics of Cu(100) surface atoms at high temperature. *Europhys. Lett.* 3:1113–1118.

Baddorf, A. P. and Plummer, E. W. 1991. Enhanced surface anharmonicity observed in vibrations on Cu(110). *Phys. Rev. Lett.* 66:2770–2773.

Ballone, P. and Rubini, S. 1996. Roughening transition of an amorphous metal surface: A molecular dynamics study. *Phys. Rev. Lett*. 77:3169–3172.

Barnett, R. N. and Landman, U. 1991. Surface premelting of Cu(110). *Phys. Rev. B* 44:3226–3239.

Barth, J. V., Brune, H., Ertl, G., and Behm, R. J. 1990. Scanning tunneling microscopy observations on the reconstructed Au(111) surface: Atomic structure, long-range superstructure, rotational domains, and surface defects. *Phys. Rev. B* 42:9307–9318.

Beaudet, Y., Lewis, L. J., and Persson, M. 1993. Anharmonic effects at Ni(100) surface. *Phys. Rev. B* 47:4127–4130.

Binnig, G., Rohrer, H., Gerber, Ch., and Weibel, E. 1983. (111) Facets as the origin of reconstructed Au(110) surfaces. *Surf. Sci.* 131:L379–L384.

Bohnen, K. P. and Ho, K. M. 1993. Structure and dynamics at metal surfaces. *Surf. Sci. Rep.* 19:99–120.

Botez, C. E., Elliot, W. C., Miceli, P. F., and Stephens, P. W. Thermal expansion of the Ag(111) surface measured by x-ray scattering. In press.

Bracco, G. 1994. Thermal roughening of copper (110) surfaces of unreconstructed fcc metals. *Phys. Low-Dim. Struc.* 8:1–22.

Bracco, G., Bruschi, L., and Tatarek, R. 1996. Anomalous linewidth behavior of the S_3 surface resonance on Ag(110). *Europhys. Lett.* 34:687–692.

Bracco, G., Malo, C., Moses, C. J., and Tatarek, R. 1993. On the primary mechanism of surface roughening: The Ag(110) case. *Surf. Sci.* 287/288:871–875.

Cao, Y. and Conrad, E. 1990a. Anomalous thermal expansion of Ni(001). *Phys. Rev. Lett.* 65:2808–2811.

Cao, Y. and Conrad, E. 1990b. Approach to thermal roughening of Ni(110): A study by high-resolution low energy electron diffraction. *Phys. Rev. Lett.* 64:447–450.

Car, R. and Parrinello, M. 1985. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* 55:2471–2474.

Chan, C. M., VanHove, M. A., Weinberg, W. H., and Williams, E. D. 1980. An R-factor analysis of several models of the reconstructed Ir(110)-(1×2) surface. *Surf. Sci.* 91:440.

Clark, B. C., Herman, R., and Wallis, R. F. 1965. Theoretical mean-square displacements for surface atoms in face-centered cubic lattices with applications to nickel. *Phy. Rev.* 139:A860–A867.

Conrad, E. H. and Engel, T. 1994. The equilibrium crystals shape and the roughening transition on metal surfaces. *Surf. Sci.* 299/300:391–404.

Coulman, D. J., Wintterlin, J., Behm, R. J., and Ertl, G. 1990. Novel mechanism for the formation of chemisorption phases: The (2×1) O-Cu(110) "add-row" reconstruction. *Phys. Rev. Lett.* 64:1761–1764.

Daw, M. S., Foiles, S. M., and Baskes, M. I. 1993. The embedded-atom method: A review of theory and applications. *Mater. Sci. Rep.* 9:251.

Dosch, H., Hoefer, T., Pisel, J., and Johnson, R. L. 1991. Synchrotron x-ray scattering from the Al(110) surface at the onset of surface melting. *Europhys. Lett.* 15:527–533.

Durukanoglu, D., Kara, A., and Rahman, T. S. 1997. Local structural and vibrational properties of stepped surfaces: Cu(211), Cu(511), and Cu(331). *Phys. Rev. B* 55:13894–13903.

Ercolessi, F., Parrinello, M., and Tosatti, E. 1988. Simulation of gold in the glue model. *Phil. Mag. A* 58:213–226.

Finnis, M. W. and Sinclair, J. E. 1984. A simple empirical N-body potential for transition metals. *Phil. Mag. A* 50:45–55.

Foiles, S. M., Baskes, M. I., and Daw, M. S. 1986. Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt and their alloys. *Phys. Rev. B* 33:7983–7991.

Frenken, J. W. M. and van der Veen, J. F. 1985. Observation of surface melting. *Phys. Rev. Lett.* 54:134–137.

Gorse, D. and Lapujoulade, J. 1985. A study of the high temperature stability of low index planes of copper by atomic beam diffraction. *Surf. Sci.* 162:847–857.

Hakkinen, H. and Manninen, M. 1992. Computer simulation of disordering and premelting at low-index faces of copper. *Phys. Rev. B* 46:1725–1742.

Hansen, J. P. and Klein, M. L. 1976. Dynamical structure factor $S(\vec{q}, \omega)$] of rare-gas solids. *Phy. Rev. B*13:878–887.

Held, G. A., Jordan-Sweet, J. L., Horn, P. M., Mak, A., and Birgeneau, R. J. 1987. X–ray scattering study of the thermal roughening of Ag(110). *Phys. Rev. Lett.* 59:2075–2078.

Herman, J. W. and Elsayed-Ali, H. E. 1992. Superheating of Pb(111). *Phys. Rev. Lett.* 69:1228–1231.

Heyraud, J. C. and Metois, J. J. 1987. *J. Cryst. Growth* 82:269.

Jacobsen, K. W., Norskov, J. K., and Puska, M. J. 1987. Interatomic interactions in the effective-medium theory. *Phys. Rev. B* 35:7423.

Jiang, Q. T., Fenter, P., and Gustafsson, T. 1991. Geometric structure and surface vibrations of Cu(001) determined by medium-energy ion scattering. *Phys. Rev. B* 44:5773–5778.

Jones, E. R., McKinney, J. T., and Webb, M. B. 1966. Surface lattice dynamics of silver. I. Low-energy electron Debye-Waller factor. *Phys. Rev.* 151:476–483.

Kürpick, U., Kara A., and Rahman, T. S. 1997. The role of lattice vibrations in adatom diffusion. *Phys. Rev. Lett.* 78:1086–1089.

Kara, A., Durukanoglu, S., and Rahman, T. S. 1997a. Vibrational dynamics and thermodynamics of Ni(977). *J. Chem. Phys.* 106:2031–2037.

Kara, A., Staikov, P., Al-Rawi, A. N., and Rahman, T. S. 1997b. Thermal expansion of Ag(111). *Phys. Rev. B* 55:R13440–R13443.

Kellogg, G. L. 1985. Direct observations of the (1×2) surface reconstruction of the Pt(110) plane. *Phys. Rev. Lett.* 55:2168–2171.

Kohn, W. and Sham, L. J. 1965. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* 140:A1133–A1138.

Lapujoulade, J. 1994. The roughening of metal surfaces. *Surf. Sci. Rep.* 20:191–249.

Lehwald, S., Szeftel, J. M., Ibach, H., Rahman, T. S., and Mills, D. L. 1983. Surface phonon dispersion of Ni(100) measured by inelastic electron scattering. *Phys. Rev. Lett.* 50:518–521.

Lewis, L. J. 1994. Thermal relaxation of Ag(111). *Phys. Rev. B* 50:17693–17696.

Loisel, B., Lapujoulade, J., and Pontikis, V. 1991. Cu(110): Disorder or enhanced anharmonicity? A computer simulation study of surface defects and dynamics. *Surf. Sci.* 256:242–252.

Maradudin, A., and Fein, A. E. 1962. Scattering of neutrons by an anharmonic crystal. *Phys. Rev.* 128:2589–2608.

Moraboit, J. M., Steiger, R. F., and Somarjai, G. A. 1969. Studies of the mean displacement of surface atoms in the (100) silver single crystals at low temperature. *Phys. Rev.* 179:638–644.

Nelson, J. S., Daw, M. S., and Sowa, E. C. 1989. Cu(111) and Ag(111) Surface-phonon spectrum: The importance of avoided crossings. *Phys. Rev. B* 40:1465–1480.

Onuferko, J. H., Woodruff, D. P., and Holland, B. W. 1979. LEED structure analysis of the Ni(100) (2×2)C(p4g) structure: A case of adsorbate-induced substrate distortion. *Surf. Sci.* 87:357–374.

Payne, M. C., Teter, M. P., Allen, D. C., Arias, T. A., and Joannopoulos, J. D. 1992. Iterative minimization techniques for *ab initio*, total-energy calculations: Molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* 64:1045–1097.

Perdereau, J., Biberian, J. P., and Rhead, G. E. 1974. Adsorption and surface alloying of lead monolayers on (111) and (110) faces of gold. *J. Phys. F: Metal Phys.* 4:798.

Rahman, T. S. and Tian, Z. 1993. Anharmonic effects at metal surfaces. *J. Elec Spectros. Rel. Phenom.* 64/65:651–663.

Rahman, T. S., Tian, Z., and Black, J. E. 1997. Surface disordering, roughening and premelting of Ag(110). *Surf. Sci.* 374:9–16.

Rodach, T., Bohnen, K. P., and Ho, K. M. 1993. First principles calculation of lattice relaxation at low index surfaces of Cu. *Surf. Sci.* 286:66–72.

Sandy, A. R., Mochrie, S. G. J., Zehner, D. M., Grubel, G., Huang, K. G., and Gibbs, D. 1992. Reconstruction of the Pt(111) surface. *Phys. Rev. Lett.* 68:2192–2195.

Statiris, P., Lu, H. C., and Gustafsson, T. 1994. Temperature dependent sign reversal of the surface contraction of Ag(111). *Phys. Rev. Lett.* 72:3574–3577.

Stoltze, P., Norskov, J. K., and Landman, U. 1989. The onset of disorder in Al(110) surfaces below the melting point. *Surf. Sci. Lett.* 220:L693–L700.

Toh, C. P., Ong, C. K., and Ercolessi, F. 1994. Molecular-dynamics study of surface relaxation, stress, and disordering of Pb(110). *Phys. Rev. B* 50:17507.

van Biejeren, H. and Nolden, I. 1987. The roughening transition. *In* Topics in Current Physics, Vol. 23 (W. Schommers and P. von Blanckenhagen, eds.) p. 259. Springer-Verlag, Berlin.

Voter, A. 1997. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* 78:3908–3911.

Wang, C. Z., Fasolino, A., and Tosatti, E. 1988. Molecular-dynamics theory of the temperature-dependent surface phonons of W(001). *Phys. Rev.* B37:2116–2122.

Yang, L. and Rahman, T. S. 1991. Enhanced anharmonicity on Cu(110). *Phys. Rev. Lett.* 67:2327–2330.

Yang, W. S., Jona, F., and Marcus, P. M. 1983. Atomic structure of Si(001) 2×1. *Phys. Rev. B* 28:2049–2059.

Yang, H. N., Lu, T. M., and Wang, G. C. 1989. High resolution low energy electron diffraction study of Pb(110) surface roughening transition. *Phys. Rev. Lett.* 63:1621–1624.

Yang, L., Rahman, T. S., and Daw, M. S. 1991. Surface vibrations of Ag(100) and Cu(100): A molecular dynamics study. *Phys. Rev. B* 44:13725–13733.

Zeppenfeld, P., Kern, K., David, R., and Comsa, G. 1989. No thermal roughening on Cu(110) up to 900° K. *Phys. Rev. Lett.* 62:63–66.

Zhou, S. J., Beazley, D. M., Lomdahl, P. S., and Holian, B. L. 1997. Large scale molecular dynamics simulation of three-dimensional ductile failure. *Phys. Rev. Lett.* 78:479–482.

TALAT S. RAHMAN
Kansas State University
Manhattan, Kansas

SIMULATION OF CHEMICAL VAPOR DEPOSITION PROCESSES

INTRODUCTION

Chemical vapor deposition (CVD) processes constitute an important technology for the manufacturing of thin solid films. Applications include semiconducting, conducting, and insulating films in the integrated circuit industry, superconducting thin films, antireflection and spectrally selective coatings on optical components, and anticorrosion and antiwear layers on mechanical tools and equipment. Compared to other deposition techniques, such as sputtering, sublimation, and evaporation, CVD is very versatile and offers good control of film structure and composition, excellent uniformity, and sufficiently high growth rates. Perhaps the most important advantage of CVD over other deposition techniques is its capability for conformal deposition—i.e., the capability of depositing films of uniform thickness on highly irregularly shaped surfaces.

In CVD processes a thin film is deposited from the gas phase through chemical reactions at the surface. Reactive gases are introduced into the controlled environment of a reactor chamber in which the substrates on which deposition takes place are positioned. Depending on the process conditions, reactions may take place in the gas phase, leading to the creation of gaseous intermediates. The energy required to drive the chemical reactions can be supplied thermally by heating the gas in the reactor (thermal CVD), but also by supplying photons in the form of, e.g., laser light to the gas (photo CVD), or through the application of an electrical discharge (plasma CVD or plasma enhanced CVD). The reactive gas species fed into the reactor and the reactive intermediates created in the gas phase diffuse toward and adsorb onto the solid surface. Here, solid-catalyzed reactions lead to the growth of the desired film. CVD processes are reviewed in several books and survey papers (e.g., Bryant, 1977; Bunshah, 1982; Dapkus, 1982; Hess et al., 1985; Jensen, 1987; Sherman, 1987; Vossen and Kern, 1991; Jensen et al., 1991; Pierson, 1992; Granneman, 1993; Hitchman and Jensen, 1993; Kodas and Hampden-Smith, 1994; Rees, 1996; Jones and O'Brien, 1997).

The application of new materials, the desire to coat and reinforce new ceramic and fibrous materials, and the tremendous increase in complexity and performance of semiconductor and similar products employing thin films

are leading to the need to develop new CVD processes and to ever increasing demands on the performance of processes and equipment. This performance is determined by the interacting influences of hydrodynamics and chemical kinetics in the reactor chamber, which in turn are determined by process conditions and reactor geometry. It is generally felt that the development of novel reactors and processes can be greatly improved if simulation is used to support the design and optimization phase.

In the last decades, various sophisticated mathematical models have been developed, which relate process characteristics to operating conditions and reactor geometry (Heritage, 1995; Jensen, 1987; Meyyappan, 1995a). When based on fundamental laws of chemistry and physics, rather than empirical relations, such models can provide a scientific basis for design, development, and optimization of CVD equipment and processes. This may lead to a reduction of time and money spent on the development of prototypes and to better reactors and processes. Besides, mathematical CVD models may also provide fundamental insights into the underlying physicochemical processes and may be of help in the interpretation of experimental data.

PRINCIPLES OF THE METHOD

In CVD, physical and chemical processes take place at greatly differing length scales. Such processes as adsorption, chemical reactions, and nucleation take place at the microscopic (i.e., molecular) scale. These phenomena determine film characteristics such as adhesion, morphology, and purity. At the mesoscopic (i.e., micrometer) scale, free molecular flow and Knudsen diffusion phenomena in small surface structures determine the conformality of the deposition. At the macroscopic (i.e., reactor) scale, gas flow, heat transfer, and species diffusion determine process characteristics such as film uniformity and reactant consumption.

Ideally, a CVD model consists of a set of mathematical equations describing the relevant macroscopic, mesoscopic, and microscopic physicochemical processes in the reactor, and relating these phenomena to microscopic, mesoscopic, and macroscopic properties of the deposited films.

The first step is the description of the processes taking place at the substrate surface. A surface chemistry model describes how reactions between free sites, adsorbed species, and gaseous species close to the surface lead to the growth of a solid film and the creation of reaction products. In order to model these surface processes, the macroscopic distribution of the gas temperatures and species concentrations at the substrate surface must be known. Concentration distributions near the surface will generally differ from the inlet conditions and depend on the hydrodynamics and transport phenomena in the reactor chamber. Therefore, the core of a CVD model is formed by a transport phenomena model, consisting of a set of partial differential equations with appropriate boundary conditions describing the gas flow and the transport of energy and species in the reactor at the macroscopic scale, complemented by a thermal radiation model, describing the radiative heat exchange between the reactor walls. A gas phase chemistry model gives the reaction paths and rate constants for the homogeneous reactions, which influence the species concentration distribution through the production/destruction of chemical species. Often, plasma discharges are used in addition to thermal heating in order to activate the chemical reactions. In such a case, the gas phase transport and chemistry models must be extended with a plasma model. The properties of the gas mixture appearing in the transport model can be predicted based on the kinetic theory of gases. Finally, the macroscopic distributions of gas temperatures and concentrations at the substrate surface serve as boundary conditions for free molecular flow models, predicting the mesoscopic distribution of species within small surface structures and pores.

The abovementioned constituent parts of a CVD simulation model are briefly described in the following subsections. Within this unit, it is impossible to present all details needed for successfully setting up CVD simulations. However, an attempt is made to give a general impression of the basic ideas and approaches in CVD modeling, and of its capabilities and limitations. Many references to available literature are provided that will assist the interested reader in further study of the subject.

Surface Chemistry

Any CVD model must incorporate some form of surface chemistry. However, very limited information is usually available on CVD surface processes. Therefore, simple lumped chemistry models are widely used, assuming one single overall deposition reaction. In certain epitaxial processes operated at atmospheric pressure and high temperature, the rate of this overall reaction is very fast compared to convection and diffusion in the gas phase: the deposition is transport limited. In that case, useful predictions on deposition rates and uniformity can be obtained by simply setting the overall reaction rate to infinity (Wahl, 1977; Moffat and Jensen, 1986; Holstein et al., 1989; Fotiadis, 1990; Kleijn and Hoogendoorn, 1991). When the deposition rate is kinetically limited by the surface reaction, however, some expression for the overall rate constant as a function of species concentrations and temperature is needed. Such lumped reaction models have been published for CVD of, e.g., polycrystalline silicon (Jensen and Graves, 1983; Roenigk and Jensen, 1985; Peev et al., 1990a; Hopfmann et al., 1991), silicon nitride (Roenigk and Jensen, 1987; Peev et al., 1990b), silicon oxide (Kalindindi and Desu, 1990), silicon carbide (Koh and Woo, 1990), and gallium arsenide (Jansen et al., 1991).

In order to arrive at such a lumped deposition model, a surface reaction mechanism can be proposed, consisting of several elementary steps. As an example, for the process of chemical vapor deposition of tungsten from tungsten hexafluoride and hydrogen [overall reaction: $WF_6(g) + 3H_2(g) \rightarrow W(\text{solid}) + 6HF(g)$] the following mechanism can be proposed:

1. $WF_6(g) + s \leftrightarrow WF_6\ (s)$
2. $WF_6(s) + 5s \leftrightarrow W(\text{solid}) + 6F(s)$

3. $H_2(g) + 2s \leftrightarrow 2H(s)$
4. $H(s) + F(s) \leftrightarrow HF(s) + s$
5. $HF(s) \leftrightarrow HF(g) + s$ (1)

where (g) denotes a gaseous species, (s) an adsorbed species, and s a free surface site. Then, one of the steps is considered to be rate limiting, leading to a rate expression of the form

$$R^s = R^s(P_1, \ldots, P_N, T_s) \tag{2}$$

relating the deposition rate R^s to the partial pressures P_i of the gas species and the surface temperature T_s. When, e.g., in the above example, the desorption of HF (step 5) is considered to be rate limiting, the growth rate can be shown to depend on the H_2 and WF_6 gas concentrations as (Kleijn, 1995)

$$R^s = \frac{c_1[H_2]^{1/2}[WF_6]^{1/6}}{1 + c_2[WF_6]^{1/6}} \tag{3}$$

The expressions in brackets represent the gas concentrations. Finally, the validity of the mechanism is judged through comparison of predicted and experimental trends in the growth rate as a function of species pressures. Thus, it can be concluded that Equation 3 compares favorably with experiments in which a 0.5 order in H_2 and a 0 to 0.2 order in WF_6 is found. The constants c_1 and c_2 are parameters, which can be fitted to experimental growth rates as a function of temperature.

An alternative approach is the so-called "reactive sticking coefficient" concept for the rate of a reaction like A(g) → B(solid) + C(g). The sticking coefficient γ_A with 0 $0 \leq \gamma_A \leq 1$) is defined as the fraction of molecules of A that are incorporated into the film upon collision with the surface. For the deposition rate we now find (Motz and Wise, 1960):

$$R^s = \frac{\gamma_A}{1 - \gamma_A/2} \times \frac{P_A}{(2\pi M_A R T_S)^{1/2}} \tag{4}$$

with M_A the molar mass of species A, P_A its partial pressure, and T_S the surface temperature. The value of γ_A (with $0 \leq \gamma_A \leq 1$) has to be fitted to experimental data again.

Lumped reaction mechanisms are unlikely to hold for strongly differing process conditions, and do not provide insight into the role of intermediate species and reactions. More fundamental approaches to surface chemistry modeling, based on chains of elementary reaction steps at the surface and theoretically predicted reaction rates, applying techniques based on bond dissociation enthalpies and transition state theory, are just beginning to emerge. The prediction of reaction paths and rate constants for heterogeneous reactions is considerably more difficult than for gas-phase reactions. Although clearly not as mature as gas-phase kinetics modeling, impressive progress in the field of surface reaction modeling has been made in the last few years (Arora and Pollard, 1991; Wang and Pollard, 1994), and detailed surface reaction models have been published for CVD of, e.g., epitaxial silicon (Coltrin et al., 1989; Giunta et al., 1990a), polycrystalline silicon (Kleijn, 1991), silicon oxide (Giunta et al., 1990b), silicon carbide (Allendorf and Kee, 1991), tungsten (Arora and Pollard, 1991; Wang and Pollard, 1993; Wang and Pollard, 1995), gallium arsenide (Tirtowidjojo and Pollard, 1988; Mountziaris and Jensen, 1991; Masi et al., 1992; Mountziaris et al., 1993), cadmium telluride (Liu et al., 1992), and diamond (Frenklach and Wang, 1991; Meeks et al., 1992; Okkerse et al., 1997).

Gas-Phase Chemistry

A gas-phase chemistry model should state the relevant reaction pathways and rate constants for the homogeneous reactions. The often applied semiempirical approach assumes a small number of participating species and a simple overall reaction mechanism in which rate expressions are fitted to experimental data. This approach can be useful in order to optimize reactor designs and process conditions, but is likely to lose its validity outside the range of process conditions for which the rate expressions have been fitted. Also, it does not provide information on the role of intermediate species and reactions.

A more fundamental approach is based on setting up a chain of elementary reactions. A large system of all plausible species and elementary reactions is constructed. Typically, such a system consists of hundreds of species and reactions. Then, in a second step, sensitivity analysis is used to eliminate reaction pathways that do not contribute significantly to the deposition process, thus reducing the system to a manageable size. Since experimental data are generally not available in the literature, reaction rates are estimated by using tools such as statistical thermodynamics, bond dissociation enthalpies, and transition-state theory and Rice-Rampsberger-Kassel-Marcus (RRKM) theory. For a detailed description of these techniques the reader is referred to standard textbooks (e.g., Slater, 1959; Robinson and Holbrook, 1972; Benson, 1976; Laidler, 1987; Steinfeld et al., 1989). Also, computational chemistry techniques (Clark, 1985; Simka et al., 1996; Melius et al., 1997) allowing for the computation of heats of formation, bond dissociation enthalpies, transition state structures, and activation energies are now being used to model CVD chemical reactions. Thus, detailed gas-phase reaction models have been published for CVD of, e.g., epitaxial silicon (Coltrin et al., 1989; Giunta et al., 1990a), polycrystalline silicon (Kleijn, 1991), silicon oxide (Giunta et al., 1990b), silicon carbide (Allendorf and Kee, 1991), tungsten (Arora and Pollard, 1991; Wang and Pollard, 1993; Wang and Pollard, 1995), gallium arsenide (Tirtowidjojo and Pollard, 1988; Mountziaris and Jensen, 1991; Masi et al., 1992; Mountziaris et al., 1993), cadmium telluride (Liu et al., 1992), and diamond (Frenklach and Wang, 1991; Meeks et al., 1992; Okkerse et al., 1997).

The chain of elementary reactions in the gas phase mechanism will consist of unimolecular reactions of the general form A → C + D, and bimolecular reactions of the general form A + B → C + D. Their forward reaction rates are k[A] and k[A][B], respectively, with k the reaction rate constant and [A] and [B] the concentrations, respectively, of species A and B.

Generally, the temperature dependence of a reaction rate constant k(T) can be expressed accurately by a (modified) Arrhenius expression:

$$k(T) = AT^{\beta}\exp\left(\frac{-E}{RT}\right) \tag{5}$$

where T is the absolute temperature, A is a pre-exponential factor, E is the activation energy, and β is an exponent accounting for possible nonideal Arrhenius behavior. At the high temperatures and low pressures used in many CVD processes, unimolecular reactions may be in the so-called pressure fall-off regime, in which the rate constants may exhibit significant pressure dependencies. This is described by defining a low-pressure and a high-pressure rate constant according to Equations 6 and 7, respectively:

$$k_0(T) = A_0T^{\beta_0}\exp\left(\frac{-E_0}{RT}\right) \tag{6}$$

$$k_\infty(T) = A_\infty T^{\beta_\infty}\exp\left(\frac{-E_\infty}{RT}\right) \tag{7}$$

combined to a pressure dependent rate constant:

$$k(P,T) = \frac{k_0(T)k_\infty(T)P}{k_\infty(T) + k_0(T)P}F_{\text{corr}}(P,T) \tag{8}$$

where T, A, E, and β have the same meaning as in Equation 5, and the subscripts zero and infinity refer to the low-pressure and high-pressure limits, respectively. In its simplest form, $F_{\text{corr}} = 1$, but more accurate correction functions have been proposed (Gilbert et al., 1983; Kee et al., 1989). The correction function $F_{\text{corr}}(P,T)$ can be predicted from RRKM theory (Robinson and Holbrook, 1972; Forst, 1973; Roenigk et al., 1987; Moffat et al., 1991a). Computer programs with the necessary algorithms are available as public domain software (e.g., Hase and Bunker, 1973).

For the prediction of bimolecular reaction rates, transition-state theory can be applied (Benson, 1976; Steinfeld et al., 1989). According to this theory, the rate constant, k, for the bimolecular reaction:

$$A + B \rightarrow X^{\ddagger} \rightarrow C \tag{9}$$

(where $X^{\ddagger}$ is the transition state) is given by:

$$k = L^{\ddagger}\frac{k_BT}{h}\frac{Q^{\ddagger}}{Q_AQ_B}\exp\left(-\frac{E_0}{k_BT}\right) \tag{10}$$

where k_B is Boltzmann's constant and h is Planck's constant. The calculation of the partition functions $Q^{\ddagger}$, Q_A, and Q_B, and the activation energy E_0, requires knowledge of the structure and vibrational frequencies of the reactants and the intermediates. Parameters for the transition state are especially hard to obtain experimentally and are usually estimated from experimental frequency factors and activation energies of similar reactions, using empirical thermochemical rules (Benson, 1976). More recently, computational (quantum) chemistry methods (Stewart, 1983; Dewar et al., 1984; Clark, 1985; Hehre et al., 1986; Melius et al., 1997; also see Practical Aspects of the Method for information about software from Biosym Technologies) have been used to obtain more precise and direct information about transition state parameters (e.g., Ho et al., 1985, 1986; Ho and Melius, 1990; Allendorf and Melius, 1992; Zachariah and Tsang, 1995; Simka et al., 1996).

The forward and reverse reaction rates of a reversible reaction are related through thermodynamic constraints. Therefore, in a self-consistent chemistry model only the forward (or reverse) rate constant should be predicted or taken from experiments. The reverse (or forward) rate constant must then be obtained from thermodynamic constraints. When $\Delta G^0(T)$ is the standard Gibbs energy change of a reaction at standard pressure, P^0, its equilibrium constant is given by:

$$K_{\text{eq}}(T) = \exp\left(-\frac{\Delta G^0(T)}{RT}\right) \tag{11}$$

and the forward and reverse rate constants k_f and k_r are related as:

$$k_r(P,T) = \frac{k_f(P,T)}{K_{\text{eq}}(T)}\left(\frac{RT}{P^0}\right)^{\Delta\nu} \tag{12}$$

where $\Delta\nu$ is the net molar change of the reaction.

Free Molecular Transport

When depositing films on structured surfaces (e.g., a porous material or the structured surface of a wafer in IC manufacturing), the morphology and conformality of the film is determined by the meso-scale behavior of gas molecules inside these structures. At atmospheric pressure, mean free path lengths of gas molecules are of the order of 0.1 μm. At the low pressures applied in many CVD processes (down to 10 Pa), the mean free path may be as large as 1 mm. As a result, the gas behavior inside small surface structures cannot be described with continuum models, and a molecular approach must be used, based on approximate solutions to the Boltzmann equation. The main focus in meso-scale CVD modeling has been on the prediction of the conformality of deposited films. A good review can be found in Granneman (1993).

A first approach has been to model CVD in small cylindrical holes as a Knudsen diffusion and heterogeneous reaction process (Raupp and Cale, 1989; Chatterjee and McConica, 1990; Hasper et al., 1991; Schmitz and Hasper, 1993). This approach is based on a one-dimensional pseudocontinuum mass balance equation for the species concentration c along the depth z of a hole with diameter r:

$$\pi r^2 D_K\frac{\partial^2 c}{\partial z^2} = 2\pi r R(c) \tag{13}$$

The Knudsen diffusion coefficient (D_K) is proportional to the diameter of the hole (Knudsen, 1934):

$$D_K = \frac{2}{3}r\left(\frac{8RT}{\pi M}\right)^{1/2} \tag{14}$$

In this equation, M represents the molar mass of the species. For zero order reaction kinetics [deposition rate $R(c) \neq f(c)$] the equations can be solved analytically. For non-zero order kinetics they must be solved numerically (Hasper et al., 1991). The great advantages of this approach are its simplicity and the fact that it can easily be extended to include both the free molecular, transition, and continuum regimes. An important disadvantage is that the approach can be used for simple geometries only.

A more fundamental approach is the use of Monte Carlo techniques (Cooke and Harris, 1989; Ikegawa and Kobayashi, 1989; Rey et al., 1991; Kersch and Morokoff, 1995). The method can easily be extended to include phenomena such as complex gas and surface chemistry, specular reflection of particles from the walls, surface diffusion (Wulu et al., 1991), and three-dimensional effects (Coronell and Jensen, 1993). The method can in principle be applied to the free molecular, transition, and near-continuum regimes, but extension to the continuum regime is prohibited by computational demands. Surface smoothing techniques and the simulation of a large number of molecules are necessary to avoid (nonphysical) local thickness variations.

A third approach is the so-called "ballistic transport reaction" model (Cale and Raupp, 1990a,b; Cale et al., 1990, 1991), which has been implemented in the EVOLVE code (see Practical Aspects of the Method). The method makes use of the analogy between diffuse reflection and absorption of photons on gray surfaces, and the scattering and adsorption of reactive molecules on the walls of the hole. The approach seems to be very successful in the prediction of film profiles in complex two- and three-dimensional geometries. It is much more flexible than the Knudsen diffusion type models and has a more sound theoretical foundation. Compared to Monte Carlo methods, the ballistic method seems to be very efficient, requiring much less computational effort. Its main limitation is probably that the method is limited to the free molecular regime and cannot be extended to the transition or continuum flow regimes.

An important issue is the coupling between macro-scale and meso-scale phenomena (Gobbert et al., 1996; Hebb and Jensen, 1996; Rodgers and Jensen, 1998). On the one hand, the boundary conditions for meso-scale models are determined by the local macro-scale gas species concentrations. On the other hand, the meso-scale phenomena determine the boundary conditions for the macro-scale model.

Plasma

The modeling of plasma-based CVD chemistry is inherently more complex than the modeling of comparable neutral gas processes. The low-temperature, partially ionized discharges are characterized by a number of interacting effects, e.g., plasma generation of active species, plasma power deposition, and loss mechanisms and surface processes at the substrates and walls (Brinkmann et al., 1995b; Meyyappan, 1995b). The strong nonequilibrium in a plasma implies that equilibrium chemistry does not apply. In addition it causes a variety of spatio-temporal inhomogeneities. Nevertheless, important progress is being made in developing adequate mathematical models for low-pressure gas discharges. A detailed treatment of plasma modeling is beyond the scope of this unit, and the reader is referred to review texts (Chapman, 1980; Birdsall and Langdon, 1985; Boenig, 1988; Graves, 1989; Kline and Kushner, 1989; Brinkmann et al., 1995b; Meyyappan, 1995b).

Various degrees of complexity can be distinguished in plasma modeling: On the one hand, engineering-oriented, so-called "SPICE" models, are conceptually simple, but include various degrees of approximation and empirical information and are only able to provide qualitative results. On the other hand, so-called particle-in-cell (PIC)/Monte Carlo collision models (Hockney and Eastwood, 1981; Birdsall and Langdon, 1985; Birdsall, 1991) are closely related to the fundamental Boltzmann equation, and are suited to investigate basic principles of discharge operation. However, they require considerable computational resources, and in practice their use is limited to one-dimensional simulations only. This also applies to so-called "fluid-dynamical" models (Park et al., 1989; Gogolides and Sawin, 1992), in which the plasma is described in terms of its macroscopic variables. This results in a relatively simple set of partial differential equations with appropriate boundary conditions for, e.g., the electron density and temperature, the ion density and average velocity, and the electrical field potential. However, these equations are numerically stiff, due to the large differences in relevant time scales which have to be resolved, which makes computations expensive.

As an intermediate approach between the above two extremes, Brinkmann and co-workers (Brinkmann et al., 1995b) have proposed the so-called effective drift diffusion model for low-pressure RF plasmas as found in plasma enhanced CVD. This regime has been shown to be physically equivalent to fluid-dynamical models, with differences in predicted plasma properties of $<10\%$. However, compared to fluid-dynamical models, savings in CPU time of 1 to 2 orders of magnitude are obtained, allowing for the application in two- and three-dimensional equipment simulation (Brinkmann et al., 1995a).

Hydrodynamics

In the macro-scale modeling of the gas flow in the reactor, the gas is usually treated as a continuum. This assumption is valid when the mean free path length of the molecules is much smaller than a characteristic dimension of the reactor geometry, i.e., in general for pressures >100 Pa and typical dimensions >1 cm. Furthermore, the gas is treated as an ideal gas and the flow is assumed to be laminar, the Reynolds number being well below values at which turbulence might be expected.

In CVD we have to deal with multicomponent gas mixtures. The composition of an N-component gas mixture can be described in terms of the dimensionless mass fractions ω_i of its constituents, which sum up to unity:

$$\sum_{i=1}^{N} \omega_i = 1 \qquad (15)$$

Their diffusive fluxes can be expressed as mass fluxes $\vec{j}_i$ with respect to the mass-averaged velocity $\vec{v}$:

$$\vec{j}_i = \rho\omega_i(\vec{v}_i - \vec{v}) \tag{16}$$

The transport of momentum, heat, and chemical species is described by a set of coupled partial differential equations (Bird et al., 1960; Kleijn and Werner, 1993; Kleijn, 1995; Kleijn and Kuijlaars, 1995). The conservation of mass is given by the continuity equation:

$$\frac{\partial\rho}{\partial t} = -\nabla\cdot(\rho\vec{v}) \tag{17}$$

where ρ is the gas density and t the time. The conservation of momentum is given for Newtonian fluids by:

$$\frac{\partial\rho\vec{v}}{\partial t} = -\nabla\cdot(\rho\vec{v}\vec{v}) + \nabla\cdot\{\mu[\nabla\vec{v} + (\nabla\vec{v})^{\dagger}] - \frac{2}{3}\mu(\nabla\cdot\vec{v})\mathbf{I}\} - \nabla p + \rho\vec{g} \tag{18}$$

where μ is viscosity, $\mathbf{I}$ the unity tensor, p is the pressure, and $\vec{g}$ the gravity vector. The transport of thermal energy can be expressed in terms of temperature T. Apart from convection, conduction, and pressure terms, its transport equation comprises a term denoting the Dufour effect (transport of heat due to concentration gradients), a term representing the transport of enthalpy through diffusion of gas species, and a term representing production of thermal energy through chemical reactions, as follows:

$$c_p\frac{\partial\rho T}{\partial t} = -c_p\nabla\cdot(\rho\vec{v}T) + \nabla\cdot(\lambda\nabla T) + \frac{DP}{Dt} + \underbrace{\nabla\cdot\left(RT\sum_{i=1}^{N}\frac{D_i^T}{M_i}\nabla(\ln x_i)\right)}_{\text{Dufour}} \underbrace{-\sum_{i=1}^{N}\vec{j}_i\cdot\nabla\frac{H_i}{M_i}}_{\text{inter-diffusion}} \underbrace{-\sum_{i=1}^{N}\sum_{k=1}^{K}H_i\nu_{ik}R_k^g}_{\text{heat of reaction}} \tag{19}$$

where c_p is the specific heat, λ the thermal conductivity, P pressure, x_i is the mole fraction of gas i, D_i^T its thermal diffusion coefficient, H_i its enthalpy, $\vec{j}_i$ its diffusive mass flux, ν_{ik} its stoichiometric coefficient in reaction k, M_i its molar mass, and R_k^g the net reaction rate of reaction k. The transport equation for the ith gas species is given by:

$$\frac{\partial\rho\omega_i}{\partial t} = \underbrace{-\nabla\cdot(\rho\vec{v}\omega_i)}_{\text{convection}} - \underbrace{\nabla\cdot\vec{j}_i}_{\text{diffusion}} + \underbrace{M_i\sum_{k=1}^{K}\nu_{ik}R_k^g}_{\text{reaction}} \tag{20}$$

where $\vec{j}_i$ represents the diffusive mass flux of species i. In an N-component gas mixture, there are $N-1$ independent species equations of the type of Equation 20, since the mass fraction must sum up to unity (see Eq. 15).

Two phenomena of minor importance in many other processes may be specifically prominent in CVD, i.e., multicomponent effects and thermal diffusion (Soret effect). The most commonly applied theory for modeling gas diffusion is Fick's Law, which, however, is valid for isothermal, binary mixtures only. In the rigorous kinetic theory of N-component gas mixtures, the following expression for the diffusive mass flux vector is found (Hirschfelder et al., 1967),

$$\frac{M}{\rho}\sum_{j=1,j\neq i}^{N}\frac{\omega_i\vec{j}_j - \omega_j\vec{j}_i}{M_j\mathcal{D}_{ij}} = \nabla\omega_i + \omega_i\frac{\nabla M}{M} - \nabla(\ln T)\frac{M}{\rho}\sum_{j=1,j\neq i}^{N}\frac{\omega_i D_j^T - \omega_j D_i^T}{M_j\mathcal{D}_{ij}} \tag{21}$$

where M is the average molar mass, $\mathcal{D}_{ij}$ is the binary diffusion coefficient of a gas pair and D_i^T is the thermal diffusion coefficient of a gas species. In general, $D_i^T > 0$ for large, heavy molecules (which therefore are driven toward cold zones in the reactor), $D_i^T < 0$ for small, light molecules (which therefore are driven toward hot zones in the reactor), and $\Sigma D_i^T = 0$. Equation 21 can be rewritten by separating the diffusive mass flux vector $\vec{j}_i$ into a flux driven by concentration gradients $\vec{j}_i^C$ and a flux driven by temperature gradients $\vec{j}_i^T$:

$$\vec{j}_i = \vec{j}_i^C + \vec{j}_i^T \tag{22}$$

with

$$\frac{M}{\rho}\sum_{j=1}^{N}\frac{\omega_i\vec{j}_j^C - \omega_j\vec{j}_i^C}{M_j\mathcal{D}_{ij}} = \nabla\omega_i + \omega_i\frac{\nabla M}{M} \tag{23}$$

and

$$\vec{j}_i^T = -D_i^T\nabla(\ln T) \tag{24}$$

Equation 23 relates the N diffusive mass flux vectors $\vec{j}_i^C$ to the N mass fractions and mass fraction gradients. In many numerical schemes however, it is desirable that the species transport equation (Eq. 20) contains a gradient-driven "Fickian" diffusion term. This can be obtained by rewriting Equation 23 as:

$$\vec{j}_i^C = \underbrace{-\rho D_i^{SM}\nabla\omega_i}_{\text{Fick term}} - \underbrace{\rho\omega_i D_i^{SM}\frac{\nabla m}{m}}_{\text{multi-component 1}} + \underbrace{m\omega_i D_i^{SM}\sum_{j=1,j\neq i}\frac{\vec{j}_j^C}{M_j\mathcal{D}_{ij}}}_{\text{multi-component 2}} \tag{25}$$

and defining a diffusion coefficient D_i^{SM}:

$$D_i^{SM} = \left(\sum_{j=1,j\neq i}^{N}\frac{x_i}{\mathcal{D}_{ij}}\right)^{-1} \tag{26}$$

The divergence of the last two terms in Equation 25 is treated as a source term. Within an iterative solution scheme, the unknown diffusive fluxes $\vec{j}_j^C$ can be taken from a previous iteration.

The above transport equations are supplemented with the usual boundary conditions in the inlets and outlets

and at the nonreacting walls. On reacting walls there will be a net gaseous mass production which leads to a velocity component normal to the wafer surface:

$$\vec{n} \cdot (\rho \vec{v}) = \sum_i M_i \sum_l \sigma_{il} R_l^s \tag{27}$$

where $\vec{n}$ is the outward-directed unity vector normal to the surface, ρ is the local density of the gas mixture, R_l^s the rate of the lth surface reaction and σ_{il} the stoichiometric coefficient of species i in this reaction. The net total mass flux of the ith species normal to the wafer surface equals its net mass production:

$$\vec{n} \cdot (\rho \omega_i \vec{v} + \vec{j_i}) = M_i \sum_l \sigma_{il} R_l^s \tag{28}$$

Kinetic Theory

The modeling of transport phenomena and chemical reactions in CVD processes requires knowledge of the thermochemical properties (specific heat, heat of formation, and entropy) and transport properties (viscosity, thermal conductivity, and diffusivities) of the gas mixture in the reactor chamber.

Thermochemical properties of gases as a function of temperature can be found in various publications (Svehla, 1962; Coltrin et al., 1986, 1989; Giunta et al., 1990a,b; Arora and Pollard, 1991) and databases (Gordon and McBride, 1971; Barin and Knacke, 1977; Barin et al., 1977; Stull and Prophet, 1982; Wagman et al., 1982; Kee et al., 1990). In the absence of experimental data, thermochemical properties may be obtained from ab initio molecular structure calculations (Melius et al., 1997).

Only for the most common gases can transport properties be found in the literature (Maitland and Smith, 1972; l'Air Liquide, 1976; Weast, 1984). The transport properties of less common gas species may be calculated from kinetic theory (Svehla, 1962; Hirschfelder et al., 1967; Reid et al., 1987; Kleijn and Werner, 1993; Kleijn, 1995; Kleijn and Kuijlaars, 1995). Assumptions have to be made for the form of the intermolecular potential energy function $\phi(r)$. For nonpolar molecules, the most commonly used intermolecular potential energy function is the Lennard-Jones potential:

$$\phi(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{29}$$

where r is the distance between the molecules, σ the collision diameter of the molecules, and ε their maximum energy of attraction. Lennard-Jones parameters for many CVD gases can be found in Svehla (1962), Coltrin et al. (1986), Coltrin et al. (1989), Arora and Pollard (1991), and Kee et al. (1991), or can be estimated from properties of the gas at the critical point or at the boiling point (Bird et al., 1960):

$$\frac{\varepsilon}{k_B} = 0.77 T_c \quad \text{or} \quad \frac{\varepsilon}{k_B} = 1.15 T_b \tag{30}$$

$$\sigma = 0.841 V_c \quad \text{or} \quad \sigma = 2.44 \frac{T_c}{P_c} \quad \text{or} \quad \sigma = 1.166 V_{b,l} \tag{31}$$

where T_c and T_b are the critical temperature and normal boiling point temperature (K), P_c is the critical pressure (atm), V_c and $V_{b,l}$ are the molar volume at the critical point and the liquid molar volume at the normal boiling point ($cm^3\ mol^{-1}$), and k_B is the Boltzmann constant. For most CVD gases, only rough estimates of Lennard-Jones parameters are available. Together with inaccuracies in the assumptions made in kinetic theory, this leads to an accuracy of predicted transport properties of typically 10% to 25%. When the transport properties of its constituent gas species are known, the properties of a gas mixture can be calculated from semiempirical mixture rules (Reid et al., 1987; Kleijn and Werner, 1993; Kleijn, 1995; Kleijn and Kuijlaars, 1995). The inaccuracy in predicted mixture properties may well be as large as 50%.

Radiation

CVD reactor walls, windows, and substrates adopt a certain temperature profile as a result of their conjugate heat exchange. These temperature profiles may have a large influence on the deposition process. This is even more true for lamp-heated reactors, such as rapid thermal CVD (RTCVD) reactors, in which the energy bookkeeping of the reactor system is mainly determined by radiative heat exchange.

The transient temperature distribution in solid parts of the reactor is described by the Fourier equation (Bird et al., 1960):

$$\rho_s c_{p,s} \frac{\partial T}{\partial t} = \nabla \cdot (\lambda_s \nabla T_s) + q''' \tag{32}$$

where q''' is the heat-production rate in the solid material, e.g., due to inductive heating, ρ_s, $c_{p,s}$, λ_s, and T_s are the solid density, specific heat, thermal conductivity, and temperature. The boundary conditions at the solid-gas interfaces take the form:

$$\vec{n} \cdot (\lambda_s \nabla T_s) = q''_{\text{conv}} + q''_{\text{rad}} \tag{33}$$

where q''_{conv} and q''_{rad} are the convective and radiative heat fluxes to the solid and $\vec{n}$ is the outward directed unity vector normal to the surface. For the interface between a solid and the reactor gases (the temperature distribution of which is known) we have:

$$q''_{\text{conv}} = -\vec{n} \cdot (\lambda_g \nabla T_g) \tag{34}$$

where λ_g and T_g are the thermal conductivity and temperature of the reactor gases. Usually, we do not have detailed information on the temperature distribution outside the reactor. Therefore, we have to use heat transfer relations like:

$$q''_{\text{conv}} = \alpha_{\text{conv}} (T_s - T_{\text{ambient}}) \tag{35}$$

to model the convective heat losses to the ambient, where α_{conv} is a heat-transfer coefficient.

The most challenging part of heat-transfer modeling is the radiative heat exchange inside the reactor chamber,

which is complicated by the complex geometry, the spectral and temperature dependence of the optical properties (Ordal et al., 1985; Palik, 1985), and the occurrence of specular reflections. An extensive treatment of all these aspects of the modeling of radiative heat exchange can be found, e.g., in Siegel and Howell (1992), Kersch and Morokoff (1995), Kersch (1995a), or Kersch (1995b).

An approach that can be used if the radiating surfaces are diffuse-gray (i.e., their absorptivity and emissivity are independent of direction and wavelength) is the so-called Gebhart absorption-factor method (Gebhart, 1958, 1971). The reactor walls are divided into small surface elements, across which a uniform temperature is assumed. Exchange factors G_{ij} between pairs $i - j$ of surface elements are evaluated, which are determined by geometrical line-of-sight factors and optical properties. The net radiative heat transfer to surface element j now equals:

$$q''_{rad,j} = \frac{1}{A_j} \sum_i G_{ij} \varepsilon_i \sigma_B T_i^4 A_i - \varepsilon_j \sigma_B T_j^4 \qquad (36)$$

where ε is the emissivity, σ_B the Stefan-Boltzmann constant, and A the surface area of the element.

In order to incorporate further refinements, such as wavelength, temperature and directional dependence of the optical properties, and specular reflections, Monte Carlo methods (Howell, 1968) are more powerful than the Gebhart method. The emissive power is partitioned into a large number of rays of energy leaving each surface, which are traced through the reactor as they are being reflected, transmitted, or absorbed at various surfaces. By choosing the random distribution functions of the emission direction and the wavelength for each ray appropriately, the total emissive properties from the surface may be approximated. By averaging over a large number of rays, the total heat exchange fluxes may be computed (Coronell and Jensen, 1994; Kersch and Morokoff, 1995; Kersch, 1995a,b).

PRACTICAL ASPECTS OF THE METHOD

In the previous section, the seven main aspects of CVD simulation (i.e., surface chemistry, gas-phase chemistry, free molecular flow, plasma physics, hydrodynamics, kinetic theory, and thermal radiation) have been discussed (see Principles of the Method). An ideal CVD simulation tool should integrate models for all these aspects of CVD. Such a comprehensive tool is not available yet. However, powerful software tools for each of these aspects can be obtained commercially, and some CVD simulation software combines several over the necessary models.

Surface Chemistry

For modeling surface processes at the interface between a solid and a reactive gas, the SURFACE CHEMKIN package (available from Reaction Design; also see Coltrin et al., 1991a,b) is undoubtedly the most flexible and powerful tool available at present. It is a suite of FORTRAN codes allowing for easily setting up surface reaction simulations. It defines a formalism for describing surface processes between various gaseous, adsorbed, and solid bulk species and performs bookkeeping on concentrations of all these species. In combination with the SURFACE PSR (Moffat et al., 1991b), SPIN (Coltrin et al., 1993a) and CRESLAF (Coltrin et al., 1993b) codes (all programs available from Reaction Design), it can be used to model the surface reactions in a perfectly stirred tank reactor, a rotating disk reactor, or a developing boundary layer flow along a reacting surface. Simple problems can be run on a personal computer in a few minutes; more complex problems may take dozens of minutes on a powerful workstation.

No models or software are available for routinely predicting surface reaction kinetics. In fact, this is as yet probably the most difficult and unsolved issue in CVD modeling. Surface reaction kinetics are estimated based on bond dissociation enthalpies, transition state theory, and analogies with similar gas phase reactions; the success of this approach largely depends on the skills and expertise of the chemist performing the analysis.

Gas-Phase Chemistry

Similarly, for modeling gas-phase reactions, the CHEMKIN package (available from Reaction Design; also see Kee et al., 1989) is the *de facto* standard modeling tool. It is a suite of FORTRAN codes allowing for easily setting up reactive gas flow problems, which computes production/destruction rates and performs bookkeeping on concentrations of gas species. In combination with the CHEMKIN THERMODYNAMIC DATABASE (Reaction Design) it allows for the self-consistent evaluation of species thermochemical data and reverse reaction rates. In combination with the SURFACE PSR (Moffat et al., 1991a), SPIN (Coltrin et al., 1993a), and CRESLAF (Coltrin et al., 1993b) codes (all programs available from Reaction Design) it can be used to model reactive flows in a perfectly stirred tank reactor, a rotating disk reactor, or a developing boundary layer flow, and it can be used together with SURFACE CHEMKIN (Reaction Design) to simulate problems with both gas and surface reactions. Simple problems can be run on a personal computer in a few minutes; more complex problems may take dozens of minutes on a powerful workstation.

Various proprietary and shareware software programs are available for predicting gas phase rate constants by means of theoretical chemical kinetics (Hase and Bunker, 1973) and for evaluating molecular and transition state structures and electronic energies (available from Biosym Technologies and Gaussian Inc.). These programs, especially the latter, require significant computing power. Once the possible reaction paths have been identified and reaction rates have been estimated, sensitivity analysis with, e.g., the SENKIN package (available from Reaction Design, also see Lutz et al., 1993) can be used to eliminate insignificant reactions and species. As in surface chemistry, setting up a reaction model that can confidently be used in predicting gas phase chemistry is still far from trivial, and the success largely depends on the skills and expertise of the chemist performing the analysis.

Free Molecular Transport

As described in the previous section (see Principles of the Method) the "ballistic transport-reaction" model is probably the most powerful and flexible approach to modeling free molecular gas transport and chemical reactions inside very small surface structures (i.e., much smaller than the gas molecules' mean free path length). This approach has been implemented in the EVOLVE code. EVOLVE 4.1a is a low-pressure transport, deposition, and etch-process simulator developed by T.S. Cale at Arizona State University, Tempe, Ariz., and Motorola Inc., with support from the Semiconductor Research Corporation, The National Science Foundation, and Motorola, Inc. It allows for the prediction of the evolution of film profiles and composition inside small two-dimensional and three-dimensional holes of complex geometry as functions of operating conditions, and requires only moderate computing power, provided by a personal computer. A Monte Carlo model for microscopic film growth has been integrated into the computational fluid dynamics (CFD) code CFD-ACE (available from CFDRC).

Plasma Physics

The modeling of plasma physics and chemistry in CVD is probably not yet mature enough to be done routinely by nonexperts. Relatively powerful and user-friendly continuum plasma simulation tools have been incorporated into some tailored CFD codes, such as Phoenics-CVD from Cham, Ltd. and CFDPLASMA (ICP) from CFDRC. They allow for two- and three-dimensional plasma modeling on powerful workstations at typical CPU times of several hours. However, the accurate modeling of plasma properties requires considerable expert knowledge. This is even more true for the relation between plasma physics and plasma enhanced chemical reaction rates.

Hydrodynamics

General purpose CFD packages for the simulation of multi-dimensional fluid flow have become available in the last two decades. These codes have mostly been based on either the finite volume method (Patankar, 1980; Minkowycz et al., 1988) or the finite element method (Taylor and Hughes, 1981; Zienkiewicz and Taylor, 1989). Generally, these packages offer easy grid generation for complex two-dimensional and three-dimensional geometries, a large variety of physical models (including models for gas radiation, flow in porous media, turbulent flow, two-phase flow, non-Newtonian liquids, etc.), integrated graphical post-processing, and menu-driven user-interfaces allowing the packages to be used without detailed knowledge of fluid dynamics and computational techniques. Obviously, CFD packages are powerful tools for CVD hydrodynamics modeling. It should, however, be realized that they have not been developed specifically for CVD modeling. As a result: (1) the input data must be formulated in a way that is not very compatible with common CVD practice; (2) many features are included that are not needed for CVD modeling, which makes the packages bulky and slow; (3) the numerical solvers are generally not very well suited for the solution of the stiff equations typical of CVD chemistry; (4) some output that is of specific interest in CVD modeling is not provided routinely; and (5) the codes do not include modeling features that are needed for accurate CVD modeling, such as gas species thermodynamic and transport property databases, solids thermal and optical property databases, chemical reaction mechanism and rate constants databases, gas mixture property calculation from kinetic theory, multicomponent ordinary and thermal diffusion, multiple chemical species in the gas phase and at the surface, multiple chemical reactions in the gas phase and at the surface, plasma physics and plasma chemistry models, and non-gray, non-diffuse wall-to-wall radiation models. The modifications required to include these features in general-purpose fluid dynamics codes are not trivial, especially when the source codes are not available. Nevertheless, promising results in the modeling of CVD reactors have been obtained with general purpose CFD codes.

A few CFD codes have been specially tailored for CVD reactor scale simulations including the following.

PHOENICS-CVD (Phoenics-CVD, 1997), a finite-volume CVD simulation tool based on the PHOENICS flow simulator by Cham Ltd. (developed under EC-ESPRIT project 7161). It includes databases for thermal and optical solid properties and thermodynamic and transport gas properties (CHEMKIN-format), models for multicomponent (thermal) diffusion, kinetic theory models for gas properties, multireaction gas-phase and surface chemistry capabilities, the effective drift diffusion plasma model and an advanced wall-to-wall thermal radiation model, including spectral dependent optical properties, semitransparent media and specular reflection. Its modeling capabilities and examples of its applications have been described in Heritage (1995).

CFD-ACE, a finite-volume CFD simulation tool by CFD Research Corporation. It includes models for multicomponent (thermal) diffusion, efficient algorithms for stiff multistep gas and surface chemistry, a wall-to-wall thermal radiation model (both gray and non-gray) including semitransparent media, and integrated Monte Carlo models for free molecular flow phenomena inside small structures. The code can be coupled to CFD-PLASMA to perform plasma CVD simulations.

FLUENT, a finite-volume CFD simulation tool by Fluent Inc. It includes models for multicomponent (thermal) diffusion, kinetic theory models for gas properties, and (limited) multistep gas-phase chemistry and simple surface chemistry.

These codes have been available for a relatively short time now and are still continuously evolving. They allow for the two-dimensional modeling of gas flow with simple chemistry on powerful personal computers in minutes. For three-dimensional simulations and simulations including, e.g., plasma, complex chemistry, or radiation effects, a powerful workstation is needed, and CPU times may be several hours. Potential users should compare the capabilities, flexibility, and user-friendliness of the codes to their own needs.

Kinetic Theory

Kinetic gas theory models for predicting transport properties of multicomponent gas mixtures have been incorporated into the PHOENICS-CVD, CVD-ACE, and FLUENT flow simulation codes. The CHEMKIN suite of codes contains a library of routines as well as databases (Kee et al., 1990) for evaluating transport properties of multicomponent gas mixture as well.

Thermal Radiation

Wall-to-wall thermal radiation models, including essential features for CVD modeling such as spectrally dependent (non-gray) optical properties, semitransparent media (e.g., quartz) and specular reflection, on, e.g., polished metal surfaces, have been incorporated in the PHOENICS-CVD and CFD-ACE flow simulation codes. In addition many stand-alone thermal radiation simulators are available, e.g., ANSYS (from Swanson Analysis Systems).

PROBLEMS

CVD simulation is a powerful tool in reactor design and process optimization. With commercially available CFD software, rather straightforward and reliable hydrodynamic modeling studies can be performed, which give valuable information on, e.g., flow recirculations, dead zones, and other important reactor design issues. Thermal radiation simulations can also be performed relatively easily, and they provide detailed insight in design parameters such as heating uniformities and peak thermal load.

However, as soon as one wishes to perform a comprehensive CVD simulation to predict issues such as film deposition rate and uniformity, conformality, and purity, several problems arise.

The first problem is that every available numerical simulation code to be used for CVD simulation has some limitations or drawbacks. The most powerful and tailored CVD simulation models—i.e., the CHEMKIN suite from Reaction Design—allows for the hydrodynamic simulation of highly idealized and simplified flow systems only, and does not include models for thermal radiation and molecular behavior in small structures. CFD codes (even the ones that have been tailored for CVD modeling, such as PHOENICS-CVD, CFD-ACE, and FLUENT) have limitations with respect to modeling stiff multi-reaction chemistry. The coupling to molecular flow models (if any) is only one-way, and incorporated plasma models have a limited range of validity and require specialist knowledge. The simulation of complex three-dimensional reactor geometries with detailed chemistry, plasma and/or radiation can be very cumbersome, and may require long CPU times.

A second and perhaps even more important problem is the lack of detailed, reliable, and validated chemistry models. Models have been proposed for some important CVD processes (see Principles of the Method), but their testing and validation has been rather limited. Also, the "translation" of published models to the input format required by various software is error-prone. In fact, the unknown chemistry of many processes is the most important bottle-neck in CVD simulation.

The CHEMKIN codes come with detailed chemistry models (including rate constants) for a range of processes. To a lesser extent, detailed chemistry models for some CVD processes have been incorporated in the databases of PHOENICS-CVD as well.

Lumped chemistry models should be used with the greatest care, since they are unlikely to hold for process conditions different from those for which they have been developed, and it is sometimes even doubtful whether they can be used in a different reactor than that for which they have been tested. Even detailed chemistry models based on elementary processes have a limited range of validity, and the use of these models in a different pressure regime is especially dangerous. Without fitting, their accuracy in predicting growth rates may well be off by 100%. The use of theoretical tools for predicting rate constants in the gas phase requires specialized knowledge and is not completely straightforward. This is even more the case for surface reaction kinetics, where theoretical tools are just beginning to be developed.

A third problem is the lack of accurate input data, such as Lennard-Jones parameters for gas property prediction, thermal and optical solid properties (especially for coated surfaces), and plasma characteristics. Some scattered information and databases containing relevant parameters are available, but almost every modeler setting up CVD simulations for a new process will find that important data are lacking.

Furthermore, the coupling between the macro-scale (hydrodynamics, plasma, and radiation) parts of CVD simulations and meso-scale models for molecular flow and deposition in small surface structures is difficult (Jensen et al., 1996; Gobbert et al., 1997), and this, in particular, is what one is interested in for most CVD modeling.

Finally, CVD modeling does not, at present, predict film structure, morphology, or adhesion. It does not predict mechanical, optical, or electrical film properties. It does not lead to the invention of new processes or the prediction of successful precursors. It does not predict optimal reactor configurations or processing conditions (although it can be used in evaluating the process performance as a function of reactor configuration and processing conditions). It does not generally lead to quantitatively correct growth rate or step coverage predictions without some fitting aided by prior experimental knowledge of the deposition kinetics.

However, in spite of all these limitations, carefully setup CVD simulations can provide reactor designers and process developers with a wealth of information, as shown in many studies (Kleijn, 1995). CVD simulations predict general trends in the process characteristics and deposition properties in relation to process conditions and reactor geometry and can provide fundamental insight into the relative importance of various phenomena. As such, it can be an important tool in process optimization and reactor design, pointing out bottlenecks in the design and issues that need to be studied more carefully. All of this leads to more efficient, faster, and less expensive process design, in which less trial and error is involved. Thus,

successful attempts have been made in using simulation, for example, to optimize hydrodynamic reactor design and eliminate flow recirculations (Evans and Greif, 1987; Visser et al., 1989; Fotiadis et al., 1990), to predict and optimize deposition rate and uniformity (Jensen and Graves, 1983; Kleijn and Hoogendoorn, 1991; Biber et al., 1992), to optimize temperature uniformity (Badgwell et al., 1994; Kersch and Morokoff, 1995), to scale up existing reactors to large wafer diameters (Badgwell et al., 1992), to optimize process operation and processing conditions with respect to deposition conformality (Hasper et al., 1991; Kristof et al., 1997), to predict the influence of processing conditions on doping rates (Masi et al., 1992), to evaluate loading effects on selective deposition rates (Holleman et al., 1993), and to study the influence of operating conditions on self-limiting effects (Leusink et al., 1992) and selectivity loss (Werner et al., 1992; Kuijlaars, 1996).

The success of these exercises largely depends on the skills and experience of the modeler. Generally, all available CVD simulation software leads to erroneous results when used by careless or inexperienced modelers.

LITERATURE CITED

Allendorf, M. and Kee, R. 1991. A model of silicon carbide chemical vapor deposition. *J. Electrochem. Soc*. 138:841–852.

Allendorf, M. and Melius, C. 1992. Theoretical study of the thermochemistry of molecules in the *Si-C-H* system. *J. Phys. Chem*. 96:428–437.

Arora, R. and Pollard, R. 1991. A mathematical model for chemical vapor deposition influenced by surface reaction kinetics: Application to low pressure deposition of tungsten. *J. Electrochem. Soc*. 138:1523–1537.

Badgwell, T., Edgar, T., and Trachtenberg, I. 1992. Modeling and scale-up of multiwafer LPCVD reactors. *AIChE Journal* 138:926–938.

Badgwell, T., Trachtenberg, I., and Edgar, T. 1994. Modeling the wafer temperature profile in a multiwafer LPCVD furnace. *J. Electrochem. Soc*. 141:161–171.

Barin, I. and Knacke, O. 1977. Thermochemical Properties of Inorganic Substances. Springer-Verlag, Berlin.

Barin, I., Knacke, O., and Kubaschewski, O. 1977. Thermochemical Properties of Inorganic Substances. Supplement. Springer-Verlag, Berlin.

Benson, S. 1976. Thermochemical Kinetics (2nd ed.). John Wiley & Sons, New York.

Biber, C., Wang, C., and Motakef, S. 1992. Flow regime map and deposition rate uniformity in vertical rotating-disk omvpe reactors. *J. Crystal Growth* 123:545–554.

Bird, R. B., Stewart, W., and Lightfood, E. 1960. Transport Phenomena. John Wiley & Sons, New York.

Birdsall, C. 1991. Particle-in-cell charged particle simulations, plus Monte Carlo collisions with neutral atoms. *IEEE Trans. Plasma Sci.* 19:65–85.

Birdsall, C. and Langdon, A. 1985. Plasma Physics via Computer Simulation. McGraw-Hill, New York.

Boenig, H. 1988. Fundamentals of Plasma Chemistry and Technology. Technomic Publishing Co., Lancaster, Pa.

Brinkmann, R., Vogg, G., and Werner, C. 1995a. Plasma enhanced deposition of amorphous silicon. *Phoenics J*. 8:512–522.

Brinkmann, R., Werner, C., and Fürst, R. 1995b. The effective drift-diffusion plasma model and its implementation into phoenics-cvd. *Phoenics J*. 8:455–464.

Bryant, W. 1977. The fundamentals of chemical vapour deposition. *J. Mat. Science* 12:1285–1306.

Bunshah, R. 1982. Deposition Technologies for Films and Coatings. Noyes Publications, Park Ridge, N.J.

Cale, T. and Raupp, G. 1990a. Free molecular transport and deposition in cylindrical features. *J. Vac. Sci. Technol. B* 8:649–655.

Cale, T. and Raupp, G. 1990b. A unified line-of-sight model of deposition in rectangular trenches. *J. Vac. Sci. Technol. B* 8:1242–1248.

Cale, T., Raupp, G., and Gandy, T. 1990. Free molecular transport and deposition in long rectangular trenches. *J. Appl. Phys.* 68:3645–3652.

Cale, T., Gandy, T., and Raupp, G. 1991. A fundamental feature scale model for low pressure deposition processes. *J. Vac. Sci. Technol. A* 9:524–529.

Chapman, B. 1980. Glow Discharge Processes. John Wiley & Sons, New York.

Chatterjee, S. and McConica, C. 1990. Prediction of step coverage during blanket CVD tungsten deposition in cylindrical pores. *J. Electrochem. Soc.* 137:328–335.

Clark, T. 1985. A Handbook of Computational Chemistry. John Wiley & Sons, New York.

Coltrin, M., Kee, R., and Miller, J. 1986. A mathematical model of silicon chemical vapor deposition. Further refinements and the effects of thermal diffusion. *J. Electrochem. Soc.* 133:1206–1213.

Coltrin, M., Kee, R., and Evans, G. 1989. A mathematical model of the fluid mechanics and gas-phase chemistry in a rotating disk chemical vapor deposition reactor. *J. Electrochem. Soc*. 136: 819–829.

Coltrin, M., Kee, R., and Rupley, F. 1991a. Surface Chemkin: A general formalism and software for analyzing heterogeneous chemical kinetics at a gas-surface interface. *Int. J. Chem. Kinet*. 23:1111–1128.

Coltrin, M., Kee, R., and Rupley, F. 1991b. *Surface Chemkin (Version 4.0)*. Technical Report SAND90-8003. Sandia National Laboratories Albuquerque, N.M./Livermore, Calif.

Coltrin, M., Kee, R., Evans, G., Meeks, E., Rupley, F., and Grcar, J. 1993a. SPIN *(Version 3.83): A* FORTRAN *program for modeling one-dimensional rotating-disk/stagnation-flow chemical vapor deposition reactors*. Technical Report SAND91-8003.UC-401 Sandia National Laboratories, Albuquerque, N.M./Livermore, Calif.

Coltrin, M., Moffat, H., Kee, R., and Rupley, F. 1993b. CRESLAF *(Version 4.0): A* FORTRAN *program for modeling laminar, chemically reacting, boundary-layer flow in cylindrical or planar channels*. Technical Report SAND93-0478.UC-401 Sandia National Laboratories Albuquerque, N.M./Livermore, Calif.

Cooke, M. and Harris, G. 1989. Monte Carlo simulation of thin-film deposition in a rectangular groove. *J. Vac. Sci. Technol. A* 7:3217–3221.

Coronell, D. and Jensen, K. 1993. Monte Carlo simulations of very low pressure chemical vapor deposition. *J. Comput. Aided Mater. Des*. 1:1–12.

Coronell, D. and Jensen, K. 1994. Monte Carlo simulation study of radiation heat transfer in the multiwafer LPCVD reactor. *J. Electrochem. Soc*. 141:496–501.

Dapkus, P. 1982. Metalorganic chemical vapor deposition. *Annu. Rev. Mater. Sci.* 12:243–269.

Dewar, M., Healy, E., and Stewart, J. 1984. Location of transition states in reaction mechanisms. *J. Chem. Soc. Faraday Trans. II* 80:227–233.

Evans, G. and Greif, R. 1987. A numerical model of the flow and heat transfer in a rotating disk chemical vapor deposition reactor. *J. Heat Transfer* 109:928–935.

Forst, W. 1973. Theory of Unimolecular Reactions. Academic Press, New York.

Fotiadis, D. 1990. Two- and Three-dimensional Finite Element Simulations of Reacting Flows in Chemical Vapor Deposition of Compound Semiconductors. Ph.D. thesis. University of Minnesota, Minneapolis, Minn.

Fotiadis, D., Kieda, S., and Jensen, K. 1990. Transport phenomena in vertical reactors for metalorganic vapor phase epitaxy: I. Effects of heat transfer characteristics, reactor geometry, and operating conditions. *J. Crystal Growth* 102:441–470.

Frenklach, M. and Wang, H. 1991. Detailed surface and gas-phase chemical kinetics of diamond deposition. *Phys. Rev. B.* 43:1520–1545.

Gebhart, B. 1958. A new method for calculating radiant exchanges. *Heating, Piping, Air Conditioning* 30:131–135.

Gebhart, B. 1971. Heat Transfer (2nd ed.). McGraw-Hill, New York.

Gilbert, R., Luther, K., and Troe, J. 1983. Theory of unimolecular reactions in the fall-off range. *Ber. Bunsenges. Phys. Chem.* 87:169–177.

Giunta, C., McCurdy, R., Chapple-Sokol, J., and Gordon, R. 1990a. Gas-phase kinetics in the atmospheric pressure chemical vapor deposition of silicon from silane and disilane. *J. Appl. Phys.* 67:1062–1075.

Giunta, C., Chapple-Sokol, J., and Gordon, R. 1990b. Kinetic modeling of the chemical vapor deposition of silicon dioxide from silane or disilane and nitrous oxide. *J. Electrochem. Soc.* 137:3237–3253.

Gobbert, M. K., Ringhofer, C. A., and Cale, T. S. 1996. Mesoscopic scale modeling of micro loading during low pressure chemical vapor deposition. *J. Electrochem. Soc.* 143(8):524–530.

Gobbert, M., Merchant, T., Burocki, L., and Cale, T. 1997. Vertical integration of CVD process models. *In* Chemical Vapor Deposition: Proceedings of the 14th International Conference and EUROCVD-II (M. Allendorf and B. Bernard, eds.) pp. 254–261. Electrochemical Society, Pennington, N.J.

Gogolides, E. and Sawin, H. 1992. Continuum modeling of radio-frequency glow discharges. I. Theory and results for electropositive and electronegative gases. *J. Appl. Phys.* 72:3971–3987.

Gordon, S. and McBride, B. 1971. *Computer Program for Calculation of Complex Chemical Equilibrium Compositions, Rocket Performance, Incident and Reflected Shocks and Chapman-Jouguet Detonations*. Technical Report SP-273 NASA. National Aeronautics and Space Administration, Washington, D.C.

Granneman, E. 1993. Thin films in the integrated circuit industry: Requirements and deposition methods. *Thin Solid Films* 228:1–11.

Graves, D. 1989. Plasma processing in microelectronics manufacturing. *AIChE Journal* 35:1–29.

Hase, W. and Bunker, D. 1973. Quantum chemistry program exchange (qcpe) 11, 234. Quantum Chemistry Program Exchange, Department of Chemistry, Indiana University, Bloomington, Ind.

Hasper, A., Kleijn, C., Holleman, J., Middelhoek, J., and Hoogendoorn, C. 1991. Modeling and optimization of the step coverage of tungsten LPCVD in trenches and contact holes. *J. Electrochem. Soc.* 138:1728–1738.

Hebb, J. B. and Jensen, K. F. 1996. The effect of multilayer patterns on temperature uniformity during rapid thermal processing. *J. Electrochem. Soc.* 143(3):1142–1151.

Hehre, W., Radom, L., Schleyer, P., and Pople, J. 1986. Ab Initio Molecular Orbital Theory. John Wiley & Sons, New York.

Heritage, J. R. (ed.) 1995. Special issue on PHOENICS-CVD and its applications. *PHOENICS J.* 8(4):402–552.

Hess, D., Jensen, K., and Anderson, T. 1985. Chemical vapor deposition: A chemical engineering perspective. *Rev. Chem. Eng.* 3:97–186.

Hirschfelder, J., Curtiss, C., and Bird, R. 1967. Molecular Theory of Gases and Liquids. John Wiley & Sons Inc., New York.

Hitchman, M. and Jensen, K. (eds.) 1993. Chemical Vapor Deposition: Principles and Applications. Academic Press, London.

Ho, P. and Melius, C. 1990. A theoretical study of the thermochemistry of sif_n and SiH_nF_m compounds and Si_2F_6. *J. Phys. Chem.* 94:5120–5127.

Ho, P., Coltrin, M., Binkley, J., and Melius, C. 1985. A theoretical study of the heats of formation of $SiH_n, SiCl_n$, and SiH_nCl_m compounds. *J. Phys. Chem.* 89:4647–4654.

Ho, P., Coltrin, M., Binkley, J., and Melius, C. 1986. A theoretical study of the heats of formation of Si_2H_n ($n = 0-6$) compounds and trisilane. *J. Phys. Chem.* 90:3399–3406.

Hockney, R. and Eastwood, J. 1981. Computer simulations using particles. McGraw-Hill, New York.

Holleman, J., Hasper, A., and Kleijn, C. 1993. Loading effects on kinetical and electrical aspects of silane-reduced low-pressure chemical vapor deposited selective tungsten. *J. Electrochem. Soc.* 140:818–825.

Holstein, W., Fitzjohn, J., Fahy, E., Golmour, P., and Schmelzer, E. 1989. Mathematical modeling of cold-wall channel CVD reactors. *J. Crystal Growth* 94:131–144.

Hopfmann, C., Werner, C., and Ulacia, J. 1991. Numerical analysis of fluid flow and non-uniformities in a polysilicon LPCVD batch reactor. *Appl. Surf. Sci.* 52:169–187.

Howell, J. 1968. Application of Monte Carlo to heat transfer problems. *In* Advances in Heat Transfer (J. Hartnett and T. Irvine, eds.), Vol. 5. Academic Press, New York.

Ikegawa, M. and Kobayashi, J. 1989. Deposition profile simulation using the direct simulation Monte Carlo method. *J. Electrochem. Soc.* 136:2982–2986.

Jansen, A., Orazem, M., Fox, B., and Jesser, W. 1991. Numerical study of the influence of reactor design on MOCVD with a comparison to experimental data. *J. Crystal Growth* 112:316–336.

Jensen, K. 1987. Micro-reaction engineering applications of reaction engineering to processing of electronic and photonic materials. *Chem. Eng. Sci.* 42:923–958.

Jensen, K. and Graves, D. 1983. Modeling and analysis of low pressure CVD reactors. *J. Electrochem. Soc.* 130:1950–1957.

Jensen, K., Mihopoulos, T., Rodgers, S., and Simka, H. 1996. CVD simulations on multiplelength scales. *In* CVD XIII: Proceedings of the 13th International Conference on Chemical Vapor Deposition (T. Besman, M. Allendorf, M. Robinson, and R. Ulrich, eds.) pp. 67–74. Electrochemical Society, Pennington, N.J.

Jensen, K. F., Einset, E., and Fotiadis, D. 1991. Flow phenomena in chemical vapor deposition of thin films. *Annu. Rev. Fluid Mech.* 23:197–232.

Jones, A. and O'Brien, P. 1997. CVD of compound semiconductors. VCH, Weinheim, Germany.

Kalindindi, S. and Desu, S. 1990. Analytical model for the low pressure chemical vapor deposition of SiO_2 from tetraethoxysilane. *J. Electrochem. Soc*. 137:624–628.

Kee, R., Rupley, F., and Miller, J. 1989. *Chemkin-II: A Fortran chemical kinetics package for the analysis of gas-phase chemical kinetics*. Technical Report SAND89-8009B.UC-706. Sandia National Laboratories, Albuquerque, N.M.

Kee, R., Rupley, F., and Miller, J. 1990. *The Chemkin thermodynamic data base*. Technical Report SAND87-8215B.UC-4. Sandia National Laboratories, Albuquerque, N.M./Livermore, Calif.

Kee, R., Dixon-Lewis, G., Warnatz, J., Coltrin, M., and Miller, J. 1991. *A FORTRAN computer code package for the evaluation of gas-phase multicomponent transport properties*. Technical Report SAND86-8246.UC-401. Sandia National Laboratories, Albuquerque, N.M./Livermore, Calif.

Kersch, A. 1995a. Radiative heat transfer modeling. *Phoenics J*. 8:421–438.

Kersch, A. 1995b. RTP reactor simulations. *Phoenics J*. 8:500–511.

Kersch, A. and Morokoff, W. 1995. Transport Simulation in Microelectronics. Birkhuser, Basel.

Kleijn, C. 1991. A mathematical model of the hydrodynamics and gas-phase reactions in silicon LPCVD in a single-wafer reactor. *J. Electrochem. Soc*. 138:2190–2200.

Kleijn, C. 1995. Chemical vapor deposition processes. *In* Computational Modeling in Semiconductor Processing (M. Meyyappan, ed.) pp. 97–229. Artech House, Boston.

Kleijn, C. and Hoogendoorn, C. 1991. A study of 2- and 3-d transport phenomena in horizontal chemical vapor deposition reactors. *Chem. Eng. Sci*. 46:321–334.

Kleijn, C. and Kuijlaars, K. 1995. The modeling of transport phenomena in CVD reactors. *Phoenics J*. 8:404–420.

Kleijn, C. and Werner, C. 1993. Modeling of Chemical Vapor Deposition of Tungsten Films. Birkhuser, Basel.

Kline, L. and Kushner, M. 1989. Computer simulations of materials processing plasma discharges. *Crit. Rev. Solid State Mater. Sci*. 16:1–35.

Knudsen, M. 1934. Kinetic Theory of Gases. Methuen and Co. Ltd., London.

Kodas, T. and Hampden-Smith, M. 1994. The chemistry of metal CVD. VCH, Weinheim, Germany.

Koh, J. and Woo, S. 1990. Computer simulation study on atmospheric pressure CVD process for amorphous silicon carbide. *J. Electrochem. Soc*. 137:2215–2222.

Kristof, J., Song, L., Tsakalis, T., and Cale, T. 1997. Programmed rate and optimal control chemical vapor deposition of tungsten. *In* Chemical Vapor Deposition: Proceedings of the 14th International Conference and EUROCVD-II (M. Allendorf and C. Bernard, eds.) pp. 1566–1573. Electrochemical Society, Pennington, N.J.

Kuijlaars, K. 1996. Detailed Modeling of Chemistry and Transport in CVD Reactors—Application to Tungsten LPCVD. Ph.D. thesis, Delft University of Technology, The Netherlands.

Laidler, K. 1987. Chemical Kinetics (3rd ed.). Harper and Row, New York.

l'Air Liquide, D. S. 1976. Encyclopdie des Gaz. Elseviers Scientific Publishing, Amsterdam.

Leusink, G., Kleijn, C., Oosterlaken, T., Janssen, G., and Radelaar, S. 1992. Growth kinetics and inhibition of growth of chemical vapor deposited thin tungsten films on silicon from tungsten hexafluoride. *J. Appl. Phys*. 72:490–498.

Liu, B., Hicks, R., and Zinck, J. 1992. Chemistry of photo-assisted organometallic vapor-phase epitaxy of cadmium telluride. *J. Crystal Growth* 123:500–518.

Lutz, A., Kee, R., and Miller, J. 1993. SENKIN: *A* FORTRAN *program for predicting homogeneous gas phase chemical kinetics with sensitivity analysis*. Technical Report SAND87-8248.UC-401. Sandia National Laboratories, Albuquerque, N.M.

Maitland, G. and Smith, E. 1972. Critical reassessment of viscosities of 11 common gases. *J. Chem. Eng. Data* 17:150–156.

Masi, M., Simka, H., Jensen, K., Kuech, T., and Potemski, R. 1992. Simulation of carbon doping of GaAs during MOVPE. *J. Crystal Growth* 124:483–492.

Meeks, E., Kee, R., Dandy, D., and Coltrin, M. 1992. Computational simulation of diamond chemical vapor deposition in premixed $C_2H_2/O_2/H_2$ and CH_4/O_2–strained flames. *Combust. Flame* 92:144–160.

Melius, C., Allendorf, M., and Coltrin, M. 1997. Quantum chemistry: A review of *ab initio* methods and their use in predicting thermochemical data for CVD processes. *In* Chemical Vapor Deposition: Proceedings of the 14th International Conference and EUROCVD-II. (M. Allendorf and C. Bernard, eds.) pp. 1–14. Electrochemical Society, Pennington, N.J.

Meyyappan, M. (ed.) 1995a. Computational Modeling in Semiconductor Processing. Artech House, Boston.

Meyyappan, M. 1995b. Plasma process modeling. *In* Computational Modeling in Semiconductor Processing (M. Meyyappan, ed.) pp. 231–324. Artech House, Boston.

Minkowycz, W., Sparrow, E., Schneider, G., and Pletcher, R. 1988. Handbook of Numerical Heat Transfer. John Wiley & Sons, New York.

Moffat, H. and Jensen, K. 1986. Complex flow phenomena in MOCVD reactors. I. Horizontal reactors. *J. Crystal Growth* 77:108–119.

Moffat, H., Jensen, K., and Carr, R. 1991a. Estimation of the Arrhenius parameters for $SiH_4 \rightleftharpoons SiH_2 + H_2$ and $\Delta H_f(SiH_2)$ by a nonlinear regression analysis of the forward and reverse reaction rate data. *J. Phys. Chem*. 95:145–154.

Moffat, H., Glarborg, P., Kee, R., Grcar, J., and Miller, J. 1991b. *SURFACE PSR: A FORTRAN Program for Modeling Well-Stirred Reactors with Gas and Surface Reactions*. Technical Report SAND91-8001.UC-401. Sandia National Laboratories, Albuquerque, N.M./Livermore, Calif.

Motz, H. and Wise, H. 1960. Diffusion and heterogeneous reaction. III. Atom recombination at a catalytic boundary. *J. Chem. Phys*. 31:1893–1894.

Mountziaris, T. and Jensen, K. 1991. Gas-phase and surface reaction mechanisms in MOCVD of GaAs with trimethyl-gallium and arsine. *J. Electrochem. Soc*. 138:2426–2439.

Mountziaris, T., Kalyanasundaram, S., and Ingle, N. 1993. A reaction-transport model of GaAs growth by metal organic chemical vapor deposition using trimethyl-gallium and tertiary-butyl-arsine. *J. Crystal Growth* 131:283–299.

Okkerse, M., Klein-Douwel, R., de Croon, M., Kleijn, C., ter Meulen, J., Marin, G., and van den Akker, H. 1997. Simulation of a diamond oxy-acetylene combustion torch reactor with a reduced gas-phase and surface mechanism. *In* Chemical Vapor Deposition: Proceedings of the 14th International Conference and EUROCVD-II (M. Allendorf and C. Bernard, eds.) pp. 163–170. Electrochemical Society, Pennington, N.J.

Ordal, M., Bell, R., Alexander, R., Long, L., and Querry, M. 1985. Optical properties of fourteen metals in the infrared and far infrared. *Appl. Optics* 24:4493.

Palik, E. 1985. Handbook of Optical Constants of Solids. Academic Press, New York.

Park, H., Yoon, S., Park, C., and Chun, J. 1989. Low pressure chemical vapor deposition of blanket tungsten using a gaseous mixture of WF_6, SiH_4 and H_2. *Thin Solid Films* 181:85–93.

Patankar, S. 1980. Numerical Heat Transfer and Fluid Flow. Hemisphere Publishing, Washington, D.C.

Peev, G., Zambov, L., and Yanakiev, Y. 1990a. Modeling and optimization of the growth of polycrystalline silicon films by thermal decomposition of silane. *J. Crystal Growth* 106:377–386.

Peev, G., Zambov, L., and Nedev, I. 1990b. Modeling of low pressure chemical vapour deposition of Si_3N_4 thin films from dichlorosilane and ammonia. *Thin Solid Films* 190:341–350.

Pierson, H. 1992. Handbook of Chemical Vapor Deposition. Noyes Publications, Park Ridge, N.J.

Raupp, G. and Cale, T. 1989. Step coverage prediction in low-pressure chemical vapor deposition. *Chem. Mater.* 1:207–214.

Rees, W. Jr. (ed.) 1996. CVD of Nonmetals. VCH, Weinheim, Germany.

Reid, R., Prausnitz, J., and Poling, B. 1987. The Properties of Gases and Liquids (2nd ed.). McGraw-Hill, New York.

Rey, J., Cheng, L., McVittie, J., and Saraswat, K. 1991. Monte Carlo low pressure deposition profile simulations. *J. Vac. Sci. Techn. A* 9:1083–1087.

Robinson, P. and Holbrook, K. 1972. Unimolecular Reactions. Wiley-Interscience, London.

Rodgers, S. T. and Jensen, K. F. 1998. Multiscale monitoring of chemical vapor deposition. *J. Appl. Phys.* 83(1):524–530.

Roenigk, K. and Jensen, K. 1987. Low pressure CVD of silicon nitride. *J. Electrochem. Soc.* 132:448–454.

Roenigk, K., Jensen, K., and Carr, R. 1987. Rice-Rampsberger-Kassel-Marcus theoretical prediction of high-pressure Arrhenius parameters by non-linear regression: Application to silane and disilane decomposition. *J. Phys. Chem.* 91:5732–5739.

Schmitz, J. and Hasper, A. 1993. On the mechanism of the step coverage of blanket tungsten chemical vapor deposition. *J. Electrochem. Soc.* 140:2112–2116.

Sherman, A. 1987. Chemical Vapor Deposition for Microelectronics. Noyes Publications, New York.

Siegel, R. and Howell, J. 1992. Thermal Radiation Heat Transfer (3rd ed.). Hemisphere Publishing, Washington, D.C.

Simka, H., Hierlemann, M., Utz, M., and Jensen, K. 1996. Computational chemistry predictions of kinetics and major reaction pathways for germane gas-phase reactions. *J. Electrochem. Soc.* 143:2646–2654.

Slater, N. 1959. Theory of Unimolecular Reactions. Cornell Press, Ithaca, N.Y.

Steinfeld, J., Fransisco, J., and Hase, W. 1989. Chemical Kinetics and Dynamics. Prentice-Hall, Englewood Cliffs, N.J.

Stewart, J. 1983. Quantum chemistry program exchange (qcpe), no. 455. Quantum Chemistry Program Exchange, Department of Chemistry, Indiana University, Bloomington, Ind.

Stull, D. and Prophet, H. (eds.). 1974–1982. JANAF thermochemical tables volume NSRDS-NBS 37. NBS, Washington D.C., second edition. Supplements by Chase, M. W., Curnutt, J. L., Hu, A. T., Prophet, H., Syverud, A. N., Walker, A. C., McDonald, R. A., Downey, J. R., Valenzuela, E. A., *J. Phys. Ref. Data.* 3, p. 311 (1974); 4, p. 1 (1975); 7, p. 793 (1978); 11, p. 695 (1982).

Svehla, R. 1962. *Estimated Viscosities and Thermal Conductivities of Gases at High Temperatures*. Technical Report R-132 NASA. National Aeronautics and Space Administration, Washington, D.C.

Taylor, C. and Hughes, T. 1981. Finite Element Programming of the Navier-Stokes Equations. Pineridge Press Ltd., Swansea, U.K.

Tirtowidjojo, M. and Pollard, R. 1988. Elementary processes and rate-limiting factors in MOVPE of GaAs. *J. Crystal Growth* 77:108–114.

Visser, E., Kleijn, C., Govers, C., Hoogendoorn, C., and Giling, L. 1989. Return flows in horizontal MOCVD reactors studied with the use of TiO particle injection and numerical calculations. *J. Crystal Growth* 94:929–946 (Erratum 96:732–735).

Vossen, J. and Kern, W. (eds.) 1991. Thin Film Processes II. Academic Press, Boston.

Wagman, D., Evans, W., Parker, V., Schumm, R., Halow, I., Bailey, S., Churney, K., and Nuttall, R. 1982. The NBS tables of chemical thermodynamic properties. *J. Phys. Chem. Ref. Data* 11 (Suppl. 2).

Wahl, G. 1977. Hydrodynamic description of CVD processes. *Thin Solid Films* 40:13–26.

Wang, Y. and Pollard, R. 1993. A mathematical model for CVD of tungsten from tungstenhexafluoride and silane. *In* Advanced Metallization for ULSI Applications in 1992 (T. Cale and F. Pintchovski, eds.) pp. 169–175. Materials Research Society, Pittsburgh.

Wang, Y.-F. and Pollard, R. 1994. A method for predicting the adsorption energetics of diatomic molecules on metal surfaces. *Surface Sci.* 302:223–234.

Wang, Y.-F. and Pollard, R. 1995. An approach for modeling surface reaction kinetics in chemical vapor deposition processes. *J. Electrochem. Soc.* 142:1712–1725.

Weast, R. (ed.) 1984. Handbook of Chemistry and Physics. CRC Press, Boca Raton, Fla.

Werner, C., Ulacia, J., Hopfmann, C., and Flynn, P. 1992. Equipment simulation of selective tungsten deposition. *J. Electrochem. Soc.* 139:566–574.

Wulu, H., Saraswat, K., and McVitie, J. 1991. Simulation of mass transport for deposition in via holes and trenches. *J. Electrochem. Soc.* 138:1831–1840.

Zachariah, M. and Tsang, W. 1995. Theoretical calculation of thermochemistry, energetics, and kinetics of high-temperature $Si_xH_yO_z$ reactions. *J. Phys. Chem.* 99:5308–5318.

Zienkiewicz, O. and Taylor, R. 1989. The Finite Element Method (4th ed.). McGraw-Hill, London.

KEY REFERENCES

Hitchman and Jensen, 1993. See above.

Extensive treatment of the fundamental and practical aspects of CVD processes, including experimental diagnostics and modeling

Meyyappan, 1995. See above.

Comprehensive review of the fundamentals and numerical aspects of CVD, crystal growth and plasma modeling. Extensive literature review up to 1993

The Phoenics Journal, Vol. 8 (4), 1995.

Various articles on theory of CVD and PECVD modeling. Nice illustrations of the use of modeling in reactor and process design

CHRIS R. KLEIJN
Delft University of Technology
Delft, The Netherlands

MAGNETISM IN ALLOYS

INTRODUCTION

The human race has used magnetic materials for well over 2000 years. Today, magnetic materials power the world, for they are at the heart of energy conversion devices, such as generators, transformers, and motors, and are major components in automobiles. Furthermore, these materials will be important components in the energy-efficient vehicles of tomorrow. More recently, besides the obvious advances in semiconductors, the computer revolution has been fueled by advances in magnetic storage devices, and will continue to be affected by the development of new multicomponent high-coercivity magnetic alloys and multilayer coatings. Many magnetic materials are important for some of their other properties which are superficially unrelated to their magnetism. Iron steels and iron-nickel (so-called "Invar", or volume INVARiant) alloys are two important examples from a long list. Thus, to understand a wide range of materials, the origins of magnetism, as well as the interplay with alloying, must be uncovered. A quantum-mechanical description of the electrons in the solid is needed for such understanding, so as to describe, on an equal footing and without bias, as many key microscopic factors as possible. Additionally, many aspects, such as magnetic anisotropy and hence permanent magnetism, need the full power of relativistic quantum electrodynamics to expose their underpinnings.

From Atoms to Solids

Experiments on atomic spectra, and the resulting highly abundant data, led to several empirical rules which we now know as Hund's rules. These rules describe the filling of the atomic orbitals with electrons as the atomic number is changed. Electrons occupy the orbitals in a shell in such a way as to maximize both the total spin and the total angular momentum. In the transition metals and their alloys, the orbital angular momentum is almost "quenched;" thus the spin Hund's rule is the most important. The quantum mechanical reasons behind this rule are neatly summarized as a combination of the Pauli exclusion principle and the electron-electron (Coulomb) repulsion. These two effects lead to the so-called "exchange" interaction, which forces electrons with the same spin states to occupy states with different spatial distribution, i.e., with different angular momentum quantum numbers. Thus the exchange interaction has its origins in minimizing the Coulomb energy locally—i.e., the intrasite Coulomb energy—while satisfying the other constraints of quantum mechanics. As we will show later in this unit, this minimization in crystalline metals can result in a competition between intrasite (local) and intersite (extended) effects—i.e. kinetic energy stemming from the curvature of the wave functions. When the overlaps between the orbitals are small, the intrasite effects dominate, and magnetic moments can form. When the overlaps are large, electrons hop from site to site in the lattice at such a rate that a local moment cannot be sustained. Most of the existing solids are characterized in terms of this latter picture. Only in a few places in the periodic table does the former picture closely reflect reality. We will explore one of these places here, namely the 3*d* transition metals.

In the 3*d* transition metals, the states that are derived from the 3*d* and 4*s* atomic levels are primarily responsible for a metal's physical properties. The 4*s* states, being more spatially extended (higher principal quantum number), determine the metal's overall size and its compressibility. The 3*d* states are more (but not totally) localized and give rise to a metal's magnetic behavior. Since the 3*d* states are not totally localized, the electrons are considered to be mobile giving rise to the name "itinerant" magnetism for such cases.

At this point, we want to emphasize that moment formation and the alignment of these moments with each other have different origins. For example, magnetism plays an important role in the stability of stainless steel, FeNiCr. Although it is not ferromagnetic (having zero net magnetization), the moments on its individual constituents have not disappeared; they are simply not aligned. Moments may exist on the atomic scale, but they might not point in the same direction, even at near-zero temperatures. The mechanisms that are responsible for the moments and for their alignment depend on different aspects of the electronic structure. The former effect depends on the gross features, while the latter depends on very detailed structure of the electronic states.

The itinerant nature of the electrons makes magnetism and related properties difficult to model in transition metal alloys. On the other hand, in magnetic insulators the exchange interactions causing magnetism can be represented rather simply. Electrons are appropriately associated with particular atomic sites so that "spin" operators can be specified and the famous Heisenberg-Dirac Hamiltonian can then be used to describe the behavior of these systems. The Hamiltonian takes the following form,

$$H = -\sum_{ij} J_{ij} \hat{S}_i \cdot \hat{S}_j \qquad (1)$$

in which J_{ij} is an "exchange" integral, measuring the size of the electrostatic and exchange interaction and $\hat{S}_i$ is the spin vector on site i. In metallic systems, it is not possible to allocate the itinerant electrons in this way and such pairwise intersite interactions cannot be easily identified. In such metallic systems, magnetism is a complicated many-electron effect to which Hund's rules contribute. Many have labored with significant effort over a long period to understand and describe it.

One common approach involves a mapping of this problem onto one involving independent electrons moving in the fields set up by all the other electrons. It is this aspect that gives rise to the spin-polarized band structure, an often used basis to explain the properties of metallic magnets. However, this picture is not always sufficient. Herring (1966), among others, noted that certain components of metallic magnetism can also be discussed using concepts of localized spins which are, strictly speaking, only relevant to magnetic insulators. Later on in this unit, we discuss how the two pictures have been

combined to explain the temperature dependence of the magnetic properties of bulk transition metals and their alloys.

In certain metals, such as stainless steel, magnetism is subtly connected with other properties via the behavior of the spin-polarized electronic structure. Dramatic examples are those materials which show a small thermal expansion coefficient below the Curie temperature, T_c, a large forced expansion in volume when an external magnetic field is applied, a sharp decrease of spontaneous magnetization and of the Curie temperature when pressure is applied, and large changes in the elastic constants as the temperature is lowered through T_c. These are the famous "Invar" materials, so called because these properties were first found to occur in the fcc alloys Fe-Ni (65% Fe), Fe-Pd, and Fe-Pt (Wassermann, 1991). The compositional order of an alloy is often intricately linked with its magnetic state, and this can also reveal physically interesting and technologically important new phenomena. Indeed, some alloys, such as $Ni_{75}Fe_{25}$, develop directional chemical order when annealed in a magnetic field (Chikazurin and Graham, 1969). Magnetic short-range correlations above T_c, and the magnetic order below, weaken and alter the chemical ordering in iron-rich Fe-Al alloys, so that a ferromagnetic $Fe_{80}Al_{20}$ alloy forms a DO_3 ordered structure at low temperatures, whereas paramagnetic $Fe_{75}Al_{25}$ forms a B2 ordered phase at comparatively higher temperatures (Stephens, 1985; McKamey et al., 1991; Massalski et al., 1990; Staunton et al., 1997). The magnetic properties of many alloys are sensitive to the local environment. For example, ordered Ni-Pt (50%) is an anti-ferromagnetic alloy (Kuentzler, 1980), whereas its disordered counterpart is ferromagnetic (MAGNETIC MOMENT AND MAGNETIZATION, MAGNETIC NEUTRON SCATTERING). The main part of this unit is devoted to a discussion of the basis underlying such magneto-compositional effects.

Since the fundamental electrostatic exchange interactions are isotropic, and do not couple the direction of magnetization to any spatial direction, they fail to give a basis for a description of magnetic anisotropic effects which lie at the root of technologically important magnetic properties, including domain wall structure, linear magnetostriction, and permanent magnetic properties in general. A description of these effects requires a relativistic treatment of the electrons' motions. A section of this unit is assigned to this aspect as it touches the properties of transition metal alloys.

PRINCIPLES OF THE METHOD

The Ground State of Magnetic Transition Metals: Itinerant Magnetism at Zero Temperature

Hohenberg and Kohn (1964) proved a remarkable theorem stating that the ground state energy of an interacting many-electron system is a unique functional of the electron density $n(\mathbf{r})$. This functional is a minimum when evaluated at the true ground-state density $n_o(\mathbf{r})$. Later Kohn and Sham (1965) extended various aspects of this theorem, providing a basis for practical applications of the density functional theory. In particular, they derived a set of single-particle equations which could include all the effects of the correlations between the electrons in the system. These theorems provided the basis of the modern theory of the electronic structure of solids.

In the spirit of Hartree and Fock, these ideas form a scheme for calculating the ground-state electron density by considering each electron as moving in an effective potential due to all the others. This potential is not easy to construct, since all the many-body quantum-mechanical effects have to be included. As such, approximate forms of the potential must be generated.

The theorems and methods of the density functional (DF) formalism were soon generalized (von Barth and Hedin, 1972; Rajagopal and Callaway, 1973) to include the freedom of having different densities for each of the two spin quantum numbers. Thus the energy becomes a functional of the particle density, $n(\mathbf{r})$, and the local magnetic density, $\mathbf{m}(\mathbf{r})$. The former is sum of the spin densities, the latter, the difference. Each electron can now be pictured as moving in an effective magnetic field, $\mathbf{B}(\mathbf{r})$, as well as a potential, $V(\mathbf{r})$, generated by the other electrons. This spin density functional theory (SDFT) is important in systems where spin-dependent properties play an important role, and provides the basis for the spin-polarized electronic structure mentioned in the introduction. The proofs of the basic theorems are provided in the originals and in the many formal developments since then (Lieb, 1983; Driezler and da Providencia, 1985).

The many-body effects of the complicated quantum-mechanical problem are hidden in the exchange-correlation functional $E_{xc}[n(\mathbf{r}), \mathbf{m}(\mathbf{r})]$. The exact solution is intractable; thus some sort of approximation must be made. The local approximation (LSDA) is the most widely used, where the energy (and corresponding potential) is taken from the uniformly spin-polarized homogeneous electron gas (see SUMMARY OF ELECTRONIC STRUCTURE METHODS and PREDICTION OF PHASE DIAGRAMS). Point by point, the functional is set equal to the exchange and correlation energies of a homogeneously polarized electron gas, ε_{xc}, with the density and magnetization taken to be the local values, $E_{xc}[n(\mathbf{r}), \mathbf{m}(\mathbf{r})] = \int \varepsilon_{xc}[n(\mathbf{r}), \mathbf{m}(\mathbf{r})]\, n(\mathbf{r})\, d\mathbf{r}$ (von Barth and Hedin, 1972; Hedin and Lundqvist, 1971; Gunnarsson and Lundqvist, 1976; Ceperley and Alder, 1980; Vosko et al., 1980).

Since the "landmark" papers on Fe and Ni by Callaway and Wang (1977), it has been established that spin-polarized band theory, within this Spin Density Functional formalism (see reviews by Rajagopal, 1980; Kohn and Vashishta, 1982; Driezler and da Providencia, 1985; Jones and Gunnarsson, 1989) provides a reliable quantitative description of magnetic properties of transition metal systems at low temperatures (Gunnarsson, 1976; Moruzzi et al., 1978; Koelling, 1981). In this modern version of the Stoner-Wohlfarth theory (Stoner, 1939; Wohlfarth, 1953), the magnetic moments are assumed to originate predominately from itinerant d electrons. The exchange interaction, as defined above, correlates the spins on a site, thus creating a local moment. In a ferromagnetic metal, these moments are aligned so that the systems possess a finite magnetization per site (see GENERATION AND MEASUREMENT OF MAGNETIC FIELDS, MAGNETIC MOMENT AND MAGNETIZATION,

and THEORY OF MAGNETIC PHASE TRANSITIONS). This theory provides a basis for the observed non-integer moments as well as the underlying many-electron nature of magnetic moment formation at $T = 0$ K.

Within the approximations inherent in LSDA, electronic structure (band theory) calculations for the pure crystalline state are routinely performed. Although most include some sort of shape approximation for the charge density and potentials, these calculations give a good representation of the electronic density of states (DOS) of these metals. To calculate the total energy to a precision of less than a few milli–electron volts and to reveal fine details of the charge and moment density, the shape approximation must be eliminated. Better agreement with experiment is found when using extensions of the LSDA. Nonetheless, the LSDA calculations are important in that the ground-state properties of the elements are reproduced to a remarkable degree of accuracy. In the following, we look at a typical LSDA calculation for bcc iron and fcc nickel.

Band theory calculations for bcc iron have been done for decades, with the results of Moruzzi et al. (1978) being the first of the more accurate LSDA calculations. The figure on p. 170 of their book (see Literature Cited) shows the electronic density of states (DOS) as a function of the energy. The density of states for the two spins are almost (but not quite) simply rigidly shifted. As typical of bcc structures, the d band has two major peaks. The Fermi energy resides in the top of d bands for the spins that are in the majority, and in the trough between the uppermost peaks. The iron moment extracted from this first-principles calculation is 2.2 Bohr magnetons per atom, which is in good agreement with experiment. Further refinements, such as adding the spin-orbit contributions, eliminating the shape approximation of the charge densities and potentials, and modifying the exchange-correlation function, push the calculations into better agreement with experiment. The equilibrium volume determined within LSDA is more delicate, with the errors being $\sim$3% about twice the amount for the typical nonmagnetic transition metal. The total energy of the ferromagnetic bcc phase was also found to be close to that of the nonmagnetic fcc phase, and only when improvements to the LSDA were incorporated did the calculations correctly find the former phase the more stable.

On the whole, the calculated properties for nickel are reproduced to about the same degree of accuracy. As seen in the plot of the DOS on p. 178 of Moruzzi et al. (1978), the Fermi energy lies above the top of the majority-spin d bands, but in the large peak in the minority-spin d bands. The width of the d band has been a matter of a great deal of scrutiny over the years, since the width as measured in photoemission experiments is much smaller than that extracted from band-theory calculations. It is now realized that the experiments measure the energy of various excited states of the metal, whereas the LSDA remains a good theory of the ground state. A more comprehensive theory of the photoemission process has resulted in a better, but by no means complete, agreement with experiment. The magnetic moment, a ground state quantity extracted from such calculations, comes out to be $\sim$0.6 Bohr magnetons per atom, close to the experimental measurements. The equilibrium volume and other such quantities are in good agreement with experiment, i.e., on the same order as for iron.

In both cases, the electronic bands, which result from the solution of the one-electron Kohn-Sham Schrödinger equations, are nearly rigidly exchange split. This rigid shift is in accord with the simple picture of Stoner-Wohlfarth theory which was based on a simple Hubbard model with a single tight-binding d band treated in the Hartree-Fock approximation. The model Hamiltonian is

$$\hat{H} = \sum_{ij,\sigma} (\varepsilon_0^\sigma \delta_{ij} + t_{ij}^\sigma) a_{i,\sigma}^\dagger a_{j,\sigma} + \frac{I}{2} \sum_{i,\sigma} a_{i,\sigma}^\dagger a_{i,\sigma} a_{i,-\sigma}^\dagger a_{i,-\sigma} \qquad (2)$$

in which $a_{i,\sigma}$ and $a_{i,\sigma}^\dagger$ are respectively the creation and annihilation operators, ε_0^σ a site energy (with spin index σ), t_{ij} a hopping parameter, inversely related to the d-band width, and I the many-body Hubbard parameter representing the intrasite Coulomb interactions. And within the Hartree-Fock approximation, a pair of operators is replaced by their average value, $\langle \ldots \rangle$, i.e., their quantum mechanical expectation value. In particular, $a_{i,\sigma}^\dagger \; a_{i,\sigma} a_{i,-\sigma}^\dagger a_{i,-\sigma} \approx a_{i\cdot\sigma}^\dagger a_{i,\sigma} \langle a_{i,-\sigma}^\dagger a_{i,-\sigma} \rangle$, where $\langle a_{i,-\sigma}^\dagger a_{i,-\sigma} \rangle = 1/2\,(\bar{n}_i - \mu_i \sigma)$. On each site, the average particle numbers are $\bar{n}_i = \bar{n}_{i,+1} + \bar{n}_{i,-1}$ and the moments are $\bar{\mu}_i = \bar{n}_{i,+1} - \bar{n}_{i,-1}$. Thus the Hartree-Fock Hamiltonian is given by

$$\hat{H} = \sum_{ij,\sigma} \left[\left(\varepsilon_0^\sigma + \frac{1}{2} I \bar{n}_i - \frac{1}{2} I \bar{\mu}_i \right) \delta_{ij} + t_{ij}^\sigma \right] a_{i,\sigma}^\dagger a_{j,\sigma} \qquad (3)$$

where $\bar{n}_i$ is the number of electrons on site i and $\bar{\mu}_i$ the magnetization. The terms $I\bar{n}_i/2$ and $I\bar{\mu}_i/2$ are the effective potential and magnetic fields, respectively. The main omission of this approximation is the neglect of the spin-flip particle-hole excitations and the associated correlations.

This rigidly exchange-split band-structure picture is actually valid only for the special cases of the elemental ferromagnetic transition metals Fe, Ni, and Co, in which the d bands are nearly filled, i.e., towards the end of the $3d$ transition metal series. Some of the effects which are to be extracted from the electronic structure of the alloys can be gauged within the framework of simple, single-d-band, tight-binding models. In the middle of the series, the metals Cr and Mn are anti-ferromagnetic; those at the end, Fe, Ni, and Co are ferromagnetic. This trend can be understood from a band-filling point of view. It has been shown (e.g. Heine and Samson, 1983) that the exchange-splitting in a nearly filled tight-binding d band lowers the system's energy and hence promotes ferromagnetism. On the other hand, the imposition of an exchange field that alternates in sign from site to site in the crystal lattice lowers the energy of the system with a half-filled d band, and hence drives anti-ferromagnetism. In the alloy analogy, almost-filled bands lead to phase separation, i.e., $k = 0$ ordering; half-filled bands lead to ordering with a zone-boundary wavevector. This latter case is the analog of antiferromagnetism. Although the electronic structure of SDF theory, which provides such good quantitative estimates of magnetic properties of metals when compared to the experimentally measured values, is somewhat more complicated than this; the gross features can be usefully discussed in this manner.

Another aspect from the calculations of pure magnetic metals—which have been reviewed comprehensively by Moruzzi and Marcus (1993), for example—that will prove topical for the discussion of 3*d* metallic alloys, is the variation of the magnetic properties of the 3*d* metals as the crystal lattice spacing is altered. Moruzzi et al. (1986) have carried out a systematic study of this phenomenon with their "fixed spin moment" (FSM) scheme. The most striking example is iron on an fcc lattice (Bagayoko and Callaway, 1983). The total energy of fcc Fe is found to have a global minimum for the nonmagnetic state and a lattice spacing of 6.5 atomic units (a.u.) or 3.44 Å. However, for an increased lattice spacing of 6.86 a.u. or 3.63 Å, the energy is minimized for a ferromagnetic state, with a magnetization per site, $\bar{\mu} \approx 1\mu_B$ (Bohr magneton). With a marginal expansion of the lattice from this point, the ferromagnetic state strengthens with $\bar{\mu} \approx 2.4\,\mu_B$. These trends have also been found by LMTO calculations for non-collinear magnetic structures (Mryasov et al., 1992). There is a hint, therefore, that the magnetic properties of fcc iron alloys are likely to be connected to the alloy's equilibrium lattice spacing and vice versa. Moreover these properties are sensitive to both thermal expansion and applied pressure. This apparently is the origin of the "low spin–high spin" picture frequently cited in the many discussions of iron Invar alloys (Wassermann, 1991). In their review article, Moruzzi and Marcus (1993) have also summarized calculations on other 3*d* metals noting similar connections between magnetic structure and lattice spacing. As the lattice spacing is increased beyond the equilibrium value, the electronic bands narrow, and thus the magnetic tendencies are enhanced. More discussion on this aspect is included with respect to Fe-Ni alloys, below.

We now consider methods used to calculate the spin-polarized electronic structure of the ferromagnetic 3*d* transition metals when they are alloyed with other metallic components. Later we will see the effects on the magnetic properties of these materials where, once again, the rigidly split band structure picture is an inappropriate starting point.

Solid-Solution Alloys

The self-consistent Korringa-Kohn-Rostoker coherent-potential approximation (KKR-CPA; Stocks et al., 1978; Stocks and Winter, 1982; Johnson et al., 1990) is a mean-field adaptation of the LSDA to systems with substitutional disorder, such as, solid-solution alloys, and this has been discussed in COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS. To describe the theory, we begin by recalling what the SDFT-LDA means for random alloys. The straightforward but computationally intractable track along which one could proceed involves solving the usual self-consistent Kohn-Sham single-electron equations for all configurations, and averaging the relevant expectation values over the appropriate ensemble of configurations to obtain the desired observables. To be specific, we introduce an occupation variable, ξ_i, which takes on the value 1 or 0; 1 if there is an A atom at the lattice site i, or 0 if the site is occupied by a B atom. To specify a configuration, we must then assign a value to these variables ξ_i at each site. Each configuration can be fully described by a set of these variables $\{\xi_i\}$. For an atom of type α on site k, the potential and magnetic field that enter the Kohn-Sham equations are not independent of its surroundings and depend on all the occupation variables, i.e., $V_{k,\alpha}(\mathbf{r}, \{\xi_i\})$, $\mathbf{B}_{k,\alpha}(\mathbf{r},\{\xi_i\})$. To find the ensemble average of an observable, for each configuration, we must first solve (self-consistently) the Kohn-Sham equations. Then for each configuration, we are able to calculate the relevant quantity. Finally, by summing these results, weighted by the correct probability factor, we find the required ensemble average.

It is impossible to implement all of the above sequence of calculations as described, and the KKR-CPA was invented to circumvent these computational difficulties. The first premise of this approach is that the occupation of a site, by an A atom or a B atom, is independent of the occupants of any other site. This means that we neglect short-range order for the purposes of calculating the electronic structure and approximate the solid solution by a random substitutional alloy. A second premise is that we can invert the order of solving the Kohn-Sham equations and averaging over atomic configurations, i.e., find a set of Kohn-Sham equations that describe an appropriate "average" medium. The first step is to replace, in the spirit of a mean-field theory, the local potential function $V_{k,\alpha}(\mathbf{r},\{\xi_i\})$ and magnetic field $\mathbf{B}_{k,\alpha}(\mathbf{r},\{\xi_i\})$ with $V_{k,\alpha}(\mathbf{r})$ and $\mathbf{B}_{k,\alpha}(\mathbf{r})$, the average over all the occupation variables except the one referring to the site k, at which the occupying atom is known to be of type α. The motion of an electron, on the average, through a lattice of these potentials and magnetic fields randomly distributed with the probability c that a site is occupied by an A atom, and 1-c by a B atom, is obtained from the solution of the Kohn-Sham equations using the CPA (Soven, 1967). Here, a lattice of identical effective potentials and magnetic fields is constructed such that the motion of an electron through this ordered array closely resembles the motion of an electron, on the average, through the disordered alloy. The CPA determines the effective medium by insisting that the substitution of a single site of the CPA lattice by either an A or a B atom produces no further scattering of the electron on the average. It is then possible to develop a spin density functional theory and calculational scheme in which the partially averaged electronic densities, $n_A(\mathbf{r})$ and $n_B(\mathbf{r})$, and the magnetization densities $\mu_A(\mathbf{r})$, $\mu_B(\mathbf{r})$, associated with the A and B sites respectively, total energies, and other equilibrium quantities are evaluated (Stocks and Winter, 1982; Johnson et al., 1986, 1990; Johnson and Pinski, 1993).

The data from both x-ray and neutron scattering in solid solutions show the existence of Bragg peaks which define an underlying "average" lattice (see Chapters 10 and 13). This symmetry is evident in the average electronic structure given by the CPA. The Bloch wave vector is still a useful quantum number, but the average Bloch states also have a finite lifetime as a consequence of the disorder. Probably the strongest evidence for accuracy of the calculated electron lifetimes (and velocities) are the results for the residual resistivity of Ag-Pd alloys (Butler and Stocks, 1984; Swihart et al., 1986).

The electron movement through the lattice can be described using multiple-scattering theory, a Green's-function method, which is sometimes called the Korringa-Kohn-Rostoker (KKR) method. In this merger of multiple scattering theory with the coherent potential approximation (CPA), the ensemble-averaged Green's function is calculated, its poles defining the averaged energy eigenvalue spectrum. For systems without disorder, such energy eigenvalues can be labeled by a Bloch wavevector, **k**, are real, and thus can be related to states with a definite momentum and have infinite lifetimes. The KKR-CPA method provides a solution for the averaged electronic Green's function in the presence of a random placement of potentials, corresponding to the random occupation of the lattice sites. The poles now occur at complex values, **k** is usually still a useful quantum number but in the presence of this disorder, (discrete) translation symmetry is not perfect, and electrons in these states are scattered as they traverse the lattice. The useful result of the KKR-CPA method is that it provides a configurationally averaged Green function, from which the ensemble average of various observables can be calculated (Faulkner and Stocks, 1980).

Recently, super-cell versions of approximate ensemble-averaging are being explored due to advances in computers and algorithms (Faulkner et al., 1997). However, strictly speaking, such averaging is limited by the size of the cell and the shape approximation for the potentials and charge density. Several interesting results have been obtained from such an approach (Abrikosov et al., 1995; Faulkner et al., 1998). Neither the single-site CPA and the super-cell approach are exact; they give comple mentary information about the electronic structure in alloys.

Alloy Electronic Structure and Slater-Pauling Curves

Before the reasons for the loss of the conventional Stoner picture of rigidly exchange-split bands can be laid out, we describe some typical features of the electronic structure of alloys. A great deal has been written on this subject, which demonstrates clearly how these features are also connected with the phase stability of the system. An insight into this subject can be gained from many books and articles (Johnson et al., 1987; Pettifor, 1995; Ducastelle, 1991; Györffy et al., 1989; Connolly and Williams, 1983; Zunger, 1994; Staunton et al., 1994).

Consider two elemental d-electron densities of states, each with approximate width W and one centered at energy ε_A, the other at ε_B, related to atomic-like d-energy levels. If $(\varepsilon_A - \varepsilon_B) \gg W$ then the alloy's densities of states will be "split band" in nature (Stocks et al., 1978) and, in Pettifor's language, an ionic bond is established as charge flows from the A atoms to the B atoms in order to equilibrate the chemical potentials. The virtual bound states associated with impurities in metals are rough examples of split band behavior. On the other hand, if $(\varepsilon_{A\ -\ \varepsilon B}) \ll W$, then the alloy's electronic structure can be categorized as "common-band"-like. Large-scale hybridization now forms between states associated with the A and B atoms. Each site in the alloy is nearly charge-neutral as an individual ion is efficiently screened by the metallic response function of the alloy (Ziman, 1964). Of course, the actual interpretation of the detailed electronic structure involving many bands is often a complicated mixture of these two models.

In either case, half-filling of the bands lowers the total energy of the system as compared to the phase-separated case (Heine and Samson, 1983; Pettifor, 1995; Ducastelle, 1991), and an ordered alloy will form at low temperatures. When magnetism is added to the problem, an extra ingredient, namely the difference between the exchange field associated with each type of atomic species, is added. For majority spin electrons, a rough measure of the degree of "split-band" or "common-band" nature of the density of states is governed by $(\varepsilon_A^{\uparrow} - \varepsilon_B^{\uparrow})/W$ and a similar measure $(\varepsilon_A^{\downarrow} - \varepsilon_B^{\downarrow}/W$ for the minority spin electrons. If the exchange fields differ to any large extent, then for electrons of one spin-polarization, the bands are common-band-like while for the others a "split-band" label may be more appropriate. The outcome is a spin-polarized electronic structure that cannot be described by a rigid exchange splitting.

Hund's rules dictate that it is frequently energetically favorable for the majority-spin d states to be fully occupied. In many cases, at the cost of a small charge transfer, this is accomplished. Nickel-rich nickel-iron alloys provide such examples (Staunton et al., 1987) as shown in Figure 1. A schematic energy level diagram is shown in Figure 2.

One of the first tasks of theories or explanations based on electronic structure calculations is to provide a simple explanation of why the average magnetic moments per atom of so many alloys, M, fall on the famous Slater-Pauling curve, when plotted against the alloys' valence electron per atom ratio. The usual Slater-Pauling curve for 3d row (Chikazumi, 1964) consists of two straight lines. The plot rises from the beginning of the 3d row, abruptly changes the sign of its gradient and then drops smoothly to zero at the end of the row. There are some important groups of compounds and alloys whose parameters do not fall on this line, but, for these systems also, there appears to be some simple pattern.

For those ferromagnetic alloys of late transition metals characterized by completely filled majority spin d states, it is easy to see why they are located on the negative-gradient straight line. The magnetization per atom, $M = N_{\uparrow} - N_{\downarrow}$, where $N_{\uparrow(\downarrow)}$, describes the occupation of the majority (minority) spin states which can be trivially re-expressed in terms of the number of electrons per atom Z, so that $M = 2N_{\uparrow} - Z$. The occupation of the s and p states changes very little across the 3d row, and thus $M = 2N_{d\uparrow} - Z + 2N_{sp\uparrow}$, which gives $M = 10 - Z + 2N_{sp\uparrow}$.

Many other systems, most commonly bcc based alloys, are not strong ferromagnets in this sense of filled majority spin d bands, but possess a similar attribute. The chemical potential (or Fermi energy at T = 0 K) is pinned in a deep valley in the minority spin density of states (Johnson et al., 1987; Kubler, 1984). Pure bcc iron itself is a case in point, the chemical potential sitting in a trough in the minority spin density of states (Moruzzi et al., 1978, p. 170). Figure 1B shows another example in an iron-rich, iron-vanadium alloy. The other major segment of the Slater-Pauling curve of a positive-gradient straight line can be explained by

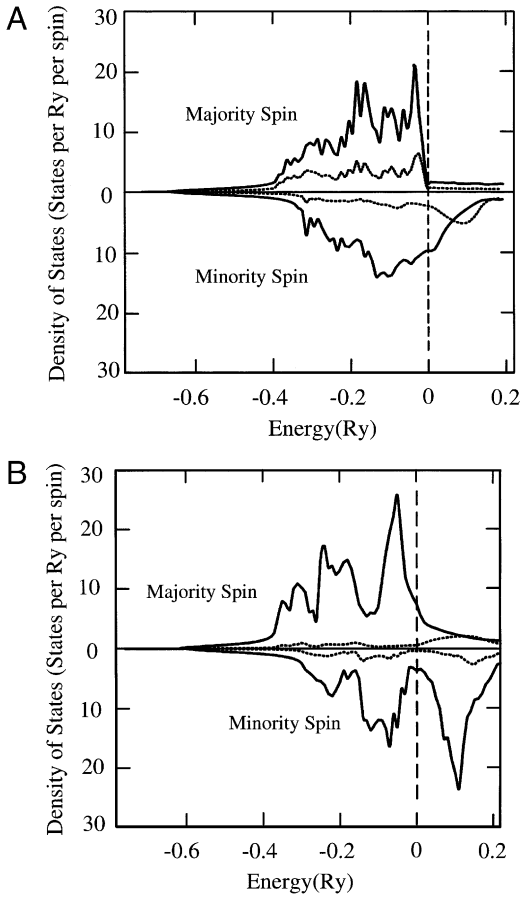


Figure 1. (**A**) The electronic density of states for ferromagnetic $Ni_{75}Fe_{25}$. The upper half displays the density of states for the majority-spin electrons, the lower half, for the minority-spin electrons. Note, in the lower half, the axis for the abscissa is inverted. These curves were calculated within the SCF-KKR-CPA, see Johnson et al. (1987). (**B**) The electronic density of states for ferromagnetic $Fe_{87}V_{13}$. The upper half displays the density of states for the majority-spin electrons; the lower half, for the minority-spin electrons. Note, in the lower half, the axis for the abscissa is inverted. These curves were calculated within the SCF-KKR-CPA (see Johnson et al., 1987).

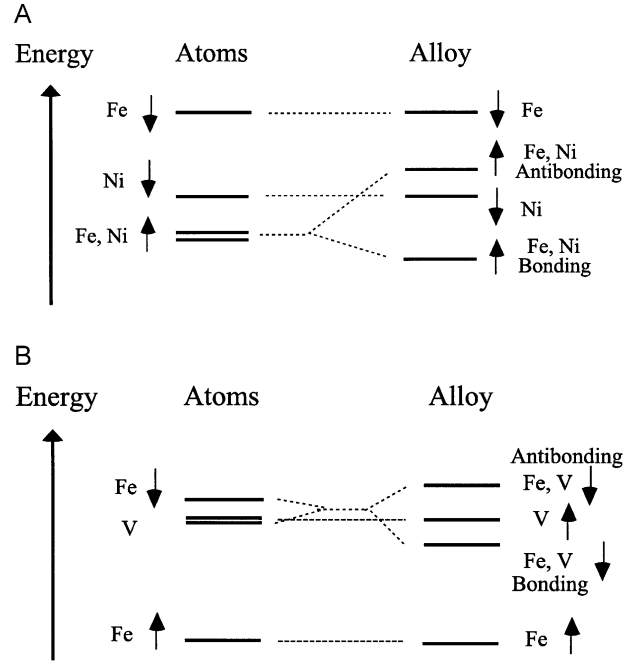


Figure 2. (**A**) Schematic energy level diagram for Ni-Fe alloys. (**B**) Schematic energy level diagram for Fe-V Alloys.

using this feature of the electronic structure. The pinning of the chemical potential in a trough of the minority spin d density of states constrains $N_{d\downarrow}$ to be fixed in all these alloys to be roughly three. In this circumstance the magnetization per atom $M = Z - 2N_{d\downarrow} - 2N_{sp\downarrow} = Z - 6 - 2N_{sp\downarrow}$. Further discussion on this topic is given by Malozemoff et al. (1984), Williams et al. (1984), Kubler (1984), Gubanov et al. (1992), and others. Later in this unit, to illustrate some of the remarks made, we will describe electronic structure calculations of three compositionally disordered alloys together with the ramifications for understanding of their properties.

Competitive and Related Techniques: Beyond the Local Spin-Density Approximation

Over the past few years, improved approximations for E_{xc} have been developed which maintain all the best features of the local approximation. A stimulus has been the work of Langreth and Mehl (1981, 1983), who supplied corrections to the local approximation in terms of the gradient of the density. Hu and Langreth (1986) have specified a spin-polarized generalization. Perdew and co-workers (Perdew and Yue, 1986; Wang and Perdew, 1991) contributed several improvements by ensuring that the generalized gradient approximation (GGA) functional satisfies some relevant sum rules. Calculations of the ground state properties of ferromagnetic iron and nickel were carried out (Bagno et al., 1989; Singh et al., 1991; Haglund, 1993) and compared to LSDA values. The theoretically estimated lattice constants from these calculations are slightly larger and are therefore more in line with the experimental values. When the GGA is used instead of LSDA, one removes a major embarrassment for LSDA calculations, namely that paramagnetic bcc iron is no longer energetically stable over ferromagnetic bcc iron. Further applications of the SDFT-GGA include one on the magnetic and cohesive properties of manganese in various crystal structures (Asada and Terakura, 1993) and another on the electronic and magnetic structure of the ordered B2 FeCo alloy (Liu and Singh, 1992). In addition, Perdew et al. (1992) have presented a comprehensive study of the GGA for a range of systems and have also given a review of the GGA (Perdew et al., 1996; Ernzerhof et al., 1996).

Notwithstanding the remarks made above, SDF theory within the local spin-density approximation (LSDA) provides a good quantitative description of the low-temperature properties of magnetic materials containing simple and transition metals, which are the main interests of this unit, and the Kohn-Sham electronic structure also gives a reasonable description of the quasi-particle

spectral properties of these systems. But it is not nearly so successful in its treatment of systems where some states are fairly localized, such as many rare-earth systems (Brooks and Johansson, 1993) and Mott insulators. Much work is currently being carried out to address the shortcomings found for these fascinating materials. Anisimov et al. (1997) noted that in exact density functional theory, the derivative of the total energy with respect to number of electrons, $\partial E/\partial N$, should have discontinuities at integral values of N, and that therefore the effective one-electron potential of the Kohn-Sham equations should also possess appropriate discontinuities. They therefore added an orbital-dependent correction to the usual LDA potentials and achieved an adequate description of the photoemission spectrum of NiO. As an example of other work in this area, Severin et al. (1993) have carried out self-consistent electronic structure calculations of rare-earth(R)-Co_2 and R-Co_2H_4 compounds within the LDA but in which the effect of the localized open $4f$ shell associated with the rare-earth atoms on the conduction band was treated by constraining the number of $4f$ electrons to be fixed. Brooks et al. (1997) have extended this work and have described crystal field quasiparticle excitations in rare earth compounds and extracted parameters for effective spin Hamiltonians. Another related approach to this constrained LSDA theory is the so-called "LSDA + U" method (Anisimov et al., 1997) which is also used to account for the orbital dependence of the Coulomb and exchange interactions in strongly correlated electronic materials.

It has been recognized for some time that some of the shortcomings of the LDA in describing the ground state properties of some strongly correlated systems may be due to an unphysical interaction of an electron with itself (Jones and Gunnarsson, 1989). If the exact form of the exchange-correlation functional $E_{\rm xc}$ were known, this self-interaction would be exactly canceled. In the LDA, this cancellation is not perfect. Several efforts improve cancellation by incorporating this self-interaction correction (SIC; Perdew and Zunger, 1981; Pederson et al., 1985). Using a cluster technique, Svane and Gunnarsson (1990) applied the SIC to transition metal oxides where the LDA is known to be particularly defective and where the GGA does not bring any significant improvements. They found that this new approach corrected some of the major discrepancies. Similar improvements were noted by Szotek et al. (1993) in an LMTO implementation in which the occupied and unoccupied states were split by a large on-site Coulomb interaction. For Bloch states extending throughout the crystal, the SIC is small and the LDA is adequate. However, for localized states the SIC becomes significant. SIC calculations have been carried out for the parent compound of the high T_c superconducting ceramic, La_2CuO_4 (Temmerman et al., 1993) and have been used to explain the γ-α transition in the strongly correlated metal, cerium (Szotek et al., 1994; Svane, 1994; Beiden et al., 1997).

Spin Density Functional Theory within the local exchange and correlation approximation also has some serious shortcomings when straightforwardly extended to finite temperatures and applied to itinerant magnetic materials of all types. In the following section, we discuss ways in which improvements to the theory have been made.

Magnetism at Finite Temperatures: The Paramagnetic State

As long ago as 1965, Mermin (1965) published the formal structure of a finite temperature density functional theory. Once again, a many-electron system in an external potential, $V^{\rm ext}$, and external magnetic field, $\mathbf{B}^{\rm ext}$, described by the (non-relativistic) Hamiltonian is considered. Mermin proved that, in the grand canonical ensemble at a given temperature T and chemical potential ν, the equilibrium particle $n(\mathbf{r})$ and magnetization $\mathbf{m}(\mathbf{r})$ densities are determined by the external potential and magnetic field. The correct equilibrium particle and magnetization densities minimize the Gibbs grand potential, Ω

$$\Omega = \int V^{\rm ext}(\mathbf{r}) n(\mathbf{r})\, d\mathbf{r} - \int \mathbf{B}^{\rm ext}(\mathbf{r}) \cdot \mathbf{m}(\mathbf{r})\, d\mathbf{r} + \frac{e^2}{2}\int\int \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}\, d\mathbf{r}' + G[n,\mathbf{m}] - \nu \int n(\mathbf{r})\, d\mathbf{r} \quad (4)$$

where G is a unique functional of charge and magnetization densities at a given T and ν. The variational principle now states that Ω is a minimum for the equilibrium, n and $\mathbf{m}$. The function G can be written as

$$G[n,\mathbf{m}] = T_{\rm s}[n,\mathbf{m}] - TS_{\rm s}[n,\mathbf{m}] + \Omega_{\rm xc}[n,\mathbf{m}] \quad (5)$$

with $T_{\rm s}$ and $S_{\rm s}$ being respectively the kinetic energy and entropy of a system of noninteracting electrons with densities n, $\mathbf{m}$, at a temperature T. The exchange and correlation contribution to the Gibbs free energy is $\Omega_{\rm xc}$. The minimum principle can be shown to be identical to the corresponding equation for a system of noninteracting electrons moving in an effective potential $\tilde{V}$

$$\tilde{V}[n,\mathbf{m}] = \left(V^{\rm ext} + e^2\int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}' + \frac{\delta\,\Omega_{\rm xc}}{\delta\, n(r)}\right)\tilde{1} + \left(\frac{\delta\,\Omega_{\rm xc}}{\delta \mathbf{m}(\mathbf{r})} - \mathbf{B}^{\rm ext}\right)\cdot\tilde{\sigma} \quad (6)$$

which satsify the following set of equations

$$\left(-\frac{\hbar^2}{2m}\tilde{1}\,\vec{\nabla}^2 + \tilde{V}\right)\varphi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}) \quad (7)$$

$$n(\mathbf{r}) = \sum_i f(\varepsilon_i - \nu)\, tr\,[\phi_i^*(\mathbf{r})\phi_i(\mathbf{r})] \quad (8)$$

$$\mathbf{m}(\mathbf{r}) = \sum_i f(\varepsilon_i - \nu)\, tr\,[\phi_i^*(\mathbf{r})\tilde{\sigma}\phi_i(\mathbf{r})] \quad (9)$$

where $f(\varepsilon - \nu)$ is the Fermi-Dirac function. Rewriting Ω as

$$\Omega = -\sum_i f(\varepsilon_i - \nu) N(\varepsilon_i) - \frac{e^2}{2}\int\int \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}\, d\mathbf{r}' + \Omega_{\rm xc} - \int d\mathbf{r}\left(\frac{\delta\Omega_{\rm xc}}{\delta n(\mathbf{r})} n(\mathbf{r}) + \frac{\delta\Omega_{\rm xc}}{\delta \mathbf{m}(\mathbf{r})}\cdot\mathbf{m}(\mathbf{r})\right) \quad (10)$$

involves a sum over effective single particle states and where *tr* represents the trace over the components of the Dirac spinors which in turn are represented by $\phi_i(\mathbf{r})$, its conjugate transpose being $\phi_i^*(\mathbf{r})$. The nonmagnetic part of the potential is diagonal in this spinor space, being propor-

tional to the 2×2 unit matrix, $\tilde{1}$. The Pauli spin matrices $\tilde{\sigma}$ provide the coupling between the components of the spinors, and thus to the spin orbit terms in the Hamiltonian. Formally, the exchange-correlation part of the Gibbs free energy can be expressed in terms of spin-dependent pair correlation functions (Rajagopal, 1980), specifically

$$\Omega_{\mathrm{xc}}[n,\mathbf{m}] = \frac{e^2}{2}\int_0^1 d\lambda \int\int \sum_{\sigma,\sigma'} \frac{n_\sigma(\mathbf{r}) n_{\sigma'}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} g_\lambda(\sigma,\sigma';\mathbf{r},\mathbf{r}')\, d\mathbf{r}\, d\mathbf{r}' \tag{11}$$

The next logical step in the implementation of this theory is to form the finite temperature extension of the local approximation (LDA) in terms of the exchange-correlation part of the Gibbs free energy of a homogeneous electron gas. This assumption, however, severely underestimates the effects of the thermally induced spin-wave excitations. The calculated Curie temperatures are much too high (Gunnarsson, 1976), local moments do not exist in the paramagnetic state, and the uniform static paramagnetic susceptibility does not follow a Curie-Weiss behavior as seen in many metallic systems.

Part of the pair correlation function $g_\lambda(\sigma,\sigma';r,r')$ is related by the fluctuation-dissipation theorem to the magnetic susceptibilities that contain the information about these excitations. These spin fluctuations interact with each other as temperature is increased. Ω_{xc} should deviate significantly from the local approximation, and, as a consequence, the form of the effective single-electron states are modified. Over the past decade or so, many attempts have been made to model the effects of the spin fluctuations while maintaining the spin-polarized single-electron basis, and hence describe the properties of magnetic metals at finite temperatures. Evidently, the straightforward extension of spin-polarized band theory to finite temperatures misses the dominant thermal fluctuation of the magnetization and the thermally averaged magnetization, **M**, can only vanish along with the "exchange-splitting" of the electronic bands (which is destroyed by particle-hole, "Stoner" excitations across the Fermi surface). An important piece of this neglected component can be pictured as orientational fluctuations of "local moments," which are the magnetizations within each unit cell of the underlying crystalline lattice and are set up by the collective behavior of all the electrons. At low temperatures, these effects have their origins in the transverse part of the magnetic susceptibility. Another related ingredient involves the fluctuations in the magnitudes of these "moments," and concomitant charge fluctuations, which are connected with the longitudinal magnetic response at low temperatures. The magnetization **M** now vanishes as the disorder of the "local moments" grows. From this broad consensus (Moriya, 1981), several approaches exist which only differ according to the aspects of the fluctuations deemed to be the most important for the materials which are studied.

Competitive and Related Techniques: Fluctuating "Local Moments"

Some fifteen years ago, work on the ferromagnetic $3d$ transition metals—Fe, Co, and Ni—could be roughly partitioned into two categories. In the main, the Stoner excitations were neglected and the orientations of the "local moments," which were assumed to have fixed magnitudes independent of their orientational environment, corresponded to the degrees of freedom over which one thermally averaged. Firstly, the picture of the Fluctuating Local Band (FLB) theory was constructed (Korenman et al., 1977a,b,c; Capellman, 1977; Korenman, 1985), which included a large amount of short-range magnetic order in the paramagnetic phase. Large spatial regions contained many atoms, each with their own moment. These moments had sizes equivalent to the magnetization per site in the ferromagnetic state at $T = 0$ K and were assumed to be nearly aligned so that their orientations vary gradually. In such a state, the usual spin-polarized band theory can be applied and the consequence of the gradual change to the orientations could be added perturbatively. Quasi-elastic neutron scattering experiments (Ziebeck et al., 1983) on the paramagnetic phases of Fe and Ni, later reproduced by Shirane et al. (1986), were given a simple though not uncontroversial (Edwards, 1984) interpretation of this picture. In the case of inelastic neutron scattering, however, even the basic observations were controversial, let alone their interpretations in terms of "spin-waves" above T_c which may be present in such a model. Realistic calculations (Wang et al., 1982) in which the magnetic and electronic structures are mutually consistent are difficult to perform. Consequently, examining the full implications of the FLB picture and systematic improvements to it has not made much headway.

The second type of approach is labeled the "disordered local moment" (DLM) picture (Hubbard, 1979; Hasegawa, 1979; Edwards, 1982; Liu, 1978). Here, the local moment entities associated with each lattice site are commonly assumed (at the outset) to fluctuate independently with an apparent total absence of magnetic short-range order (SRO). Early work was based on the Hubbard Hamiltonian. The procedure had the advantage of being fairly straightforward and more specific than in the case of FLB theory. Many calculations were performed which gave a reasonable description of experimental data. Its drawbacks were its simple parameter-dependent basis and the fact that it could not provide a realistic description of the electronic structure, which must support the important magnetic fluctuations. The dominant mechanisms therefore might not be correctly identified. Furthermore, it is difficult to improve this approach systematically.

Much work has focused on the paramagnetic state of body-centered cubic iron. It is generally agreed that "local moments" exist in this material for all temperatures, although the relevance of a Heisenberg Hamiltonian to a description of their behavior has been debated in depth. For suitable limits, both the FLB and DLM approaches can be cast into a form from which an effective classical Heisenberg Hamiltonian can be extracted

$$-\sum_{ij} J_{ij}\hat{e}_i \cdot \hat{e}_j \tag{12}$$

The "exchange interaction" parameters J_{ij} are specified in terms of the electronic structure owing to the itinerant

nature of the electrons in this metal. In the former FLB model, the lattice Fourier transform of the J_{ij}'s

$$L(\mathbf{q}) = \sum_{ij} J_{ij}(\exp(i\mathbf{q}\cdot\mathbf{R}_{ij}) - 1) \tag{13}$$

is equal to $Av\mathbf{q}^2$, where v is the unit cell volume and A is the Bloch wall stiffness, itself proportional to the spin wave stiffness constant D (Wang et al., 1982). Unfortunately the J_{ij}'s determined from this approach turn out to be too short-ranged to be consistent with the initial assumption of substantial magnetic SRO above T_c.

In the DLM model for iron, the interactions, J_{ij}'s, can be obtained from consideration of the energy of an interacting electron system in which the local moments are constrained to be oriented along directions $\hat{e}_i$ and $\hat{e}_j$ on sites i and j, averaging over all the possible orientations on the other sites (Oguchi et al., 1983; Györffy et al., 1985), albeit in some approximate way. The J_{ij}'s calculated in this way are suitably short-ranged and a mutual consistency between the electronic and magnetic structures can be achieved.

A scenario between these two limiting cases has been proposed (Heine and Joynt, 1988; Samson, 1989). This was also motivated by the apparent substantial magnetic SRO above T_c in Fe and Ni, deduced from neutron scattering data, and emphasized how the orientational magnetic disorder involves a balance in the free energy between energy and entropy. This balance is delicate, and it was shown that it is possible for the system to disorder on a scale coarser than the atomic spacing and for the magnetic and electronic structures. The length scale is, however, not as large as that initially proposed by the FLB theory.

"First-Principles" Theories

These pictures can be put onto a "first-principles" basis by grafting the effects of these orientational spin fluctuations onto SDF theory (Györffy et al., 1985; Staunton et al., 1985; Staunton and Györffy, 1992). This is achieved by making the assumption that it is possible to identify and to separate fast and slow motions. On a time scale long in comparison with an electronic hopping time but short when compared with a typical spin fluctuation time, the spin orientations of the electrons leaving a site are sufficiently correlated with those arriving so that a non-zero magnetization exists when the appropriate quantity is averaged on this time scale. These are the "local moments" which can change their orientations $\{\hat{e}_i\}$ slowly with respect to the time scale, whereas their magnitudes $\{\mu_i(\{\hat{e}_j\})\}$ fluctuate rapidly. Note that, in principle, the magnitude of a moment on a site depends on its orientational environment.

The standard SDF theory for studying electrons in spin-polarized metals can be adapted to describe the states of the system for each orientational configuration $\{\hat{e}_i\}$ in a similar way as in the case of noncollinear magnetic systems (Uhl et al., 1992; Sandratskii and Kubler, 1993; Sandratskii, 1998). Such a description holds the possibility to yield the magnitudes of the local moments $\mu_k = \mu_k(\{\hat{e}_j\})$ and the electronic Grand Potential for the constrained system $\Omega(\{\hat{e}_i\})$. In the implementation of this theory, the moments for bcc Fe and fictitious bcc Co are fairly independent of their orientational environment, whereas for those in fcc Fe, Co, and Ni, the moments are further away from being local quantities.

The long time averages can be replaced by ensemble averages with the Gibbsian measure $P(\{\hat{e}_j\}) = e^{-\beta\Omega(\{\hat{e}_j\})}/Z$, where the partition function is

$$Z = \prod_i \int d\hat{e}_i e^{-\beta\Omega(\{\hat{e}_j\})} \tag{14}$$

where β is the inverse of $k_B T$ with Boltzmann's constant k_B. The thermodynamic free energy, which accounts for the entropy associated with the orientational fluctuations as well as creation of electron-hole pairs, is given by $F = -k_B T \ln Z$. The role of a classical "spin" (local moment) Hamiltonian, albeit a highly complicated one, is played by $\Omega(\{\hat{e}_i\})$.

By choosing a suitable reference "spin" Hamiltonian $\Omega(\{\hat{e}_i\})$ and expanding about it using the Feynman-Peierls' inequality (Feynman, 1955), an approximation to the free energy is obtained

$$F \leq F_0 + \langle \Omega - \Omega_0 \rangle_0 = \tilde{F}$$

with

$$F_0 = -k_B T \ln\left[\prod_i \int d\hat{e}_i e^{-\beta\Omega_0(\{\hat{e}_i\})}\right] \tag{15}$$

and

$$\langle X \rangle_0 = \frac{\prod_i \int d\hat{e}_i X e^{-\beta\Omega_0}}{\prod_i \int d\hat{e}_i e^{-\beta\Omega_0}} = \prod_i \int d\hat{e}_i P_0(\{\hat{e}_i\}) X(\{\hat{e}_i\}) \tag{16}$$

With Ω_0 expressed as

$$\Omega_0 = \sum_i \omega_i^{(1)}(\hat{e}_i) + \sum_{i\neq j} \omega_{ij}^{(2)}(\hat{e}_i, \hat{e}_j) + \cdots \tag{17}$$

a scheme is set up that can in principle be systematically improved. Minimizing $\tilde{F}$ to obtain the best estimate of the free energy gives $\omega_i^{(1)}$, $\omega_{ij}^{(2)}$ etc., as expressions involving restricted averages of $\Omega(\{\hat{e}_i\})$ over the orientational configurations.

A mean-field-type theory, which turns out to be equivalent to a "first principles" formulation of the DLM picture, is established by taking the first term only in the equation above. Although the SCF-KKR-CPA method (Stocks et al., 1978; Stocks and Winter, 1982; Johnson et al. 1990) was developed originally for coping with compositional disorder in alloys, using it in explicit calculations for bcc Fe and fcc Ni gave some interesting results. The average magnitude of the local moments, $\langle \mu_i(\{\hat{e}_j\}) \rangle_{\hat{e}_i} = \mu_i(\hat{e}_i) = \bar{\mu}$, in the paramagnetic phase of iron was $1.91\mu_B$. (The total magnetization is zero since $\langle \mu_i(\{\hat{e}_j\}) \rangle = 0$. This value is roughly the same magnitude as the magnetization per atom in

the low temperature ferromagnetic state. The uniform, paramagnetic susceptibility, $\chi(T)$, followed a Curie-Weiss dependence upon temperature as observed experimentally, and the estimate of the Curie temperature T_c was found to be 1280 K, also comparing well with the experimental value of 1040 K. In nickel, however, $\bar{\mu}$ was found to be zero and the theory reduced to the conventional LDA version of the Stoner model with all its shortcomings.

This mean field DLM picture of the paramagnetic state was improved by including the effects of correlations between the local moments to some extent. This was achieved by incorporating the consequences of Onsager cavity fields into the theory (Brout and Thomas, 1967; Staunton and Györffy, 1992). The Curie temperature T_c for Fe is shifted downward to 1015 K and the theory gives a reasonable description of neutron scattering data (Staunton and Györffy, 1992). This approach has also been generalized to alloys (Ling et al., 1994a,b). A first application to the paramagnetic phase of the "spin-glass" alloy $Cu_{85}Mn_{15}$ revealed exponentially damped oscillatory magnetic interactions in agreement with extensive neutron scattering data and was also able to determine the underlying electronic mechanisms. An earlier application to fcc Fe showed how the magnetic correlations change from anti-ferromagnetic to ferromagnetic as the lattice is expanded (Pinski et al., 1986). This study complemented total energy calculations for fcc Fe for both ferromagnetic and antiferromagnetic states at absolute zero for a range of lattice spacings (Moruzzi and Marcus, 1993).

For nickel, the theory has the form of the static, high-temperature limit of Murata and Doniach (1972), Moriya (1979), and Lonzarich and Taillefer (1985), as well as others, to describe itinerant ferromagnets. Nickel is still described in terms of exchange-split spin-polarized bands which converge as T_c is approached but where the spin fluctuations have drastically renormalized the exchange interaction and lowered T_c from ~3000 K (Gunnarsson, 1976) to 450 K. The neglect of the dynamical aspects of these spin fluctuations has led to a slight overestimation of this renormalization, but $\chi(T)$ again shows Curie-Weiss behavior as found experimentally, and an adequate description of neutron scattering data is also provided (Staunton and Györffy, 1992). Moreover, recent inverse photoemission measurements (von der Linden et al., 1993) have confirmed the collapse of the "exchange-splitting" of the electronic bands of nickel as the temperature is raised towards the Curie temperature in accord with this Stoner-like picture, although spin-resolved, resonant photoemission measurements (Kakizaki et al., 1994) indicate the presence of spin fluctuations.

The above approach is parameter-free, being set up in the confines of SDF theory, and represents a fairly well defined stage of approximation. But there are still some obvious shortcomings in this work (as exemplified by the discrepancy between the theoretically determined and experimentally measured Curie constants). It is worth highlighting the key omission, the neglect of the dynamical effects of the spin fluctuations, as emphasized by Moriya (1981) and others.

Competitive and Related Technique for a "First-Principles" Treatment of the Paramagnetic States of Fe, Ni, and Co

Uhl and Kubler (1996) have also set up an *ab initio* approach for dealing with the thermally induced spin fluctuations, and they also treat these excitations classically. They calculate total energies of systems constrained to have spin-spiral $\{\hat{e}_i\}$ configurations with a range of different propagation vectors $\mathbf{q}$ of the spiral, polar angles θ, and spiral magnetization magnitudes μ using the non-collinear fixed spin moment method. A fit of the energies to an expression involving $\mathbf{q}$, θ, and μ is then made. The Feynman-Peierls inequality is also used where a quadratic form is used for the "reference Hamiltonian," H_0. Stoner particle-hole excitations are neglected. The functional integrations involved in the description of the statistical mechanics of the magnetic fluctuations then reduce to Gaussian integrals. Similar results to Staunton and Györffy (1992) have been obtained for bcc Fe and for fcc Ni. Uhl and Kubler (1997) have also studied Co and have recently generalized the theory to describe magnetovolume effects. Face-centered cubic Fe and Mn have been studied alongside the "Invar" ordered alloy, Fe_3Pt.

One way of assessing the scope of validity of these sorts of *ab initio* theoretical approaches, and the severity of the approximations employed, is to compare their underlying electronic bases with suitable spectroscopic measurements.

"Local Exchange Splitting"

An early prediction from a "first principles" implementation of the DLM picture was that a "local-exchange" splitting should be evident in the electronic structure of the paramagnetic state of bcc iron (Györffy et al., 1983; Staunton et al., 1985). Moreover, the magnitude of this splitting was expected to vary sharply as a function of wave-vector and energy. At some wave-vectors, if the "bands" did not vary much as a function of energy, the local exchange splitting would be roughly of the same size as the rigid exchange splitting of the electronic bands of the ferromagnetic state, whereas at other points where the "bands" have greater dispersion, the splitting would vanish entirely. This local exchange-splitting is responsible for local moments.

Photoemission (PES) experiments (Kisker et al., 1984, 1985) and inverse photoemission (IPES) experiments (Kirschner et al., 1984) observed these qualitative features. The experiments essentially focused on the electronic structure around the Γ and H points for a range of energies. Both the Γ'_{25} and Γ'_{12} states were interpreted as being exchange-split, whereas the H'_{25} state was not, although all were broadened by the magnetic disorder. Among the DLM calculations of the electronic structure for several wave-vectors and energies (Staunton et al., 1985), those for the Γ and H points showed the Γ'_{12} state as split and both the Γ'_{25} and H'_{25} states to be substantially broadened by the local moment disorder, but not locally exchange split. Haines et al. (1985, 1986) used a tight-binding model to describe the electronic structure, and employed the recursion method to average over various orientational configurations. They concluded that a

modest degree of SRO is compatible with spectroscopic measurements of the Γ'_{25} d state in paramagnetic iron. More extensive spectroscopic data on the paramagnetic states of the ferromagnetic transition metals would be invaluable in developing the theoretical work on the important spin fluctuations in these systems.

As emphasized in the introduction to this unit, the state of magnetic order in an alloy can have a profound effect upon various other properties of the system. In the next subsection we discuss its consequence upon the alloy's compositional order.

Interrelation of Magnetism and Atomic Short Range Order

A challenging problem to study in metallic alloys is the interplay between compositional order and magnetism and the dependence of magnetic properties on the local chemical environment. Magnetism is frequently connected to the overall compositional ordering, as well as the local environment, in a subtle and complicated way. For example, there is an intriguing link between magnetic and compositional ordering in nickel-rich Ni-Fe alloys. $Ni_{75}Fe_{25}$ is paramagnetic at high temperatures; it becomes ferromagnetic at ~900 K, and then, at at temperature just 100 K cooler, it chemically orders into the Ni_3Fe $L1_2$ phase. The Fe-Al phase diagram shows that, if cooled from the melt, paramagnetic $Fe_{80}Al_{20}$ forms a solid solution (Massalski et al., 1990). The alloy then becomes ferromagnetic upon further cooling to 935 K, and then forms an apparent DO_3 phase at ~670 K. An alloy with just 5% more aluminum orders instead into a B2 phase directly from the paramagnetic state at roughly 1000 K, before ordering into a DO_3 phase at lower temperatures. In this subsection, we examine this interrelation between magnetism and compositional order. It is necessary to deal with the statistical mechanics of thermally induced compositional fluctuations to carry out this task. COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS has described this in some detail (see also Györffy and Stocks, 1983; Györffy et al., 1989; Staunton et al., 1994; Ling et al., 1994b), so here we will simply recall the salient features and show how magnetic effects are incorporated.

A first step is to construct (formally) the grand potential for a system of interacting electrons moving in the field of a particular distribution of nuclei on a crystal lattice of an A_cB_{1-c} alloy using SDF theory. (The nuclear diffusion times are very long compared with those associated with the electrons' movements and thus the compositional and electronic degrees of freedom decouple.) For a site i of the lattice, the variable ξ_i is set to unity if the site is occupied by an A atom and zero if a B atom is located on it. In other words, an Ising variable is specified. A configuration of nuclei is denoted $\{\xi_i\}$ and the associated electronic grand potential is expressed as $\Omega(\{\xi_i\})$. Averaging over the compositional fluctuations with measure

$$P(\{\xi_i\}) = \frac{\exp(-\beta\Omega(\{\xi_i\}))}{\prod_i \sum_{\xi_i} \exp(-\beta\Omega(\{\xi_i\}))} \tag{18}$$

gives an expression for the free energy of the system at temperature T

$$F(\{\xi_i\}) = -k_B T \ln\left[\prod_i \sum_{\xi_i} \exp(-\beta\Omega\{\xi_i\})\right] \tag{19}$$

In essence, $\Omega(\{\xi_i\})$ can be viewed as a complicated concentration-fluctuation Hamiltonian determined by the electronic "glue" of the system. To proceed, some reasonable approximation needs to be made (see review provided in COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS. A course of action, which is analogous with our theory of spin fluctuations in metals at finite T, is to expand about a suitable reference Hamiltonian Ω_0 and to make use of the Feynman-Peierls inequality (Feynman, 1955). A mean field theory is set up with the choice

$$\Omega_0 = \sum_i V_i^{\text{eff}} \cdot \xi_i \tag{20}$$

in which $V_i^{\text{eff}} = \langle\Omega\rangle^0_{Ai} - \langle\Omega\rangle^0_{Bi}$ where $\langle\Omega\rangle^0_{A(B)i}$ is the Grand Potential averaged over all configurations with the restriction that an A(B) nucleus is positioned on the site i. These partial averages are, in principle, accessible from the SCF-KKR-CPA framework and indeed this mean field picture has a satisfying correspondence with the single-site nature of the coherent potential approximation to the treatment of the electronic behavior. Hence at a temperature T, the chance of finding an A atom on a site i is given by

$$c_i = \frac{\exp(-\beta(V_i^{\text{eff}} - \nu))}{(1 + \exp(-\beta(V_i^{\text{eff}} - \nu))} \tag{21}$$

where ν is the chemical potential difference which preserves the relative numbers of A and B atoms overall.

Formally, the probability of occupation can vary from site to site, but it is only the case of a homogeneous probability distribution $c_i = c$ (the overall concentration) that can be tackled in practice. By setting up a response theory, however, and using the fluctuation-dissipation theorem, it is possible to write an expression for the compositional correlation function and to investigate the system's tendency to order or phase segregate.

If a field, which couples to the occupation variables $\{\xi_i\}$ and varies from site-to-site, is applied to the high temperature homogeneously disordered system, it induces an inhomogeneous concentration distribution $\{c + \delta c_i\}$. As a result, the electronic charge rearranges itself (Staunton et al., 1994; Treglia et al., 1978) and, for those alloys which are magnetic in the compositionally disordered state, the magnetization density also changes, i.e. $\{\delta\mu_i^{\text{A}}\}$, $\{\delta\mu_i^{\text{B}}\}$. A theory for the compositional correlation function has been developed in terms of the SCF-KKR-CPA framework (Györffy and Stocks, 1983) and is discussed at length in COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS. In reciprocal "concentration-wave" vector space (Khachaturyan, 1983), this has the Ornstein-Zernicke form

$$\alpha(\mathbf{q}) = \frac{\beta c(1-c)}{(1 - \beta c(1-c)S^{(2)}(\mathbf{q}))} \tag{22}$$

in which the Onsager cavity fields have been incorporated (Brout and Thomas, 1967; Staunton and Györffy, 1992;

Staunton et al., 1994) ensuring that the site-diagonal part of the fluctuation dissipation theorem is satisfied. The key quantity $S^{(2)}(\mathbf{q})$ is the direct correlation function and is determined by the electronic structure of the disordered alloy.

In this way, an alloy's tendency to order depends crucially on the magnetic state of the system and upon whether or not the electronic structure is spin-polarized. If the system is paramagnetic, then the presence of "local moments" and the resulting "local exchange splitting" will have consequences. In the next section, we describe three case studies where we show the extent to which an alloy's compositional structure is dependent on whether the underlying electronic structure is "globally" or "locally" spin-polarized, i.e., whether the system is quenched from a ferromagnetic or paramagnetic state. We look at nickel-iron alloys, including those in the "Invar" concentration range, iron-rich Fe-V alloys, and finally gold-rich AuFe alloys.

The value of $\mathbf{q}$ for which $S^{(2)}(q)$, the direct correlation function, has its greatest value signifies the wavevector for the static concentration wave to which the system is unstable at a low enough temperature. For example, if this occurs at $q = 0$, phase segregation is indicated, whilst for a $A_{75}B_{25}$ alloy a maximum value at $\mathbf{q} = (1, 0, 0)$ points to an $L1_2(Cu_3Au)$ ordered phase at low temperatures. An important part of $S^{(2)}(\mathbf{q})$ derives from an electronic state filling effect and ties in neatly with the notion that half-filled bands promote ordered structures whilst nearly filled or nearly empty states are compatible with systems that cluster when cooled (Ducastelle, 1991; Heine and Samson, 1983). This propensity can be totally different depending on whether the electronic structure is spin-polarized or not, and hence whether the compositionally disordered state is ferromagnetic or paramagnetic as is the case for nickel-rich $Ni_{75}Fe_{25}$, for example (Staunton et al., 1987). The remarks made earlier in this unit about bonding in alloys and spin-polarization are clearly relevant here. For example, majority spin electrons in strongly ferromagnetic alloys like $Ni_{75}Fe_{25}$, which completely occupy the majority spin d states "see" very little difference between the two types of atomic site (Fig. 2) and hence contribute little to $S^{(2)}(\mathbf{q})$ and it is the filling of the minority-spin states which determine the eventual compositional structure. A contrasting picture describes those alloys, usually bcc-based, in which the Fermi energy is pinned in a valley in the minority density of states (Fig. 2, panel B) and where the ordering tendency is largely governed by the majority-spin electronic structure (Staunton et al., 1990).

For a ferromagnetic alloy, an expression for the lattice Fourier transform of the magneto-compositional cross-correlation function $\vartheta_{ik} = \langle \mu_i \xi_k \rangle - \langle \mu_i \rangle \langle \xi_k \rangle$ can be written down and evaluated (Staunton et al., 1990; Ling et al., 1995a). Its lattice Fourier transform turns out to be a simple product involving the compositional correlation function, $\vartheta(\mathbf{q}) = \alpha(\mathbf{q})\gamma(\mathbf{q})$, so that ϑ_{ik} is a convolution of $\gamma_{ik} = \delta\langle \mu_i \rangle / \delta c_k$ and α_{kj}. The quantity γ_{ik} has components

$$\gamma_{ik} = (\mu^{\mathrm{A}} - \mu^{\mathrm{B}})\delta_{ik} + c\frac{\delta \mu_i^{\mathrm{A}}}{\delta c_k} + (1-c)\frac{\delta \mu_i^{\mathrm{B}}}{\delta c_k} \qquad (23)$$

The last two quantities, $\delta\mu_i^{\mathrm{A}}/\delta c_k$ and $\delta\mu_i^{\mathrm{B}}/\delta c_k$, can also be evaluated in terms of the spin-polarized electronic structure of the disordered alloy. They describe the changes to the magnetic moment μ_i on a site i in the lattice occupied by either an A or a B atom when the probability of occupation is altered on another site k. In other words, γ_{ik} quantifies the chemical environment effect on the sizes of the magnetic moments. We studied the dependence of the magnetic moments on their local environments in FeV and FeCr alloys in detail from this framework (Ling et al., 1995a). If the application of a small external magnetic field is considered along the direction of the magnetization, expressions dependent upon the electronic structure for the magnetic correlation function can be similarly found. These are related to the static longitudinal susceptibility $\chi(\mathbf{q})$.

The quantities $\alpha(\mathbf{q})$, $\vartheta(\mathbf{q})$, and $\chi(\mathbf{q})$ can be straightforwardly compared with information obtained from x-ray (Krivoglaz, 1969; also see Chapter 10, section b) and neutron scattering (Lovesey, 1984; also see MAGNETIC NEUTRON SCATTERING), nuclear magnetic resonance (NUCLEAR MAGNETIC RESONANCE IMAGING), and Mössbauer spectroscopy (MOSSBAUER SPECTROMETRY) measurements. In particular, the cross-sections obtained from diffuse polarized neutron scattering can be written

$$\left.\frac{d\sigma}{d\omega}\right|_{\zeta} = \frac{d\sigma^{\mathrm{N}}}{d\omega} + \zeta\frac{d\sigma}{d\omega} + \frac{d\sigma^{\mathrm{M}}}{d\omega} \qquad (24)$$

where $\zeta = +1(-1)$ if the neutrons are polarized (anti-) parallel to the magnetization (see MAGNETIC NEUTRON SCATTERING). The nuclear component $d\sigma^{\mathrm{N}}/d\omega$ is proportional to the compositional correlation function, $\alpha(\mathbf{q})$ (closely related to the Warren-Cowley short-range order parameters). The magnetic component $d\sigma^{\mathrm{M}}/d\omega$ is proportional to $\chi(\mathbf{q})$. Finally $d\sigma^{\mathrm{NM}}/d\omega$ describes the magneto-compositional correlation function $\gamma(\mathbf{q})\alpha(\mathbf{q})$ (Marshall, 1968; Cable and Medina, 1976). By interpreting such experimental measurements by such calculations, electronic mechanisms which underlie the correlations can be extracted (Staunton et al., 1990; Cable et al., 1989).

Up to now, everything has been discussed with respect to spin-polarized but non-relativistic electronic structure. We now touch briefly on the relativistic extension to this approach to describe the important magnetic property of magnetocrystalline anisotropy.

MAGNETIC ANISOTROPY

At this stage, we recall that the fundamental "exchange" interactions causing magnetism in metals are intrinsically isotropic, i.e., they do not couple the direction of magnetization to any spatial direction. As a consequence they are unable to provide any sort of description of magnetic anisotropic effects which lie at the root of technologically important magnetic properties such as domain wall structure, linear magnetostriction, and permanent magnetic properties in general. A fully relativistic treatment of the electronic effects is needed to get a handle on these phenomena. We consider that aspect in this subsection. In a solid with

an underlying lattice, symmetry dictates that the equilibrium direction of the magnetization be along one of the cystallographic directions. The energy required to alter the magnetization direction is called the magnetocrystalline anisotropy energy (MAE). The origin of this anisotropy is the interaction of magnetization with the crystal field (Brooks, 1940) i.e., the spin-orbit coupling.

Competitive and Related Techniques for Calculating MAE

Most present-day theoretical investigations of magnetocrystalline anisotropy use standard band structure methods within the scalar-relativistic local spin-density functional theory, and then include, perturbatively, the effects from spin-orbit coupling, a relativistic effect. Then by using the force theorem (Mackintosh and Anderson, 1980; Weinert et al., 1985), the difference in total energy of two solids with the magnetization in different directions is given by the difference in the Kohn-Sham single-electron energy sums. In practice, this usually refers only to the valence electrons, the core electrons being ignored.

There are several investigations in the literature using this approach for transition metals (e.g. Gay and Richter, 1986; Daalderop et al., 1993), as well as for ordered transition metal alloys (Sakuma, 1994; Solovyev et al., 1995) and layered materials (Guo et al., 1991; Daalderop et al., 1993; Victora and MacLaren, 1993), with varying degrees of success. Some controversy surrounds such perturbative approaches regarding the method of summing over all the "occupied" single-electron energies for the perturbed state which is not calculated self-consistently (Daalderop et al., 1993; Wu and Freeman, 1996). Freeman and coworkers (Wu and Freeman, 1996) argued that this "blind Fermi filling" is incorrect and proposed the state-tracking approach in which the occupied set of perturbed states are determined according to their projections back to the occupied set of unperturbed states. More recently, Trygg et al. (1995) included spin-orbit coupling self-consistently in the electronic structure calculations, although still within a scalar-relativistic theory. They obtained good agreement with experimental magnetic anisotropy constants for bcc Fe, fcc Co, and hcp Co, but failed to obtain the correct magnetic easy axis for fcc Ni.

Practical Aspects of the Method

The MAE in many cases is of the order of μeV, which is several (as many as 10) orders of magnitude smaller than the total energy of the system. With this in mind, one has to be very careful in assessing the precision of the calculations. In many of the previous works, fully relativistic approaches have not been used, but it is possible that only a fully relativistic framework may be capable of the accuracy needed for reliable calculations of MAE. Moreover either the total energy or the single-electron contribution to it (if using the force theorem) has been calculated separately for each of the two magnetization directions and then the MAE obtained by a straight subtraction of one from the other. For this reason, in our work, some of which we outline below, we treat relativity and magnetization (spin polarization) on an equal footing. We also calculate the energy difference directly, removing many systematic errors.

Strange et al. (1989a, 1991) have developed a relativistic spin-polarized version of the Korringa-Kohn-Rostoker (SPR-KKR) formalism to calculate the electronic structure of solids, and Ebert and coworkers (Ebert and Akai, 1992) have extended this formalism to disordered alloys by incorporating coherent-potential approximation (SPR-KKR-CPA). This formalism has successfully described the electronic structure and other related properties of disordered alloys (see Ebert, 1996 for a recent review) such as magnetic circular x-ray dichroism (X-RAY MAGNETIC CIRCULAR DICHROISM), hyperfine fields, magneto-optic Kerr effect (SURFACE MAGNETO-OPTIC KERR EFFECT). Strange et al. (1989a, 1989b) and more recently Staunton et al. (1992) have formulated a theory to calculate the MAE of elemental solids within the SPR-KKR scheme, and this theory has been applied to Fe and Ni (Strange et al., 1989a, 1989b). They have also shown that, in the nonrelativistic limit, MAE will be identically equal to zero, indicating that the origin of magnetic anisotropy is indeed relativistic.

We have recently set up a robust scheme (Razee et al., 1997, 1998) for calculating the MAE of compositionally disordered alloys and have applied it to Ni_cPt_{1-c} and Co_cPt_{1-c} alloys and we will describe our results for the latter system in a later section. Full details of our calculational method are found elsewhere (Razee et al., 1997) and we give a bare outline here only. The basis of the magnetocrystalline anisotropy is the relativistic spin-polarized version of density functional theory (see e.g. MacDonald and Vosko, 1979; Rajagopal, 1978; Ramana and Rajagopal, 1983; Jansen, 1988). This, in turn, is based on the theory for a many electron system in the presence of a "spin-only" magnetic field (ignoring the diamagnetic effects), and leads to the relativistic Kohn-Sham-Dirac single-particle equations. These can be solved using spin-polarized, relativistic, multiple scattering theory (SPR-KKR-CPA).

From the key equations of the SPR-KKR-CPA formalism, an expression for the magnetocrystalline anisotropy energy of disordered alloys is derived starting from the total energy of a system within the local approximation of the relativistic spin-polarized density functional formalism. The change in the total energy of the system due to the change in the direction of the magnetization is defined as the magnetocrystalline anisotropy energy, i.e., $\Delta E = E[n(\mathbf{r}),\ \mathbf{m}(\mathbf{r},\hat{e}_1)] - E[n(\mathbf{r}),\ \mathbf{m}(\mathbf{r},\hat{e}_2)]$, with $\mathbf{m}(\mathbf{r},\hat{e}_1)$, $\mathbf{m}(\mathbf{r},\hat{e}_2)$ being the magnetization vectors pointing along two directions $\hat{e}_1$ and $\hat{e}_1$ respectively; the magnitudes are identical. Considering the stationarity of the energy functional and the local density approximation, the contribution to ΔE is predominantly from the single-particle term in the total energy. Thus, now we have

$$\Delta E = \int^{\varepsilon_{F1}} \varepsilon n(\varepsilon, \hat{e}_1)\, d\varepsilon - \int^{\varepsilon_{F2}} \varepsilon n(\varepsilon, \hat{e}_2)\, d\varepsilon \qquad (25)$$

where ε_{F1} and ε_{F2} are the respective Fermi levels for the two orientations. This expression can be manipulated into one involving the integrated density of states and where a cancellation of a large part has taken place, i.e.,

$$\Delta E = -\int^{\varepsilon_{F1}} d\varepsilon (N(\varepsilon, \hat{e}_1) - N(\varepsilon, \hat{e}_2)) - \frac{1}{2} N(\varepsilon_{F2}, \hat{e}_2)(\varepsilon_{F1} - \varepsilon_{F2})^2 + O(\varepsilon_{F1} - \varepsilon_{F2})^3 \qquad (26)$$

In most cases, the second term is very small compared to the first term. This first term must be evaluated accurately, and it is convenient to use the Lloyd formula for the integrated density of states (Staunton et al., 1992; Gubanov et al., 1992).

MAE of the Pure Elements Fe, Ni, and Co

Several groups including ours have estimated the MAE of the magnetic 3*d* transition metals. We found that the Fermi energy for the [001] direction of magnetization calculated within the SPR-KKR-CPA is ~1 to 2 mRy above the scalar relativistic value for all the three elements (Razee et al., 1997). We also estimated the order of magnitude of the second term in the equation above for these three elements, and found that it is of the order of 10^{-2} μeV, which is one order of magnitude smaller than the first term. We compared our results for bcc Fe, fcc Co, and fcc Ni with the experimental results, as well as the results of previous calculations (Razee et al., 1997). Among previous calculations, the results of Trygg et al. (1995) are closest to the experiment, and therefore we gauged our results against theirs. Their results for bcc Fe and fcc Co are in good agreement with the experiment if orbital polarization is included. However, in case of fcc Ni, their prediction of the magnitude of MAE, as well as the magnetic easy axis, is not in accord with experiment, and even the inclusion of orbital polarization fails to improve the result. Our results for bcc Fe and fcc Co are also in good agreement with the experiment, predicting the correct easy axis, although the magnitude of MAE is somewhat smaller than the experimental value. Considering that in our calculations orbital polarization is not included, our results are quite satisfactory. In case of fcc Ni, we obtain the correct easy axis of magnetization, but the magnitude of MAE is far too small compared to the experimental value, but in line with other calculations. As noted earlier, in the calculation of MAE, the convergence with regard to the Brillouin zone integration is very important. The Brillouin zone integrations had to be done with much care.

DATA ANALYSIS AND INITIAL INTERPRETATION

The Energetics and Electronic Origins for Atomic Long- and Short-Range Order in NiFe Alloys

The electronic states of iron and nickel are similar in that for both elements the Fermi energy is placed near or at the top of the majority-spin *d* bands. The larger moment in Fe as compared to Ni, however, manifests itself via a larger exchange-splitting. To obtain a rough idea of the electronic structures of Ni_cFe_{1-c} alloys, we imagine aligning the Fermi energies of the electronic structures of the pure elements. The atomic-like *d* levels of the two, marking the center of the bands, would be at the same energy for the majority spin electrons, whereas for the minority spin electrons, the levels would be at rather different energies, reflecting the differing exchange fields associated with each sort of atom (Fig. 2). In Figure 1, we show the density of states of $Ni_{75}Fe_{25}$ calculated by the SCF-KKR-CPA, and we interpreted along those lines. The majority spin density of states possesses very sharp structure, which indicates that in this compositionally disordered alloy majority spin electrons "see" very little difference between the two types of atom, with the DOS exhibiting "common-band" behavior. For the minority spin electrons the situation is reversed. The density of states becomes "split-band"-like owing to the large separation of levels (in energy) and due to the resulting compositional disorder. As pointed out earlier, the majority spin *d* states are fully occupied, and this feature persists for a wide range of concentrations of fcc Ni_cFe_{1-c} alloys: for *c* greater than 40%, the alloys' average magnetic moments fall nicely on the negative gradient slope of the Slater-Pauling curve. For concentrations less than 35%, and prior to the Martensitic transition into the bcc structure at around 25% (the famous "Invar" alloys), the Fermi energy is pushed into the peak of majority-spin *d* states, propelling these alloys away from the Slater-Pauling curve. Evidently the interplay of magnetism and chemistry (Staunton et al., 1987; Johnson et al., 1989) gives rise to most of the thermodynamic and concentration-dependent properties of Ni-Fe alloys.

The ferromagnetic DOS of fcc Ni-Fe, given in Figure 1A, indicates that the majority-spin *d* electrons cannot contribute to chemical ordering in Ni-rich Ni-Fe alloys, since the states in this spin channel are filled. In addition, because majority-spin *d* electrons "see" little difference between Ni and Fe, there can be no driving force for chemical order or for clustering (Staunton et al., 1987; Johnson et al., 1989). However, the difference in the exchange splitting of Ni and Fe leads to a very different picture for minority-spin *d* electrons (Fig. 2). The bonding-like states in the minority-spin DOS are mostly Ni, whereas the antibonding-like states are predominantly Fe. The Fermi level of the electrons lies between these bonding and anti-bonding states. This leads to the Cu-Au-type atomic short-range order and to the long-range order found in the region of $Ni_{75}Fe_{25}$ alloys. As the Ni concentration is reduced, the minority-spin bonding states are slowly depopulated, reducing the stability of the alloy, as seen in the heats of formation (Johnson and Shelton, 1997). Ultimately, when enough electrons are removed (by adding more iron), the Fermi level enters the majority-spin *d* band and the anomalous behavior of Ni-Fe alloys occurs: increases in resistivity and specific heat, collapse of moments (Johnson et al., 1987), and competing magnetic states (Johnson et al., 1989; Abrikosov et al., 1995).

Moment Alignment Versus Moment Formation in fcc Fe. Before considering the last aspect, that of competing magnetic states and their connection to volume effects, it is instructive to consider the magnetic properties of Fe on an fcc lattice, even though it exists only at high temperatures. Moruzzi and Marcus (1993) have reviewed the calculations of the energetics and moments of fcc Fe in both antiferromagnetic (AFM) and ferromagnetic (FM) states for a range of lattice spacings. Here we refer to a comparison with the DLM paramagnetic state (PM; Pinski et al., 1986; Johnson et al., 1989). For large volumes (lattice spacings), the FM state has large moments and is lowest in energy. At small volumes, the PM state is lowest in energy and is the global energy minimum. At intermediate

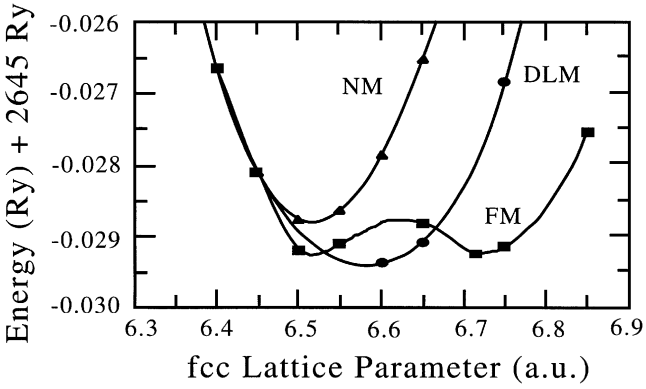


Figure 3. The volume dependence of the total energy of various magnetic states of $Ni_{25}Fe_{75}$. The total energy of the states of fcc $Ni_{25}Fe_{75}$ with the designations FM (moments aligned), the DLM (moments disordered), and NM (zero moments) are plotted as a function of the fcc lattice parameter. See Johnson et al. (1989), and Johnson and Shelton (1997).

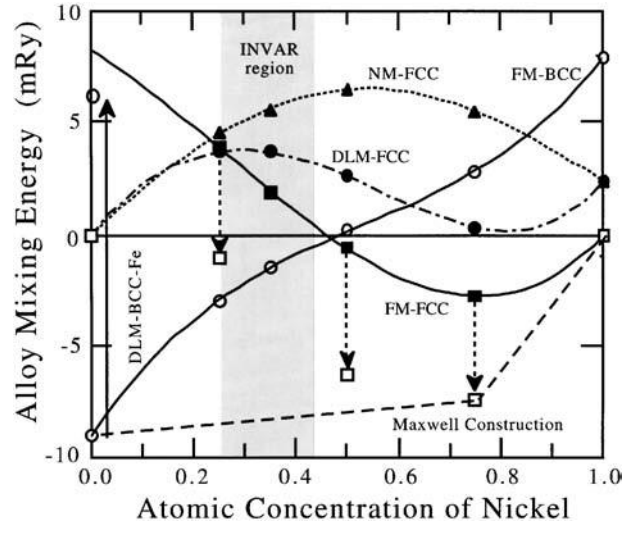


Figure 4. The concentration dependence of the total energy of various magnetic states of Ni-Fe Alloys. The total energy of the some magnetic states of Ni-Fe alloys are plotted as a function of concentration. Note that the Maxwell construction indicates that the ordered fcc phases, $Fe_{50}Ni_{50}$ and $Fe_{75}Ni_{25}$, are metastable. Adapted from Johnson and Shelton (1997).

volumes, however, the AFM and PM states have similar-size moments and energies, although at a value of the lattice constant of 6.6 a.u. or 3.49 Å, the local moments in the PM state collapse. These results suggest that the Fe-Fe magnetic correlations on an fcc lattice are extremely sensitive to volume and evolve from FM to AFM as the lattice is compressed. This suggestion was confirmed by explicit calculations of the magnetic correlations in the PM state (Pinski et al., 1986).

In Figure 9 of Johnson et al. (1989), the energetics of fcc $Ni_{35}Fe_{65}$ were a particular focal point. This alloy composition is well within the Invar region, near to the magnetic collapse, and exhibiting the famous negative thermal expansion. The energies of four magnetic states—i.e., non-magnetic (NM), ferromagnetic (FM), paramagnetic (PM), represented by the disordered local moment state (DLM), and anti-ferromagnetic (AFM)—were within 1.5 mRy, or 250 K of each other (Fig. 3). The $Ni_{35}Fe_{65}$ calculations were a subset of many calculations that were done for various Ni compositions and magnetic states. As questions still remained regarding the true equilibrium phase diagram of Ni-Fe, Johnson and Shelton (1997) calculated the heats of formation, ΔE_f, or mixing energies, for various Ni compositions and for several magnetic fcc and bcc Ni-Fe phases relative to the pure endpoints, NM-fcc Fe and FM-fcc Ni.

For the NM-fcc Ni-rich alloys, they found the function ΔE_f (as a function of composition) to be positive and convex everywhere, indicating that these alloys should cluster. While this argument is not always true, we have shown that the calculated ASRO for NM-fcc Ni-Fe does indeed show clustering (Staunton et al., 1987; Johnson et al., 1989). This was a consequence of the absence of exchange-splitting in a Stoner paramagnet and filling of unfavorable antibonding *d*-electron states. This, at best, would be a state seen only at extreme temperatures, possibly near melting. Thermochemical measurements at high temperatures in Ni-rich, Ni-Fe alloys appear to support this hypothesis (Chuang et al., 1986).

In the past, the NM (possessing zero local moments) state has been used as an approximate PM state, and the energy difference between the FM and NM state seems to reflect well the observed non-symmetric behavior of the Curie temperature when viewed as a function of Ni concentration. However, this is fortuitous agreement, and the lack of exchange-splitting in the NM state actually suppresses ordering.

As shown in figure 2 of Johnson and Shelton (1997) and in Figure 4 of this unit, the PM-DLM state, with its local exchange-splitting on the Fe sites, is lower in energy, and therefore a more relevant (but still approximate) PM state. Even in the Invar region, where the energy differences are very small, the exchange-splitting has important consequences for ording.

While the DLM state is much more representative of the PM state, it does not contain any of magnetic short-range order (MSRO) that exists above the Curie temperature. This shortcoming of the model is relevant because the ASRO calculated from this approximate PM state yields very weak ordering (spinodal-ordering temperature below 200 K) for $Ni_{75}Fe_{25}$, which is not, however, of $L1_2$ type. The ASRO calculated for fully-polarized FM $Ni_{75}Fe_{25}$ is $L1_2$-like, with a spinodal around 475 K, well below the actual chemical-ordering temperature of 792 K (Staunton et al., 1987; Johnson et al., 1989). Recent diffuse scattering measurements by Jiang et al. (1996) find weak $L1_2$-like ASRO in Ni_3Fe samples quenched from 1273 K, which is above the Curie temperature of 800 K. It appears that some degree of magnetic order (both short- or long-range) is required for the ASRO to have $\mathbf{k} = (1,0,0)$ wavevector instabilities (or $L1_2$ type chemical ordering tendencies). Nonetheless, the local exchange splitting in the DLM state, which exists only on the Fe sites (the Ni moments are quenched), does lead to weak ordering, as compared

to the tendency to phase-separate that is found when local exchange splitting is absent in the NM case.

Importantly, this indicates that sample preparation (whether above or below the Curie point) and the details of the measuring procedure (e.g., if data is taken *in situ* or after quench) affect what is measured. Two time scales are important: roughly speaking, the electron hopping time is 10^{-15} sec, whereas the chemical hopping time (or diffusion) is 10^{-3} to 10^{+6} sec. Now we consider diffuse-scattering experiments. For samples prepared in the Ni-rich alloys, but below the Curie temperature, it is most likely that a smaller difference would be found from data taken *in situ* or on quenched samples, because the (global) FM exchange-split state has helped establish the chemical correlations in both cases. On the other hand, in the Invar region, the Curie temperature is much lower than that for Ni-rich alloys and lies in the two-phase region. Samples annealed in the high-temperature, PM, fcc solid-solution phase and then quenched should have (at best) very weak ordering tendencies. The electronic and chemical degrees of freedom respond differently to the quench. Jiang et al. (1996) have recently measured ASRO versus composition in Ni-Fe system using anomalous x-ray scattering techniques. No evidence for ASRO is found in the Invar region, and the measured diffuse intensity can be completely interpreted in terms of static-displacement (size-effect) scattering. These results are in contrast to those found in the 50% and 75% Ni samples annealed closer to, but above, the Curie point and before being quenched.

The calculated ASRO intensities in 35%, 50%, and 75% Ni FM alloys are very similar in magnitude and show the Cu-Au ordering tendencies. Figure 2 of Johnson and Shelton (1997) shows that the Cu-Au-type $T = 0$ K ordering energies lie close to one another. While this appears to contradict the experimental findings (Jiang et al., 1996), recall that the calculated ASRO for PM-DLM Ni_3Fe shows ordering to be suppressed. The scattering data obtained from the Invar alloy was from a sample quenched well above the Curie temperature. Theory and experiment may then be in agreement: the ASRO is very weak, allowing size-effect scattering to dominate. Notably, volume fluctuations and size effects have been suggested as being responsible for, or at least contributing to, many of the anomalous Invar properties (Wassermann, 1991; Mohn et al., 1991; Entel et al., 1993, 1998). In all of our calculations, including the ASRO ones, we have ignored lattice distortions and kept an ideal lattice described by only a single-lattice parameter.

From anomalous x-ray scattering data, Jiang et al. (1996) find that for the differing alloy compositions in the fcc phase, the Ni-Ni nearest-neighbor (NN) distance follows a linear concentration dependence (i.e., Vegard's rule), the Fe-Fe NN distance is almost independent of concentration, and the Ni-Fe NN distance is actually smaller than that of Ni-Ni. This latter measurement is obviously contrary to hard-sphere packing arguments. Basically, Fe-Fe like to have larger "local volume" to increase local moments, and for Ni-Fe pairs the bonding is promoted (smaller distance) with a concomitant increase in the local Ni moment. Indeed, experiment and our calculations find about a 5% increase in the average moment upon chemical ordering in Ni_3Fe. These small local displacements in the Invar region actively contribute to the diffuse intensity (discussed above) when the ASRO is suppressed in the PM phase.

The Negative Thermal Expansion Effect. While many of the thermodynamic behaviors and anomalous properties of Ni-Fe Invar have been explained, questions remain regarding the origin of the negative thermal expansion. It is difficult to incorporate the displacement fluctuations (thermal phonons) on the same footing as magnetic and compositional fluctuations, especially within a first-principles approach. Progress on this front has been made by Mohn et al. (1991) and others (Entel et al., 1993, 1998; Uhl and Kubler, 1997). Recently, a possible cause of the negative thermal-expansion coefficient in Ni-Fe has been given within an effective Grüneisen theory (Abrikosov et al., 1995). Yet, this explanation is not definitive because the effect of phonons was not considered, i.e., only the electronic part of the Grüneisen constant was calculated.

For example, at 35% Ni, we find within the ASA calculations that the Ni and Fe moments respectively, are 0.62 μ_B and 2.39 μ_B for the $T = 0$ K FM state, and 0.00 μ_B and 1.56 μ_B in the DLM state, in contrast to a NM state (zero moments). From neutron-scattering data (Collins, 1966), the PM state contains moments of $\sim$1.42 μ_B on iron, similar to that found in the DLM calculations (Johnson et al., 1989).

Now we move on to consider a purely electronic explanation. In Figure 3, we show a plot of energy versus lattice parameter for a 25% Ni alloy in the NM, PM, and FM states. The FM curve has a double-well feature, i.e., two solutions, one with a large lattice parameter with high moments; the other, at a smaller volume has smaller moments. For the spin-restricted NM calculation (i.e. zero moments), a significant energy difference exists, even near low-spin FM minimum. The FM moments at smaller lattice constants are smaller than 0.001 Bohr magnetons, but finite. As Abrikosov et al. (1995) discuss, this double solution of the energy-versus-lattice parameter of the $T = 0$ K FM state produces an anomaly in the Grüneisen constant that leads to a negative thermal expansion effect. They argue that this is the only possible electronic origin of a negative thermal expansion coefficient. However, if temperature effects are considered—in particular, thermally induced phonons and local moment disorder—then it is not clear that this double-solution behavior is relevant near room temperature, where the lattice measurements are made.

Specifically, calculations of the heats of formation as in Figure 4 indicate that already at $T = 0$ K, neglecting the large entropy of such a state, the DLM state (or an AFM state) is slightly more energetically stable than the FM state at 25% Ni, and is intermediate to the NM and FM states at 35% Ni. Notice that the energy differences for 25% Ni are $\sim$0.5 mRy. Because of the high symmetry of the DLM state, in contrast to the FM case, a double-well feature is not seen in the energy-versus-volume curve (see Fig. 3). As Ni content is increased from 25%, the

low-spin solution rises in energy relative to high-spin solution before vanishing for alloys with more than 35% Ni (see figure 9 of Johnson et al., 1989). Thus, there appears to be less of a possibility of having a negative thermal expansion from this double-solution electronic effect as the temperature is raised, since thermal effects disorder the orientations of the moments (i.e. magnetization versus T should become Brillouin-like) and destroy, or lessen, this double-well feature. In support of this argument, consider that the Invar alloys do have signatures like spin-glasses—e.g., magnetic susceptibility—and the DLM state at $T = 0$ K could be supposed to be an approximate uncorrelated spin-glass (see discussion in Johnson et al., 1989). Thus, at elevated temperatures, both electronic and phonon effects contribute in some way, or, as one would think intuitively, phonons dominate.

The data from figure 3 and figure 9 of Johnson et al. (1989) show that a small energy is associated with orientation of disordering moments at 35% (an energy gain at 25% Ni) and that this yields a volume contraction of ~2%, from the high-spin FM state to (low-spin) DLM-PM state. On the other hand, raising the temperature gives rise to a lattice expansion due to phonon effects of ~1% to 2%. Therefore, the competition of these two effects lead to a small, perhaps negative, thermal expansion. This can only occur in the Invar region (for compositions greater than 25% and less than 40% Ni) because here states are sufficiently close in energy, with the DLM state being higher in energy. A Maxwell construction including these four states rules out the low-spin FM solution. A more quantitative explanation remains. Only by merging the effects of phonons with the magnetic disorder at elevated temperatures can one balance the expansion due to the former with the contraction due to the latter, and form a complete theory of the INVAR effect.

Magnetic Moments and Bonding in FeV Alloys

A simple schematic energy level diagram is shown in Figure 2B for Fe_cV_{1-c}. The d energy levels of Fe are exchange-split, showing that it is energetically favorable for pure bcc iron to have a net magnetization. Exchange-splitting is absent in pure vanadium. As in the case of the Ni_cFe_{1-c} alloys, we assume charge neutrality and align the two Fermi energies. The vanadium d levels lie much more closer in energy to the minority-spin d levels of iron than to its majority-spin ones. Upon alloying the two metals in a bcc structure, the bonding interactions have a larger effect on the minority-spin levels than those of the majority spin, owing to the smaller energy separation. In other words, Fe induces an exchange-splitting on the V sites to lower the kinetic energy which results in the formation of bonding and anti-bonding minority-spin alloy states. More minority-spin V-related d states are occupied than majority-spin d states, with the consequence of a moment on the vanadium sites anti-parallel to the larger moment on the Fe sites. The moments are not sustained for concentrations of iron less than 30%, since the Fe-induced exchange-splitting on the vanadium sites diminishes along with the average number of Fe atoms surrounding a vanadium site in the alloy. As for the majority-spin levels, well separated in energy, "split bands" form, i.e., states which reside mostly on one constituent or the other. In Figure 1B, we show the spin-polarized density of states of an iron-rich FeV alloy determined by the SCF-KKR-CPA method, where all these features can be identified.

Since the Fe and V majority-spin d states are well separated in energy, we expect a very smeared DOS in the majority-spin channel, due to the large disorder that the majority-spin electrons "see" as they travel through the lattice. On the other hand, the minority (spin-down) electron DOS should have peaks associated with the lower-energy, bonding states, as well as other peaks associated with the higher-energy, antibonding states. Note that the majority-spin DOS is very smeared due to chemical disorder, and the minority-spin DOS is much sharper, with the bonding states fully occupied and the antibonding states unoccupied. Note that the vertical line indicates the Fermi level, or chemical potential, of the electrons, below which the states are occupied. The Fermi level lies in this trough of the minority density of states for almost the entire concentration range. As discussed earlier, it is this positioning of the Fermi level holding the minority-spin electrons at a fixed number which gives rise to the mechanism for the straight 45° line on the left hand side of the Slater-Pauling curve. In general, the DOS depends on the underlying symmetry of the lattice and the subtle interplay between bonding and magnetism. Once again, we emphasize that the rigidly-split spin densities of states seen in the ferromagnetic elemental metals clearly do not describe the electronic structure in alloys. The variation of the moments on the Fe and V sites, as well as the average moments per site versus concentration as described by SCF-KKR-CPA calculations, are in good agreement with experimental measurement (Johnson et al., 1987).

ASRO and Magnetism in FeV

In this subsection we describe our investigation of the atomic short-range order in iron-vanadium alloys at (or rapidly quenched from) temperatures T_0 above any compositional ordering temperature. For these systems we find the ASRO to be rather insensitive to whether T_0 is above or below the alloy's magnetic Curie temperatures T_c, owing to the presence of "local exchange-splitting" in the electronic structure of the paramagnetic state. Iron-rich FeV alloys have several attributes that make them suitable systems in which to investigate both ASRO and magnetism. Firstly, their Curie temperatures (~1000 K) lie in a range where it is possible to compare and contrast the ASRO set up in both the ferromagnetic and paramagnetic states. The large difference in the coherent neutron scattering lengths, $b_{Fe} - b_V \approx 10$ fm, together with the small size difference, make them good candidates for neutron diffuse scattering experimental analyses.

In figure 1 of Cable et al. (1989), the neutron-scattering cross-sections as displayed along three symmetry directions measured in the presence of a saturating magnetic field for a $Fe_{87}V_{13}$ single crystal quenched a ferromagnetically ordered state. The structure of the curves is attributed to nuclear scattering connected with the ASRO,

$c(1-c)(b_{Fe}-b_V)^2\alpha(\mathbf{q})$. The most intense peaks occur at (1,0,0) and (1,1,1), indicative of a β-CuZn(B2) ordering tendency. Substantial intensity lies in a double peak structure around (1/2,1/2,1/2). We showed (Staunton et al., 1990, 1997) how our ASRO calculations for ferromagnetic $Fe_{87}V_{13}$ could reproduce all the details of the data. With the chemical potential being pinned in a trough of the minority-spin density of states (Fig. 1B), the states associated with the two different atomic species are substantially hybridized. Thus, the tendency to order is governed principally by the majority-spin electrons. These split-band states are roughly half-filled to produce the strong ordering tendency. The calculations also showed that part of the structure around (1/2,1/2,1/2) could be traced back to the majority-spin Fermi surface of the alloy.

By fitting the direct correlation function $S^{(2)}(\mathbf{q})$ in terms of real-space parameters

$$S^{(2)}(\mathbf{q}) = S_0^{(2)} + \sum_n \sum_{i\in n} S_n^{(2)} \exp(i\mathbf{q}\cdot\mathbf{R}_i) \qquad (27)$$

we found the fit is dominated by the first two parameters which determine the large peak at 1,0,0. However, the fit also showed a long-range component that was derived from the Fermi-surface effect. The real-space fit of data produced by Cable et al. (1989) showed large negative values for the first two shells, also followed by a weak long-ranged tail.

Cable et al. (1989) claimed that the effective temperature, for at least part of the sample, was indeed below its Curie temperature. To investigate this aspect, we carried out calculations for the ASRO of paramagnetic (DLM) $Fe_{87}V_{13}$ (Staunton et al., 1997). Once again, we found the largest peaks to be located at (1,0,0) and (1,1,1) but a careful scrutiny found less structure around (1/2,1/2,1/2) than in the ferromagnetic alloy. The ordering correlations are also weaker in this state. For the paramagnetic DLM state, the local exchange-splitting also pushes many antibonding states above the chemical potential ν (see Fig. 5). This happens although ν is no longer wedged in a valley in the density of states. The compositional ordering mechanism is similar to, although weaker than, that of the ferromagnetic alloy. The real space fit of $S^{(2)}(\mathbf{q})$ also showed a smaller long-ranged tail. Evidently the "local-moment" spin fluctuation disorder has broadened the alloy's Fermi surface and diminished its effect upon the ASRO.

Figure 3 of Pierron-Bohnes et al. (1995) shows measured neutron diffuse scattering intensities from $Fe_{80}V_{20}$ in its paramagnetic state at 1473 K and 1133 K (the Curie temperature is 1073 K) for scattering vectors in both the (1,0,0) and (1,−1,0) planes, following a standard correction for instrumental background and multiple scattering. Maximal intensity lies near (1,0,0) and (1,1,1) without subsidiary structure about (1/2, 1/2,1/2). Our calculations of the ASRO of paramagnetic $Fe_{80}V_{20}$, very similar to those of $Fe_{87}V_{13}$, are consistent with these features.

We also studied the type and extent of magnetic correlations in the paramagnetic state. Ferromagnetic correlations were shown which grow in intensity as T is reduced. These lead to an estimate of $T_c = 980$ K, which agrees well

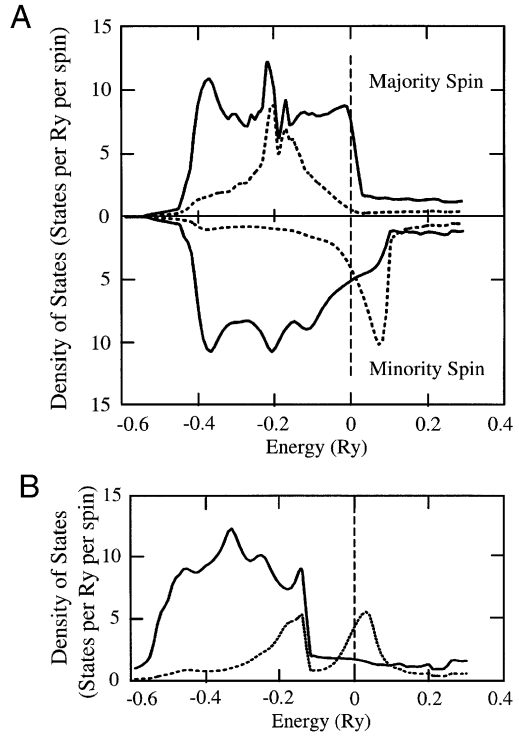


Figure 5. (**A**) The local electronic density of states for $Fe_{87}V_{13}$ with the moment directions being disordered. The upper half displays the density of states for the majority-spin electrons, the lower half, for the minority-spin electrons. Note that in the lower half the axis for the abscissa is inverted. These curves were calculated within the SCF-KKR-CPA (see Staunton et al., 1997). The solid line indicated contributions on the iron sites; the dashed line, the vanadiums sites. (**B**) The total electronic density of states for $Fe_{87}V_{13}$ with the moment directions being disordered. These curves were calculated within the SCF-KKR-CPA, see Johnson et al. (1989), and Johnson and Shelton (1997). The solid line indicates contributions on the iron sites; the dashed line, the vanadium sites.

with the measured value of 1073 K. (The calculated T_c for $Fe_{87}V_{13}$ of 1075 K also compares well with the measured value of 1180 K.) We also examined the importance of modeling the paramagnetic alloy in terms of local moments by repeating the calculations of ASRO, assuming a Stoner paramagnetic (NM) state in which there are no local moments and hence zero exchange splitting of the electronic structure, local or otherwise. The maximum intensity is now found at about (1/2,1/2,0) in striking contrast to both the DLM calculations and the experimental data.

In summary, we concluded that experimental data on FeV alloys are well interpreted by our calculations of ASRO and magnetic correlations. ASRO is evidently strongly affected by the local moments associated with the iron sites in the paramagnetic state, leading to only small differences between the topologies of the ASRO established in samples quenched from above and below T_c. The principal difference is the growth of structure around (1/2,1/2,1/2) for the ferromagnetic state. The ASRO strengthens quite sharply as the system orders magnetically, and it would be interesting if an *in situ*,

polarized-neutron, scattering experiment could be carried out to investigate this.

The ASRO of Gold-Rich AuFe Alloys: Dependence Upon Magnetic State

In sharp contrast to FeV, this study shows that magnetic order, i.e., alignment of the local moments, has a profound effect upon the ASRO of AuFe alloys. In Chapter 18 we discussed the electronic hybridization (size) effect which gives rise to the $\mathbf{q} = \{1,0,0\}$ ordering in NiPt. This is actually a more ubiquitous effect than one may at first imagine. In this subsection we show that the observed $\mathbf{q} = (1,1/2,0)$ short-range order in paramagnetic AuFe alloys that have been fast quenched from high temperature results partially from such an effect. Here we point out how magnetic effects also have an influence upon this unusual $\mathbf{q} = (1,1/2,0)$ short-range order (Ling et al., 1995b). We note that there has been a lengthy controversy over whether these alloys form a Ni_4Mo-type, or (1,1/2,0) special-point ASRO when fast-quenched from high temperatures, or whether the observed x-ray and neutron diffuse scattering intensities (or electron micrograph images) around (1,1/2,0) are merely the result of clusters of iron atoms arranged so as to produce this unusual type of ASRO. The issue was further complicated by the presence of intensity peaks around small $\mathbf{q} = (0,0,0)$ in diffuse x-ray scattering measurements and electron micrographs of some heat-treated AuFe alloys.

The uncertainty about the ASRO in these alloys arises from their strong dependence on thermal history. For example, when cooled from high temperatures, AuFe alloys in the concentration range of 10% to 30% Fe first form solid solutions on an underlying fcc lattice at around 1333 K. Upon further cooling below 973 K, α-Fe clusters begin to precipitate, coexisting with the solid solution and revealing their presence in the form of subsidiary peaks at $\mathbf{q} = (0,0,0)$ in the experimental scattering data. The number of α-Fe clusters formed within the fcc AuFe alloy, however, depends strongly on its thermal history and the time scale of annealing (Anderson and Chen, 1994; Fratzl et al., 1991). The miscibility gap appears to have a profound effect on the precipitation of α-Fe clusters, with the maximum precipitation occurring if the alloys had been annealed in the miscibility gap, i.e., between 573 and 773 K (Fratzl et al., 1991). Interestingly, all the AuFe crystals that reveal $\mathbf{q} = (0,0,0)$ correlations have been annealed at temperatures below both the experimental and our theoretical spinodal temperatures. On the other hand, if the alloys were homogenized at high temperatures outside the miscibility gap and then fast quenched, no αFe nucleation was found.

We have modeled the paramagnetic state of Au-Fe alloys in terms of disordered local moments in accord with the theoretical background described earlier. We calculated both $\alpha(\mathbf{q})$ and $\chi(\mathbf{q})$ in DLM-paramagnetic $Au_{75}Fe_{25}$ and for comparison have also investigated the ASRO in ferromagnetic $Au_{75}Fe_{25}$ (Ling et al., 1995b). Our calculations of $\alpha(\mathbf{q})$ for $Au_{75}Fe_{25}$ in the paramagnetic state show peaks at (1,1/2,0) with a spinodal ordering temperature of 780 K. This is in excellent agreement with experiment. Remarkably, as the temperature is lowered below 1600 K the peaks in $\alpha(\mathbf{q})$ shift to the (1,0.45,0) position with a gradual decrease towards (1,0,0) (Ling et al., 1995b). This streaking of the (1,1/2, 0) intensities along the (1,1,0) direction is also observed in electron micrograph measurements (van Tendeloo et al., 1985). The magnetic SRO in this alloy is found to be clearly ferromagnetic, with $\chi(\mathbf{q})$ peaking at (0,0,0). As such, we explored the ASRO in the "fictitious" FM alloy and find that $\alpha(\mathbf{q})$ shows peaks at (1,0,0).

Next, we show that the special point ordering in paramagnetic $Au_{75}Fe_{25}$ has its origins in the inherent "locally exchange-split" electronic structure of the disordered alloy. This is most easily understood from the calculated compositionally averaged densities of states (DOS), shown in Figure 5. Note that the double peak in the paramagnetic DLM Fe density of states in Figure 5A arises from the "local" exchange splitting, which sets up the "local moments" on the Fe sites. Similar features exist in the DOS of DLM $Fe_{87}V_{13}$. Within the DLM picture of the paramagnetic phase, it is important to note that this local DOS is obtained from the local axis of quantization on a given site due to the direction of the moment. All compositional and moment orientations contributing to the DOS must be averaged over, since moments point randomly in all directions. In comparison to a density of states in a ferromagnetic alloy, which has one global axis of quantization, the peaks in the DLM density of states are reminiscent of the more usual FM exchange splitting in Fe, as shown in Figure 5B.

What is evident from the DOS is that the chemical potential in the paramagnetic DLM state is located in an "antibonding"-like, exchange-split Fe peak. In addition, the "hybridized" bonding states that are created below the Fe d band are due to interaction with the wider-band Au (just as in NiPt). As a result of these two electronic effects, one arising from hybridization and the other from electronic exchange-splitting, a competition arises between (1,0,0)-type ordering from the t_{2g} hybridization states well below the Fermi level and (0,0,0)-type "ordering" (i.e., clustering) from the filling of unfavorable antibonding states. Recall again that the filling of bonding-type states favors chemical ordering, while the filling of antibonding-type states opposes chemical ordering, i.e., favors clustering. The competition between (1,0,0) and (0,0,0) type ordering from the two electronic effects yields a (1,1/2,0)-type ASRO. In this calculation, we can check this interpretation by artificially changing the chemical potential (or Fermi energy at $T = 0$ K) and then perform the calculation at a slightly different band-filling, or e/a. As the Fermi level is lowered below the higher-energy, exchange-split Fe peak, we find that the ASRO rapidly becomes (1,0,0)-type, simply because the unfavorable antibonding states are being depopulated and thus the clustering behavior suppressed.

As we have already stated, the ferromagnetic alloy exhibits (1,0,0)-type ASRO. In Figure 5B, at the Fermi level, the large antibonding, exchange-split, Fe peak is absent in the majority-spin manifold of the DOS, although it remains in the minority-spin manifold DOS. In other words, half of the states that were giving rise to the clustering behavior have been removed from consideration.

This happens because of the global exchange-splitting in the FM alloy; that is, a larger exchange-splitting forms and the majority-spin states become filled. Thus, rather than changing the ASRO by changing the electronic band-filling, one is able to alter the ASRO by changing the distribution of electronic states via the magnetic properties. Because the paramagnetic susceptibility $\chi(\mathbf{q})$ suggests that the local moments in the PM state are ferromagnetically correlated (Ling et al., 1995b), the alloy already is susceptible to FM ordering. This can be readily accomplished, for example, by magnetic-annealing the $Au_{75}Fe_{25}$ when preparing them at high temperatures, i.e. by placing the samples *in situ* into a strong magnetic field to align the moments. After the alloy is thermally annealed, the chemical response of the alloy is dictated by the electronic DOS in the FM disordered alloy, rather than that of the PM alloy, with the resulting ASRO being of (1,0,0)-type.

In summary, we have described two competing electronic mechanisms responsible for the unusual (1,1/2,0) ordering propensity observed in fast-quenched gold-rich AuFe alloys. This special point ordering we find to be determined by the inherent nature of the disordered alloy's electronic structure. Because the magnetic correlations in paramagnetic $Au_{75}Fe_{25}$ are found to be clearly ferromagnetic, we proposed that AuFe alloys grown in a magnetic field after homogenization at high temperature in the field, and then fast quenching, will produce a novel (1,0,0)-type ASRO in these crystals (Ling et al., 1995b).

We now move on and describe our studies of magnetocrystalline anisotropy in compositionally disordered alloys and hence show the importance of relativistic spin-orbit coupling upon the spin-polarized electronic structure.

Magnetocrystalline Anisotropy of Co_cPt_{1-c} Alloys

Co_cPt_{1-c} alloys are interesting for many reasons. Large magnetic anisotropy (MAE; Hadjipanayis and Gaunt, 1979; Lin and Gorman, 1992) and large magneto-optic Kerr effect (SURFACE MAGNETO-OPTIC KERR EFFECT) signals compared to the Co/Pt multilayers in the whole range of wavelengths (820 to 400 nm; Weller et al., 1992, 1993) make these alloys potential magneto-optical recording materials. The chemical stability of these alloys, a suitable Curie temperature, and the ease of manufacturing enhance their usefulness in commercial applications. Furthermore, study of these alloys may lead to an improved understanding of the fundamental physics of magnetic anisotropy; the spin-polarization in the alloys being induced by the presence of Co whereas a large spin-orbit coupling effect can be associated with the Pt atoms.

Most experimental work on Co-Pt alloys has been on the ordered tetragonal phase, which has a very large magnetic anisotropy ~400 μeV, and magnetic easy axis along the *c* axis (Hadjipanayis and Gaunt, 1979; Lin and Gorman, 1992). We are not aware of any experimental work on the bulk disordered fcc phase of these alloys. However, some results have been reported for disordered fcc phase in the form of thin films (Weller et al., 1992, 1993; Suzuki et al., 1994; Maret et al., 1996; Tyson et al., 1996). It is found that the magnitude of MAE is more than one order of magnitude smaller than that of the bulk ordered phase, and that the magnetic easy axis varies with film thickness. From these data we can infer that a theoretical study of the MAE of the bulk disordered alloys provides insight into the mechanism of magnetic anisotropy in the ordered phase as well as in thin films. We investigated the magnetic anisotropy of disordered fcc phase of Co_cPt_{1-c} alloys for $c = 0.25$, 0.50, and 0.75 (as well as the pure elements Fe, Ni, and Co, and also Ni_cPt_{1-c}). In our calculations, we used self-consistent potentials from spin-polarized scalar-relativistic KKR-CPA calculations and predicted that the easy axis of magnetization is along the ⟨111⟩ direction of the crystal for all the three compositions, and the anisotropy is largest for $c = 0.50$.

In this first calculation of the MAE of disordered alloys we started with atomic sphere potentials generated from the self-consistent spin-polarized scalar relativistic KKR-CPA for Co_cPt_{1-c} alloys and constructed spin-dependent potentials. We recalculated the Fermi energy within the SPR-KKR-CPA method for magnetization along the ⟨001⟩ direction. This was necessary since earlier studies on the MAE of the 3*d* transition metal magnets were found to be quite sensitive to the position of the Fermi level (Daalderop et al., 1993; Strange et al., 1991). For all the three compositions of the alloy, the difference in the Fermi energies of the scalar relativistic and fully relativistic cases were of the order of 5 mRy, which is quite large compared to the magnitude of MAE. The second term in the expression above for the MAE was indeed small in comparison with the first, which needed to be evaluated very accurately. Details of the calculation can be found elsewhere (Razee et al., 1997).

In Figure 6, we show the MAE of disordered fcc-Co_cPt_{1-c} alloys for $c = 0.25$, 0.5, and 0.75 as a function of temperature between 0 K and 1500 K. We note that for all the three compositions, the MAE is positive at all temperatures, implying that the magnetic easy axis is always along the ⟨111⟩ direction of the crystal, although the magnitude of MAE decreases with increasing temperature. The magnetic easy axis of fcc Co is also along the ⟨111⟩ direction but the magnitude of MAE is smaller. Thus, alloying

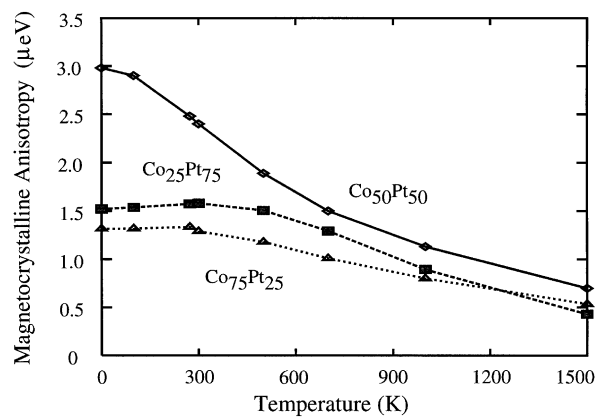


Figure 6. Magneto-anisotropy energy of disordered fcc-Co_cPt_{1-c} alloys for $c = 0.25$, 0.5, and 0.75 as a function of temperature. Adapted from Razee et al. (1997).

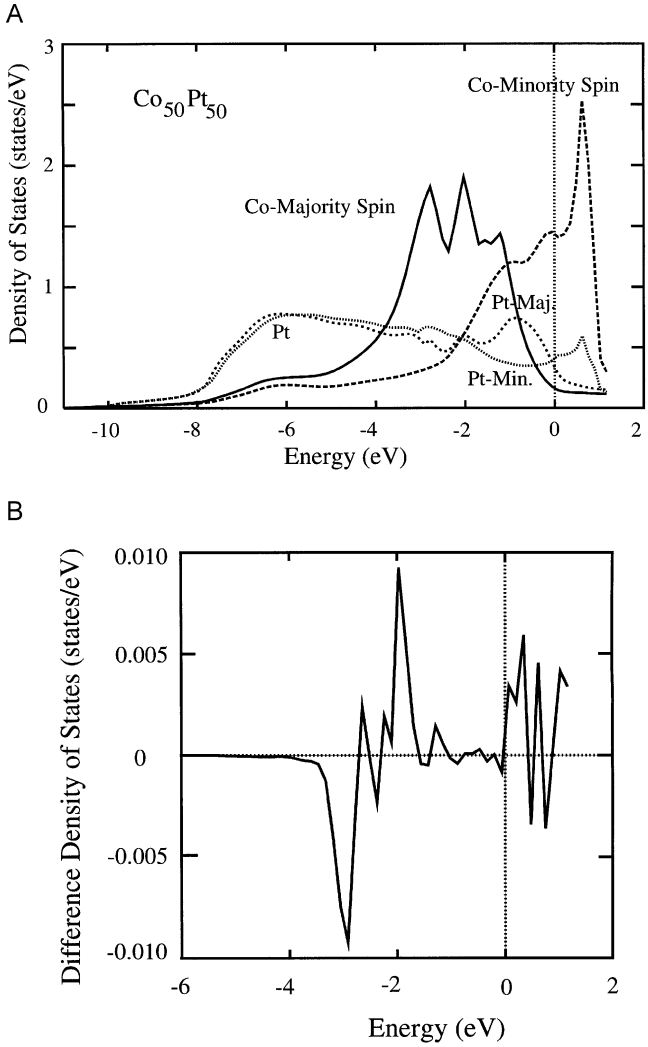


Figure 7. (**A**) The spin-resolved density of states on Co and Pt Atoms in the $Co_{0.50}Pt_{0.50}$ alloy magnetized along the ⟨001⟩ direction. (**B**) The density of states difference between the two magnetization directions for $Co_{0.50}Pt_{0.50}$. Adapted from Razee et al. (1997).

with Pt does not alter the magnetic easy axis. The equiatomic composition has the largest MAE, which is ~3.0 μeV at 0 K. In these alloys, one component (Co) has a large magnetic moment but weak spin-orbit coupling, while the other component (Pt) has strong spin-orbit coupling but small magnetic moment. Adding Pt to Co results in a monotonic decrease in the average magnetic moment of the system with the spin-orbit coupling becoming stronger. At $c = 0.50$, both the magnetic moment as well as the spin-orbit coupling are significant; for other compositions either the magnetic moment or the spin-orbit coupling is weaker. This trade-off between spin-polarization and spin-orbit coupling is the main reason for the MAE being largest around this equiatomic composition.

In finer detail, the magnetocrystalline anisotropy of a system can be understood in terms of its electronic structure. In Figure 7A, we show the spin-resolved density of states on Co and Pt atoms in the $Co_{0.50}Pt_{0.50}$ alloy magnetized along the ⟨001⟩ direction. The Pt density of states is rather structureless, except around the Fermi energy where there is spin-splitting due to hybridization with Co *d* bands. When the direction of magnetization is oriented along the ⟨111⟩ direction of the crystal, the electronic structure also changes due to redistribution of the electrons, but the difference is quite small in comparison with the overall density of states. So in Figure 7B, we have plotted the density of states difference for the two magnetization directions. In the lower part of the band, which is Pt-dominated, the difference between the two is small, whereas it is quite oscillatory in the upper part dominated by Co *d*-band complex. There are also spikes at energies where there are peaks in the Co-related part of the density of states. Due to the oscillatory nature of this curve, the magnitude of MAE is quite small; the two large peaks around 2 eV and 3 eV below the Fermi energy almost cancel each other, leaving only the smaller peaks to contribute to the MAE. Also, due to this oscillatory behavior, a shift in the Fermi level will alter the magnitude as well as the sign of the MAE. This curve also tells us that states far removed from the Fermi level (in this case, 4eV below the Fermi level) can also contribute to the MAE, and not just the electrons near the Fermi surface.

In contrast to what we have found for the disordered fcc phase of Co_cPt_{1-c} alloys, in the ordered tetragonal CoPt alloy the MAE is quite large (~400 μeV), two orders of magnitude greater than what we find for the disordered $Co_{0.50}Pt_{0.50}$ alloy. Moreover, the magnetic easy axis is along the *c* axis (Hadjipanayis and Gaunt, 1979). Theoretical calculations of MAE for ordered tetragonal CoPt alloy (Sakuma, 1994; Solovyev et al., 1995), based on scalar relativistic methods, do reproduce the correct easy axis but overestimate the MAE by a factor of 2. Currently, it is not clear whether it is the atomic ordering or the loss of cubic symmetry of the crystal in the tetragonal phase which is responsible for the altogether different magnetocrystalline anisotropies in disordered and ordered CoPt alloys. A combined effect of the two is more likely; we are studying the effect of atomic short-range order on the magnetocrystalline anisotropy of alloys.

PROBLEMS AND CONCLUSIONS

Magnetism in transition metal materials can be described in quantitative detail by spin-density functional theory (SDFT). At low temperatures, the magnetic properties of a material are characterized in terms of its spin-polarized electronic structure. It is on this aspect of magnetic alloys that we have concentrated. From this basis, the early Stoner-Wohlfarth picture of rigidly exchange-split, spin-polarized bands is shown to be peculiar to the elemental ferromagnets only. We have identified and shown the origins of two commonly occurring features of ferromagnetic alloy electronic structures, and the simple structure of the Slater-Pauling curve for these materials (average magnetic moment versus electron per atom ratio), can be traced back to the spin-polarized electronic structure.

The details of the electronic basis of the theory can, with care, be compared to results from modern spectroscopic

experiment. Much work is ongoing to make this comparison as rigorous as possible. Indeed, our understanding of metallic magnets and their scope for technological application are developing via the growing sophistication of some experiments, together with improvements in quantitative theory.

Although SDFT is "first-principled," most applications resort to the local approximation (LSDA) for the many electron exchange and correlation effects. This approximation is widely used and delivers good results in many calculations. It does have shortcomings, however, and there are many efforts aimed at trying to improve it. We have referred to some of this work, mentioning the "generalized gradient approximation" GGA and the "self-interaction correction" SIC in particular.

The LDA in magnetic materials fails when it is straightforwardly adapted to high temperatures. This failure can be redressed by a theory that includes the effects of thermally induced magnetic excitations, but which still maintains the spin-polarized electronic structure basis of standard SDFT. "Local moments," which are set up by the collective behavior of all the electrons, and are associated with atomic sites, change their orientations on a time scale which is long compared to the time that itinerant *d* electrons take to progress from site to site. Thus, we have a picture of electrons moving through a lattice of effective magnetic fields set up by particular orientations of these "local moments." At high temperatures, the orientations are thermally averaged, so that in the paramagnetic state there is zero magnetization overall. Although not spin-polarized "globally"—i.e., when averaged over all orientational configurations—the electronic structure is modified by the local-moment fluctuations, so that "local spin-polarization" is evident. We have described a mean field theory of this approach and have described its successes for the elemental ferromagnetic metals and for some iron alloys. The dynamical effects of these spin fluctuations in a first-principles theory remain to be included.

We have also emphasized how the state of magnetic order of an alloy can have a major effect on various other properties of the system, and we have dealt at length with its effect upon atomic short-range order by describing case studies of NiFe, FeV, and AuFe alloys. We have linked the results of our calculations with details of "globally" and "locally" spin-polarized electronic structure. The full consequences of lattice displacement effects have yet to be incorporated. We have also discussed the relativistic generalization of SDFT and covered its implication for the magnetocrystalline anisotropy of disordered alloys, with specific illustrations for CoPt alloys.

In summary, the magnetic properties of transition metal alloys are fundamentally tied up with the behavior of their electronic "glues." As factors like composition and temperature are varied, the underlying electronic structure can change and thus modify an alloy's magnetic properties. Likewise, as the magnetic order transforms, the electronic structure is affected and this, in turn, leads to changes in other properties. Here we have focused upon the effect on ASRO, but much could also have been written about the fascinating link between magnetism and elastic properties—"Invar" phenomena being a particularly dramatic example. The electronic mechanisms that thread all these properties together are very subtle, both to understand and to uncover. Consequently, it is often required that a study be attempted that is parameter-free as far as possible, so as to remove any pre-existing bias. This calculational approach can be very fruitful, provided it is followed alongside suitable experimental measurements as a check of its correctness.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation (U.S.), the Engineering and Physical Sciences Research Council (U.K.), and the Department of Energy (U.S.) at the Fredrick Seitz Material Research Lab at the University of Illinois under grants DEFG02-ER9645439 and DE-AC04-94AL85000.

LITERATURE CITED

Abrikosov, I. A., Eriksson, O., Soderlind, P., Skriver, H. L., and Johansson, B., 1995. Theoretical aspects of the Fe_cNi_{1-c} Invar alloy. *Phys. Rev. B.* 51:1058–1066.

Anderson, J. P. and Chen, H., 1994. Determination of the short-range order structure of Au-25 at pct Fe using wide-angle diffuse synchrotron x-ray-scattering. *Metallurgical and Materials Transactions A* 25A:1561.

Anisimov,V. I., Aryasetiawan, F., and Liechtenstein, A. I., 1997. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+U method. *J. Phys.: Condens. Matter* 9:767–808.

Asada, T. and Terakura, K., 1993. Generalized-gradient-approximation study of the magnetic and cohesive properties of bcc, fcc, and hcp Mn. *Phys. Rev. B* 47:15992–15995.

Bagayoko, D. and Callaway, J., 1983. Lattice-parameter dependence of ferromagnetism in bcc and fcc iron. *Phys. Rev. B* 28:5419–5422.

Bagno, P., Jepsen, O., and Gunnarsson, O., 1989. Ground-state properties of 3rd-row elements with nonlocal density functionals. *Phys. Rev. B* 40:1997–2000.

Beiden, S. V., Temmerman, W. M., Szotek, Z., and Gehring, G. A., 1997. Self-interaction free relativistic local spin density approximation: equivalent of Hund's rules in γ-Ce. *Phys. Rev. Lett.* 79:3970–3973.

Brooks, H., 1940. Ferromagnetic Anisotropy and the Itinerant Electron Model. *Phys. Rev.* 58:B909.

Brooks, M. S. S. and Johansson, B., 1993. Density functional theory of the ground state properties of rare earths and actinides *In* Handbook of Magnetic Materials (K. H. J. Buschow, ed.). p. 139. Elsevier/North Holland, Amsterdam.

Brooks, M. S. S., Eriksson, O., Wills, J. M., and Johansson, B., 1997. Density functional theory of crystal field quasiparticle excitations and the ab-initio calculation of spin hamiltonian parameters. *Phys. Rev. Lett.* 79:2546.

Brout, R. and Thomas, H., 1967. Molecular field theory: the Onsager reaction field and the spherical model. *Physics* 3:317.

Butler, W. H. and Stocks, G. M., 1984. Calculated electrical conductivity and thermopower of silver-palladium alloys. *Phys. Rev. B* 29:4217.

Cable, J. W. and Medina, R. A., 1976. Nonlinear and nonlocal moment disturbance effects in Ni-Cr alloys. *Phys. Rev. B* 13:4868.

Cable, J. W., Child, H. R., and Nakai, Y., 1989. Atom-pair correlations in Fe-13.5-percent-V. *Physica* 156 & 157 B:50.

Callaway, J. and Wang, C. S., 1977. Energy bands in ferromagnetic iron. *Phys. Rev. B* 16:2095–2105.

Capellman, H., 1977. Theory of itinerant ferromagnetism in the 3-d transition metals. *Z. Phys. B* 34:29.

Ceperley, D. M. and Alder, B. J., 1980. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.* 45:566–569.

Chikazumi, S., 1964. Physics of Magnetism. Wiley, New York.

Chikazurin, S. and Graham, C. D., 1969. Directional order. *In* Magnetism and Metallurgy, (A. E. Berkowitz and E. Kneller, eds.). vol. II, pp. 577–619. Academic Press, Inc, New York.

Chuang, Y.-Y., Hseih, K.-C., and Chang,Y. A, 1986. A thermodynamic analysis of the phase equilibria of the Fe-Ni system above 1200K. *Metall. Trans. A* 17:1373.

Collins, M. F., 1966. Paramagnetic scattering of neutrons by an iron-nickel alloy. *J. Appl. Phys.* 37:1352.

Connolly, J. W. D. and Williams, A. R., 1983. Density-functional theory applied to phase transformations in transition-metal alloys. *Phys. Rev. B* 27:RC5169.

Daalderop, G. H. O., Kelly, P. J., and Schuurmans, M. F. H., 1993. Comment on state-tracking first-principles determination of magnetocrystalline anisotropy. *Phys. Rev. Lett.* 71:2165.

Driezler, R. M. and da Providencia, J. (eds.) 1985. Density Functional Methods in Physics. Plenum, New York.

Ducastelle, F. 1991. Order and Phase Stability in Alloys. Elsevier/North-Holland, Amsterdam.

Ebert, H. 1996. Magneto-optical effects in transition metal systems. *Rep. Prog. Phys.* 59:1665.

Ebert, H. and Akai, H., 1992. Spin-polarised relativistic band structure calculations for dilute and doncentrated disordered alloys, *In* Applications of Multiple Scattering Theory to Materials Science, Mat. Res. Soc. Symp. Proc. 253, (W.H. Butler, P.H. Dederichs, A. Gonis, and R.L. Weaver, eds.). pp. 329. Materials Research Society Press, Pittsburgh.

Edwards, D. M. 1982. The paramagnetic state of itinerant electron-systems with local magnetic-moments. I. Static properties. *J. Phys. F: Metal Physics* 12:1789–1810.

Edwards, D. M. 1984. On the dynamics of itinerant electron magnets in the paramagnetic state. *J.Mag.Magn.Mat.* 45:151–156.

Entel, P., Hoffmann, E., Mohn, P., Schwarz, K., and Moruzzi, V. L. 1993. First-principles calculations of the instability leading to the INVAR effect. *Phys. Rev. B* 47:8706–8720.

Entel, P., Kadau, K., Meyer, R., Herper, H. C. Acet, M., and Wassermann, E.F. 1998. Numerical simulation of martensitic transformations in magnetic transition-metal alloys. *J. Mag. and Mag. Mat.* 177:1409–1410.

Ernzerhof, M., Perdew, J. P., and Burke, K. 1996. Density functionals: where do they come from, why do they work? *Topics in Current Chemistry* 180:1–30.

Faulkner, J. S. and Stocks, G. M., 1980. Calculating properties with the coherent-potential approximation. *Phys. Rev. B* 21:3222.

Faulkner, J. S., Wang, Y., and Stocks, G. M. 1997. Coulomb energies in alloys. *Phys. Rev. B* 55:7492.

Faulkner, J. S., Moghadam, N. Y., Wang, Y., and Stocks, G. M. 1998. Evaluation of isomorphous models of alloys. *Phys. Rev. B* 57:7653.

Feynman, R. P., 1955. Slow electrons in a polar crystal. *Phys. Rev.* 97:660.

Fratzl, P., Langmayr, F., and Yoshida, Y. 1991. Defect-mediated nucleation of alpha-iron in Au-Fe alloys. *Phys. Rev. B* 44:4192.

Gay, J. G. and Richter, R., 1986. Spin anisotropy of ferromagnetic-films. *Phys. Rev. Lett.* 56:2728.

Gubanov, V. A., Liechtenstein, A. I., and Postnikov, A. V. 1992. Magnetism and Electronic Structure of Crystals, Springer Series in Solid State Physics Vol. 98. Springer-Verlag, New York.

Gunnarsson, O. 1976. Band model for magnetism of transition metals in the spin-density-functional formalism. *J. Phys. F: Metal Physics* 6:587–606.

Gunnarsson, O. and Lundqvist, B. I., 1976. Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism. *Phys. Rev. B* 13:4274–4298.

Guo, G. Y., Temmerman, W. M., and Ebert, H. 1991. 1st-principles determination of the magnetization direction of Fe monolayer in noble-metals. *J. Phys. Condens. Matter* 3:8205.

Györffy, B. L. and Stocks, G. M. 1983. Concentration waves and fermi surfaces in random metallic alloys. *Phys. Rev. Lett.* 50:374.

Györffy, B. L., Kollar, J., Pindor, A. J., Staunton, J. B., Stocks, G. M., and Winter, H. 1983. *In* The Proceedings from Workshop on 3d Metallic Magnetism, Grenoble, France, March, 1983, pp. 121–146.

Györffy, B. L., Pindor, A. J., Staunton, J. B., Stocks, G. M., and Winter, H. 1985. A first-principles theory of ferromagnetic phase-transitions in metals. *J. Phys. F: Metal Physics* 15:1337–1386.

Györffy, B. L., Johnson, D. D., Pinski, F. J., Nicholson, D. M., Stocks, G. M. 1989. The electronic structure and state of compositional order in metallic alloys. *In* Alloy Phase Stability (G. M. Stocks and A. Gonis, eds.). NATO-ASI Series Vol. 163. Kluwer Academic Publishers, Boston.

Hadjipanayis, G. and Gaunt, P. 1979. An electron microscope study of the structure and morphology of a magnetically hard PtCo alloy. *J. Appl. Phys.* 50:2358.

Haglund, J. 1993. Fixed-spin-moment calculations on bcc and fcc iron using the generalized gradient approximation. *Phys. Rev. B* 47:566–569.

Haines, E. M., Clauberg, R., and Feder, R. 1985. Short-range magnetic order near the Curie temperature in iron from spin-resolved photoemission. *Phys. Rev. Lett.* 54:932.

Haines, E. M., Heine, V., and Ziegler, A. 1986. Photoemission from ferromagnetic metals above the Curie temperature. 2. Cluster calculations for Ni. *J. Phys. F: Metal Physics* 16:1343.

Hasegawa, H. 1979. Single-site functional-integral approach to itinerant-electron ferromagnetism. *J. Phys. Soc. Jap.* 46:1504.

Hedin, L. and Lundqvist, B. I. 1971. Explicit local exchange-correlation potentials. *J. Phys. C: Solid State Physics* 4:2064–2083.

Heine,V. and Joynt, R. 1988. Coarse-grained magnetic disorder above T_c in iron. *Europhys. Letts.* 5:81–85.

Heine,V. and Samson, J. H. 1983. Magnetic, chemical and structural ordering in transition-metals. *J. Phys. F: Metal Physics* 13:2155–2168.

Herring, C. 1966. Exchange interactions among itinerant electrons *In* Magnetism IV (G.T. Rado and H. Suhl, eds.). Academic Press, Inc., New York.

Hohenberg, P. and Kohn, W. 1964. Inhomogeneous electron gas. *Phys. Rev.* 136:B864–B872.

Hu, C. D. and Langreth, D. C. 1986. Beyond the random-phase approximation in nonlocal-density-functional theory. *Phys. Rev. B* 33:943–959.

Hubbard, J. 1979. Magnetism of iron. II. *Phys. Rev. B* 20:4584.

Jansen, H. J. F. 1988. Magnetic-anisotropy in density-functional theory. *Phys. Rev. B* 38:8022–8029.

Jiang, X., Ice, G. E., Sparks, C. J., Robertson, L., and Zschack, P. 1996. Local atomic order and individual pair displacements of $Fe_{46.5}Ni_{53.5}$ and $Fe_{22.5}Ni_{77.5}$ from diffuse x-ray scattering studies. *Phys. Rev. B* 54:3211.

Johnson, D. D. and Pinski, F. J. 1993. Inclusion of charge correlations in the calculation of the energetics and electronic structure for random substitutional alloys. *Phys. Rev. B* 48:11553.

Johnson, D. D. and Shelton, W. A. 1997. The energetics and electronic origins for atomic long- and short-range order in Ni-Fe invar alloys. *In* The Invar Effect: A Centennial Symposium, (J.Wittenauer, ed.). p. 63. The Minerals Metals, and Materials Society, Warrendale, Pa.

Johnson, D. D., Pinski, F. J., and Staunton, J. B. 1987. The Slater-Pauling curve: First-principles calculations of the moments of $Fe_{1-c}Ni_c$ and $V_{1-c}Fe_c$. *J. Appl. Phys.* 61:3715–3717.

Johnson, D. D., Pinski, F. J., Staunton, J. B., Györffy, B. L, and Stocks, G. M. 1989. Theoretical insights into the underlying mechanisms responsible for their properties. *In* Physical Metallurgy of Controlled Expansion "INVAR-type" Alloys (K. C. Russell and D. Smith, eds.). The Minerals, Metals, and Materials Society, Warrendale, Pa.

Johnson, D. D., Nicholson, D. M., Pinski, F. J., Stocks, G. M., and Györffy, B.L. 1986. Density-functional theory for random alloys: Total energy with the coherent potential approximation. *Phys. Lett.* 56:2096.

Johnson, D. D., Nicholson, D. M., Pinski, F. J., Stocks, G. M., and Györffy, B. L. 1990. Total energy and pressure calculations for random substitutional alloys. *Phys. Rev. B* 41:9701.

Jones, R. O. and Gunnarsson, O. 1989. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.* 61:689–746.

Kakizaki, A., Fujii, J., Shimada, K., Kamata, A., Ono, K., Park, K. H., Kinoshita, T., Ishii, T., and Fukutani, H. 1994. Fluctuating local magnetic-moments in ferromagnetic Ni observed by the spin-resolved resonant photoemission. *Phys. Rev. Lett.* 17: 2781–2784.

Khachaturyan, A. G. 1983. Theory of structural transformations in solids. John Wiley & Sons, New York.

Kirschner, J., Globl, M., Dose, V., and Scheidt, H. 1984. Wave-vector dependent temperature behavior of empty bands in ferromagnetic iron. *Phys. Rev. Lett.* 53:612–615.

Kisker, E., Schroder, K., Campagna, M., and Gudat, W. 1984. Temperature-dependence of the exchange splitting of Fe by spin- resolved photoemission spectroscopy with synchrotron radiation. *Phys. Rev. Lett.* 52:2285–2288.

Kisker, E., Schroder, K., Campagna, M., and Gudat, W. 1985. Spin-polarised, angle-resolved photoemission study of the electronic structure of Fe(100) as a function of temperature. *Phys. Rev. B* 31:329–339.

Koelling, D. D. 1981. Self-consistent energy-band calculations. *Rep. Prog. Phys.* 44:139–212.

Kohn, W. and Sham, L. J. 1965. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* 140:A1133.

Kohn,W. and Vashishta, P. 1982. Physics of solids and liquids. (B.I. Lundqvist and N. March, eds.). Plenum, New York.

Korenman, V. 1985. Theories of itinerant magnetism. *J. Appl. Phys.* 57:3000–3005.

Korenman, V., Murray, J. L., and Prange, R. E. 1977a. Local-band theory of itinerant ferromagnetism. I. Fermi-liquid theory. *Phys. Rev. B* 16:4032.

Korenman, V., Murray, J. L., and Prange, R. E. 1977b. Local-band theory of itinerant ferromagnetism. II. Spin waves. *Phys. Rev. B* 16:4048.

Korenman, V., Murray, J. L., and Prange, R. E. 1977c. Local-band theory of itinerant ferromagnetism. III. Nonlinear Landau-Lifshitz equations. *Phys. Rev. B* 16:4058.

Krivoglaz, M. 1969. Theory of x-ray and thermal-neutron scattering by real crystals. Plenum Press, New York.

Kubler, J. 1984. First principle theory of metallic magnetism. *Physica B* 127:257–263.

Kuentzler, R. 1980. Ordering effects in the binary T-Pt alloys. *In* Physics of Transition Metals. 1980, Institute of Physics Conference Series no.55 (P. Rhodes, ed.). pp. 397–400. Institute of Physics, London.

Langreth, D. C. and Mehl, M. J. 1981. Easily implementable nonlocal exchange-correlation energy functionals. *Phys. Rev. Lett.* 47:446.

Langreth, D. C. and Mehl, M. J. 1983. Beyond the local density approximation in calculations of ground state electronic properties. *Phys. Rev. B* 28:1809.

Lieb, E. 1983. Density functionals for coulomb-systems. *Int. J. of Quantum Chemistry* 24:243.

Lin, C. J. and Gorman, G. L. 1992. Evaporated CoPt alloy films with strong perpendicular magnetic anisotropy. *Appl. Phys. Lett.* 61:1600.

Ling, M. F., Staunton, J. B., and Johnson, D. D. 1994a. Electronic mechanisms for magnetic interactions in a Cu-Mn spin-glass. *Europhys. Lett.* 25:631–636.

Ling, M. F., Staunton, J. B., and Johnson, D. D. 1994b. A first-principles theory for magnetic correlations and atomic short-range order in paramagnetic alloys. I. *J. Phys.: Condens. Matter* 6:5981–6000.

Ling, M. F., Staunton, J. B., and Johnson, D. D. 1995a. All-electron, linear response theory of local environment effects in magnetic, metallic alloys and multilayers. *J. Phys.: Condensed Matter* 7:1863–1887.

Ling, M. F., Staunton, J. B., Pinski, F. J., and Johnson, D. D. 1995b. Origin of the {1,1/2,0} atomic short range order in Au-rich Au-Fe alloys. *Phys. Rev. B: Rapid Communications* 52:3816– 3819.

Liu, A. Y. and Singh, D. J. 1992. General-potential study of the electronic and magnetic structure of FeCo. *Phys. Rev. B* 46: 11145–11148.

Liu, S. H. 1978. Quasispin model for itinerant magnetism—effects of short-range order. *Phys. Rev. B* 17:3629–3638.

Lonzarich, G. G. and Taillefer, L. 1985. Effect of spin fluctuations on the magnetic equation of state of ferromagnetic or nearly ferromagnetic metals. *J.Phys.C: Solid State Physics* 18:4339.

Lovesey, S. W. 1984. Theory of neutron scattering from condensed matter. 1. Nuclear scattering, International series of monographs on physics. Clarendon Press, Oxford.

MacDonald, A. H. and Vosko, S. H. 1979. A relativistic density functional formalism. *J. Phys. C: Solid State Physics* 12:6377.

Mackintosh, A.R. and Andersen, O. K. 1980. The electronic structure of transition metals. *In* Electrons at the Fermi Surface (M. Springford, ed.). pp. 149–224. Cambridge University Press, Cambridge.

Malozemoff, A. P., Williams, A. R., and Moruzzi, V. L. 1984. Band-gap theory of strong ferromagnetism: Application to concentrated crystalline and amorphous Fe-metalloid and Co-metalloid alloys. *Phys. Rev. B* 29:1620–1632.

Maret, M., Cadeville, M.C., Staiger, W., Beaurepaire, E., Poinsot, R., and Herr, A. 1996. Perpendicular magnetic anisotropy in Co_xPt_{1-x} alloy films. *Thin Solid Films* 275:224.

Marshall, W. 1968. Neutron elastic diffuse scattering from mixed magnetic systems. *J. Phys. C: Solid State Physics* 1:88.

Massalski, T. B., Okamoto, H., Subramanian, P. R., and Kacprzak, L. 1990. Binary alloy phase diagrams. American Society for Metals, Metals Park, Ohio.

McKamey, C. G., DeVan, J. H., Tortorelli, P. F., and Sikka, V. K. 1991. A review of recent developments in Fe_3Al-based alloys. *Journal of Materials Research* 6:1779–1805.

Mermin, N. D. 1965. Thermal properties of the inhomogeneous electron gas. *Phys. Rev.* 137:A1441.

Mohn, P., Schwarz, K., and Wagner, D. 1991. Magnetoelastic anomalies in Fe-Ni invar alloys. *Phys. Rev. B* 43:3318.

Moriya, T. 1979. Recent progress in the theory of itinerant electron magnetism. *J. Mag. Magn. Mat.* 14:1.

Moriya, T. 1981. Electron Correlations and Magnetism in Narrow Band Systems. Springer, New York.

Moruzzi, V. L. and Marcus, P. M. 1993. Energy band theory of metallic magnetism in the elements. *In* Handbook of Magnetic Materials, vol. 7 (K. Buschow, ed.) pp. 97. Elsevier/North Holland, Amsterdam.

Moruzzi, V. L., Janak, J. F., and Williams, A. R. 1978. Calculated Electronic Properties of Metals. Pergamon Press, Elmsford, N.Y.

Moruzzi, V. L., Marcus, P. M., Schwarz, K., and Mohn, P. 1986. Ferromagnetic phases of BCC and FCC Fe, Co and Ni. *Phys. Rev. B* 34:1784.

Mryasov, O. N., Gubanov, V. A., and Liechtenstein, A. I. 1992. Spiral-spin-density-wave states in fcc iron: Linear-muffin-tin-orbitals band-structure approach. *Phys. Rev. B* 21:12330–12336.

Murata, K. K. and Doniach, S. 1972. Theory of magnetic fluctuations in itinerant ferromagnets. *Phys. Rev. Lett.* 29:285.

Oguchi, T., Terakura, K., and Hamada, N. 1983. Magnetism of iron above the Curie-temperature. *J. Phys. F.: Metal Physics* 13:145–160.

Pederson, M. R., Heaton, R. A., and Lin, C. C. 1985. Density-functional theory with self-interaction correction: application to the lithium molecule. *J. Chem. Phys.* 82:2688.

Perdew, J. P. and Yue, W. 1986. Accurate and simple density functional for the electronic exchange energy: Generalised gradient approximation. *Phys. Rev. B* 33:8800.

Perdew, J. P. and Zunger, A. 1981. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* 23:5048–5079.

Perdew, J. P., Chevary, J. A., Vosko, S. H., Jackson, K. A., Pedersen, M., Singh, D. J., and Fiolhais, C. 1992. Atoms, molecules, solids and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* 46:667.

Perdew, J. P., Burke, K., and Ernzerhof, M. 1996. GGA made simple. *Phys. Rev. Lett.* 77:3865–68.

Pettifor, D. G. 1995. Bonding and Structure of Molecules and Solids. Oxford University Press, Oxford.

Pierron-Bohnes, V., Kentzinger, E., Cadeville, M. C., Sanchez, J. M., Caudron, R., Solal, F., and Kozubski, R. 1995. Experimental determination of pair interactions in a $Fe_{0.804}V_{0.196}$ single crystal. *Phys. Rev. B* 51:5760.

Pinski, F. J., Staunton, J. B., Györffy, B. L., Johnson, D. D., and Stocks, G. M. 1986. Ferromagnetism versus antiferromagnetism in face-centered-cubic iron. *Phys. Rev. Lett.* 56:2096–2099.

Rajagopal, A. K. 1978. Inhomogeneous relativistic electron gas. *J. Phys. C: Solid State Physics* 11:L943.

Rajagopal, A. K. 1980. Spin density functional formalism. *Adv. Chem. Phys.* 41:59.

Rajagopal, A. K. and Callaway, J. 1973. Inhomogeneous electron gas. *Phys. Rev. B* 7:1912.

Ramana, M. V. and Rajagopal, A. K. 1983. Inhomogeneous relativistic electron-systems: A density-functional formalism. *Adv.Chem.Phys.* 54:231–302.

Razee, S. S. A., Staunton, J. B., and Pinski, F. J. 1997. First-principles theory of magneto-crystalline anisotropy of disordered alloys: Application to cobalt-platinum. *Phys. Rev. B* 56: 8082.

Razee, S. S. A., Staunton, J. B., Pinski, F. J., Ginatempo, B., and Bruno, E. 1998. Magnetic anisotropies in NiPt and CoPt alloys. *J. Appl. Phys.* In press.

Sakuma, A. 1994. First principle calculation of the magnetocrystalline anisotropy energy of FePt and CoPt ordered alloys. *J. Phys. Soc. Japan* 63:3053.

Samson, J. 1989. Magnetic correlations in paramagnetic iron. *J. Phys.: Condens. Matter* 1:6717–6729.

Sandratskii, L. M. 1998. Non-collinear magnetism in itinerant electron systems: Theory and applications. *Adv. Phys.* 47:91.

Sandratskii, L. M. and Kubler, J. 1993. Local magnetic moments in BCC Co. *Phys. Rev. B* 47:5854.

Severin, L., Gasche, T., Brooks, M. S. S., and Johansson, B. 1993. Calculated Curie temperatures for RCo_2 and RCo_2H_4 compounds. *Phys. Rev. B* 48:13547.

Shirane, G., Boni, P., and Wicksted, J. P. 1986. Paramagnetic scattering from Fe(3.5 at-percent Si): Neutron measurements up to the zone boundary. *Phys. Rev. B* 33:1881–1885.

Singh, D. J., Pickett, W. E., and Krakauer, H. 1991. Gradient-corrected density functionals: Full-potential calculations for iron. *Phys. Rev. B* 43:11628–11634.

Solovyev, I. V., Dederichs, P. H., and Mertig, I. 1995. Origin of orbital magnetization and magnetocrystalline anisotropy in TX ordered alloys (where T = Fe, Co and X = Pd, Pt). *Phys. Rev. B* 52:13419.

Soven, P. 1967. Coherent-potential model of substitutional disordered alloys. *Phys. Rev.* 156:809–813.

Staunton, J. B., and Györffy, B. L. 1992. Onsager cavity fields in itinerant-electron paramagnets. *Phys. Rev. Lett.* 69:371–374.

Staunton, J. B., Györffy, B. L., Pindor, A. J., Stocks, G. M., and Winter, H. 1985. Electronic-structure of metallic ferromagnets above the Curie-temperature. *J. Phys. F.: Metal Physics* 15:1387–1404.

Staunton, J. B., Johnson, D. D., and Györffy, B. L. 1987. Interaction between magnetic and compositional order in Ni-rich Ni_cFe_{1-c} alloys. *J. Appl. Phys.* 61:3693–3696.

Staunton, J. B., Johnson, D. D., and Pinski, F. J. 1990. Theory of compositional and magnetic correlations in alloys: Interpretation of a diffuse neutron-scattering experiment on an iron-vanadium single-crystal. *Phys. Rev. Lett.* 65:1259–1262.

Staunton, J. B., Matsumoto, M., and Strange, P. 1992. Spin-polarized relativistic KKR *In* Applications of Multiple Scattering Theory to Materials Science, Mat. Res. Soc. 253, (W. H. Butler, P. H. Dederichs, A. Gonis, and R. L. Weaver, eds.). pp. 309. Materials Research Society, Pittsburgh.

Staunton, J. B., Johnson, D. D., and Pinski, F. J. 1994. Compositional short-range ordering in metallic alloys: band-filling, charge-transfer, and size effects from a first-principles, all-electron, Landau-type theory. *Phys. Rev. B* 50:1450.

Staunton, J. B., Ling, M. F., and Johnson, D. D. 1997. A theoretical treatment of atomic short-range order and magnetism in iron-rich b.c.c. alloys. *J. Phys. Condensed Matter* 9:1281–1300.

Stephens, J. R. 1985. The B2 aluminides as alternate materials. *In* High Temperature Ordered Intermetallic Alloys, vol. 39. (C. C. Koch, C. T. Liu, and N. S. Stolhoff, eds.). pp. 381. Materials Research Society, Pittsburgh.

Stocks, G. M. and Winter, H. 1982. Self-consistent-field-Korringa-Kohn-Rostoker-Coherent-Potential approximation for random alloys. *Z. Phys. B* 46:95–98.

Stocks, G. M., Temmerman, W. M., and Györffy, B. L. 1978. Complete solution of the Korringa-Kohn-Rostoker coherent-potential-approximation equations: Cu-Ni alloys. *Phys. Rev. Lett.* 41:339.

Stoner, E. C. 1939. Collective electron ferromagnetism II. Energy and specific heat. *Proc. Roy. Soc. A* 169:339.

Strange, P., Ebert, H., Staunton, J. B., and Györffy, B. L. 1989a. A relativistic spin-polarized multiple-scattering theory, with applications to the calculation of the electronic-structure of condensed matter. *J. Phys. Condens. Matter* 1:2959.

Strange, P., Ebert, H., Staunton, J. B., and Györffy, B. L. 1989b. A first-principles theory of magnetocrystalline anisotropy in metals. *J. Phys. Condens. Matter* 1:3947.

Strange, P., Staunton, J. B., Györffy, B. L., and Ebert, H. 1991. First principles theory of magnetocrystalline anisotropy *Physica B* 172:51.

Suzuki, T., Weller, D., Chang, C. A., Savoy, R., Huang, T. C., Gurney, B., and Speriosu, V. 1994. Magnetic and magneto-optic properties of thick face-centered-cubic Co single-crystal films. *Appl. Phys. Lett.* 64:2736.

Svane, A. 1994. Electronic structure of cerium in the self-interaction corrected local spin density approximation. *Phys. Rev. Lett.* 72:1248–1251.

Svane, A. and Gunnarsson, O. 1990. Transition metal oxides in the self-interaction corrected density functional formalism. *Phys. Rev. Lett.* 65:1148.

Swihart, J. C., Butler, W. H., Stocks, G. M., Nicholson, D. M., and Ward, R. C. 1986. First principles calculation of residual electrical resistivity of random alloys. *Phys. Rev. Lett.* 57:1181.

Szotek, Z., Temmerman, W. M., and Winter, H. 1993. Application of the self-interaction correction to transition metal oxides. *Phys. Rev. B* 47:4029.

Szotek, Z., Temmerman, W. M., and Winter, H. 1994. Self-interaction corrected, local spin density description of the $\gamma-\alpha$ transition in Ce. *Phys. Rev. Lett.* 72:1244–1247.

Temmerman, W. M., Szotek, Z., and Winter, H. 1993. Self-interaction-corrected electronic structure of La_2CuO_4. *Phys. Rev. B* 47:11533–11536.

Treglia, G., Ducastelle, F., and Gautier, F. 1978. Generalized perturbation theory in disordered transition metal alloys: Application to the self-consistent calculation of ordering energies. *J. Phys. F: Met. Phys.* 8:1437–1456.

Trygg, J., Johansson, B., Eriksson, O., and Wills, J. M. 1995. Total energy calculation of the magnetocrystalline anisotropy energy in the ferromagnetic 3d metals. *Phys. Rev. Lett.* 75: 2871.

Tyson, T. A., Conradson, S. D., Farrow, R. F. C., and Jones, B. A. 1996. Observation of internal interfaces in Pt_xCo_{1-x} ($x = 0.7$) alloy films: A likely cause of perpendicular magnetic anisotropy. *Phys. Rev. B* 54:R3702.

Uhl, M. and Kubler, J. 1996. Exchange-coupled spin-fluctuation theory: Application to Fe, Co and Ni. *Phys. Rev. Lett.* 77:334.

Uhl, M. and Kubler, J. 1997. Exchange-coupled spin-fluctuation theory: Calculation of magnetoelastic properties. *J. Phys.: Condensed Matter* 9:7885.

Uhl, M., Sandratskii, L. M., and Kubler, J. 1992. Electronic and magnetic states of γ-Fe. *J. Mag. Magn. Mat.* 103:314–324.

van Tendeloo, G., Amelinckx, S., and de Fontaine, D. 1985. On the nature of the short-range order in {1,1/2,0} alloys. *Acta. Crys. B* 41:281.

Victora, R.H. and MacLaren, J. M. 1993. Theory of magnetic interface anisotropy. *Phys. Rev. B* 47:11583.

von Barth, U., and Hedin, L. 1972. A local exchange-correlation potential for the spin polarized case: I. *J. Phys. C: Solid State Physics* 5:1629–1642.

von der Linden, Donath, M., and Dose, V. 1993. Unbiased access to exchange splitting of magnetic bands using the maximum entropy method. *Phys. Rev. Lett.* 71:899–902.

Vosko, S. H., Wilk, L., and Nusair, M. 1980. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.* 58: 1200–1211.

Wang, Y. and Perdew, J. P. 1991. Correlation hole of the spin-polarised electron gas with exact small wave-vector and high density scaling. *Phys. Rev. B* 44:13298.

Wang, C. S., Prange, R. E., Korenman, V. 1982. Magnetism in iron and nickel. *Phys. Rev. B* 25:5766–5777.

Wassermann, E. F. 1991. The INVAR problem. *J. Mag. Magn. Mat.* 100:346–362.

Weinert, M., Watson, R. E., and Davenport, J. W. 1985. Total-energy differences and eigenvalue sums. *Phys. Rev. B* 32:2115.

Weller, D., Brandle, H., Lin, C. J., and Notary, H. 1992. Magnetic and magneto-optical properties of cobalt-platinum alloys with perpendicular magnetic anisotropy. *Appl. Phys. Lett.* 61:2726.

Weller, D., Brandle, H., and Chappert, C. 1993. Relationship between Kerr effect and perpendicular magnetic anisotropy in $Co_{1-x}Pt_x$ and $Co_{1-x}Pd_x$ alloys. *J. Magn. Magn. Mater.* 121:461.

Williams, A. R., Malozemoff, A. P., Moruzzi, V. L., and Matsui, M. 1984. Transition between fundamental magnetic behaviors revealed by generalized Slater-Pauling construction. *J. Appl. Phys.* 55:2353–2355.

Wohlfarth, E. P. 1953. The theoretical and experimental status of the collective electron theory of ferromagnetism. *Rev. Mod. Phys.* 25:211.

Wu, R. and Freeman, A. J. 1996. First principles determinations of magnetostriction in transition metals. *J. Appl. Phys.* 79:6209.

Ziebeck, K. R. A., Brown, P. J., Deportes, J., Givord, D., Webster, P. J., and Booth, J. G. 1983. Magnetic correlations in metallic magnetics at finite temperatures. *Helv. Phys. Acta.* 56:117–130.

Ziman, J. M. 1964. The method of neutral pseudo-atoms in the theory of metals. *Adv. Phys.* 13:89.

Zunger, A. 1994. First-principles statistical mechanics of semiconductor alloys and intermetallic compounds. *In* Statics and Dynamics of Alloy Phase Transformations, NATO-ASI Series (A. Gonis and P. E. A.Turchi, eds.). pp. 361–420. Plenum Press, New York.

F. J. PINSKI
University of Cincinnati
Cincinnati, Ohio

J. B. STAUNTON
S. S. A. RAZEE
University of Warwick
Coventry, U.K.

D. D. JOHNSON
University of Illinois
Urbana-Champaign, Illinois

KINEMATIC DIFFRACTION OF X RAYS

INTRODUCTION

Diffraction by x rays, electrons, or neutrons has enjoyed great success in crystal structure determination (e.g., the structures of DNA, high-T_c superconductors, and reconstructed silicon surfaces). For a perfectly ordered crystal, diffraction results in arrays of sharp Bragg reflection spots periodically arranged in reciprocal space. Analysis of the Bragg peak locations and their intensities leads to the identification of crystal lattice type, symmetry group, unit cell dimensions, and atomic configuration within a unit cell. On the other hand, for crystals containing lattice defects such as dislocations, precipitates, local ordered domains, surface, and interfaces, diffuse intensities are produced in addition to Bragg peaks. The distribution and magnitude of diffuse intensities are dependent on the type of imperfection present and the x-ray energy used in a diffraction experiment.

Diffuse scattering is usually weak, and thus more difficult to measure, but it is rich in structure information that often cannot be obtained by other experimental means. Since real crystals are generally far from perfect, many properties exhibited by them are therefore determined by the lattice imperfections present. Consequently, understanding of the atomic structures of these lattice imperfections (e.g., atomic short-range order, extended vacancy defect complexes, phonon properties, composition fluctuation, charge density waves, static displacements, and superlattices) and of the roles these imperfections play (e.g., precipitation hardening, residual stresses, phonon softening, and phase transformations) is of paramount importance if these materials properties are to be exploited for optimal use.

This unit addresses the fundamental principles of diffraction based upon the kinematic diffraction theory for x rays. (Nevertheless, the diffraction principles described in this unit may be extended to kinematic diffraction events involving thermal neutrons or electrons.) The accompanying DYNAMICAL DIFFRACTION is concerned with dynamic diffraction theory, which applies to diffraction from single crystals of high quality so that multiple scattering becomes significant and kinematic diffraction theory becomes invalid. In practice, most x-ray diffraction experiments are carried out on crystals containing a sufficiently large number of defects that kinematic theory is generally applicable.

This unit is divided into two major sections. In the first section, the fundamental principles of kinematic diffraction of x rays will be discussed and a systematic treatment of theory will be given. In the second section, the practical aspects of the method will be discussed; specific expressions for kinematically diffracted x-ray intensities will be described and used to interpret diffraction behavior from real crystals containing lattice defects. Neither specific diffraction techniques and analysis nor sample preparation methods will be described in this unit. Readers may refer to X-RAY TECHNIQUES for experimental details and specific applications.

PRINCIPLES OF THE METHOD

Overview of Scattering Processes

When a stream of radiation (e.g., photons or neutrons) strikes matter, various interactions can take place, one of which is the scattering process that may be best described using the wave properties of radiation. Depending on the energy, or wavelength, of the incident radiation, scattering may occur on different levels—at the atomic, molecular, or microscopic scale. While some scattering events are noticeable in our daily routines (e.g., scattering of visible light off the earth's atmosphere to give a blue sky and scattering from tiny air bubbles or particles in a glass of water to give it a hazy appearance), others are more difficult to observe directly with human eyes, especially for those scattering events that involve x rays or neutrons.

X rays are electromagnetic waves or photons that travel at the speed of light. They are no different from visible light, but have wavelengths ranging from a few hundredths of an angstrom (Å) to a few hundred angstroms. The conversion from wavelength to energy for all photons is given in the following equation with wavelength λ in angstroms and energy in kilo-electron volts (keV):

$$\lambda(\text{Å}) = \frac{c}{\nu} = \frac{12.40(\text{Å}\cdot\text{keV})}{E(\text{keV})} \qquad (1)$$

in which c is the speed of light (3×10^{18} Å/s) and ν is the frequency. It is customary to classify x rays with a wavelength longer than a few angstroms as "soft x rays" as opposed to "hard x rays" with shorter wavelengths ($\lesssim$1 Å) and higher energies ($\gtrsim$ keV).

In what follows, a general scattering theory will be presented. We shall concentrate on the kinematic scattering theory, which involves the following assumptions:

1. The traveling wave model is utilized so that the x-ray beam may be represented by a plane wave formula.
2. The source-to-specimen and the specimen-to-detector distances are considered to be far greater than the distances separating various scattering centers. Therefore, both the incident and the scattering beam can be represented by a set of parallel rays with no divergence.
3. Interference between x-ray beams scattered by elements at different positions is a result of superposition of those scattered traveling waves with different paths.
4. No multiple scattering is allowed: that is, the once-scattered beam inside a material will not rescatter. (This assumption is most important since it separates kinematic scattering theory from dynamic scattering theory.)
5. Only the elastically scattered beam is considered; conservation of x-ray energy applies.

The above assumptions form the basis of the kinematic scattering/diffraction theory; they are generally valid

assumptions in the most widely used methods for studying scattering and diffraction from materials. In some cases, such as diffraction from perfect or nearly perfect single crystals, dynamic scattering theory must be employed to explain the nature of the diffraction events (DYNAMICAL DIFFRACTION). In other cases, such as Compton scattering, where energy exchanges occur in addition to momentum transfers, inelastic scattering theories must be invoked.

While the word "scattering" refers to a deflection of beam from its original direction by the scattering centers that could be electrons, atoms, molecules, voids, precipitates, composition fluctuations, dislocations, and so on, the word "diffraction" is generally defined as the constructive interference of coherently scattered radiation from regularly arranged scattering centers such as gratings, crystals, superlattices, and so on. Diffraction generally results in strong intensity in specific, fixed directions in reciprocal (momentum) space, which depend on the translational symmetry of the diffracting system. Scattering, however, often generates weak and diffuse intensities that are widely distributed in reciprocal space. A simple picture may be drawn to clarify this point. For instance, interaction of radiation with an amorphous substance is a "scattering" process that reveals broad and diffuse intensity maxima, whereas with a crystal it is a "diffraction" event, as sharp and distinct peaks appear. Sometimes the two words are interchangeable, as the two events may occur concurrently or indistinguishably.

Elementary Kinematic Scattering Theory

In Figure 1, an incoming plane wave P_1, traveling in the direction specified by the unit vector $\mathbf{s}_0$, interacts with the specimen, and the scattered beam, another plane wave P_2, travels along the direction $\mathbf{s}$, again a unit vector. Taking into consideration increments of volume within the specimen, ΔV_1, waves scattered from different increments of volume will interfere with each other: that is, their instantaneous amplitudes will be additive (a list of symbols used is contained in the Appendix). Since the variation of amplitude with time will be sinusoidal, it is necessary to keep track of the phase of the wave scattered from individual volume elements. Therefore the scattered wave along $\mathbf{s}$ is made up of components scattered from the individual volume elements, the path differences traveled by each individual ray traveling from P_1 to P_2. In reference to an arbitrary point O in the specimen, the path difference between a ray scattered from the volume element ΔV_1 and that from O is

$$\Delta r_1 = \mathbf{r}_1 \cdot \mathbf{s} - \mathbf{r}_1 \cdot \mathbf{s}_0 = \mathbf{r}_1 \cdot (\mathbf{s} - \mathbf{s}_0) \tag{2}$$

Thus, the difference in phase between waves scattered from the two points will be proportional to the difference in distances that the two waves travel from P_1 to P_2—a path difference equal to the wavelength λ, corresponding to a phase difference ϕ of 2π radians:

$$\frac{\phi}{2\pi} = \frac{\Delta r_1}{\lambda} \tag{3}$$

In general, the phase of the wave scattered from the jth increment of volume ΔV_j, relative to the phase of the wave scattered from the origin O, will thus be

$$\phi_j = \frac{2\pi(\mathbf{s} - \mathbf{s}_0) \cdot \mathbf{r}_j}{\lambda} \tag{4}$$

Equation 4 may be expressed by $\phi_j = 2\pi \mathbf{K} \cdot \mathbf{r}_j$, where $\mathbf{K}$ is the scattering vector (Fig. 2),

$$\mathbf{K} = \frac{\mathbf{s} - \mathbf{s}_0}{\lambda} \tag{5}$$

The phase of the scattered radiation is then expressed by the plane wave $e_j^{i\phi}$, and the resultant amplitude is obtained by summing over the complex amplitudes scattered from each incremental scattering center:

$$A = \sum_j f_j e^{2\pi i \mathbf{K} \cdot \mathbf{r}_j} \tag{6}$$

where f_j is the scattering power, or scattering length, of the jth volume element (this scattering power will be further discussed a little later). For a continuous medium viewed on a larger scale, as is the case in small-angle scattering,

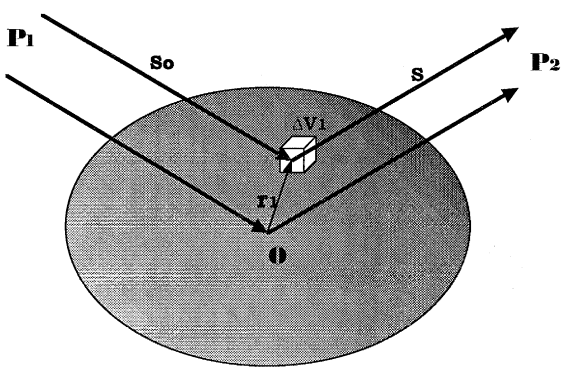


Figure 1. Schematics showing a diffracting element ΔV_1 at a distance $\mathbf{r}_1$ from an arbitrarily chosen origin O in the crystal. The incident and the diffraction beam directions are indicated by the unit vectors, $\mathbf{s}_0$ and $\mathbf{s}$, respectively.

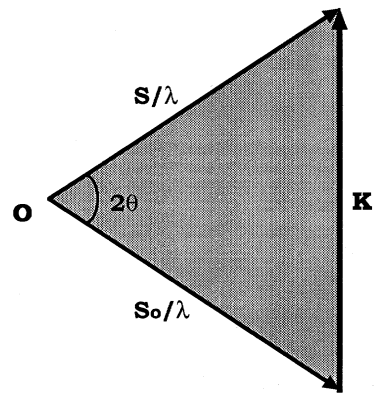


Figure 2. The diffraction condition is determined by the incident and the scattering beam direction unit vectors normalized against the specified wavelength (λ). The diffraction vector $\mathbf{K}$ is defined as the difference of the two vectors $\mathbf{s}/\lambda$ and $\mathbf{s}_0/\lambda$. The diffraction angle, 2θ, is defined by these two vectors as well.

the summation sign in Equation 6 may be replaced by an integral over the entire volume of the irradiated specimen.

The scattered intensity, $I(\mathbf{K})$, written in absolute units, which are commonly known as electron units, is proportional to the square of the amplitude in Equation 6:

$$I(\mathbf{K}) = AA^* = \left| \sum_j f_j e^{2\pi i \mathbf{K}\cdot\mathbf{r}_j} \right|^2 \tag{7}$$

Diffraction from a Crystal

For a crystalline material devoid of defects, the atomic arrangement may be represented by a primitive unit cell with lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ that display a particular set of translational symmetries. Since every unit cell is identical, the above summation over the diffracting volume within a crystal can be replaced by the summation over a single unit cell followed by a summation over the unit cells contained in the diffraction volume:

$$I(\mathbf{K}) = \left| \sum_j^{\text{u.c.}} f_j e^{2\pi i \mathbf{K}\cdot\mathbf{r}_j} \right|^2 \left| \sum_n e^{2\pi i \mathbf{K}\cdot\mathbf{r}_n} \right|^2 = |F(\mathbf{K})|^2 |G(\mathbf{K})|^2 \tag{8}$$

The first term, known as the structure factor, $F(\mathbf{K})$, is a summation of all scattering centers within one unit cell (u.c.). The second term defines the interference function, $G(\mathbf{K})$, which is a Fourier transformation of the real-space point lattice. The vector $\mathbf{r}_n$ connects the origin to the nth lattice point and is written as:

$$\mathbf{r}_n = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \tag{9}$$

where n_1, n_2, and n_3 are integers. Consequently, the single summation for the interference function may be replaced by a triple summation over n_1, n_2, and n_3:

$$|G(\mathbf{K})|^2 = \left| \sum_{n_1}^{N_1} e^{2\pi i \mathbf{K}\cdot n_1\mathbf{a}_1} \right|^2 \left| \sum_{n_2}^{N_2} e^{2\pi i \mathbf{K}\cdot n_2\mathbf{a}_2} \right|^2 \left| \sum_{n_3}^{N_3} e^{2\pi i \mathbf{K}\cdot n_3\mathbf{a}_3} \right|^2 \tag{10}$$

where N_1, N_2, and N_3 are numbers of unit cells along the three lattice vector directions, respectively. For large N_i, Equation 10 reduces to

$$|G(\mathbf{K})|^2 = \frac{\sin^2(\pi\mathbf{K}\cdot N_1\mathbf{a}_1)}{\sin^2(\pi\mathbf{K}\cdot\mathbf{a}_1)} \frac{\sin^2(\pi\mathbf{K}\cdot N_2\mathbf{a}_2)}{\sin^2(\pi\mathbf{K}\cdot\mathbf{a}_2)} \frac{\sin^2(\pi\mathbf{K}\cdot N_3\mathbf{a}_3)}{\sin^2(\pi\mathbf{K}\cdot\mathbf{a}_3)} \tag{11}$$

A general display of the above interference function is shown in Figure 3. First, the function is a periodic one. Maxima occur at specific $\mathbf{K}$ locations followed by a series of secondary maxima with much reduced amplitudes. It is noted that the larger the N_i the sharper the peak, because the width of the peak is inversely proportional to N_i while the peak height equals N_i^2. When N_i approaches

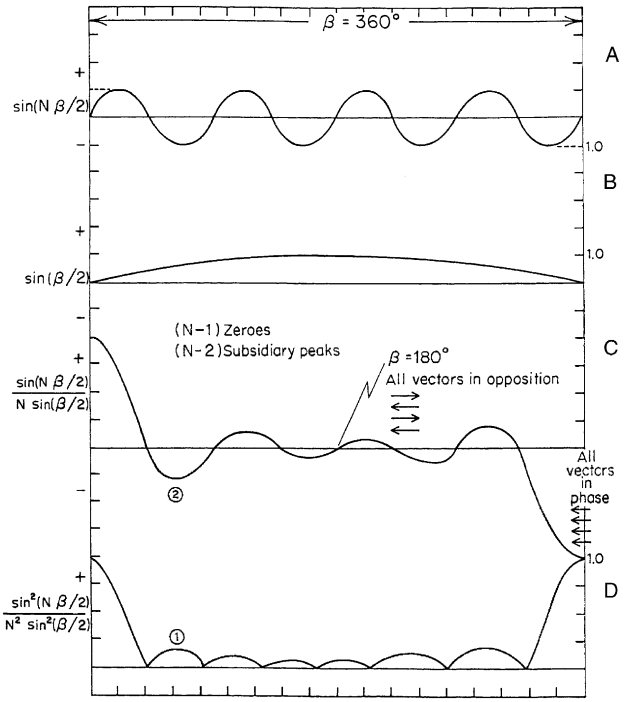


Figure 3. Schematic drawing of the interference function for $N = 8$ showing periodicity with angle β. The amplitude of the function equals N^2 while the width of the peak is proportional to $1/N$, where N represents the number of unit cells contributing to diffraction. There are $N - 1$ zeroes in (**D**) and $N - 2$ subsidary maxima besides the two large ones at $\beta = 0°$ and 360°. Curves (**C**) and (**D**) have been normalized to unity. After Buerger (1960).

infinity, the interference function is a delta function with a value N_i. Therefore, when $N_{i\to\infty}$, Equation 11, becomes

$$|G(\mathbf{K})|^2 = N_1N_2N_3 = N_v \tag{12}$$

where N_v is the total number of unit cells in the diffracting volume.

For diffraction to occur from such a three-dimensional (3D) crystal, the following three conditions must be satisfied simultaneously to give constructive interference, that is, to have significant values for $G(\mathbf{K})$

$$\mathbf{K}\cdot\mathbf{a}_1 = h, \quad \mathbf{K}\cdot\mathbf{a}_2 = k, \quad \mathbf{K}\cdot\mathbf{a}_3 = l \tag{13}$$

where h, k, and l are integers. These are three conditions known as Laue conditions. Obviously, for diffraction from a lower-dimensional crystal, one or two of the conditions are removed. The Laue conditions, Equation 13, indicate that scattering is described by sets of planes spaced h/a_1, k/a_2, and l/a_3 apart and perpendicular to $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, respectively. Therefore, diffraction from a one-dimensional (1D) crystal with a periodicity a would result in sheets of intensities perpendicular to the crystal direction and separated by a distance $1/a$. For a two-dimensional (2D) crystal, the diffracted intensities would be distributed along rods normal to the crystal plane. In three dimensions

(3D), the Laue conditions define arrays of points that form the reciprocal lattice.

The reciprocal lattice may be defined by means of three reciprocal space lattice vectors that are, in turn, defined from the real-space primitive unit cell vectors as in Equation 14:

$$\mathbf{b}_i = \frac{\mathbf{a}_j \times \mathbf{a}_k}{V_a} \tag{14}$$

where i, j, and k are permutations of three integers, 1, 2, and 3, and V_a is the volume of the primitive unit cell constructed by $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. There exists an orthonormal relationship between the real-space and the reciprocal space lattice vectors, as in Equation 15:

$$\mathbf{a}_i \cdot \mathbf{b}_j = \delta_{ij} \tag{15}$$

where δ_{ij} is the Kronecker delta function, which is defined as

$$\begin{aligned} \delta_{ij} &= 1 \qquad \text{for} \quad i = j \\ &= 0 \qquad \text{for} \quad i \neq j \end{aligned} \tag{16}$$

A reciprocal space vector $\mathbf{H}$ can thus be expressed as a summation of reciprocal space lattice vectors:

$$\mathbf{H} = h\,\mathbf{b}_1 + k\,\mathbf{b}_2 + l\,\mathbf{b}_3 \tag{17}$$

where h, k, and l are integers. The magnitude of this vector, $\mathbf{H}$, can be shown to be equal to the inverse of the interplanar spacing, d_{hkl}. It can also be shown that the vector $\mathbf{H}$ satisfies the three Laue conditions (Equation 13). Consequently, the interference function in Equation 11 would have significant values when the following condition is satisfied

$$\mathbf{K} = \mathbf{H} \tag{18}$$

This is the vector form of Bragg's law.

As shown in Equation 18 and Figure 2, when the scattering vector $\mathbf{K}$, as defined according to the incident and the diffracted beam directions and the associated wavelength, matches one of the reciprocal space lattice vectors $\mathbf{H}$, the interference function will have significant value, thereby showing constructive interference—Bragg diffraction. It can be shown by taking the magnitudes of the two vectors $\mathbf{H}$ and $\mathbf{K}$ that the familiar scalar form of the Bragg's law is recovered:

$$2d_{hkl} \sin\theta = n\lambda \qquad n = 1, 2, \ldots \tag{19}$$

By combining Equations 12 and 8, we now conclude that when Bragg's law is met (i.e., when $\mathbf{K} = \mathbf{H}$), the diffracted intensity becomes

$$I(\mathbf{K}) = N_v |F(\mathbf{K})|^2 \tag{20}$$

Structure Factor

The structure factor, designated by the symbol F, is obtained by adding together all the waves scattered from one unit cell; it therefore displays the interference effect among all scattering centers within one unit cell. Certain extinction conditions may appear for a combination of h, k, and l values as a result of the geometrical arrangement of atoms or molecules within the unit cell. If a unit cell contains N atoms, with fractional coordinates x_i, y_i, and z_i for the ith atom in the unit cell, then the structure factor for the hkl reflection is given by

$$F(hkl) = \sum_i^N f_i e^{2\pi i(hx_i + ky_i + lz_i)} \tag{21}$$

where the summation extends over all the N atoms of the unit cell. The parameter F is generally a complex number and expresses both the amplitude and phase of the resultant wave. Its absolute value gives the magnitude of diffracting power as given in Equation 20.

Some examples of structure-factor calculations are given as follows:

1. For all primitive cells with one atom per lattice point, the coordinates for this atom are 0 0 0. The structure factor is

$$F = f \tag{22}$$

2. For a body-centered cell with two atoms of the same kind, their coordinates are 0 0 0 and $\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}$, and the structure factor is

$$F = f\left(1 + e^{\pi i(h+k+l)}\right) \tag{23}$$

This expression may be evaluated for any combination of h, k, and l integers. Therefore,

$$\begin{aligned} F &= 2f \qquad \text{when } (h+k+l) \text{ is even} \\ F &= 0 \qquad \text{when } (h+k+l) \text{ is odd} \end{aligned} \tag{24}$$

3. Consider a face-centered cubic (fcc) structure with identical atoms at $x, y, z = 0\ 0\ 0$, $\frac{1}{2}\ \frac{1}{2}\ 0$, $\frac{1}{2}\ 0\ \frac{1}{2}$, and $0\ \frac{1}{2}\ \frac{1}{2}$. The structure factor is

$$\begin{aligned} F &= f\left(1 + e^{\pi i(h+k)} + e^{\pi i(k+l)} + e^{\pi i(l+h)}\right) \\ &= 4f \qquad \text{for } h, k, l \text{ all even or all odd} \\ &= 0 \qquad \text{for mixed } h, k, \text{ and } l \end{aligned} \tag{25}$$

4. Zinc blend (ZnS) has a common structure that is found in many Group III-V compounds such as GaAs and InSb and there are four Zn and four S atoms per fcc unit cell with the coordinates shown below:

$$\begin{aligned} \text{Zn:} \quad & 0\ 0\ 0, \ \ \frac{1}{2}\,\frac{1}{2}\,0, \ \ \frac{1}{2}\ 0\,\frac{1}{2}, \ \ \text{and} \ \ 0\,\frac{1}{2}\,\frac{1}{2} \\ \text{S:} \quad & \frac{1}{4}\,\frac{1}{4}\,\frac{1}{4}, \ \ \frac{3}{4}\,\frac{3}{4}\,\frac{1}{4}, \ \ \frac{3}{4}\,\frac{1}{4}\,\frac{3}{4}, \ \ \text{and} \ \ \frac{1}{4}\,\frac{3}{4}\,\frac{3}{4} \end{aligned}$$

The structure factor may be reduced to

$$F = \left(f_{\text{Zn}} + f_{\text{S}} e^{\frac{\pi i}{2}(h+k+l)}\right)\left(1 + e^{\pi i(h+k)} + e^{\pi i(k+l)} + e^{\pi i(l+h)}\right) \tag{26}$$

The second term is equivalent to the fcc conditions as in Equation 25, so h, k, and l must be unmixed integers. The first term further modifies the structure factor to yield

$$\begin{aligned} F &= f_{Zn} + f_S \quad \text{when} \quad h+k+l = 4n \text{ and } n \text{ is integer} \\ &= f_{Zn} + if_S \quad \text{when} \quad h+k+l = 4n+1 \\ &= f_{Zn} - f_S \quad \text{when} \quad h+k+l = 4n+2 \\ &= f_{Zn} - if_S \quad \text{when} \quad h+k+l = 4n+3 \end{aligned} \tag{27}$$

Scattering Power and Scattering Length

X rays are electromagnetic waves; they interact readily with electrons in an atom. In contrast, neutrons scatter most strongly from nuclei. This difference in contrast origin results in different scattering powers between x rays and neutrons even from the same species (see Chapter 13).

For a stream of unpolarized, or randomly polarized, x rays scattered from one electron, the scattered intensity, I_e, is known as the Thomson scattering per electron:

$$I_e = \frac{I_0 e^4}{m^2 r^2 c^4}\left(\frac{1+\cos^2 2\theta}{2}\right) \tag{28}$$

where I_e is the incident beam flux, e is the electron charge, m is the electron mass, c is the speed of light, r is the distance from the scattering center to the detector position, and 2θ is the scattering angle (Fig. 2). The factor $(1 + \cos^2 2\theta)/2$ is often referred to as the polarization factor. If the beam is fully or partially polarized, the total polarization factor will naturally be different. For instance, for synchrotron storage rings, x rays are linearly polarized in the plane of the ring. Therefore, if the diffraction plane containing vectors $\mathbf{s}_0$ and $\mathbf{s}$ in Figure 2 is normal to the storage ring plane, the polarization is unchanged during scattering.

Scattering of x rays from atoms is predominantly from the electrons in the atom. Because electrons in an atom do not assume a fixed position but rather are described by a wave function that satisfies the Schrodinger equation in quantum mechanics, the scattering power for x rays from an atom may be expressed by an integration of all waves scattered from these electrons as represented by an electron density function, $\rho(\mathbf{r})$,

$$f(\mathbf{K}) = \int_{\text{atom}} \rho(\mathbf{r}) e^{2\pi i \mathbf{K}\cdot\mathbf{r}} dV_{\mathbf{r}} \tag{29}$$

where $dV_{\mathbf{r}}$ is the volume increment and the integration is taken over the entire volume of the atom. The quantity f in Equation 29 is the scattering amplitude of an atom relative to that for a single electron. It is commonly known as the atomic scattering factor for x rays. The magnitude of f for different atomic species can be found in many text and reference books.

There are dispersion corrections to be made to f. These include a real and an imaginary component: the real part is related to the bonding nature of the negatively charged electrons with the positively charged nucleus, whereas the imaginary part concerns the absorption effect. Thus the true atomic scattering factor should be written

$$f = f_0 + \Delta f' + i\Delta f'' \tag{30}$$

Tabulated values for these correction terms, often referred to as the Honl corrections, can be found in the International Table for X-ray Crystallography (1996) or other references.

In conclusion, the intensity expressions shown in Equations 7, 8, and 20 are written in electron units, an absolute unit independent of incident beam flux and polarization factor. These intensity expressions represent the fundamental forms of kinematic diffraction. Applications of these fundamental diffraction principles to several specific examples of scattering and diffraction will be discussed in the following section.

PRACTICAL ASPECTS OF THE METHOD

Lattice defects may be classified as follows: (1) intrinsic defects, such as phonons and magnetic spins; (2) point defects, such as vacancies, substitutional, and interstitial solutes; (3) linear defects, such as dislocations, 1D superlattices, and charge density waves; (4) planar defects, such as twins, grain boundaries, surfaces, and interfaces; and (5) volume defects, such as voids, inclusions, precipitate particles, and magnetic clusters. In this section, kinematically scattered x-ray diffuse intensity expressions will be presented to correlate to lattice defects. Specific examples include: (1) thermal diffuse scattering from phonons, (2) short-range ordering or clustering in binary alloys, (3) surface/interface diffraction for reconstruction and interface structure, and (4) small-angle x-ray scattering from nanometer-sized particles dispersed in an otherwise uniform matrix. Not included in the discussion is the most fundamental use of the Bragg peak intensities for the determination of crystal structure from single crystals and for the analysis of lattice parameter, particle size distribution, preferred orientation, residual stress, and so on, from powder specimens. Discussion of these topics may be found in X-RAY POWDER DIFFRACTION and in many excellent books [e.g., Azaroff and Buerger (1958), Buerger (1960), Cullity (1978), Guinier (1994), Klug and Alexander (1974), Krivoglaz (1969), Noyan and Cohen (1987), Schultz (1982), Schwartz and Cohen (1987), and Warren (1969)].

Thermal Diffuse Scattering (TDS)

At any finite temperature, atoms making up a crystal do not stay stationary but rather vibrate in an cooperative manner; this vibrational amplitude usually becomes bigger at higher temperatures. Because of the periodic nature of crystals and the interconnectivity of an atomic network coupled by force constants, the vibration of an atom at a given position is related to the vibrations of others via atomic displacement waves (known as phonons) traveling through a crystal. The displacement of each atom is the sum total of the effects of these waves. Atomic vibration is considered one “imperfection” or “defect” that is intrinsic

to the crystal and is present at all times. The scattering process for phonons is basically inelastic, and involves energy transfer as well as momentum transfer. However, for x rays the energy exchange in such an inelastic scattering process is only a few hundredths of an electron volt, much too small compared to the energy of the x-ray photon used (typically in the neighborhood of thousands of electron volt) to allow them to be conveniently separated from the elastically scattered x rays in a normal diffraction experiment. As a result, thermal diffuse x-ray scattering may be treated in either a quasielastic or elastic manner. Such is not the case with thermal neutron scattering since energy resolution in this case is sufficient to separate the inelastic scattering due to phonons from other elastic parts. In this section, we shall discuss thermal diffuse x-ray scattering only.

The development of the scattering theory of the effect of thermal vibration on the x-ray diffraction in crystals is associated primarily with the Debye (1913a,b,c, 1913–1914), Waller (1923), Faxen (1918, 1923), and James (1948). The whole subject was brought together for the first time in a book by James (1948). Warren (1969), who adopted the approach of James, has written a comprehensive chapter on this subject on which this section is based. What follows is a short summary of the formulations used in the thermal diffuse x-ray scattering analysis. Examples of TDS applications may be found in Warren (1969) and in papers by Dvorack and Chen (1983) and by Takesue et al. (1997).

The most familiar effect of temperature vibration is the reduction of the Bragg reflections by the well-known Debye-Waller factor. This effect may be seen from the structure factor calculation:

$$F(\mathbf{K}) = \sum_{m}^{\text{u.c.}} f_m e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m} \tag{31}$$

where the upper limit, u.c., means summation over the unit cell.

Let $\mathbf{r}_m = \mathbf{r}_m^0 + \mathbf{u}_m(t)$, where $\mathbf{r}_m^0$ represents the average location of the mth atom and $\mathbf{u}_m$ is the dynamic displacement, a function of time t. Thus,

$$F(\mathbf{K}) = \sum_{m}^{\text{u.c.}} f_m e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m^0} e^{2\pi i \mathbf{K}\cdot\mathbf{u}_m} \tag{32}$$

As a first approximation, the second exponential term in Equation 32 may be expanded into a Taylor series up to the second-order terms:

$$e^{2\pi i \mathbf{K}\cdot\mathbf{u}_m} \approx 1 + 2\pi i \mathbf{K}\cdot\mathbf{u}_m - 2\pi^2(\mathbf{K}\cdot\mathbf{u}_m)^2 + \cdots \tag{33}$$

A time average may be performed for Equation 33, as a typical TDS experiment measuring interval is much longer than the phonon vibrational period, so that

$$\langle e^{2\pi i \mathbf{K}\cdot\mathbf{u}_m}\rangle \approx 1 - 2\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_m)^2\rangle + \cdots \tag{34}$$

In arriving at Equation 34, the linear average of the displacement field is set to zero, as is true for a random thermal vibration. Thus,

$$\langle e^{2\pi i \mathbf{K}\cdot\mathbf{u}_m}\rangle \approx e^{-2\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_m)^2\rangle} = e^{-M_m} \tag{35}$$

where $M_m = 2\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_m)^2\rangle$, known as the Debye-Waller temperature factor for mth atom.

Therefore, the total scattering intensity that is proportional to the square of the structure factor reduces to

$$I(\mathbf{K}) \propto \langle|F(\mathbf{K})|^2\rangle \approx \left|\sum_{m}^{\text{u.c.}} f_m e^{-M_m} e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m^0}\right|^2 \tag{36}$$

It now becomes obvious that thermal vibrations of atoms reduce the x-ray scattering intensities by the effect of the Debye-Waller temperature factor, exp(–M), in which M is proportional to the mean-squared displacement of a vibrating atom and is 2θ dependent. The effect of the Debye-Waller factor is to decrease the amplitude of a given Bragg reflection but to keep the diffraction profile unaltered. The above approximation assumed that each individual atom vibrates independently from others; this is naturally incorrect, as correlated vibrations of atoms by way of lattice waves (phonons) are present in crystals. This cooperative motion of atoms must be included in the TDS treatment. A more rigorous approach, in accord with the TDS treatment of Warren (1969), is now described for a cubic crystal with one atom per unit cell.

Starting with a general intensity equation expressed in terms of electron units and defining the time-dependent dynamic displacement vector $\mathbf{u}_m$, one obtains

$$\begin{aligned} I(\mathbf{K}) &= \left\langle \left| \sum_{m}^{\text{u.c.}} f_m e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m} \right|^2 \right\rangle \\ &= \left\langle \sum_{m}^{\text{u.c.}} f_m e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m} \sum_{n}^{\text{u.c.}} f_n e^{-2\pi i \mathbf{K}\cdot\mathbf{r}_n} \right\rangle \\ &= \left\langle \left| \sum_{m}^{\text{u.c.}} \sum_{n}^{\text{u.c.}} f_m f_n^* e^{2\pi i \mathbf{K}\cdot\mathbf{r}_m n} \right|^2 \right\rangle \\ &= \sum_{m}^{\text{u.c.}} \sum_{n}^{\text{u.c.}} |f_m f_n^*| e^{2\pi i \mathbf{K}\cdot\mathbf{r}_{mn}^0} \langle e^{2\pi i \mathbf{K}\cdot\mathbf{u}_{mn}}\rangle \end{aligned} \tag{37}$$

in which $\mathbf{r}_{mn} = \mathbf{r}_m - \mathbf{r}_n$, $\mathbf{r}_{mn}^0 = \mathbf{r}_m^0 - \mathbf{r}_n^0$, and $\mathbf{u}_{mn} = \mathbf{u} - \mathbf{u}_n$. Therefore, coupling between atoms is kept in the term $\mathbf{u}_{mn}$. Again, the approximation is applied with the assumption that a small vibrational amplitude is considered, so that a Taylor expansion may be used and the linear average set to zero:

$$\begin{aligned} \langle e^{2\pi i \mathbf{K}\cdot\mathbf{u}_{mn}}\rangle &\approx 1 - 2\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_{mn})^2\rangle \approx e^{-2\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_{mn})^2\rangle} \\ &= e^{\langle P_{mn}^2\rangle/2} \end{aligned} \tag{38}$$

where

$$\langle P_{mn}^2\rangle \equiv 4\pi^2\langle(\mathbf{K}\cdot\mathbf{u}_{mn})^2\rangle \tag{39}$$

The coupling between atomic vibrations may be expressed by traveling sinusoidal lattice waves, the concept of "phonons." Each lattice wave may be represented by a wave vector $\mathbf{g}$ and a frequency ω_{gj}, in which the j subscript denotes the jth component ($j = 1, 2, 3$) of the $\mathbf{g}$ lattice wave. Therefore, the total dynamic displacement of the nth atom is the sum of all lattice waves as seen in Equation 40

$$\mathbf{u}_n = \sum_{g,j} \mathbf{u}_n(g,j) \tag{40}$$

where

$$\mathbf{u}_n(g,j) = a_{gj}\mathbf{e}_{gj}\cos(\omega_{gj}t - 2\pi\mathbf{g}\cdot\mathbf{r}_n^0 - \delta_{gj}) \tag{41}$$

and a_{gj} is the vibrational amplitude; $\mathbf{e}_{gj}$ is the unit vector of the vibrating direction, that is, the polarization vector, for the gj wave; $\mathbf{g}$ is the propagation wave vector; δ_{gj} is an arbitrary phase factor; ω_{gj} is the frequency; and t is the time.

Thus, Equation 39 may be rewritten

$$\langle P_{mn}^2\rangle = 4\pi^2\left\langle\left\{\left[\sum_{gj}\mathbf{K}\cdot a_{gj}\mathbf{e}_{gj}\cos(\omega_{gj}t - 2\pi\mathbf{g}\cdot\mathbf{r}_m^0 - \delta_{gj})\right] - \left[\sum_{g'j'}\mathbf{K}\cdot a_{g'j'}\mathbf{e}_{g'j'}\cos\left(\omega_{g'j'}t - 2\pi\mathbf{g}'\cdot\mathbf{r}_n^0 - \delta_{g'j'}\right)\right]\right\}^2\right\rangle \tag{42}$$

After some mathematical manipulation, Equation 42 reduces to

$$\langle P_{mn}^2\rangle = \sum_{gj}\left\{(2\pi\mathbf{K}\cdot\mathbf{e}_{gj})^2\langle a_{gj}^2\rangle[1 - \cos(2\pi\mathbf{g}\cdot\mathbf{r}_{mn}^0)]\right\} \tag{43}$$

Defining $G_{gj} = \frac{1}{2}(2\pi\mathbf{K}\cdot\mathbf{e}_{gj})^2\langle a_{gj}^2\rangle$ and $\sum_{gj} G_{gj} = 2M$ causes the scattering equation for a single element system to reduce to

$$I(\mathbf{K}) = \sum_m^{\text{u.c.}}\sum_n^{\text{u.c.}} |fe^{-M}|^2 e^{2\pi i\mathbf{K}\bullet\mathbf{r}_{mn}^0}\cdot e^{\sum_{gj}G_{gj}\cos(2\pi\mathbf{g}\cdot\mathbf{r}_{mn}^0)} \tag{44}$$

where the first term in the product is equivalent to Equation 36, which represents scattering from the average lattice—that is, Bragg reflections—modified by the Debye-Waller temperature factor. The phonon coupling effect is contained in the second term of the product. The Debye-Waller factor $2M$ is the sum of G_{gj}, which is given by

$$2M \equiv \sum_{gj} G_{gj} = \sum_{gj}\frac{1}{2}(2\pi\mathbf{K}\cdot\mathbf{e}_{gj})^2\langle a_{gj}^2\rangle = \left(\frac{4\pi\sin\theta}{\lambda}\right)^2\left[\sum_{gj}\frac{1}{2}\langle a_{gj}^2\rangle\cos^2(\mathbf{K},\mathbf{e}_{gj})\right] \tag{45}$$

where the term in brackets is the mean-square displacement projected along the diffraction vector $\mathbf{K}$ direction.

Again, assuming small vibrational amplitude, the second term in the product of Equation 44 may be expanded into a series:

$$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots \tag{46}$$

Therefore, Equation 44 becomes

$$I(\mathbf{K}) \approx \sum_m\sum_n |fe^{-M}|^2 e^{2\pi i\mathbf{K}\cdot\mathbf{r}_{mn}^0}\cdot\left[1 + \sum_{gj}G_{gj}\cos(2\pi\mathbf{g}\cdot\mathbf{r}_{mn}^0) + \frac{1}{2}\sum_{gj}\sum_{g'j'}G_{gj}G_{g'j'}\cos(2\pi\mathbf{g}\cdot\mathbf{r}_{mn}^0)\times\cos(2\pi\mathbf{g}'\cdot\mathbf{r}_{mn}^0) + \cdots\right] \tag{47}$$

The first term, the zeroth-order thermal effect, in Equation 47 is the Debye-Waller factor−modified Bragg scattering followed by the first-order TDS, the second-order TDS, and so on. The first-order TDS is a one-phonon scattering process by which one phonon will interact with the x ray resulting in an energy and momentum exchange. The second-order TDS involves the interaction of one photon with two phonons. The expression for first-order TDS may be further simplified and related to lattice dynamics; this is described in this section. Higher-order TDS (for which force constants are required) usually become rather difficult to handle. Fortunately, they become important only at high temperatures (e.g., near and above the Debye temperature).

The first-order TDS intensity may be rewritten as follows:

$$I_1^{\text{TDS}}(\mathbf{K}) = \frac{1}{2}f^2e^{-2M}\sum_{gj}G_{gj}\left[\sum_m\sum_n e^{2\pi i(\mathbf{K}+\mathbf{g})\cdot\mathbf{r}_{mn}^0} + \sum_m\sum_n e^{2\pi i(\mathbf{K}-\mathbf{g})\cdot\mathbf{r}_{mn}^0}\right] \tag{48}$$

To obtain Equation 48, the following equivalence was used

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2} \tag{49}$$

The two double summations in the square bracket are in the form of the 3D interference function, the same as $G(\mathbf{K})$ in Equation 11, with wave vectors $\mathbf{K} + \mathbf{g}$ and $\mathbf{K} - \mathbf{g}$, respectively. We understand that the interference function has a significant value when its vector argument, $\mathbf{K} + \mathbf{g}$ and $\mathbf{K} - \mathbf{g}$ in this case, equals to a reciprocal lattice vector, $\mathbf{H}(hkl)$. Consequently, the first-order TDS reduces to

$$I_1^{\text{TDS}}(\mathbf{K}) = \frac{1}{2}f^2e^{-2M}\sum_{gj}G_{gj}[G(\mathbf{K}+\mathbf{g}) + G(\mathbf{K}-\mathbf{g})] = \frac{1}{2}N_v^2f^2e^{-2M}\sum_j G_{gj} \tag{50}$$

when $\mathbf{K} \pm \mathbf{g} = \mathbf{H}$, and N_v is the total number of atoms in the irradiated volume of the crystal.

Approximations may be applied to G_{gj} to relate it to more meaningful and practical parameters. For example, the mean kinetic energy of lattice waves is

$$\frac{1}{2}\left\langle \sum_n m \frac{d\mathbf{u}_n}{dt} \right\rangle^2 \tag{50a}$$

in which the displacement term $\mathbf{u}_n$ has been given in Equations 40 and 41, and m is the mass of a vibrating atom. If we take a first derivative of Equation 40 with respect to time (t), the kinetic energy (K.E.) becomes

$$\text{K.E.} = \frac{1}{4} mN \sum_{gj} \omega_{gj}^2 \langle a_{gj}^2 \rangle \tag{51}$$

The total energy of lattice waves is the sum of the kinetic and potential energies. For a harmonic oscillator, which is assumed in the present case, the total energy is equal to two times the kinetic energy. That is,

$$E_{\text{total}} = 2[\text{K.E.}] = \frac{1}{2} mN \sum_{gj} \omega_{gj}^2 \langle a_{gj}^2 \rangle = \sum_{gj} \langle E_{gj} \rangle \tag{52}$$

At high temperatures, the phonon energy for each gj component may be approximated by

$$\langle E_{gj} \rangle \approx kT \tag{53}$$

where k is the Boltzman constant. Thus, from Equation 52 we have

$$\langle a_{gj}^2 \rangle = \frac{2\langle E_{gj} \rangle}{mN\omega_{gj}^2} \approx \frac{2kT}{mN\omega_{gj}^2} \tag{54}$$

Substituting Equation 54 for the term $\langle a_{gj}^2 \rangle$ in Equation 50, we obtain the following expression for the first-order TDS intensity

$$I_1^{\text{TDS}}(\mathbf{K}) = f^2 e^{-2M} \frac{NkT}{m} \left(\frac{4\pi \sin\theta}{\lambda} \right)^2 \sum_{j=1}^{3} \frac{\cos^2(\mathbf{K}, \mathbf{e}_{gj})}{\omega_{gj}^2} \tag{55}$$

in which the scattering vector satisfies $\mathbf{K} \pm \mathbf{g} = \mathbf{H}$ and the cosine function is determined based upon the angle spanned by the scattering vector $\mathbf{K}$ and the phonon eigenvector $\mathbf{e}_{gj}$. In a periodic lattice, there is no need to consider elastic waves with a wavelength less than a certain minimum value because there are equivalent waves with a longer wavelength. The concept of Brillouin zone is applied to restrict the range of $\mathbf{g}$.

The significance of a measurement of the first-order TDS at various positions in reciprocal space may be observed in Figure 4, which represents the $hk0$ section of the reciprocal space of a body-centered cubic (bcc) crystal. At point P, the first-order TDS intensity is due only to elastic waves with the wave vector equal to $\mathbf{g}$, and hence only to waves propagating in the direction of $\mathbf{g}$. There are generally three independent waves for a given $\mathbf{g}$, and even in the general case, one is approximately longitudinal and the other two are approximately transverse waves. The cosine term appearing in Equation 55 may be considered as a geometrical extinction factor, which can further modify the contribution from the various elastic waves with the wave vector $\mathbf{g}$. Through appropriate strategy, it is possible to separate the phonon wave contribution from different branches. One such example may be found in Dvorack and Chen (1983).

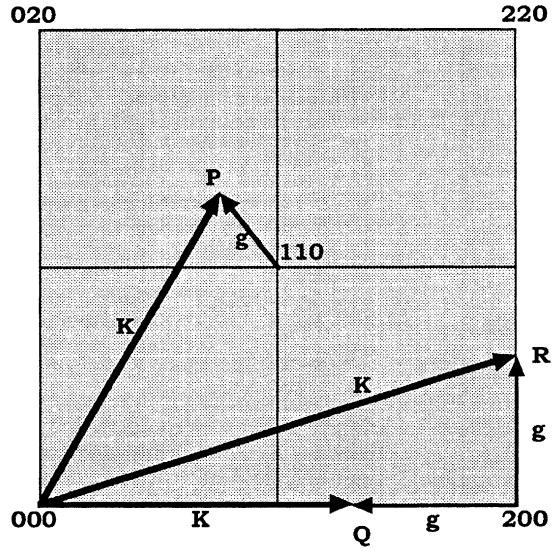


Figure 4. The ($hk0$) section of the reciprocal space corresponding to a bcc single crystal. At the general point P, there is a contribution from three phonon modes to the first-order TDS. At position Q, there is a contribution only from [100] longitudinal waves. At point R, there is a contribution from both longitudinal and transverse [100] waves.

From Equation 55, it is seen that the first-order TDS may be calculated for any given reciprocal lattice space location $\mathbf{K}$, so long as the eigenvalues, ω_{gj}, and the eigenvectors, $\mathbf{e}_{gj}$, of phonon branches are known for the system. In particular, the lower branch, or the lower-frequency phonon branches, contribute most to the TDS since the TDS intensity is inversely proportional to the square of the phonon frequencies. Quite often, the TDS pattern can be utilized to study soft-mode behavior or to identify the soft modes. The TDS intensity analysis is seldom carried out to determine the phonon dispersion curves, although such an analysis is possible (Dvorack and Chen, 1983); it requires making the measurements with absolute units and separating TDS intensities from different phonon branches. Neutron inelastic scattering techniques are much more common when it comes to determination of the phonon dispersion relationships. With the advent of high-brilliance synchrotron radiation facilities with milli-electron volt or better energy resolution, it is now possible to perform inelastic x-ray scattering experiments.

The second- and higher-order TDS might be appreciable for crystal systems showing soft modes, or close to or above the Debye temperature. The contribution of

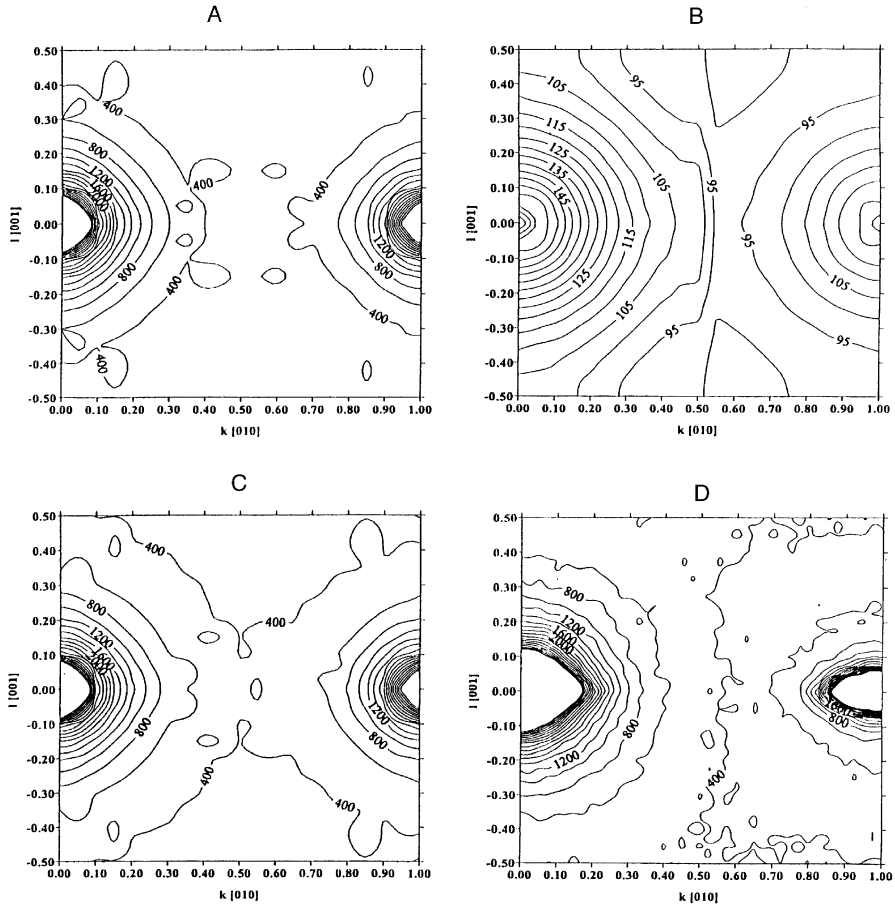


Figure 5. Equi-intensity contour maps, on the (100) plane of a cubic $BaTiO_3$ single crystal at 200°C. Calculated first-order TDS in (**A**), second-order TDS in (**B**), and the sum of (**A**) and (**B**) in (**C**), along with observed TDS intensities in (**D**).

second-order TDS represents the interaction between two phonon wave vectors with x rays and it can be calculated if the phonon dispersion relationship is known. The higher-order TDS can be significant and must be accounted for in diffuse scattering analysis in some cases. Figure 5 shows the calculated first- and second-order TDS along with the measured intensities for a $BaTiO_3$ single crystal in its paraelectric cubic phase (Takesue et al., 1997). The calculated TDS pattern shows the general features present in the observed data, but a discrepancy exists near the Brillouin zone center where measured TDS is higher than the calculation. This discrepancy is attributed to the overdamped phonon modes that are known to exist in $BaTiO_3$ due to anharmonicity.

Local Atomic Arrangement—Short-Range Ordering

A solid solution is thermodynamically defined as a single phase existing over a range of composition and temperature; it may exist over the full composition range of a binary system, be limited to a range near one of the pure constituents, or be based on some intermetallic compounds. It is, however, not required that the atoms be distributed randomly on the lattice sites; some degree of atomic ordering or segregation is the rule rather than the exception. The local atomic correlation in the absence of long-range order is the focus of interest in the present context.

The mere presence of a second species of atom, called solute atoms, requires that scattering from a solid solution produce a component of diffuse scattering throughout reciprocal space, in addition to the fundamental Bragg reflections. This component of diffuse scattering is modulated by the way the solute atoms are dispersed on and about the lattice sites, and hence contains a wealth of information. An elegant theory has evolved that allows one to treat this problem quantitatively within certain approximations, as have related techniques for visualizing and characterizing real-space, locally ordered atomic structure. More recently, it has been shown that pairwise interaction energies can be obtained from diffuse scattering studies on alloys at equilibrium. These energies offer great promise in allowing one to do realistic kinetic Ising modeling to understand how, for example, supersaturated solid solutions decompose.

An excellent, detailed review of the theory and practice of the diffuse scattering method for studying local atomic order, predating the quadratic approximation, has been given by Sparks and Borie (1966). More recent reviews on this topic were described by Chen et al. (1979) and by Epperson et al. (1994). In this section, the scattering principles for the extraction of pairwise interaction energies

are outlined for a binary solid solution showing local order. Readers may find more detailed experimental procedures and applications in XAFS SPECTROMETRY. This section is written in terms of x-ray experiments, since x rays have been used for most local order diffuse scattering investigations to date; however, neutron diffuse scattering is in reality a complementary method.

Within the kinematic approximation, the coherent scattering from a binary solid solution alloy with species A and B is given in electron units by

$$I_{\text{eu}}(\mathbf{K}) = \sum_p \sum_q f_p f_q e^{i\mathbf{K}\cdot(\mathbf{R}_p - \mathbf{R}_q)} \tag{56}$$

where f_p and f_q are the atomic scattering factors of the atoms located at sites p and q, respectively, and $(\mathbf{R}_p - \mathbf{R}_q)$ is the instantaneous interatomic vector. The interatomic vector can be $\boldsymbol{\beta}$ written as

$$(\mathbf{R}_p - \mathbf{R}_q) = \langle \mathbf{R}_p - \mathbf{R}_q \rangle + (\boldsymbol{\delta}_p - \boldsymbol{\delta}_q) \tag{57}$$

where $\boldsymbol{\delta}_p$ and $\boldsymbol{\delta}_q$ are vector displacements from the average lattice sites. The $\langle\rangle$ brackets indicate an average over time and space. Thus

$$I_{\text{eu}}(\mathbf{K}) = \sum_p \sum_q f_p f_q e^{i\mathbf{K}(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} e^{\langle i\mathbf{K}\cdot(\mathbf{R}_p - \mathbf{R}_q)\rangle} \tag{58}$$

In essence, the problem in treating local order diffuse scattering is to evaluate the factor $f_p f_q e^{i\mathbf{K}(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)}$ taking into account all possible combinations of atom pairs: AA, AB, BA, and BB.

The modern theory and the study of local atomic order diffuse scattering had their origins in the classical work by Cowley (1950) in which he set the displacement to zero. Experimental observations by Warren et al. (1951), however, soon demonstrated the necessity of accounting for this atomic displacement effect, which tends to shift the local order diffuse maxima from positions of cosine symmetry in reciprocal space.

Borie (1961) showed that a linear approximation of the exponential containing the displacements allowed one to separate the local order and static atomic displacement contributions by making use of the fact that the various components of diffuse scattering have different symmetry in reciprocal space. This approach was extended to a quadratic approximation of the atomic displacement by Borie and Sparks (1971). All earlier diffuse scattering measurements were made using this separation method. Tibbals (1975) later argued that the theory could be cast so as to allow inclusion of the reciprocal space variation of the atomic scattering factors. This is included in the state-of-art formulation by Auvray et al. (1977), which is outlined here.

Generally, for a binary substitutional alloy one can write

$$\begin{aligned}\langle f_p f_q e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)}\rangle &= X_{\text{A}} P_{pq}^{\text{AA}} f_{\text{A}}^2 \langle e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{A}} - \boldsymbol{\delta}_q^{\text{A}})}\rangle \\ &+ X_{\text{A}} P_{pq}^{\text{BA}} f_{\text{A}} f_{\text{B}} \langle e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{A}})}\rangle \\ &+ X_{\text{B}} P_{pq}^{\text{AB}} f_{\text{A}} f_{\text{B}} \langle e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{A}} - \boldsymbol{\delta}_q^{\text{B}})}\rangle + X_{\text{B}} P_{pq}^{\text{BB}} f_{\text{B}}^2 \langle e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{B}})}\rangle\end{aligned} \tag{59}$$

where X_{A} and X_{B} are atom fractions of species A and B, respectively, and P_{pq}^{AB} is the conditional probability of finding an A atom at site p provided there is a B atom at site q, and so on. There are certain relationships among the conditional probabilities for a binary substitutional solid solution:

$$X_{\text{A}} P_{pq}^{\text{BA}} = X_{\text{B}} P_{pq}^{\text{AB}} \tag{60}$$

$$P_{pq}^{\text{AA}} + P_{pq}^{\text{BA}} = 1 \tag{61}$$

$$P_{pq}^{\text{BB}} + P_{pq}^{\text{AB}} = 1 \tag{62}$$

If one also introduces the Cowley-Warren (CW) order parameter (Cowley, 1950),

$$\alpha_{pq} = 1 - \frac{P_{pq}^{\text{BA}}}{X_{\text{B}}} \tag{63}$$

Equation 59 reduces to

$$\begin{aligned}\langle f_p f_q e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)}\rangle &= (X_{\text{A}}^2 - X_{\text{A}} X_{\text{B}} \alpha_{pq}) f_{\text{A}}^2 \langle e^{i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{A}} - \boldsymbol{\delta}_q^{\text{A}})}\rangle \\ &+ 2X_{\text{A}} X_{\text{B}} (1 - \alpha_{pq}) f_{\text{A}} f_{\text{B}} \langle e^{\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{A}})}\rangle \\ &+ (X_{\text{B}}^2 + X_{\text{A}} X_{\text{B}} \alpha_{pq}) f_{\text{B}}^2 \langle e^{\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{B}})}\rangle\end{aligned} \tag{64}$$

If one makes series expansions of the exponentials and retains only quadratic and lower-order terms, it follows that

$$\begin{aligned}I_{\text{eu}}(\mathbf{K}) = &\sum_p \sum_q (X_{\text{A}} f_{\text{A}} + X_{\text{B}} f_{\text{B}})^2 e^{i\mathbf{K}\cdot\mathbf{R}_p q} \\ &+ \sum_p \sum_q X_{\text{A}} X_{\text{B}} (f_{\text{A}} - f_{\text{B}}) 2\alpha_{pq} e^{i\mathbf{K}\cdot\mathbf{R}_{pq}} \\ &+ \sum_p \sum_q \Big[(X_{\text{A}}^2 + X_{\text{A}} X_{\text{B}} \alpha_{pq}) f_{\text{A}}^2 \langle i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{A}} - \boldsymbol{\delta}_q^{\text{A}})\rangle \\ &+ 2X_{\text{A}} X_{\text{B}} (1 - \alpha_{pq}) f_{\text{A}} f_{\text{B}} \langle i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{A}})\rangle \\ &+ (X_{\text{B}}^2 + X_{\text{A}} X_{\text{B}} \alpha^p q) f_{\text{B}}^2 \langle i\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{B}})\rangle \Big] e^{i\mathbf{k}\cdot\mathbf{R}_{pq}} \\ &- \frac{1}{2} \sum_p \sum_q \Big[(X_{\text{A}}^2 + X_{\text{A}} X_{\text{B}} \alpha_{pq}) f_{\text{A}}^2 \Big\langle \Big[\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{A}} - \boldsymbol{\delta}_q^{\text{A}})\Big]^2 \Big\rangle \\ &+ 2X_{\text{A}} X_{\text{B}} (1 - \alpha_{pq}) f_{\text{A}} f_{\text{B}} \Big\langle \Big[\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{A}})\Big]^2 \Big\rangle \\ &+ (X_{\text{B}}^2 + X_{\text{A}} X_{\text{B}} \alpha_{pq}) f_{\text{B}}^2 \Big\langle \Big[\mathbf{K}\cdot(\boldsymbol{\delta}_p^{\text{B}} - \boldsymbol{\delta}_q^{\text{B}})\Big]^2 \Big\rangle \Big] e^{i\mathbf{K}\cdot\mathbf{R}_{pq}}\end{aligned} \tag{65}$$

where $e^{i\mathbf{K}\cdot\mathbf{R}_{pq}}$ denotes $e^{\langle i\mathbf{K}\cdot(\mathbf{R}_p - \mathbf{R}_q)\rangle}$. The first double summation represents the fundamental Bragg reflections for the average lattice. The second summation is the atomic-order modulated Laue monotonic, the term of primary interest here. The third sum is the so-called first-order atomic displacements, and it is purely static in nature. The final double summation is the second-order atomic displacements and contains both static and dynamic contributions. A detailed derivation would show that the second-order displacement series does not converge to zero. Rather, it represents a loss of intensity by the Bragg reflections; this is how TDS and Huang scattering originate. Henceforth, we shall use the term second-order displacement

scattering to denote this component, which is redistributed away from the Bragg positions. Note in particular that the second-order displacement component represents additional intensity, whereas the first-order size effect scattering represents only a redistribution that averages to zero. However, the quadratic approximation may not be adequate to account for the thermal diffuse scattering in a given experiment, especially for elevated temperature measurements or for systems showing a soft phonon mode. The experimental temperature in comparison to Debye temperature of the alloy is a useful guide for judging the adequacy of the quadratic approximation.

For cubic alloys that exhibit only local ordering (i.e., short-range ordering or clustering), it is convenient to replace the double summations by N times single sums over lattice sites, now specified by triplets of integers (lmn), which denote occupied sites in the lattice; N is the number of atoms irradiated by the x-ray beam. One can express the average interatomic vector as

$$\langle \mathbf{R}_{lmn} \rangle = l\mathbf{a}_1 + m\mathbf{a}_2 + n\mathbf{a}_3 \tag{66}$$

where $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are orthogonal vectors parallel to the cubic unit cell edges. The continuous variables in reciprocal space (h_1, h_2, h_3) are related to the scattering vector by

$$\mathbf{K} = 2\pi\left(\frac{\mathbf{S}-\mathbf{S}_0}{\lambda}\right) = 2\pi(h_1\mathbf{b}_1 + h_2\mathbf{b}_2 + h_3\mathbf{b}_3) \tag{67}$$

where $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ are the reciprocal space lattice vectors as defined in Equation 14. The coordinate used here is that conventionally employed in diffuse scattering work and is chosen in order that the occupied sites can be specified by a triplet of integers. Note that the 200 Bragg position becomes 100, and so on, in this notation. It is also convenient to represent the vector displacements in terms of components along the respective real-space axes as

$$(\boldsymbol{\delta}_p^{\mathrm{A}} - \boldsymbol{\delta}_q^{\mathrm{A}}) \equiv \boldsymbol{\delta}_{pq}^{\mathrm{AA}} = X_{lmn}^{\mathrm{AA}}\mathbf{a}_1 + Y_{lmn}^{\mathrm{AA}}\mathbf{a}_2 + Z_{lmn}^{\mathrm{AA}}\mathbf{a}_3 \tag{68}$$

If one invokes the symmetry of the cubic lattice and simplifies the various expressions, the coherently scattered diffuse intensity that is observable becomes, in the quadratic approximation of the atomic displacements,

$$\begin{aligned}\frac{I_D(h_1,h_2,h_3)}{NX_{\mathrm{A}}X_{\mathrm{B}}(f_{\mathrm{A}}-f_{\mathrm{B}})^2} = {} & \sum_l\sum_m\sum_n \alpha_{lmn}\cos 2\pi(h_1 l + h_2 m + h_3 n) \\ & + h_1\eta Q_x^{\mathrm{AA}} + h_1\xi Q_x^{\mathrm{BB}} + h_2\eta Q_y^{\mathrm{AA}} \\ & + h_2\xi Q_y^{\mathrm{BB}} + h_3\eta Q_z^{\mathrm{AA}} + h_3\xi Q_z^{\mathrm{BB}} \\ & + h_1^2\eta^2 R_x^{\mathrm{AA}} + 2h_1^2\eta\xi R_x^{\mathrm{AB}} + h_1^2\xi^2 R_x^{\mathrm{BB}} \\ & + h_2^2\eta^2 R_y^{\mathrm{AA}} + 2h_2^2\eta\xi R_y^{\mathrm{AB}} + h_2^2\xi^2 R_y^{\mathrm{BB}} \\ & + h_3^2\eta^2 R_z^{\mathrm{AA}} + 2h_3^2\eta\xi R_z^{\mathrm{AB}} + h_3^2\xi^2 R_z^{\mathrm{BB}} \\ & + h_1h_2\eta^2 S_{xy}^{\mathrm{AA}} + 2h_1h_2\eta\xi S_{xy}^{\mathrm{AB}} + h_1h_2\xi^2 S_{xy}^{\mathrm{BB}} \\ & + h_1h_3\eta^2 S_{xz}^{\mathrm{AA}} + 2h_1h_3\eta\xi S_{xz}^{\mathrm{AB}} + h_1h_3\xi^2 S_{xz}^{\mathrm{BB}} \\ & + h_2h_3\eta^2 S_{yz}^{\mathrm{AA}} + 2h_2h_3\eta\xi S_{yz}^{\mathrm{AB}} + h_2h_3\xi^2 S_{yz}^{\mathrm{BB}}\end{aligned} \tag{69}$$

where

$$\eta = \frac{f_{\mathrm{A}}}{f_{\mathrm{A}} - f_{\mathrm{B}}} \tag{70}$$

and

$$\xi = \frac{f_{\mathrm{B}}}{f_{\mathrm{A}} - f_{\mathrm{B}}} \tag{71}$$

The Q_i functions, which describe the first-order size effects scattering component, result from simplifying the third double summation in Equation 65 and are of the form

$$\begin{aligned}Q_x^{\mathrm{AA}} = {} & -2\pi\sum_l\sum_m\sum_n\left(\frac{X_{\mathrm{A}}}{X_{\mathrm{B}}} + \alpha_{lmn}\right)\langle X_{lmn}^{\mathrm{AA}}\rangle \\ & \times \sin 2\pi h_1 l \cos 2\pi h_2 m \cos 2\pi h_3 n\end{aligned} \tag{72}$$

where $\langle X_{lmn}^{\mathrm{AA}}\rangle$ is the mean component of displacement, relative to the average lattice, in the x direction of the A atom at site lmn when the site at the local origin is also occupied by an A-type atom.

The second-order atomic displacement terms obtained by simplification of the fourth double summation in Equation 65 are given by expressions of the type

$$\begin{aligned}R_x^{\mathrm{AA}} = {} & -4\pi^2\sum_l\sum_m\sum_n\left(\frac{X_{\mathrm{A}}}{X_{\mathrm{B}}} + \alpha_{lmn}\right)\langle X_o^{\mathrm{A}}X_{lmn}^{\mathrm{A}}\rangle \\ & \times \cos 2\pi h_1 l \cos 2\pi h_2 m \cos 2\pi h_3 n\end{aligned} \tag{73}$$

and

$$\begin{aligned}S_{xy}^{\mathrm{AB}} = {} & -8\pi^2\sum_l\sum_m\sum_n\left(\frac{X_{\mathrm{A}}}{X_{\mathrm{B}}} + \alpha_{lmn}\right)\langle X_o^{\mathrm{A}}Y_{lmn}^{\mathrm{A}}\rangle \\ & \times \sin 2\pi h_1 l \sin 2\pi h_2 m \cos 2\pi h_3 n\end{aligned} \tag{74}$$

In Equations 73 and 74, the terms in angle brackets represent correlations of atomic displacements. For example, $\langle X_o^{\mathrm{A}}Y_{lmn}^{\mathrm{A}}\rangle$ represents the mean component of displacement in the Y direction of an A-type atom a vector distance lmn from an A-type atom at the local origin.

The first summation in Equation 69 (I_{SRO}) contains the statistical information about the local atomic ordering of primary interest; it is a 3D Fourier cosine series whose coefficients are the CW order parameters. The first term in this series ($\alpha 000$) is a measure of the integrated local order diffuse intensity, and, provided the data are normalized by the Laue monotonic unit $[X_{\mathrm{A}}X_{\mathrm{B}}(f_{\mathrm{A}}-f_{\mathrm{B}})^2]$, should have the value of unity.

A schematic representation of the various contributions due to I_{SRO}, Q, and R/S components to the total diffuse intensity along an $[h00]$ direction is shown in Figure 6 for a system showing short-range clustering. As one can see, beside sharp Bragg peaks, there are I_{SRO} components concentrated near the fundamental Bragg peaks due to local clustering, the oscillating diffuse intensity due to static displacements (Q), and TDS-like intensity (R and S) near the tail of the fundamental reflections. Each of these diffuse-intensity components can be separated and

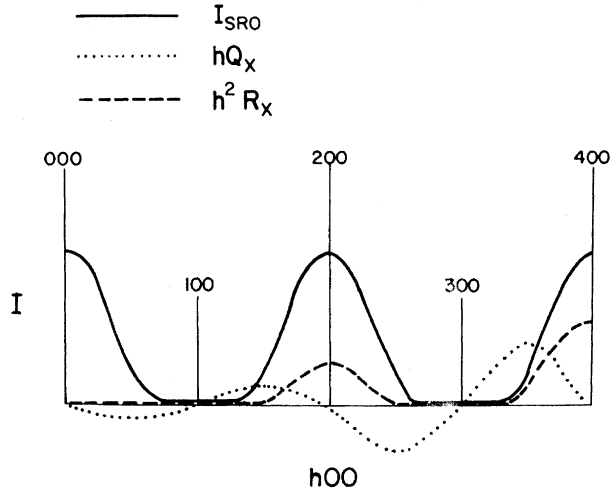


Figure 6. Schematic representation of the various contributions to diffuse x-ray scattering (l) along an [h00] direction in reciprocal space, from an alloy with short-range ordering and displacement. Fundamental Bragg reflections have all even integers; other sharp peak locations represent the superlattice peak when the system becomes ordered.

analyzed to reveal the local structure and associated static displacement fields.

The coherent diffuse scattering thus consists of 25 components for the cubic binary substitutional alloy, each of which possesses distinct functional dependence on the reciprocal space variables. This fact permits the components to be separated. To effect the separation, intensity measurements are made at a set of reciprocal lattice points referred to as the "associated set." These associated points follow from a suggestion of Tibbals (1975) and are selected according to crystallographic symmetry rules such that the corresponding 25 functions (I_{SRO}, Q, R, and S) in Equation 69 have the same absolute value. Note, however, that the intensities at the associated points need not be the same, because the functions are multiplied by various combinations of h_i, η, and ξ. Extended discussion of this topic has been given by Schwartz and Cohen (1987). An associated set is defined for each reciprocal lattice point in the required minimum volume for the local order component of diffuse scattering, and the corresponding intensities must be measured in order that the desired separation can be carried out.

The theory outlined above has heretofore been used largely for studying short-range-ordering alloys (preference for unlike nearest neighbors); however, the theory is equally valid for alloy systems that undergo clustering (preference for like nearest neighbors), or even a combination of the two. If clustering occurs, local order diffuse scattering will be distributed near the fundamental Bragg positions, including the zeroth-order diffraction; that is, small-angle scattering (SAS) will be observed. Because of the more localized nature of the order diffuse scattering, analysis is usually carried out with rather a different formalism; however, Hendricks and Borie (1965) considered some important aspects using the atomistic approach and the CW formalism.

In some cases, both short-range ordering and clustering may coexist, as in the example by Anderson and Chen (1994), who utilized synchrotron x rays to investigate the short-range-order structure of an Au–25 at.% Fe single crystal at room temperature. Two heat treatments were investigated: a 400°C aging treatment for 2 days and a 440°C treatment for 5 days, both preceded by solution treatment in the single-phase field and water quenched to room temperature. Evolution of SRO structure with aging was determined by fitting two sets of Cowley-Warren SRO parameters to a pair of 140,608-atom models. The microstructures, although quite disordered, showed a trend with aging for an increasing volume fraction of an Fe-enriched and an Fe-depleted environment—indicating that short-range ordering and clustering coexist in the system. The Fe-enriched environment displayed a preference for Fe segregation to the {110} and {100} fcc matrix planes. A major portion of the Fe-depleted environment was found to contain elements (and variations of these elements) of the $D1a$ ordered superstructure. The SRO contained in the Fe-depleted environment may best be described in terms of the standing wave packet model. This model was the first study to provide a quantitative real-space view of the atomic arrangement of the spin-glass system Au-Fe.

Surface/Interface Diffraction

Surface science is a subject that has grown enormously in the last few decades, partly because of the availability of new electron-based tools. X-ray diffraction has also contributed to many advances in the field, particularly when synchrotron radiation is used. Interface science, on the other hand, is still in its infancy as far as structural analysis is concerned. Relatively crude techniques, such as dissolution and erosion of one-half of an interface, exist but have limited application. Surfaces and interfaces may be considered as a form of defect because the uniform nature of a bulk crystal is abruptly terminated so that the properties of the surfaces and interfaces often differ significantly from the bulk.

In spite of the critical role that they play in such diverse sciences as catalysis, tribology, metallurgy, and electronic devices and the expected richness of the 2D physics of melting, magnetism, and related phase transitions, only a few surface structures are known, most of those are known only semiquantitatively (e.g., their symmetry; Somorjai, 1981). Our inability in many cases to understand atomic structure and to make the structure/properties connection in the 2D region of surfaces and interfaces has significantly inhibited progress in understanding this rich area of science.

X-ray diffraction has been an indispensable tool in 3D materials structure characterization despite the relatively low-scattering cross-section of x-ray photons compared with electrons. But the smaller number of atoms involved at surfaces and interfaces has made structural experiments at best difficult and in most cases impossible. The advent of high-intensity synchrotron radiation sources has definitely facilitated surface/interface x-ray diffraction. The nondestructive nature of the technique together

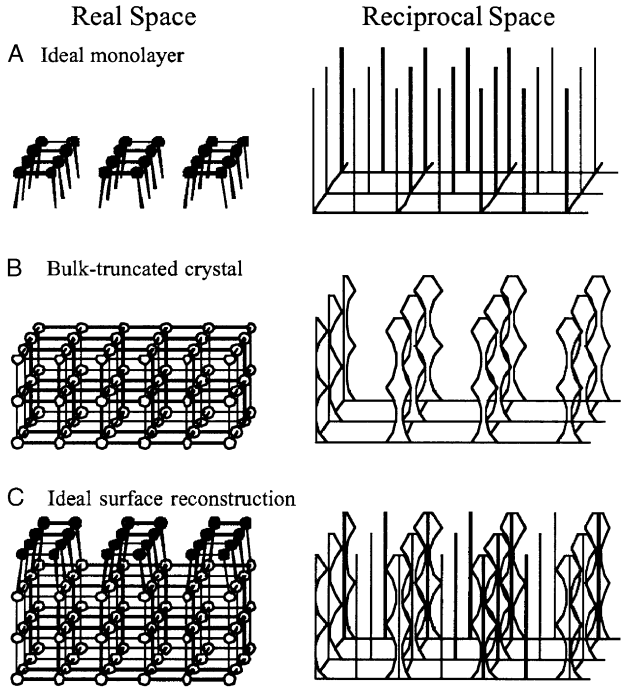


Figure 7. Real-space and reciprocal space views of an ideal crystal surface reconstruction. (**A**) A single monolayer with twice the periodicity in one direction producing featureless 2D Bragg rods whose periodicity in reciprocal space is one-half in one direction. The grid in reciprocal space corresponds to a bulk (1×1) cell. (**B**) A (1×1) bulk-truncated crystal and corresponding crystal truncation rods (CTRs). (**C**) An ideal reconstruction combining features from (**A**) and (**B**); note the overlap of one-half the monolayer or surface rods with the bulk CTRs. In general, 2D Bragg rods arising from a surface periodicity unrelated to the bulk (1×1) cell in size of orientation will not overlap with the CTRs.

with its high penetration power and negligible effect due to multiple scattering should make x-ray diffraction a premier method for quantitative surface and interface structural characterization (Chen, 1996). Up to this point we have considered diffraction from 3D crystals based upon the fundamental kinematic scattering theory laid out in the section on Diffraction from a Crystal. For diffraction from surfaces or interfaces, modifications need to be made to the intensity formulas that we shall discuss below.

Schematic pictures after Robinson and Tweet (1992) illustrating 2D layers existing at surfaces and interfaces are shown in Figure 7; there are three cases for consideration. Figure 7A is the case where an ideal 2D monolayer exists, free from interference of any other atoms. This case is hard to realize in nature. The second case is more realistic and is the one that most surface scientists are concerned with: the case of a truncated 3D crystal on top of which lies a 2D layer. This top layer could have a structure of its own, or it could be a simple continuation of the bulk structure with minor modifications. This top layer could also be of a different element or elements from the bulk. The surface structure may sometimes involve arrangement of atoms in more than one atomic layer, or may be less than one monolayer thick.

The term "interface" usually refers to the case when two bulk media of the same or different material are in contact, as Figure 7C shows. Either one or both may be crystalline, and therefore interfaces include grain boundaries as well. Rearrangement of atoms at interfaces may occur, giving rise to unique 2D diffraction patterns. By and large, the diffraction principles for scattering from surfaces or interfaces are considered identical. Consequently, the following discussion applies to both cases.

Rods from a 2D Diffraction. Diffraction from 2D structures in the above three cases can be described using Equations 8, 9, 10, 11, and 12 and Equation 20. If we take $\mathbf{a}_3$ to be along the surface/interface normal, the isolated monolayer is a 2D crystal with $N_3 = 1$. Consequently, one of the Laue conditions is relaxed, that is, there is no constraint on the magnitude of $\mathbf{K} \cdot \mathbf{a}_3$, which means the diffraction is independent of $\mathbf{K} \cdot \mathbf{a}_3$, the component of momentum transfer perpendicular to the surface. As a result, in 3D reciprocal space the diffraction pattern from this 2D structure consists of rods perpendicular to the surface, as depicted in Figure 7A. Each rod is a line of scattering extending out to infinity along the surface-normal direction, but is sharp in the other two directions parallel to the surface. For the surface of a 3D crystal, the diffuse rods resulting from the scattering of the 2D surface structure will connect the discrete Bragg peaks of the bulk. If surface/ interface reconstruction occurs, new diffuse rods will occur; these do not always run through the bulk Bragg peaks, as in the case shown in Figure 7C.

The determination of a 2D structure can, in principle, be made by following the same methods that have been developed for 3D crystals. The important point here is that one has to scan across the diffuse rods, that is, the scattering vector $\mathbf{K}$ must lie in the plane of the surface—the commonly known "in-plane" scan. Only through measurements such as these can the total integrated intensities, after resolution function correction and background subtraction, be utilized for structure analysis. The grazing-incidence x-ray diffraction technique is thus developed to accomplish this goal SURFACE X-RAY DIFFRACTION. Other techniques such as the specular reflection, standing-wave method can also be utilized to aid in the determination of surface structure, surface roughness, and composition variation.

Figure 7C represents schematically the diffraction pattern from the corresponding structure consisting of a 2D reconstructed layer on top of a 3D bulk crystal. We have simply superimposed the 3D bulk crystal diffraction pattern in the form of localized Bragg peaks (dots) with the Bragg diffraction rods deduced from the 2D structure. One should be reminded that extra reflections, that is, extra rods, could occur if the 2D surface structure differs from that of the bulk. For a 2D structure involving one layer of atoms and one unit cell in thickness, the Bragg diffraction rods, if normalized against the decaying nature of the atomic scattering factors, are flat in intensity and extend to infinity in reciprocal space. When the 2D surface structure has a thickness of more than one unit cell, a pseudo-2D structure or a very thin layer is of concern, and the Bragg diffraction rods will no longer be flat in their intensity profiles but instead fade away monotonically

from the zeroth-order plane normal to the sample surface in reciprocal space. The distance to which the diffraction rods extends is inversely dependent on the thickness of the thin layer.

Crystal Truncation Rods. In addition to the rods originating from the 2D structure, there is one other kind of diffuse rod that contributes to the observed diffraction pattern that has a totally different origin. This second type of diffuse rod has its origin in the abrupt termination of the underlying bulk single-crystal substrate, the so-called crystal truncation rods, CTRs. This contribution further complicates the diffraction pattern, but is rich in information concerning the surface termination sequence, relaxation, and roughness; therefore, it must be considered. The CTR intensity profiles are not flat but vary in many ways that are determined by the detailed atomic arrangement and static displacement fields near surfaces, as well as by the topology of the surfaces. The CTR intensity lines are always perpendicular to the surface of the substrate bulk single crystal and run through all Bragg peaks of the bulk and the surface. Therefore, for an inclined surface normal that is not parallel to any crystallographic direction, CTRs do not connect all Bragg peaks, as shown in Figure 7B.

Let us consider the interference function, Equation 11, along the surface normal, $\mathbf{a}_3$, direction. The numerator, $\sin^2(\pi \mathrm{K} \cdot N_3 \mathbf{a}_3)$, is an extremely rapid varying function of $\mathbf{K}$, at least for large N_3, and is in any case smeared out in a real experiment because of finite resolution. Since it is always positive, we can approximate it by its average value of $\frac{1}{2}$. This gives a simpler form for the limit of large N_3 that is actually independent of N_3:

$$|G_3(\mathbf{K})|^2 = \frac{1}{2\sin^2(\pi \mathbf{K} \cdot \mathbf{a}_3)} \tag{75}$$

Although the approximation is not useful at any of the Bragg peaks defined by the three Laue conditions, it does tell us that the intensity in between Bragg peaks is actually nonzero along the surface normal direction, giving rise to the CTRs. Another way of looking at CTRs comes from convolution theory. From the kinematic scattering theory presented earlier, we understand that the scattering cross-section is the product of two functions, the structure factor $F(\mathbf{K})$ and the interference function $G(\mathbf{K})$, expressed in terms of the reciprocal space vector $\mathbf{K}$. This implies, in real space, that the scattering cross-section is related to a convolution of two real-space structural functions: one defining the positions of all atoms within one unit cell and the other covering all lattice points. For an abruptly terminated crystal at a well-defined surface, the crystal is semi-infinite, which can be represented by a product of a step function with an infinite lattice. The diffraction pattern is then, by Fourier transformation, the convolution of a reciprocal lattice with the function $(2\pi \mathbf{K} \cdot \mathbf{a}_3)^{-1}$. It was originally shown by von Laue (1936) and more recently by Andrews and Cowley (1985), in a continuum approximation, that the external surface can thus give rise to streaks emanating from each Bragg peak of the bulk, perpendicular to the terminating crystal surface. This is what we now call the CTRs.

It is important to make the distinction between CTRs passing through bulk reciprocal lattice points and those due to an isolated monolayer 2D structure at the surface. Both can exist together in the same sample, especially when the surface layer does not maintain lattice correspondence with the bulk crystal substrate. To illustrate the difference and similarity of the two cases, the following equations may be used to represent the rod intensities of two different kinds:

$$I_{2\mathrm{D}} = I_0 N_1 N_2 |F(\mathbf{K})|^2 \tag{76}$$

$$I_{\mathrm{CTR}} = I_0 N_1 N_2 |F(\mathbf{K})|^2 \frac{1}{2\sin^2(\pi \mathbf{K} \cdot \mathbf{a}_3)} \tag{77}$$

The two kinds of rod have the same order-of-magnitude intensity in the "valley" far from the Bragg peaks at $\mathbf{K} \cdot \mathbf{a}_3 = l$. The actual intensity observed in a real experiment is several orders of magnitude weaker than the Bragg peaks. For the 2D rods, integrated intensities at various (hk) reflections can be measured and Fourier inverted to reveal the real-space structure of the 2D ordering. Patterson function analysis and difference Patterson function analysis are commonly utilized, along with least-squares fitting to obtain the structure information. For the CTRs, the stacking sequences and displacement of atomic layers near the surface, as well as the surface roughness factor, and so on, can be modeled through the calculation of the structure factor in Equation 77. Experimental techniques and applications of surface/interface diffraction techniques to various materials problems may be found in SURFACE X-RAY DIFFRACTION. Some of our own work may be found in the studies of buried semiconductor surfaces by Aburano et al. (1995) and by Hong et al. (1992a,b, 1993, 1996) and in the determination of the terminating stacking sequence of *c*-plane sapphire by Chung et al. (1997).

Small-Angle Scattering

The term "small-angle scattering" (SAS) is somewhat ambiguous as long as the sample, type of radiation, and incident wavelength are not specified. Clearly, Bragg reflections of all crystals when investigated with high-energy radiation (e.g., γ rays) occur at small scattering angles (small 2θ) simply because the wavelength of the probing radiation is short. Conversely, crystals with large lattice constants could lead to small Bragg angles for a reasonable wavelength value of the radiation used. These Bragg reflections, although they might appear at small angles, can be treated in essentially the same way as the large-angle Bragg reflections with their origins laid out in all previous sections. However, in the more specific sense of the term, SAS is a scattering phenomenon related to the scattering properties at small scattering vectors $\mathbf{K}$ (with magnitudes $K = 2 \sin \theta/\lambda$), or, in other words, diffuse scattering surrounding the direct beam. It is this form of diffuse SAS that is the center of discussion in this section.

SAS is produced by the variation of scattering length density over distances exceeding the normal interatomic distances in condensed systems. Aggregates of small

particles (e.g., carbon black and catalysts) in air or vacuum, particles or macromolecules in liquid or solid solution (e.g., polymers and precipitates in alloys), and systems with smoothly varying concentration (or scattering length density) profiles (e.g., macromolecules, glasses, and spinodally decomposed systems) can be investigated with SAS methods. SAS intensity appears at low **K** values, that is, **K** should be small compared with the smallest reciprocal lattice vector in crystalline substances. Because the scattering intensity is related to the Fourier transform properties, as shown in Equation 7, it follows that measurements at low K will not allow one to resolve structural details in real space over distances smaller than $d_{\min} \approx \pi/K_{\max}$, where $K_{\max}$ is the maximum value accessible in the SAS experiment. If, for example, $K_{\max} = 0.2$ Å^{-1}, then $d_{\min} = 16$ Å, and the discrete arrangement of scattering centers in condensed matter can in most cases be replaced by a continuous distribution of scattering length, averaged over volumes of about d_{min}^3. Consequently, summations over discrete scattering sites as represented in Equation 7 and the subsequent ones can be replaced by integrals.

If we replace the scattering length f_j by a locally averaged scattering length density $\rho(\mathbf{r})$, where $\mathbf{r}$ is a continuously variable position vector, Equation 7 can be rewritten

$$I(\mathbf{K}) = \left| \int_V \rho(\mathbf{r}) e^{2\pi i \mathbf{K}\cdot\mathbf{r}} d^3\mathbf{r} \right|^2 \tag{78}$$

where the integration extends over the sample volume V. The scattering length density may vary over distances of the order $d_{\min}$ as indicated earlier, and it is sometimes useful to express

$$\rho(\mathbf{r}) = \Delta\rho(\mathbf{r}) + \rho_0 \tag{79}$$

where ρ_0 is averaged over a volume larger than the resolution volume of the instrument (determined by the minimum observable value of K). Therefore, by discounting the Bragg peak, the diffuse intensity originating from inhomogeneities is

$$I(\mathbf{K}) = \left| \int_V \Delta\rho(\mathbf{r}) e^{2\pi i \mathbf{K}\cdot\mathbf{r}} d^3\mathbf{r} \right|^2 \tag{80}$$

Two-Phase Model. Let the sample contain N_p particles with a homogeneous scattering length density ρ_p, and let these particles be embedded in a matrix of homogeneous scattering length density ρ_m. From Equation 80, one obtains for the SAS scattering intensity per atom:

$$I_a(\mathbf{K}) = \frac{1}{N} |\rho_p - \rho_m|^2 \left| \int_V e^{2\pi i \mathbf{K}\cdot\mathbf{r}} d^3\mathbf{r} \right|^2 \tag{81}$$

where N is the total number of atoms in the scattering volume and the integral extends over the volume V occupied by all particles in the irradiation sample. In the most general case, the above integral contains spatial and orientational correlations among particles, as well as effects due to size distributions. For a monodispersed system free of particle correlation, the single-particle form factor is

$$F_p(\mathbf{K}) = \frac{1}{V_p} \int_{V_p} e^{2\pi i \mathbf{K}\cdot\mathbf{r}} d^3\mathbf{r} \tag{82}$$

where V_p is the particle volume so that $F(0) = 1$; we can write for N_p identical particles

$$I_a(\mathbf{K}) = \frac{N_p V_p^2}{N} |\rho_p - \rho_m|^2 |F_p(\mathbf{K})|^2 \tag{83}$$

The interference (correlation) term in Equation 81 that we have neglected to arrive at in Equation 83, is the Fourier transform $\Psi(\mathbf{K})$ of the static pair correlation function,

$$\Psi(\mathbf{K}) = \frac{1}{N_p} \sum_{i=j} e^{2\pi i \mathbf{K}\cdot(\mathbf{r}_i - \mathbf{r}_j)} \tag{84}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the position vectors of the centers of particles labeled i and j. This function will only be zero for all nonzero **K** values if the interparticle distance distribution is completely random, as is approximately the case in very dilute systems. Equation 84 is also valid for oriented anisotropic particles if they are all identically oriented. In the more frequent cases of a random orientational distribution or discrete but multiple orientations of anisotropic particles, the appropriate averages of $F_p(\mathbf{K})^2$ have to be used.

Scattering Functions for Special Cases. Many different particle form factors have been calculated by Guinier and Fournet (1955), some of which are reproduced as follows for the isotropic and uncorrelated distribution, i.e., spherically random distribution of identical particles.

Spheres. For a system of noninteracting identical spheres of radius R_s, the form factor is

$$F_s(KR_S) = \frac{3[\sin(2\pi KR_S) - 2\pi KR_S \cos(2\pi KR_S)]}{(2\pi K)^3 R_S^3} \tag{85}$$

Ellipsoids. Ellipsoids of revolution of axes $2a$, $2a$, and $2av$ yield the following form factor:

$$|F_e(K)|^2 = \int_0^{2\pi} |F_s|^2 \left(\frac{2\pi Kav}{\sqrt{\sin^2\alpha + v^2\cos^2\alpha}} \right) \cos\beta \, d\beta \tag{86}$$

where F_s is the function in Equation 85 and $\alpha = \tan^{-1}(v \tan\beta)$.

Cylinders. For a cylinder of diameter $2a$ and height $2H$, the form factor becomes

$$|F_c(K)|^2 = \int_0^{\frac{\pi}{2}} \frac{\sin^2(2\pi KH\cos\beta)}{(2\pi K)^2 H^2 \cos^2\beta} \frac{4J_1^2(2\pi Ka\sin\beta)}{(2\pi K)^2 a^2 \sin^2\beta} \sin\beta \, d\beta \tag{87}$$

where J_1 is the first-order spherical Bessel function. Porod (see Guinier and Fournet, 1955) has given an

approximation form for Equation 87 valid for $KH \gg 1$ and $a \ll H$ which, in the intermediate range $Ka < 1$, reduces to

$$|F_c(K)|^2 \approx \frac{\pi}{4\pi KH} e^{-(2\pi K)^2 a^2/4} \tag{88}$$

For infinitesimally thin rods of length $2H$, one can write

$$|F_{\text{rod}}(K)|^2 \approx \frac{S_i(4\pi KH)}{2\pi KH} - \frac{\sin^2(2\pi KH)}{(2\pi K)^2 H^2} \tag{89}$$

where

$$S_i(x) = \int_0^x x \frac{\sin t}{t} dt \tag{90}$$

For the case $KH \gg 1$, Equation 87 reduces to

$$|F_{\text{rod}}(K)|^2 \approx \frac{1}{4KH} \tag{91}$$

For flat disks (i.e., when $H \ll a$), the scattering function for $KH \ll 1$ is

$$|F_{\text{disk}}(K)|^2 = \frac{2}{(2\pi K)^2 a^2} \left\{ 1 - \frac{1}{2\pi Ka} J_1(4\pi Ka) \right\} \tag{92}$$

where J_1 is the Bessel function. For $KH < 1 \ll Ka$, Equation 92 reduces to

$$|F_{\text{disk}}(K)|^2 \approx \frac{2}{(2\pi K)^2 a^2} e^{-4\pi^2 K^2 H^2/3} \tag{93}$$

The expressions given above are isotropic averages of particles of various shapes. When preferred alignment of particles occurs, modification to the above expressions must be made.

General Properties of the SAS Function. Some general behavior of the scattering functions shown above are described.

Extrapolation to K = 0. If the measured scattering curve can be extrapolated to the origin of reciprocal space (i.e., $K = 0$) one obtains, from Equation 83, a value for the factor $V_p^2 N_p (\rho_p - \rho_m)^2/N$, which, for the case of $N_p = 1$, $\rho_m = 0$, and $\rho_p = Nf/V_p$, reduces to

$$I_a(K) = Nf^2 \tag{94}$$

For a system of scattering particles with known contrast and size, Equation 94 will yield N, the total number of atoms in the scattering volume. In the general case of unknown V_p, N_p, and $(\rho_p - \rho_m)$, the results at K = 0 have to be combined with information obtained from other parts of the SAS curve.

Guinier Approximation. Guinier has shown that at small values of Ka, where a is a linear dimension of the particles, the scattering function is approximately related to a simple geometrical parameter called the radius of gyration, R_G. For small angles, $K \approx 2\theta/\lambda$, the scattering amplitude, as shown in Equation 80, may be expressed by a Taylor's expansion up to the quadratic term:

$$\begin{aligned} A &= \int_v \Delta\rho\, e^{2\pi i \mathbf{K}\cdot\mathbf{r}} d^3\mathbf{r} \\ &\approx \Delta\rho \left\{ \int_v d^3\mathbf{r} + \int_v 2\pi i \mathbf{K}\cdot\mathbf{r} d^3\mathbf{r} - \frac{4\pi^2}{2} \int_v (\mathbf{K}\cdot\mathbf{r})^2 d^3\mathbf{r} \right\} \end{aligned} \tag{95}$$

If one chooses the center of gravity of a diffracting object as its origin , the second term is zero. The first term is the volume V of the object times $\Delta\rho$. The third integral is the second moment of the diffracting object, related to R_G.

$$K^2 R_G^2 = \frac{1}{V} \int_v (\mathbf{K}\cdot\mathbf{r})^2 d^3\mathbf{r} \tag{96}$$

Thus the scattering amplitude in Equation 95 becomes

$$A \approx \Delta\rho \left\{ V - \frac{4\pi^2}{2} V R_G^2 K^2 \right\} \approx \Delta\rho V e^{-2\pi^2 R_G^2 K^2} \tag{97}$$

and for the SAS intensity for n independent, but identical, objects:

$$I(K) = A^* \cdot A \approx n(\Delta\rho)^2 V^2 e^{-4\pi^2 R_G^2 K^2} \tag{98}$$

Equation 98 implies that for the small-angle approximation, that is, for small K or small 2θ, the intensity can be approximated by a Gaussian function versus K^2. By plotting $\ln I(K)$ versus K^2 (known as the Guinier plot), a linear relationship is expected at small K with its slope proportional to R_G, which is also commonly referred as the Guinier radius.

The radius of gyration of a homogeneous particle has been defined in Equation 96. For a sphere of radius R_s, $R_G = (\frac{3}{5})^{1/2} R_s$, and the Gaussian form of the SAS intensity function, as shown in Equation 98, coincides with the correct expression, Equations 83 and 85, up to the term proportional to K^4. Subsequent terms in the two series expansions are in fair agreement and corresponding terms have the same sign. For this case, the Guinier approximation is acceptable over a wide range of KR_G. For the oblate rotational ellipsoid with $\upsilon = 0.24$ and the prolate one with $\upsilon = 1.88$, the Guinier approximation coincides with the expansion of the scattering functions even up to K^6. In general, the concept of the radius of gyration is applicable to particles of any shapes, but the K range, where this parameter can be identified, may vary with different shapes.

Porod Approximation. For homogeneous particles with sharp boundaries and a surface area A_p, Porod (see Guinier and Fournet, 1995) has shown that for large K

$$I(\mathbf{K}) \approx \frac{2\pi A_p}{V_p^2 (2\pi K)^4} \tag{99}$$

describes the average decrease of the scattering function. Damped oscillations about this average curve may occur in systems with very uniform particle size.

Integrated Intensity (The Small-Angle Invariant). Integration of SAS intensity over all K values yields an invariant, Q. For a two-phase model system, this quantity is

$$Q = V^2 C_p (1 - C_p)(\rho_p - \rho_m)^2 \qquad (100)$$

where C_p is the volume fraction of the dispersed particles. What is noteworthy here is that this quantity enables one to determine either C_p or $(\rho_p - \rho_m)$ if the other is known. Generally, the scattering contrast $(\rho_p - \rho_m)$ is known or can be estimated and thus measurement of the invariant permits a determination of the volume fraction of the dispersed particles.

Interparticle Interference Function. We have deliberately neglected interparticle interference terms (cf. Equation 84), to obtain Equation 83; its applicability is therefore restricted to very dilute systems, typically $N_p V_p < 0.01$. As long as the interparticle distance remains much larger than the particle size, it will be possible to identify single-particle scattering properties in a somewhat restricted range, as interference effects will affect the scattering at lower K values only. However, in dense systems one approaches the case of macromolecular liquids, and both single-particle as well as interparticle effects must be realized over the whole K range of interest.

For randomly oriented identical particles of arbitrary shape, interference effects can be included by writing (cf. Equations 83 and 84)

$$I(\mathbf{K}) \propto \{\overline{|F_p(\mathbf{K})|}^2 - \overline{|F_p(\mathbf{K})|}^2 + \overline{|F_p(\mathbf{K})|}^2 W_i(\mathbf{K})\} \qquad (101)$$

where the bar indicates an average of all directions of $\mathbf{K}$, and the interference function

$$W_i(\mathbf{K}) = \overline{\Psi(\mathbf{K})} + 1 \qquad (102)$$

with the function Ψ given in Equation 84. The parameter $W_i(\mathbf{K})$ is formally identical to the liquid structure factor, and there is no fundamental difference in the treatment between the two. It is possible to introduce thermodynamic relationships if one defines an interaction potential for the scattering particles. For applications in the solid state, hard-core interaction potentials with an adjustable interaction range exceeding the dimensions of the particle may be used to rationalize interparticle interference effects. It is also possible to model interference effects by assuming a specific model, or using a statistical approach.

As $W_i(\mathbf{K}) \to 0$ for large $\mathbf{K}$, interference between particles is most prominently observed at the lower $\mathbf{K}$ values of the SAS curve. For spherical particles, the first two terms in Equation 101 are equal and the scattering cross-section, or the intensity, becomes

$$I(\mathbf{K}) = C_1 \overline{|F_s(KR_s)|}^2 W_i(\mathbf{K}, C_s) \qquad (103)$$

where C_1 is a constant factor appearing in Equation 83, F_s is the single-particle scattering form factor of a sphere, and W_i is the interference function for rigid spheres of different concentration $C_s = N_p V_p / V$. The interference effects become progressively more important with increasing C_s. At large C_s values, the SAS curve shows a peak characteristic of the interference function used.

When particle interference modifies the SAS profile, the linear portion of the Guinier plot usually becomes inaccessible at small $\mathbf{K}$. Therefore, any straight line found in a Guinier plot at relatively large $\mathbf{K}$ is accidental and gives less reliable information about particle size.

Size Distribution. Quite frequently, the size distribution of particles also complicates the interpretation of SAS patterns, and the single-particle characteristics such as R_G, L_p, V_p, A_p, and so on, defined previously for identical particles, will have to be replaced by appropriate averages over the size distribution function. In many cases, both particle interference and size distribution may appear simultaneously so that the SAS profile will be modified by both effects.

Simple expressions for the scattering from a group of nonidentical particles can only be expected if interparticle interference is neglected. By generalizing Equation 83, one can write for the scattering of a random system of nonidentical particles without orientational correlation:

$$I(\mathbf{K}) = \frac{1}{N} \sum_v V_{pv}^2 N_{pv} \Delta\rho_v^2 \overline{|F_{pv}(\mathbf{K})|}^2 \qquad (104)$$

where v is a label for particles with a particular size parameter. The bar indicates orientational averaging. If the Guinier approximation is valid for even the largest particles in a size distribution, an experimental radius of gyration determined from the lower-$\mathbf{K}$ end of the scattering curve in Guinier representation will correspond to the largest sizes in the distribution. The Guinier plot will show positive curvature similar to the scattering function of nonspherical particles. There is obviously no unique way to deduce the size distribution of particles of unknown shape from the measured scattering profile, although it is much easier to calculate the cross-section for a given model. For spherical particles, several attempts have been made to obtain the size distribution function or certain characteristics of it experimentally, but even under these simplified conditions wide distributions are difficult to determine.

ACKNOWLEDGMENTS

This chapter is dedicated to Professor Jerome B. Cohen of Northwestern University, who passed away suddenly on November 7, 1999. The author received his education on crystallography and diffraction under the superb teaching of Professor Cohen. The author treasures his over 27 years of collegial interaction and friendship with Jerry.

The author also wishes to acknowledge J. B. Cohen, J. E. Epperson, J. P. Anderson, H. Hong, R. D. Aburano, N. Takesue, G. Wirtz, and T. C. Chiang for their direct or

indirect discussions, collaborations, and/or teaching over the past 25 years. The preparation of this unit is supported in part by the U.S. Department of Energy, Office of Basic Energy Science, under contract No. DEFH02-96ER45439, and in part by the state of Illinois Board of Higher Education, under a grant number NWU98 IBHE HECA through the Frederick Seitz Materials Research Laboratory at the University of Illinois at Urbana-Champaign.

LITERATURE CITED

Aburano, R. D., Hong, H., Roesler, J. M., Chung, K., Lin, D.-S., Chen, H., and Chiang, T.-C. 1995. Boundary structure determination of Ag/Si(111) interfaces by X-ray diffraction. *Phys. Rev. B* 52(3):1839–1847.

Anderson, J. P. and Chen, H. 1994. Determination of the short-range order structure of Au-25At. Pct. Fe using wide-angle diffuse synchrotron X-ray scattering. *Metall. Mater. Trans.* 25A:1561–1573.

Auvray, X., Georgopoulos, J., and Cohen, J. B. 1977. The structure of G.P.I. zones in Al-1.7AT.%Cu. *Acta Metall.* 29:1061–1075.

Azaroff, L. V. and Buerger, M. J. 1958. The Powder Method in X-ray Crystallography. McGraw-Hill, New York.

Borie, B. S. 1961. The separation of short range order and size effect diffuse scattering. *Acta Crystallogr.* 14:472–474.

Borie, B. S. and Sparks, Jr., C. J. 1971. The interpretation of intensity distributions from disordered binary alloys. *Acta Crystallogr.* A27:198–201.

Buerger, M. J. 1960. Crystal Structure Analysis. John Wiley & Sons, New York.

Chen, H. 1996. Review of surface/interface X-ray diffraction. *Mater. Chem. Phys*. 43:116–125.

Chen, H., Comstock, R. J., and Cohen, J. B. 1979. The examination of local atomic arrangements associated with ordering. *Annu. Rev. Mater. Sci.* 9:51–86.

Chung, K. S., Hong, H., Aburano, R. D., Roesler, J. M., Chiang, T. C., and Chen, H. 1997. Interface structure of Cu thin films on C-plane sapphire using X-ray truncation rod analysis. *In* Proceedings of the Symposium on Applications of Synchrotron Radiation to Materials Science III. Vol. 437. San Francisco, Calif.

Cowley, J. M. 1950. X-ray measurement of order in single crystal of Cu_3 Au. *J. Appl. Phys.* 21:24–30.

Cullity, B. D. 1978. Elements of X-ray Diffraction. Addison Wesley, Reading, Mass.

Debye, P. 1913a. Uber den Einfluts der Warmebewegung auf die Interferenzerscheinungen bei Rontgenstrahlen. *Verh. Deutsch. Phys. Ges.* 15:678–689.

Debye, P. 1913b. Uber die Interasitatsvertweilung in den mit Rontgenstrahlen erzeugten Interferenzbildern. *Verh. Deutsch. Phys. Ges.* 15:738–752.

Debye, P. 1913c. Spektrale Zerlegung der Rontgenstrahlung mittels Reflexion und Warmebewegung. *Verh. Deutsch. Phys. Ges.* 15:857–875.

Debye, P. 1913–1914. Interferenz von Rontgenstrahlen und Warmebewegung. *Ann. Phys. Ser.* 4, 43:49.

Dvorack, M. A. and Chen, H. 1983. Thermal diffuse x-ray scattering in β-phase Cu-Al-Ni alloy. *Scr. Metall.* 17:131–134.

Epperson, J. E., Anderson, J. P., and Chen, H. 1994. The diffuse-scattering method for investigating locally ordered binary solid solution. *Metal. Mater. Trans.* 25A:17–35.

Faxen, H. 1918. Die bei Interferenz von Rontgenstrahlen durch die Warmebewegung entstehende zerstreute Strahlung. *Ann. Phys.* 54:615–620.

Faxen, H. 1923. Die bei Interferenz von Rontgenstrahlen infolge der Warmebewegung entstehende Streustrahlung. *Z. Phys.* 17:266–278.

Guinier, A. 1994. X-ray Diffraction in Crystals, Imperfect crystals, and amorphous bodies. Dover Pub. Inc., New York.

Guinier, A. and Fournet, G. 1955. Small-Angle Scattering of X-rays. John Wiley & Sons, New York.

Hendricks, R. W. and Borie, B. S. 1965. On the Determination of the Metastable Miscibility Gap From Integrated Small-Angle X-Ray Scattering Data. *In* Proc. Symp. On Small Angle X-Ray Scattering (H. Brumberger, ed.) pp. 319–334. Gordon and Breach, New York.

Hong, H., Aburano, R. D., Chung, K., Lin, D.-S., Hirschorn, E. S., Chiang, T.-C., and Chen, H. 1996. X-ray truncation rod study of Ge(001) surface roughening by molecular beam homoepitaxial growth. *J. Appl. Phys.* 79:6858–6864.

Hong, H., Aburano, R. D., Hirschorn, E. S., Zschack, P., Chen, H., and Chiang, T. C. 1993. Interaction of (1×2)-reconstructed Si(100) and Ag(110):Cs surfaces with C_{60} overlayers. *Phys. Rev. B* 47:6450–6454.

Hong, H., Aburano, R. D., Lin, D. S., Chiang, T. C., Chen, H., Zschack, P., and Specht, E. D. 1992b. Change of Si(111) surface reconstruction under noble metal films. *In* MRS Proceeding Vol. 237 (K. S. Liang, M. P. Anderson, R. J. Bruinsma and G. Scoles, eds.) pp. 387–392. Materials Research Society, Warrendale, Pa.

Hong, H., McMahon, W. E., Zschack, P., Lin, D. S., Aburano, R. D., Chen, H., and Chiang, T.C. 1992a. C60 Encapsulation of the Si(111)-(7×7) Surface. *Appl. Phys. Lett.* 61(26):3127–3129.

International Table for Crystallography 1996. International Union of Crystallography: Birmingham, England.

James, R. W. 1948. Optical Principles of the Diffraction of X-rays. G. Bell and Sons, London.

Klug, H. P. and Alexander, L. E. 1974. X-ray Diffraction Procedures. John Wiley & Sons, New York.

Krivoglaz, M. A. 1969. Theory of X-ray and Thermal-Neutron Scattering by Real Crystals. Plenum, New York.

Noyan, I. C. and Cohen, J. B. 1987. Residual Stress: Measurement by Diffraction and Interpretation. Springer-Verlag, New York.

Robinson, I. K. and Tweet, D. J. 1992. Surface x-ray diffraction. *Rep. Prog. Phys.* 55:599–651.

Schultz, J. M. 1982. Diffraction for Materials Science. Prentice-Hall, Englewood Cliffs, N.J.

Schwartz, L. H. and Cohen, J. B. 1987. Diffraction from Materials. Springer-Verlag, New York.

Somorjai, G. A. 1981. Chemistry in Two Dimensions: Surfaces. Cornell University Press, Ithaca, N.Y.

Sparks, C. J. and Borie, B. S. 1966. Methods of analysis for diffuse X-ray scatterling modulated by local order and atomic displacements. *In* Local Atomic Arrangement Studied by X-ray Diffraction (J. B. Cohen and J. E. Hilliard, eds.) pp. 5–50. Gordon and Breach, New York.

Takesue, N., Kubo, H., and Chen, H. 1997. Thermal diffuse X-ray scattering study of anharmonicity in cubic barium titanate. *J. Nucl. Instr. Methods Phys. Res.* B133:28–33.

Tibbals, J. E. 1975. The separation of displacement and substitutional disorder scattering: a correction from structure-factor ratio variation. *J. Appl. Crystallogr.* 8:111–114.

von Laue, M. 1936. Die autsere Form der Kristalle in ihrem Einfluts auf die Interferenzerscheinungen an Raumgittern. *Ann. Phys.* 5(26):55–68.

Waller, J. 1923. Zur Frage der Einwirkung der Warmebewegung auf die Interferenz von Rontgenstrahlen. *Z. Phys.* 17:398–408.

Warren, B. E. 1969. X-Ray Diffraction. Addison-Wesley, Reading, Mass.

Warren, B. E., Averbach, B. L., and Roberts, B. W. 1951. Atomic size effect in the x-ray scattering in alloys. *J. Appl. Phys.* 22(12):1493–1496.

KEY REFERENCES

Cullity, 1978. See above.

Purpose of this book is to acquaint the reader who has little or no previous knowledge of the subject with the theory of x-ray diffraction, the experimental methods involved, and the main applications.

Guinier, 1994. See above.

Begins with the general theory of diffraction, and then applies this theory to various atomic structures, amorphous bodies, crystals, and imperfect crystals. Author has assumed that the reader is familiar with the elements of crystallography and x-ray diffraction. Should be especially useful for solid-state physicists, metallorgraphers, chemists, and even biologists.

International Table for Crystallography 1996. See above.

Purpose of this series is to collect and critically evaluate modern, advanced tables and texts on well-established topics that are relevant to crystallographic research and for applications of crystallographic methods in all sciences concerned with the structure and properties of materials.

James, 1948. See above.

Intended to provide an outline of the general optical principles underlying the diffraction of x rays by matter, which may serve as a foundation on which to base subsequent discussions of actual methods and results. Therefore, all details of actual techniques, and of their application to specific problems have been considered as lying beyond the scope of the book.

Klug, and Alexander, 1974. See above.

Contains details of many x-ray diffraction experimental techniques and analysis for powder and polycrystalline materials. Serves as textbook, manual, and teacher to plant workers, graduate students, research scientists, and others who seek to work in or understand the field.

Schultz, 1982. See above.

The thrust of this book is to convince the reader of the universality and utility of the scattering method in solving structural problems in materials science. This textbook is aimed at teaching the fundamentals of scattering theory and the broad scope of applications in solving real problems. It is intended that this book be augmented by additional notes dealing with experimental practice.

Schwartz, and Cohen, 1987. See above.

Covers an extensive list of topics with many examples. It deals with crystallography and diffraction for both perfect and imperfect crystals and contains an excellent set of advanced problem solving home works. Not intended for beginners, but serves the purpose of being an excellent reference for materials scientists who wish to seek solutions by means of diffraction techniques.

Warren, 1969. See above.

The emphasis of this book is a rigorous development of the basic diffraction theory. The treatment is carried far enough to relate to experimentally observable quantities. The main part of this book is devoted to the application of x-ray diffraction methods to both crystalline and amorphous materials, and to both perfect and imperfect crystals. This book is not intended for beginners.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

$\mathbf{s}_0$	incident x-ray direction
$\mathbf{s}$	scattered x-ray direction
$\mathbf{r}_n$	vector from origin to nth lattice point
$\mathbf{u}_n(t)$	time-dependent dynamic displacement vector
$\mathbf{K}$	scattering vector, reciprocal lattice space location
f	scattering power, length, or amplitude (of an atom relative to that of a single electron)
$I(\mathbf{K})$	scattering intensity
$F(\mathbf{K})$	structure factor
$G(\mathbf{K})$	interference function
N_i	number of unit cell i
δ_{ij}	Kronecker delta function
δ_{gj}	arbitrary phase factor
δ_p, δ_q	vector displacements
$\mathbf{H}$	reciprocal space vector, reciprocal lattice vector
I_e	Thomson scattering per electron
$\rho(\mathbf{r})$	electron density function, locally averaged scattering length intensity
M	Debye-Waller temperature factor
$\mathbf{g}$	lattice wave, propogation wave vector
a_{gj}	vibrational amplitude for the gj wave
$\mathbf{e}_{gj}$	polarization vector for the gj wave
k	Boltzmann's constant
ω_{gj}	eigenvalue of the phonon branches
$\mathbf{e}_{gj}$	eigenvector of the phonon branches
α_{pq}	Cowley-Warren parameter
$\mathbf{R}_{lmn}$	interatomic vector
Q	first-order size effects scattering component
R, S	second-order atomic displacement terms
R_G	radius of gyration
C_p	volume fraction of the dispersed particles

HAYDN CHEN
University of Illinois at Urbana-Champaign
Urbana, Illinois

DYNAMICAL DIFFRACTION

INTRODUCTION

Diffraction-related techniques using x rays, electrons, or neutrons are widely used in materials science to provide basic structural information on crystalline materials. To

describe a diffraction phenomenon, one has the choice of two theories: kinematic or dynamical.

Kinematic theory, described in KINEMATIC DIFFRACTION OF X RAYS, assumes that each x-ray photon, electron, or neutron scatters only once before it is detected. This assumption is valid in most cases for x rays and neutrons since their interactions with materials are relatively weak. This single-scattering mechanism is also called the first-order Born approximation or simply the Born approximation (Schiff, 1955; Jackson, 1975). The kinematic diffraction theory can be applied to a vast majority of materials studies and is the most commonly used theory to describe x-ray or neutron diffraction from crystals that are imperfect.

There are, however, practical situations where the higher-order scattering or multiple-scattering terms in the Born series become important and cannot be neglected. This is the case, for example, with electron diffraction from crystals, where an electron beam interacts strongly with electrons in a crystal. Multiple scattering can also be important in certain application areas of x-ray and neutron scattering, as described below. In all these cases, the simplified kinematic theory is not sufficient to evaluate the diffraction processes and the more rigorous dynamical theory is needed where multiple scattering is taken into account.

Application Areas

Dynamical diffraction is the predominant phenomenon in almost all electron diffraction applications, such as low-energy electron diffraction (LOW-ENERGY ELECTRON DIFFRACTION) and reflection high-energy electron diffraction. For x rays and neutrons, areas of materials research that involve dynamical diffraction may include the situations discussed in the next six sections.

Strong Bragg Reflections. For Bragg reflections with large structure factors, the kinematic theory often overestimates the integrated intensities. This occurs for many real crystals such as minerals and even biological crystals such as proteins, since they are not ideally imperfect. The effect is usually called the extinction (Warren, 1969), which refers to the extra attenuation of the incident beam in the crystal due to the loss of intensity to the diffracted beam. Its characteristic length scale, extinction length, depends on the structure factor of the Bragg reflection being measured. One can further categorize extinction effects into two types: primary extinction, which occurs within individual mosaic blocks in a mosaic crystal, and secondary extinction, which occurs for all mosaic blocks along the incident beam path. Primary extinction exists when the extinction length is shorter than the average size of mosaic blocks and secondary extinction occurs when the extinction length is less than the absorption length in the crystal.

Large Nearly Perfect Crystals and Multilayers. It is not uncommon in today's materials preparation and crystal growth laboratories that one has to deal with large nearly perfect crystals. Often, centimeter-sized perfect semiconductor crystals such as GaAs and Si are used as substrate materials, and multilayers and superlattices are deposited using molecular-beam or chemical vapor epitaxy. Bulk crystal growers are also producing larger high-quality crystals by advancing and perfecting various growth techniques. Characterization of these large nearly perfect crystals and multilayers by diffraction techniques often involves the use of dynamical theory simulations of the diffraction profiles and intensities. Crystal shape and its geometry with respect to the incident and the diffracted beams can also influence the diffraction pattern, which can only be accounted for by dynamical diffraction.

Topographic Studies of Defects. X-ray diffraction topography is a useful technique for studying crystalline defects such as dislocations in large-grain nearly perfect crystals (Chikawa and Kuriyama, 1991; Klapper, 1996; Tanner, 1996). With this technique, an extended highly collimated x-ray beam is incident on a specimen and an image of one or several strong Bragg reflections are recorded with high-resolution photographic films. Examination of the image can reveal micrometer (μm)-sized crystal defects such as dislocations, growth fronts, and fault lines. Because the strain field induced by a defect can extend far into the single-crystal grain, the diffraction process is rather complex and a quantitative interpretation of a topographic image frequently requires the use of dynamical theory and its variation on distorted crystals developed by Takagi (1962, 1969) and Taupin (1964).

Internal Field-Dependent Diffraction Phenomena. Several diffraction techniques make use of the secondary excitations induced by the wave field inside a crystal under diffraction conditions. These secondary signals may be x-ray fluorescence (X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS) or secondary electrons such as Auger (AUGER ELECTRON SPECTROSCOPY) or photoelectrons. The intensities of these signals are directly proportional to the electric field strength at the atom position where the secondary signal is generated. The wave field strength inside the crystal is a sensitive function of the crystal orientation near a specular or a Bragg reflection, and the dynamical theory is the only theory that provides the internal wave field amplitudes including the interference between the incident and the diffracted waves or the standing wave effect (Batterman, 1964). As a variation of the standing wave effect, the secondary signals can be diffracted by the crystal lattice and form standing wave-like diffraction profiles. These include Kossel lines for x-ray fluorescence (Kossel et al., 1935) and Kikuchi (1928) lines for secondary electrons. These effects can be interpreted as the optical reciprocity phenomena of the standing wave effect.

Multiple Bragg Diffraction Studies. If a single crystal is oriented in such a way that more than one reciprocal node falls on the Ewald sphere of diffraction, a simultaneous multiple-beam diffraction will occur. These

simultaneous reflections were first discovered by Renninger (1937) and are often called Renninger reflections or detour reflections (Umweganregung, "detour" in German). Although the angular positions of the simultaneous reflections can be predicted from simple geometric considerations in reciprocal space (Cole et al., 1962), a theoretical formalism that goes beyond the kinematic theory or the first-order Born approximation is needed to describe the intensities of a multiple-beam diffraction (Colella, 1974). Because of interference among the simultaneously excited Bragg beams, multiple-beam diffraction promises to be a practical solution to the phase problem in diffraction-based structural determination of crystalline materials, and there has been a great renewed interest in this research area (Shen, 1998; 1999a,b; Chang et al., 1999).

Grazing-Incidence Diffraction. In grazing-incidence diffraction geometry, either the incident beam, the diffracted beam, or both has an incident or exit angle, with respect to a well-defined surface, that is close to the critical angle of the diffracting crystal. Full treatment of the diffraction effects in a grazing-angle geometry involves Fresnel specular reflection and requires the concept of an evanescent wave that travels parallel to the surface and decays exponentially as a function of depth into the crystal. The dynamical theory is needed to describe the specular reflectivity and the evanescent wave-related phenomena. Because of its surface sensitivity and adjustable probing depth, grazing-incidence diffraction of x rays and neutrons has evolved into an important technique for materials research and characterization.

Brief Literature Survey

Dynamical diffraction theory of a plane wave by a perfect crystal was originated by Darwin (1914) and Ewald (1917), using two very different approaches. Since then the early development of the dynamical theory has primarily been focused on situations involving only an incident beam and one Bragg-diffracted beam, the so-called two-beam case. Prins (1930) extended Darwin's theory to take absorption into account, and von Laue (1931) reformulated Ewald's approach and formed the backbone of modern-day dynamical theory. Reviews and extensions of the theory have been given by Zachariasen (1945), James (1950), Kato (1952), Warren (1969), and Authier (1970). A comprehensive review of the Ewald–von Laue theory has been provided by Batterman and Cole (1964) in their seminal article in *Review of Modern Physics*. More recent reviews can be found in Kato (1974), Cowley (1975), and Pinsker (1978). Updated and concise summaries of the two-beam dynamical theory have been given recently by Authier (1992, 1996). A historical survey of the early development of the dynamical theory was given in Pinsker (1978).

Contemporary topics in dynamical theory are mainly focused in the following four areas: multiple-beam diffraction, grazing-incidence diffraction, internal fields and standing waves, and special x-ray optics. These modern developments are largely driven by recent interests in rapidly emerging fields such as synchrotron radiation, x-ray crystallography, surface science, and semiconductor research.

Dynamical theory of x rays for multiple-beam diffraction, with two or more Bragg reflections excited simultaneously, was considered by Ewald and Heno (1968). However, very little progress was made until Colella (1974) developed a computational algorithm that made multiple-beam x-ray diffraction simulations more tractable. Recent interests in its applications to measure the phases of structure factors (Colella, 1974; Post, 1977; Chapman et al., 1981; Chang, 1982) have made multiple-beam diffraction an active area of research in dynamical theory and experiments. Approximate theories of multiple-beam diffraction have been developed by Juretschke (1982, 1984, 1986), Hoier and Marthinsen (1983), Hümmer and Billy (1986), Shen (1986, 1999b,c), and Thorkildsen (1987). Reviews on multiple-beam diffraction have been given by Chang (1984, 1992, 1998), Colella (1995), and Weckert and Hümmer (1997).

Since the pioneer experiment by Marra et al. (1979), there has been an enormous increase in the development and use of grazing-incidence x-ray diffraction to study surfaces and interfaces of solids. Dynamical theory for the grazing-angle geometry was soon developed (Afanasev and Melkonyan, 1983; Aleksandrov et al., 1984) and its experimental verifications were given by Cowan et al. (1986), Durbin and Gog (1989), and Jach et al. (1989). Meanwhile, a semikinematic theory called the distorted-wave Born approximation was used by Vineyard (1982) and by Dietrich and Wagner (1983, 1984). This theory was further developed by Dosch et al. (1986) and Sinha et al. (1988), and has become widely utilized in grazing-incidence x-ray scattering studies of surfaces and near-surface structures. The theory has also been extended to explain standing-wave-enhanced and nonspecular scattering in multilayer structures (Kortright and Fischer-Colbrie, 1987), and to include phase-sensitive scattering in diffraction from bulk crystals (Shen, 1999b,c).

Direct experimental proof of the x-ray standing wave effect was first achieved by Batterman (1964) by observing x-ray fluorescence profiles while the diffracting crystal was rotated through a Bragg reflection. While earlier works were mainly on locating impurity atoms in bulk semiconductor materials (Batterman, 1969; Golovchenko et al., 1974; Anderson et al., 1976), more recent research activities focus on determinations of atom locations and distributions in overlayers above crystal surfaces (Golovchenko et al., 1982; Funke and Materlik, 1985; Durbin et al., 1986; Patel et al., 1987; Bedzyk et al., 1989), in synthetic multilayers (Barbee and Warburton, 1984; Kortright and Fischer-Colbrie, 1987), in long-period overlayers (Bedzyk et al., 1988; Wang et al., 1992), and in electrochemical solutions (Bedzyk et al., 1986). Recent reviews on x-ray standing waves are given by Patel (1996) and Lagomarsino (1996).

The rapid increase in synchrotron radiation-based materials research in recent years has spurred new developments in x-ray optics (Batterman and Bilderback, 1991; Hart, 1996). This is especially true in the areas of x-ray wave guides for producing submicron-sized beams (Bilderback et al., 1994; Feng et al., 1995), and x-ray phase plates and polarization analyzers used for studies on magnetic materials (Golovchenko et al., 1986; Mills, 1988; Belyakov

and Dmitrienko, 1989; Hirano et al., 1991; Batterman, 1992; Shen and Finkelstein, 1992; Giles et al., 1994; Yahnke et al., 1994; Shastri et al., 1995). Recent reviews on polarization x-ray optics have been given by Hirano et al. (1995), Shen (1996a), and Malgrange (1996).

An excellent collection of articles on these and other current topics in dynamical diffraction can be found in *X-ray and Neutron Dynamical Diffraction Theory and Applications* (Authier et al., 1996).

Scope of This Unit

Given the wide range of topics in dynamical diffraction, the main purpose of this unit is not to cover every detail but to provide readers with an overview of basic concepts, formalisms, and applications. Special attention is paid to the difference between the more familiar kinematic theory and the more complex dynamical approach. Although the basic dynamical theory is the same for x rays, electrons, and neutrons, we will focus mainly on x rays since much of the original terminology was founded in x-ray dynamical diffraction. The formalism for x rays is also more complex—and thus more complete—because of the vector-field nature of electromagnetic waves. For reviews on dynamical diffraction of electrons and neutrons, we refer the readers to an excellent textbook by Cowley (1975), Moodie et al. (1997), and a recent article by Schlenker and Guigay (1996).

We will start in the Basic Principles section with the fundamental equations and concepts in dynamical diffraction theory, which are derived from classical electrodynamics. Then, in the Two-Beam Diffraction section, we move onto the widely used two-beam approximation, essentially following the description of Batterman and Cole (1964). The two-beam theory deals only with the incident beam and one strongly diffracted Bragg beam, and the multiple scattering between them; multiple scattering due to other Bragg reflections are ignored. This theory provides many basic concepts in dynamical diffraction, and is very useful in visualizing the unique physical phenomena in dynamical scattering.

A full multiple-beam dynamical theory, developed by Colella (1974), takes into account all multiple-scattering effects and surface geometries as well as giving the most complete description of the diffraction processes of x rays, electrons, or neutrons in a perfect crystal. An outline of this theory is summarized in the Multiple-Beam Diffraction section. Also included in that section is an approximate formalism, given by Shen (1986), based on second-order Born approximations. This theory takes into account only double scattering in a multiple-scattering regime yet provides a useful picture of the physics of multiple-beam interactions. Finally, an approximate yet more accurate multiple-beam theory (Shen, 1999b) based on an expanded distorted-wave approximation is presented, which can provide accurate accounts of three-beam interference profiles in the so-called reference-beam diffraction geometry (Shen, 1998).

In the Grazing-Angle Diffraction section, the main results for grazing-incidence diffraction are described using the dynamical treatment. Of particular importance is the concept of evanescent waves and its applications. Also described in this section is a so-called distorted-wave Born approximation, which uses dynamical theory to evaluate specular reflections but treats surface diffraction and scattering within the kinematic regime. This approximate theory is useful in structural studies of surfaces and interfaces, thin films, and multilayered heterostructures.

Finally, because of limited space, a few topics are not covered in this unit. One of these is the theory by Takagi and Taupin for distorted perfect crystals. We refer the readers to the original articles (Takagi, 1962, 1969; Taupin, 1964) and to recent publications by Bartels et al. (1986) and by Authier (1996).

BASIC PRINCIPLES

There are two approaches to the dynamical theory. One, based on work by Darwin (1914) and Prins (1930), first finds the Fresnel reflectance and transmittance for a single atomic plane and then evaluates the total wave fields for a set of parallel atomic planes. The diffracted waves are obtained by solving a set of difference equations similar to the ones used in classical optics for a series of parallel slabs or optical filters. Although it had not been widely used for a long time due to its computational complexity, Darwin's approach has gained more attention in recent years as a means to evaluate reflectivities for multilayers and superlattices (Durbin and Follis, 1995), for crystal truncation effects (Caticha, 1994), and for quasicrystals (Chung and Durbin, 1995).

The other approach, developed by Ewald (1917) and von Laue (1931), treats wave propagation in a periodic medium as an eigenvalue problem and uses boundary conditions to obtain Bragg-reflected intensities. We will follow the Ewald–von Laue approach since many of the fundamental concepts in dynamical diffraction can be visualized more naturally by this approach and it can be easily extended to situations involving more than two beams.

In the early literature of dynamical theory (for two beams), the mathematical forms for the diffracted intensities from general absorbing crystals appear to be rather complicated. The main reason for these complicated forms is the necessity to separate out the real and imaginary parts in dealing with complex wave vectors and wave field amplitudes before the time of computers and powerful calculators. Today these complicated equations are not necessary and numerical calculations with complex variables can be easily performed on a modern computer. Therefore, in this unit, all final intensity equations are given in compact forms that involve complex numbers. In the author's view, these forms are best suited for today's computer calculations. These simpler forms also allow readers to gain physical insights rather than being overwhelmed by tedious mathematical notations.

Fundamental Equations

The starting point in the Ewald–von Laue approach to dynamical theory is that the dielectric function $\varepsilon(\mathbf{r})$ in a

crystalline material is a periodic function in space, and therefore can be expanded in a Fourier series:

$$\varepsilon(\mathbf{r}) = \varepsilon_0 + \delta\varepsilon(\mathbf{r}) \quad \text{with} \quad \delta\varepsilon(\mathbf{r}) = -\Gamma \sum_{\mathrm{H}} F_{\mathrm{H}}\, e^{-i\mathbf{H}\cdot\mathbf{r}} \qquad (1)$$

where $\Gamma = r_e\lambda^2/(\pi V_c)$ and $r_e = 2.818 \times 10^{-5}$ Å is the classical radius of an electron, λ is the x-ray wavelength, V_c is the unit cell volume, and $-\Gamma F_{\mathrm{H}}$ is the coefficient of the **H** Fourier component with F_{H} being the structure factor. All of the Fourier coefficients are on the order of 10^{-5} to 10^{-6} or smaller at x-ray wavelengths, $\delta\varepsilon(\mathbf{r}) \ll \varepsilon_0 = 1$, and the dielectric function is only slightly less than unity.

We further assume that a monochromatic plane wave is incident on a crystal, and the dielectric response is of the same wave frequency (elastic response). Applying Maxwell's equations and neglecting the magnetic interactions, we obtain the following equation for the electric field **E** and the displacement vector **D**:

$$(\nabla^2 + k_0^2)\mathbf{D} = -\nabla \times \nabla \times (\mathbf{D} - \varepsilon_0\mathbf{E}) \qquad (2)$$

where $\mathbf{k}_0$ is the wave vector of the monochromatic wave in vacuum, $k_0 = |\mathbf{k}_0| = 2\pi/\lambda$. For treatment involving magnetic interactions, we refer to Durbin (1987). If we assume an isotropic relation between $\mathbf{D}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$, $\mathbf{D}(\mathbf{r}) = \varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r})$, and $\delta\varepsilon(\mathbf{r}) \ll \varepsilon_0$, we have

$$(\nabla^2 + k_0^2)\mathbf{D} = -\nabla \times \nabla \times (\delta\varepsilon\mathbf{D}) \qquad (3)$$

We now use the periodic condition, Equation 1, and substitute for the wave field **D** in Equation 3 a series of Bloch waves with wave vectors $\mathbf{K}_{\mathrm{H}} = \mathbf{K}_0 + \mathbf{H}$,

$$\mathbf{D}(\mathbf{r}) = \sum_{\mathrm{H}} \mathbf{D}_{\mathrm{H}} e^{-i\mathbf{K}_{\mathrm{H}}\cdot\mathbf{r}} \qquad (4)$$

where **H** is a reciprocal space vector of the crystal. For every Fourier component (Bloch wave) **H**, we arrive at the following equation:

$$[(1 - \Gamma F_0)k_0^2 - K_{\mathrm{H}}^2]\mathbf{D}_{\mathrm{H}} = -\Gamma \sum_{\mathrm{G}\neq\mathrm{H}} F_{\mathrm{H-G}}\mathbf{K}_{\mathrm{H}} \times (\mathbf{K}_{\mathrm{H}} \times \mathbf{D}_{\mathrm{G}}) \quad (5)$$

where *H–G* is the difference reciprocal space vector between H and G, the terms involving Γ^2 have been neglected and $\mathbf{K}_{\mathrm{H}} \cdot \mathbf{D}_{\mathrm{H}}$ are set to zero because of the transverse wave nature of the electromagnetic radiation. Equation 5 forms a set of fundamental equations for the dynamical theory of x-ray diffraction. Similar equations for electrons and neutrons can be found in the literature (e.g., Cowley, 1975).

Dispersion Surface

A solution to the eigenvalue equation (Equation 5) gives rise to all the possible wave vectors $\mathbf{K}_{\mathrm{H}}$ and wave field amplitude ratios inside a diffracting crystal. The loci of the possible wave vectors form a multiple-sheet three-dimensional (3D) surface in reciprocal space. This surface is called the dispersion surface, as given by Ewald (1917).

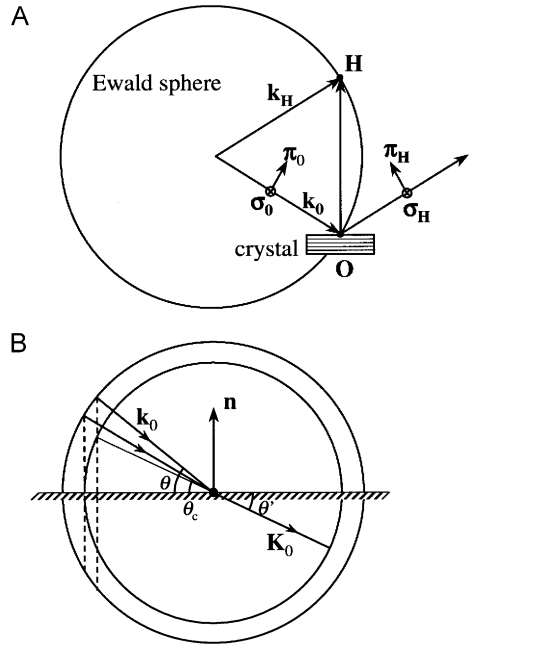


Figure 1. (**A**) Ewald sphere construction in kinematic theory and polarization vectors of the incident and the diffracted beams. (**B**) Dispersion surface in dynamical theory for a one-beam case and boundary conditions for total external reflection.

The introduction of the dispersion surface is the most significant difference between the kinematic and the dynamical theories. Here, instead of a single Ewald sphere (Fig. 1A), we have a continuous distribution of "Ewald spheres" with their centers located on the dispersion surface, giving rise to all possible traveling wave vectors inside the crystal.

As an example, we assume that the crystal orientation is far from any Bragg reflections, and thus only one beam, the incident beam $\mathbf{K}_0$, would exist in the crystal. For this "one-beam" case, Equation 5 becomes

$$[(1 - \Gamma F_0)k_0^2 - K_0^2]\mathbf{D}_0 = 0 \qquad (6)$$

Thus, we have

$$K_0 = k_0/(1 + \Gamma F_0)^{1/2} \cong k_0(1 - \Gamma F_0/2) \qquad (7)$$

which shows that the wave vector $\mathbf{K}_0$ inside the crystal is slightly shorter than that in vacuum as a result of the average index of refraction, $n = 1 - \Gamma F_0'/2$ where F_0' is the real part of F_0 and is related to the average density ρ_0 by

$$\rho_0 = \frac{\pi\Gamma F_0'}{r_e\lambda^2} \qquad (8)$$

In the case of absorbing crystals, $\mathbf{K}_0$ and F_0 are complex variables and the imaginary part, F_0'' of F_0, is related to the average linear absorption coefficient μ_0 by

$$\mu_0 = k_0\Gamma F_0'' = 2\pi\Gamma F_0''/\lambda \qquad (9)$$

Equation 7 shows that the dispersion surface in the one-beam case is a refraction-corrected sphere centered around the origin in reciprocal space, as shown in Figure 1B.

Boundary Conditions

Once Equation 5 is solved and all possible waves inside the crystal are obtained, the necessary connections between wave fields inside and outside the crystal are made through the boundary conditions. There are two types of boundary conditions in classical electrodynamics (Jackson, 1974). One states that the tangential components of the wave vectors have to be equal on both sides of an interface (Snell's law):

$$\mathbf{k}_t = \mathbf{K}_t \tag{10}$$

Throughout this unit, we use the convention that outside vacuum wave vectors are denoted by **k** and internal wave vectors are denoted by **K**, and the subscript t stands for the tangential component of the vector.

To illustrate this point, we again consider the simple one-beam case, as shown in Figure 1B. Suppose that an x-ray beam $\mathbf{k}_0$ with an angle θ is incident on a surface with **n** being its surface normal. To locate the proper internal wave vector $\mathbf{K}_0$, we follow along **n** to find its intersection with the dispersion surface, in this case, the sphere with its radius defined by Equation 7. However, we see immediately that this is possible only if θ is greater than a certain incident angle θ_c, which is the critical angle of the material. From Figure 1B, we can easily obtain that $\cos\theta_c = K_0/k_0$, or for small angles, $\theta_c = (\Gamma F_0)^{1/2}$. Below θ_c no traveling wave solutions are possible and thus total external reflection occurs.

The second set of boundary conditions states that the tangential components of the electric and magnetic field vectors, **E** and $\mathbf{H} = \hat{\mathbf{k}} \times \mathbf{E}$ ($\hat{\mathbf{k}}$ is a unit vector along the propagation direction), are continuous across the boundary. In dynamical theory literature, the eigenequations for dispersion surfaces are expressed in terms of either the electric field vector **E** or the electric displacement vector **D**. These two choices are equivalent, since in both cases a small longitudinal component on the order of ΓF_0 in the **E**-field vector is ignored, because its inclusion only contributes a term of Γ^2 in the dispersion equation. Thus **E** and **D** are interchangeable under this assumption and the boundary conditions can be expressed as the following:

$$\mathbf{D}_t^{in} = \mathbf{D}_t^{out} \tag{11a}$$

$$(\hat{\mathbf{k}} \times \mathbf{D}^{in})_t = (\hat{\mathbf{K}} \times \mathbf{D}^{out})_t \tag{11b}$$

In dynamical diffraction, the boundary condition, Equation 10, or Snell's law selects which points are excited on the dispersion surface or which waves actually exist inside the crystal for a given incident condition. The conditions, Equation 11a and Equation 11b, on the field vectors are then used to evaluate the actual internal field amplitudes and the diffracted wave intensities outside the crystal.

Dynamical theory covers a wide range of specific topics, which depend on the number of beams included in the dispersion equation, Equation 5, and the diffraction geometry of the crystal. In certain cases, the existence of some beams can be predetermined based on the physical law of energy conservation. In these cases, only Equation 11a is needed for the field boundary condition. Such is the case of conventional two-beam diffraction, as discussed in the Internal Fields section. However, both sets of conditions in Equation 11 are needed for general multiple-beam cases and for grazing-angle geometries.

Internal Fields

One of the important applications of dynamical theory is to evaluate the wave fields inside the diffracting crystal, in addition to the external diffracted intensities. Depending on the diffraction geometry, an internal field can be a periodic standing wave as in the case of a Bragg diffraction, an exponentially decayed evanescent wave as in the case of a specular reflection, or a combination of the two. Although no detectors per se can be put inside a crystal, the internal field effects can be observed in one of the following two ways.

The first is to detect secondary signals produced by an internal field, which include x-ray fluorescence (X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS), Auger electrons (AUGER ELECTRON SPECTROSCOPY), and photoelectrons. These inelastic secondary signals are directly proportional to the internal field intensity and are incoherent with respect to the internal field. Examples of this effect include the standard x-ray standing wave techniques and depth-sensitive x-ray fluorescence measurements under total external reflection.

The other way is to measure the elastic scattering of an internal field. In most cases, including the standing wave case, an internal field is a traveling wave along a certain direction, and therefore can be scattered by atoms inside the crystal. This is a coherent process, and the scattering contributions are added on the level of amplitudes instead of intensities. An example of this effect is the diffuse scattering of an evanescent wave in studies of surface or near-surface structures.

TWO-BEAM DIFFRACTION

In the two-beam approximation, we assume only one Bragg diffracted wave $\mathbf{K}_H$ is important in the crystal, in addition to the incident wave $\mathbf{K}_0$. Then, Equation 5 reduces to the following two coupled vector equations:

$$\begin{cases} [(1-\Gamma F_0)k_0^2 - K_0^2]\,\mathbf{D}_0 = -\Gamma F_{\bar{H}} \mathbf{K}_0 \times (\mathbf{K}_0 \times \mathbf{D}_H) \\ [(1-\Gamma F_0)k_0^2 - K_H^2]\,\mathbf{D}_H = -\Gamma F_{\bar{H}} \mathbf{K}_H \times (\mathbf{K}_H \times \mathbf{D}_0) \end{cases} \tag{12}$$

The wave vectors $\mathbf{K}_0$ and $\mathbf{K}_H$ define a plane that is usually called the scattering plane. If we use the coordinate system shown in Figure 1A, we can decompose the wave field amplitudes into σ and π polarization directions. Now the equations for the two polarization states decouple and can be solved separately

$$\begin{cases} [(1-\Gamma F_0)k_0^2 - K_0^2]\,D_{0\sigma,\pi} - k_0^2 \Gamma F_{\bar{H}} P D_{H\sigma,\pi} = 0 \\ -k_0^2 \Gamma F_H P D_{0\sigma,\pi} + [(1-\Gamma F_0)k_0^2 - K_H^2]\,D_{H\sigma,\pi} = 0 \end{cases} \tag{13}$$

where $P = \sigma_H \cdot \sigma_0 = 1$ for σ polarization and $P = \pi_H \cdot \pi_0 = \cos(2\theta_b)$ for π polarization, with θ_B being the Bragg angle. To seek nontrivial solutions, we set the determinant of Equation 13 to zero and solve for K_0:

$$\begin{vmatrix} (1-\Gamma F_0)k_0^2 - K_0^2 & -k_0^2\Gamma F_H^- P \\ -k_0^2\Gamma F_H P & (1-\Gamma F_0)k_0^2 - K_H^2 \end{vmatrix} = 0 \qquad (14)$$

where K_H^2 is related to $\mathbf{K}_0$ through Bragg's law, $K_H^2 = |\mathbf{K}_0 + \mathbf{H}|^2$. Solution of Equation 14 defines the possible wave vectors in the crystal and gives rise to the dispersion surface in the two-beam case.

Properties of Dispersion Surface

To visualize what the dispersion surface looks like in the two-beam case, we define two parameters ξ_0 and ξ_H, as described in James (1950) and Batterman and Cole (1964):

$$\begin{aligned} \xi_0 &\equiv [K_0^2 - (1-\Gamma F_0)k_0^2]/2k_0 = K_0 - k_0(1-\Gamma F_0/2) \\ \xi_H &\equiv [K_H^2 - (1-\Gamma F_0)k_0^2]/2k_0 = K_H - k_0(1-\Gamma F_0/2) \end{aligned} \qquad (15)$$

These parameters represent the deviations of the wave vectors inside the crystal from the average refraction-corrected values given by Equation 7. This also shows that in general the refraction corrections for the internal incident and diffracted waves are different. With these deviation parameters, the dispersion equation, Equation 14, becomes

$$\xi_0\xi_H = \frac{1}{4}k_0^2\Gamma^2P^2F_H^-F_H \qquad (16)$$

Hyperboloid Sheets. Since the right-hand side of Equation 16 is a constant for a given Bragg reflection, the dispersion surface given by this equation represents two sheets of hyperboloids in reciprocal space, for each polarization state P, as shown in Figure 2A. The hyperboloids have their diameter point, Q, located around what would be the center of the Ewald sphere (determined by Bragg's law) and asymptotically approach the two spheres centered at the origin **O** and at the reciprocal node **H**, with a refraction-corrected radius $k_0(1-\Gamma F_0/2)$. The two corresponding spheres in vacuum (outside crystal) are also shown and their intersection point is usually called the Laue point, L. The dispersion surface branches closer to the Laue point are called the α branches ($\alpha\sigma$, $\alpha\pi$), and those further from the Laue point are called the β branches ($\beta\sigma$, $\beta\pi$). Since the square-root value of the right-hand side constant in Equation 16 is much less than k_0, the gap at the diameter point is on the order of 10^{-5} compared to the radius of the spheres. Therefore, the spheres can be viewed essentially as planes in the vicinity of the diameter point, as illustrated in Figure 2B. However, the curvatures have to be considered when the Bragg reflection is in the grazing-angle geometry (see the section Grazing-Angle Diffraction).

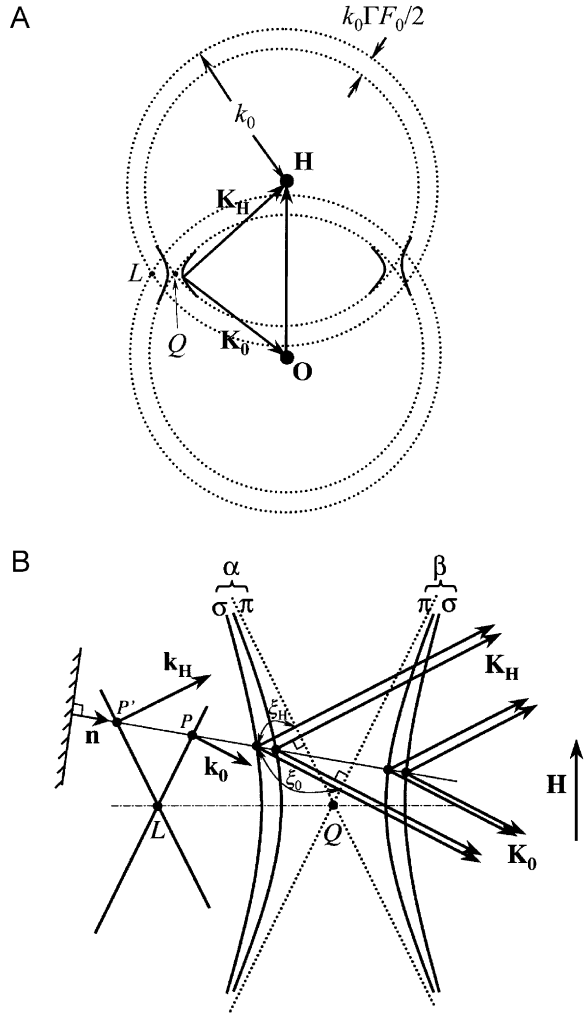


Figure 2. Dispersion surface in the two-beam case. (**A**) Overview. (**B**) Close-up view around the intersection region.

Wave Field Amplitude Ratios. In addition to wave vectors, the eigenvalue equation, Equation 13, also provides the ratio of the wave field amplitudes inside the crystal for each polarization. In terms of ξ_0 and ξ_H, the amplitude ratio is given by

$$D_H/D_0 = -2\xi_0/k_0\Gamma PF_H^- = -k_0\Gamma PF_H/2\xi_H \qquad (17)$$

Again, the actual ratio in the crystal depends entirely on the tie points selected by the boundary conditions. Around the diameter point, ξ_0 and ξ_H have similar lengths and thus the field amplitudes D_H and D_0 are comparable. Away from the exact Bragg condition, only one of ξ_0 and ξ_H has an appreciable size. Thus either D_0 or D_H dominates according to their asymptotic spheres.

Boundary Conditions and Snell's Law. To illustrate how tie points are selected by Snell's law in the two-beam case, we consider the situation in Figure 2B where a crystal surface is indicated by a shaded line. We start with an incident condition corresponding to an incident vacuum

wave vector $\mathbf{k}_0$ at point P. We then construct a surface normal passing through P and intersecting four tie points on the dispersion surface. Because of Snell's law, the wave fields associated with these four points are the only permitted waves inside the crystal. There are four waves for each reciprocal node, **O** or **H**; altogether a total of eight waves may exist inside the crystal in the two-beam case. To find the external diffracted beam, we follow the same surface normal to the intersection point P', and the corresponding wave vector connecting P' to the reciprocal node **H** would be the diffracted beam that we can measure with a detector outside the crystal.

Depending on whether or not a surface normal intercepts both α and β branches at the same incident condition, a diffraction geometry is called either the Laue transmission or the Bragg reflection case. In terms of the direction cosines γ_0 and γ_H of the external incident and diffracted wave vectors, $\mathbf{k}_0$ and $\mathbf{k}_H$, with respect to the surface normal **n**, it is useful to define a parameter b:

$$b \equiv \gamma_0/\gamma_H \equiv \mathbf{k}_0 \cdot \mathbf{n}/\mathbf{k}_H \cdot \mathbf{n} \tag{18}$$

where $b > 0$ corresponds to the Laue case and $b < 0$ the Bragg case. The cases with $b = \pm 1$ are called the symmetric Laue or Bragg cases, and for that reason b is often called the asymmetry factor.

Poynting Vector and Energy Flow. The question about the energy flow directions in dynamical diffraction is of fundamental interests to scientists who use x-ray topography to study defects in perfect crystals. Energy flow of an electromagnetic wave is determined by its time-averaged Poynting vector, defined as

$$\mathbf{S} = \frac{c}{8\pi}(\mathbf{E} \times \mathbf{H}^*) = \frac{c}{8\pi}|\mathbf{D}|^2\hat{\mathbf{K}} \tag{19}$$

where c is the speed of light, $\hat{\mathbf{K}}$ is a unit vector along the propagation direction, and terms on the order of Γ or higher are ignored. The total Poynting vector $\mathbf{S}_T$ at each tie point on each branch of the dispersion surfaces is the vector sum of those for the **O** and **H** beams

$$\mathbf{S}_T = \frac{c}{8\pi}(D_0^2\hat{\mathbf{K}}_0 + D_H^2\hat{\mathbf{K}}_H) \tag{20}$$

To find the direction of $\mathbf{S}_T$, we consider the surface normal **v** of the dispersion branch, which is along the direction of the gradient of the dispersion equation, Equation 16:

$$\begin{aligned}\mathbf{v} = \nabla(\xi_0\xi_H) = \xi_0\nabla\xi_H + \xi_H\nabla\xi_0 = \frac{\xi_0}{\xi_H}\hat{\mathbf{K}}_H + \frac{\xi_H}{\xi_0}\hat{\mathbf{K}}_0 \\ \propto D_0^2\hat{\mathbf{K}}_0 + D_H^2\hat{\mathbf{K}}_H \propto \mathbf{S}_T\end{aligned} \tag{21}$$

where we have used Equation 17 and assumed a negligible absorption ($|F_{\bar{H}} = |F_H|$). Thus we conclude that $\mathbf{S}_T$ is parallel to **v**, the normal to the dispersion surface. In other words, the total energy flow at a given tie point is always normal to the local dispersion surface. This important theorem is generally valid and was first proved by Kato (1960). It follows that the energy flow inside the crystal is parallel to the atomic planes at the full excitation condition, that is, the diameter points of the hyperboloids.

Special Dynamical Effects

There are significant differences in the physical diffraction processes between kinematic and dynamical theory. The most striking observable results from the dynamical theory are Pendellösung fringes, anomalous transmission, finite reflection width for semi-infinite crystals, x-ray standing waves, and x-ray birefringence. With the aid of the dispersion surface shown in Figure 2, these effects can be explained without formally solving the mathematical equations.

Pendellösung. In a Laue case, the α and β tie points across the diameter gap of the hyperbolic dispersion surfaces are excited simultaneously at a given incident condition. The two sets of traveling waves associated with the two branches can interfere with each other and cause oscillations in the diffracted intensity as the thickness of the crystal changes on the order of $2\pi/\Delta K$, where ΔK is simply the gap at the diameter point. These intensity oscillations are termed Pendellösung fringes and the quantity $2\pi/\Delta K$ is called the Pendellösung period. From the geometry shown in Figure 2B, it is straightforward to show that the diameter gap is given by

$$\Delta K = k_0\Gamma|P|\sqrt{F_H F_{\bar{H}}}/\cos\theta_B \tag{22}$$

where θ_B is the internal Bragg angle. As an example, at 10 keV, for Si(111) reflection, $\Delta K = 2.67 \times 10^{-5}$ Å^{-1}, and thus the Pendellösung period is equal to 23 μm.

Pendellösung interference is a unique diffraction phenomenon for the Laue geometry. Both the diffracted wave (**H** beam) and the forward-diffracted wave (**O** beam) are affected by this effect. The intensity oscillations for these two beams are 180° out of phase to each other, creating the effect of energy flow swapping back and forth between the two directions as a function of depth into the crystal surface. For more detailed discussions of Pendellösung fringes we refer to a review by Kato (1974).

We should point out that Pendellösung fringes are entirely different in origin from interference fringes due to crystal thickness. The thickness fringes are often observed in reflectivity measurements on thin film materials and can be mostly accounted for by a finite size effect in the Fraunhofer diffraction. The period of thickness fringes depends only on crystal thickness, not on the strength of the reflection, while the Pendellösung period depends only on the reflection strength, not on crystal thickness.

Anomalous Transmission. The four waves selected by tie points in the Laue case have different effective absorption coefficients. This can be understood qualitatively from the locations of the four dispersion surface branches relative to the vacuum Laue point L and to the average refraction-corrected point Q. The β branches are further from L and are on the more refractive side of Q. Therefore the waves associated with the β branches have larger than average refraction and absorption. The α branches, on the other

hand, are located closer to L and are on the less refractive side of Q. Therefore the waves on the α branches have less than average refraction and absorption. For a relatively thick crystal in the Laue diffraction geometry, the α waves would effectively be able to pass through the thickness of the crystal more easily than would an average wave. What this implies is that if the intensity is not observed in the transmitted beam at off-Bragg conditions, an anomalously "transmitted" intense beam can actually appear when the crystal is set to a strong Bragg condition. This phenomenon is called anomalous transmission; it was first observed by Borrmann (1950) and is also called the Borrmann effect.

If the Laue crystal is sufficiently thick, then even the $\alpha\pi$ wave may be absorbed and only the $\alpha\sigma$ wave will remain. In this case, the Laue-diffracting crystal can be used as a linear polarizer since only the σ-polarized x rays will be transmitted through the crystal.

Darwin Width. In Bragg reflection geometry, all the excited tie points lie on the same branch of the dispersion surface at a given incident angle. Furthermore, no tie points can be excited at the center of a Bragg reflection, where a gap exists at the diameter point of the dispersion surfaces. The gap indicates that no internal traveling waves exist at the exact Bragg condition and total external reflection is the only outlet of the incident energy if absorption is ignored. In fact, the size of the gap determines the range of incident angles at which the total reflection would occur. This angular width is usually called the Darwin width of a Bragg reflection in perfect crystals. In the case of symmetric Bragg geometry, it is easy to see from Figure 2 that the full Darwin width is

$$w = \frac{\Delta K}{k_0 \sin \theta_B} = \frac{2\Gamma|P|\sqrt{F_H F_{\overline{H}}}}{\sin 2\theta_B} \tag{23}$$

Typical values for w are on the order of a few arc-seconds.

The existance of a finite reflection width w, even for a semi-infinite crystal, may seem to contradict the mathematical theory of Fourier transforms that would give rise to a zero reflection width if the crystal size is infinite. In fact, this is not the case. A more careful examination of the situation shows that because of the extinction the incident beam would never be able to see the whole "infinite" crystal. Thus the finite Darwin width is a direct result of the extinction effect in dynamical theory and is needed to conserve the total energy in the physical system.

X-ray Standing Waves. Another important effect in dynamical diffraction is the x-ray standing waves (XSWs) (Batterman, 1964). Inside a diffracting crystal, the total wave field intensity is the coherent sum of the **O** and **H** beams and is given by (σ polarization)

$$|\mathbf{D}|^2 = |\mathbf{D}_0 e^{-i\mathbf{K}_0\cdot\mathbf{r}} + \mathbf{D}_H e^{-i\mathbf{K}_H\cdot\mathbf{r}}|^2 = |\mathbf{D}_0|^2 \times \left|1 + \frac{D_H}{D_0} e^{-i\mathbf{H}\cdot\mathbf{r}}\right|^2 \tag{24}$$

Equation 24 represents a standing wave field with a spatial period of $2\pi/|\mathbf{H}|$, which is simply the d spacing of the Bragg reflection. The field amplitude ratio D_H/D_0 has well-defined phases at α and β branches of the dispersion surface. According to Equation 17 and Figure 2, we see that the phase of D_H/D_0 is $\pi + \alpha_H$ at the α branch, since ξ_H is positive, and is α_H at the β branch, since ξ_H is negative, where α_H is the phase of the structure factor F_H and can be set to zero by a proper choice of real-space origin. Thus the α mode standing wave has its nodes on the atomic planes and the β mode standing wave has its antinodes on the atomic planes.

In Laue transmission geometry, both the α and the β modes are excited simultaneously in the crystal. However, the β mode standing wave is attenuated more strongly because its peak field coincides with the atomic planes. This is the physical origin of the Borrmann anomalous absorption effect.

The standing waves also exist in Bragg geometry. Because of its more recent applications in materials studies, we will devote a later segment (Standing Waves) to discuss this in more detail.

X-ray Birefringence. Being able to produce and to analyze a generally polarized electromagnetic wave has long benefited scientists and researchers in the field of visible-light optics and in studying optical properties of materials. In the x-ray regime, however, such abilities have been very limited because of the weak interaction of x rays with matter, especially for production and analysis of circularly polarized x-ray beams. The situation has changed significantly in recent years. The growing interest in studying magnetic and anisotropic electronic materials by x-ray scattering and spectroscopic techniques have initiated many new developments in both the production and the analyses of specially polarized x rays. The now routinely available high-brightness synchrotron radiation sources can provide naturally collimated x rays that can be easily manipulated by special x-ray optics to generate x-ray beams with polarization tunable from linear to circular. Such optics are usually called x-ray phase plates or phase retarders.

The principles of most x-ray phase plates are based on the linear birefringence effect near a Bragg reflection in perfect or nearly perfect crystals due to dynamical diffraction (Hart, 1978; Belyakov and Dmitrienko, 1989). As illustrated in Figure 2, close to a Bragg reflection **H**, the lengths of the wave vectors for the σ and the π polarizations are slightly different. The difference can cause a phase shift Δ between the σ and the π wave fields to accumulate through the crystal thickness t: $\Delta = (K_\sigma - K_\pi)t$. When the phase shift Δ reaches $\pm 90^\circ$, circularly polarized radiation is generated, and such a device is called a quarter-wave phase plate or retarder (Mills, 1988; Hirano et al., 1991; Giles et al., 1994). In addition to these transmission-type phase retarders, a reflection-type phase plate also has been proposed and studied (Brummer et al., 1984; Batterman, 1992; Shastri et al., 1995), which has the advantage of being thickness independent. However, it has been demonstrated that the Bragg transmission-type phase retarders are more robust to incident beam divergences and thus are very practical x-ray circular polarizers. They have been used for measurements of magnetic dichroism in hard permanent

magnets and other magnetic materials (Giles et al., 1994; Lang et al., 1995). Recent reviews on x-ray polarizers and phase plates can be found in articles by Hart (1991), Hirano et al. (1995), Shen (1996a), and Malgrange (1996).

Solution of the Dispersion Equation

So far we have confined our discussions to the physical effects that exist in dynamical diffraction from perfect crystals and have tried to avoid the mathematical details of the solutions to the dispersion equation, Equation 11 or 12. As we have shown, considerable physical insight concerning the diffraction processes can be gained without going into mathematical details. To obtain the diffracted intensities in dynamical theory, however, the mathematical solutions are unavoidable. In the summary of these results that follows, we will keep the formulas in a general complex form so that absorption effects are automatically taken into account.

The key to solving the dispersion equations (Equation 14 or 16) is to realize that the internal incident beam $\mathbf{K}_0$ can only differ from the vacuum incident beam $\mathbf{k}_0$ by a small component K_{0n} along the surface normal direction of the incident surface, which in turn is linearly related to ξ_0 or ξ_H. The final expression reduces to a quadratic equation for ξ_0 or ξ_H, and solving for ξ_0 or ξ_H alone results in the following (Batterman and Cole, 1964):

$$\xi_0 = \frac{1}{2}k_0|P|\Gamma\sqrt{|b|F_H F_{\bar{H}}}\left[\eta \pm (\eta^2 + b/|b|)^{1/2}\right] \qquad (25)$$

where η is the reduced deviation parameter normalized to the Darwin width

$$\eta \equiv \frac{2b}{w\sqrt{|b|}}(\Delta\theta - \Delta\theta_0) \qquad (26)$$

$\Delta\theta = \theta - \theta_B$ is the angular deviation from the vacuum Bragg angle θ_B, and $\Delta\theta_0$ is the refraction correction

$$\Delta\theta_0 \equiv \frac{\Gamma F_0(1 - 1/b)}{2\sin 2\theta_B} \qquad (27)$$

The dual signs in Equation 25 correspond to the α and β branches of the dispersion surface. In the Bragg case, $b < 0$ so the correction $\Delta\theta_0$ is always positive—that is, the θ value at the center of a reflection is always slightly larger than θ_B given by the kinematic theory. In the Laue case, the sign of $\Delta\theta_0$ depends on whether $b > 1$ or $b < 1$. In the case of absorbing crystals, both η and $\Delta\theta_0$ can be complex, and the directional properties are represented by the real parts of these complex variables while their imaginary parts are related to the absorption given by F_0'' and w.

Substituting Equation 25 into Equation 17 yields the wave field amplitude ratio inside the crystal as a function of η

$$\frac{D_H}{D_0} = -\frac{|P|}{P}\sqrt{|b|F_H/F_{\bar{H}}}[\eta \pm (\eta^2 + b/|b|)^{1/2}] \qquad (28)$$

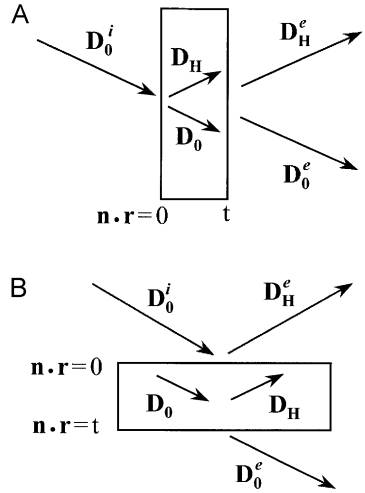


Figure 3. Boundary conditions for the wave fields outside the crystal in (**A**) Laue case and (**B**) Bragg case.

Diffracted Intensities

We now employ boundary conditions to evaluate the diffracted intensities.

Boundary Conditions. In the Laue transmission case (Fig. 3A), assuming a plane wave with an infinite cross-section, the field boundary conditions are given by the following equations:

$$\text{Entrance surface: } \begin{cases} \mathbf{D}_0^i = \mathbf{D}_{0\alpha} + \mathbf{D}_{0\beta} \\ 0 = \mathbf{D}_{H\alpha} + \mathbf{D}_{H\beta} \end{cases} \qquad (29)$$

$$\text{Exit surface: } \begin{cases} \mathbf{D}_0^e = \mathbf{D}_{0\alpha}e^{-i\mathbf{K}_{0\alpha}\cdot\mathbf{r}} + \mathbf{D}_{0\beta}e^{-i\mathbf{K}_{0\beta}\cdot\mathbf{r}} \\ \mathbf{D}_H^e = \mathbf{D}_{H\alpha}e^{-i\mathbf{K}_{H\alpha}\cdot\mathbf{r}} + \mathbf{D}_{H\beta}e^{-i\mathbf{K}_{H\beta}\cdot\mathbf{r}} \end{cases} \qquad (30)$$

In the Bragg reflection case (Fig. 3B), the field boundary conditions are given by

$$\text{Entrance surface: } \begin{cases} \mathbf{D}_0^i = \mathbf{D}_{0\alpha} + \mathbf{D}_{0\beta} \\ \mathbf{D}_H^e = \mathbf{D}_{H\alpha} + \mathbf{D}_{H\beta} \end{cases} \qquad (31)$$

$$\text{Back surface: } \begin{cases} \mathbf{D}_0^e = \mathbf{D}_{0\alpha}e^{-i\mathbf{K}_{0\alpha}\cdot\mathbf{r}} + \mathbf{D}_{0\beta}e^{-i\mathbf{K}_{0\beta}\cdot\mathbf{r}} \\ 0 = \mathbf{D}_{H\alpha}e^{-i\mathbf{K}_{H\alpha}\cdot\mathbf{r}} + \mathbf{D}_{H\beta}e^{-i\mathbf{K}_{H\beta}\cdot\mathbf{r}} \end{cases} \qquad (32)$$

In either case, there are six unknowns, $\mathbf{D}_{0\alpha}$, $\mathbf{D}_{0\beta}$, $\mathbf{D}_{H\alpha}$, $\mathbf{D}_{H\beta}$, $\mathbf{D}_0^e$, $\mathbf{D}_H^e$, and three pairs of equations, Equations 28, 29, 30, or Equations 28, 31, 32, for each polarization state. Our goal is to express the diffracted waves $\mathbf{D}_H^e$ outside the crystal as a function of the incident wave $\mathbf{D}_0^i$.

Intensities in the Laue Case. In the Laue transmission case, we obtain, apart from an insignificant phase factor,

$$\mathbf{D}_H^e = \mathbf{D}_0^i e^{-\mu_0 t/4(1/\gamma_0 + 1/\gamma_H)}\sqrt{\left|\frac{bF_H}{F_{\bar{H}}}\right|}\frac{\sin(A\sqrt{\eta^2+1})}{\sqrt{\eta^2+1}} \qquad (33)$$

where A is the effective thickness (complex) that relates to real thickness t by (Zachariasen, 1945)

$$A \equiv \frac{\pi\Gamma|P|t\sqrt{F_{\mathrm{H}}F_{\overline{\mathrm{H}}}}}{\lambda\sqrt{|\gamma_0\gamma_{\mathrm{H}}|}} \tag{34}$$

The real part of A is essentially the ratio of the crystal thickness to the Pendellösung period.

A quantity often measured in experiments is the total power P_{H} in the diffracted beam, which is equal to the diffracted intensity multiplied by the cross-section area of the beam. The power ratio P_{H}/P_0 of the diffracted beam to the incident beam is given by the intensity ratio, $|\mathbf{D}_{\mathrm{H}}^{\mathrm{e}}/\mathbf{D}_0^{\mathrm{i}}|^2$ multiplied by the area ratio, $1/|b|$, of the beam cross-sections

$$\frac{P_{\mathrm{H}}}{P_0} = \frac{1}{|b|}\left|\frac{\mathbf{D}_{\mathrm{H}}^{\mathrm{e}}}{\mathbf{D}_0^{\mathrm{i}}}\right|^2 = e^{-\mu_0 t/2[1/\gamma_0+1/\gamma_{\mathrm{H}}]}\left|\frac{F_{\mathrm{H}}}{F_{\overline{\mathrm{H}}}}\right| \times \frac{|\sin(A\sqrt{\eta^2+1})|^2}{|\eta^2+1|} \tag{35}$$

A plot of P_{H}/P_0 versus η is usually called the rocking curve. Keeping in mind that η can be a complex variable due essentially to F_0'', Equation 35 is a general expression that is valid for both nonabsorbing and absorbing crystals. A few examples of the rocking curves in the Laue case for nonabsorbing crystals are shown in Figure 4A.

For thick nonabsorbing crystals, A is large ($A \gg 1$) so the $\sin^2$ oscillations tend to average to a value equal to $\frac{1}{2}$. Thus, Equation 35 reduces to a simple Lorentzian shape

$$\frac{P_{\mathrm{H}}}{P_0} = \frac{1}{2(\eta^2+1)} \tag{36}$$

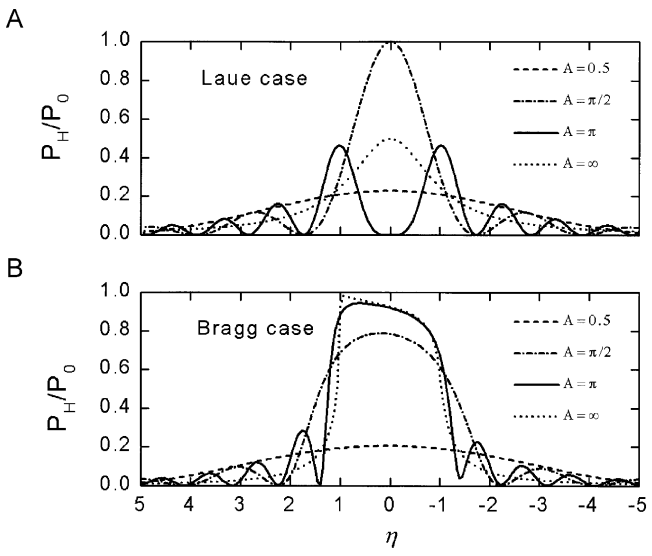


Figure 4. Diffracted intensity P_{H}/P_0 in (**A**) nonabsorbing Laue case, and (**B**) absorbing Bragg case, for several effective thicknesses. The Bragg reflection in (**B**) is for GaAs(220) at a wavelength 1.48 Å.

For thin nonabsorbing crystals ($A \ll 1$), we rewrite Equation 35 in the following form:

$$\frac{P_{\mathrm{H}}}{P_0} = \left[\frac{\sin(A\sqrt{\eta^2+1})}{\sqrt{\eta^2+1}}\right]^2 \approx \left[\frac{\sin(A\eta)}{\eta}\right]^2 \tag{37}$$

This approximation (Equation 37) can be realized by expanding the quantities in the square brackets on both sides to third power and neglecting the A^3 term since $A \ll 1$. We see that in this thin-crystal limit, dynamical theory gives the same result as kinematic theory. The condition $A \ll 1$ can be restated as the crystal thickness t is much less than the Pendellösung period.

Intensities in the Bragg Case. In the Bragg reflection case, the diffracted wave field is given by

$$\mathbf{D}_{\mathrm{H}}^{\mathrm{e}} = \mathbf{D}_0^{\mathrm{i}}\sqrt{\left|\frac{bF_{\mathrm{H}}}{F_{\overline{\mathrm{H}}}}\right|}\frac{1}{\eta + i\sqrt{\eta^2-1}\cot(A\sqrt{\eta^2-1})} \tag{38}$$

The power ratio P_{H}/P_0 of the diffracted beam to the incident, often called the Bragg reflectivity, is

$$\frac{P_{\mathrm{H}}}{P_0} = \left|\frac{F_{\mathrm{H}}}{F_{\overline{\mathrm{H}}}}\right|\frac{1}{|\eta + i\sqrt{\eta^2-1}\cot(A\sqrt{\eta^2-1})|^2} \tag{39}$$

In the case of thick crystals ($A \gg 1$), Equation 39 reduces to

$$\frac{P_{\mathrm{H}}}{P_0} = \left|\frac{F_{\mathrm{H}}}{F_{\overline{\mathrm{H}}}}\right|\frac{1}{|\eta \pm \sqrt{\eta^2-1}|^2} \tag{40}$$

The choice of the signs is such that a smaller value of P_{H}/P_0 is retained.

On the other hand, for semi-infinite crystals ($A \gg 1$), we can go back to the boundary conditions, Equations 31 and 32, and ignore the back surface altogether. If we then apply the argument that only one of the two tie points on each branch of the dispersion surface is physically feasible in the Bragg case because of the energy flow conservation, we arrive at the following simple boundary condition:

$$\mathbf{D}_0^{\mathrm{i}} = \mathbf{D}_0 \qquad \mathbf{D}_{\mathrm{H}}^{\mathrm{e}} = \mathbf{D}_{\mathrm{H}} \tag{41}$$

By using Equations 41 and 28, the diffracted power can be expressed by

$$\frac{P_{\mathrm{H}}}{P_0} = \left|\frac{F_{\mathrm{H}}}{F_{\overline{\mathrm{H}}}}\right|\left|\eta \pm \sqrt{\eta^2-1}\right|^2 \tag{42}$$

Again the sign in front of the square root is chosen so that P_{H}/P_0 is less than unity. The result is obviously identical to Equation 40.

Far away from the Bragg condition, $\eta \gg 1$, Equation 40 shows that the reflected power decreases as $1/\eta^2$. This asymptotic form represents the "tails" of a Bragg reflection (Andrews and Cowley, 1985), which are also called the crystal truncation rod in kinematic theory (Robinson,

1986). In reciprocal space, the direction of the tails is along the surface normal since the diffracted wave vector can only differ from the Bragg condition by a component normal to the surface or interface. More detailed discussions of the crystal truncation rods in dynamical theory can be found in Colella (1991), Caticha (1993, 1994), and Durbin (1995).

Examples of the reflectivity curves, Equation 39, for a GaAs crystal with different thicknesses in the symmetric Bragg case are shown in Figure 4B. The oscillations in the tails are entirely due to the thickness of the crystal. These modulations are routinely observed in x-ray diffraction profiles from semiconductor thin films on substrates and can be used to determine the thin-film thickness very accurately (Fewster, 1996).

Integrated Intensities. The integrated intensity R_H^η in the reduced η units is given by integrating the diffracted power ratio P_H/P_0 over the entire η range.

For nonabsorbing crystals in the Laue case, in the limiting cases of $A \ll 1$ and $A \gg 1$, R_H^η can be calculated analytically as (Zachariasen, 1945)

$$R_H^\eta = \int_{-\infty}^{\infty} \left(\frac{P_H}{P_0}\right) d\eta = \begin{cases} \pi A, & A \ll 1 \\ \pi/2, & A \gg 1 \end{cases} \tag{43}$$

For intermediate values of A or for absorbing crystals, the integral can only be calculated numerically. A general plot of R_H^η versus A in the nonabsorbing case is shown in Figure 5 as the dashed line.

For nonabsorbing crystals in the Bragg case, Equation 39 can be integrated analytically (Darwin, 1922) to yield

$$R_H^\eta = \int_{-\infty}^{\infty} \left(\frac{P_H}{P_0}\right) d\eta = \pi \tanh(A) = \begin{cases} \pi A, & A \ll 1 \\ \pi, & A \gg 1 \end{cases} \tag{44}$$

A plot of the integrated power in the symmetric Bragg case is shown in Figure 5 as the solid curve. Both curves in Figure 5 show a linear behavior for small A, which is consistent with kinematic theory.

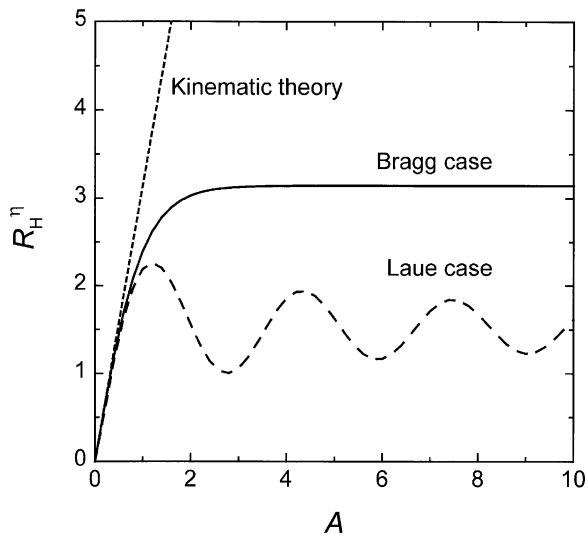


Figure 5. Comparison of integrated intensities in the Laue case and the Bragg case with the kinematic theory.

If we use the definitions of η and A, we obtain that the integrated power R_H^θ over the incident angle θ in the limit of $A \ll 1$ is given by

$$R_H^\theta = \int_{-\infty}^{\infty} \left(\frac{P_H}{P_0}\right) d\theta = \frac{w}{2} R_H^\eta = \frac{\pi}{2} wA = \frac{r_e^2 \lambda^3 P^2 |F_H|^2 t}{V_c \sin 2\theta_B} \tag{45}$$

which is identical to the integrated intensity in the kinematic theory for a small crystal (Warren, 1969). Thus in some sense kinematic theory is a limiting form of dynamical theory, and the departures of the integrated intensities at larger A values (Fig. 5) is simply the effect of primary extinction. In the thick-crystal limit $A \gg 1$, the θ-integrated intensity R_H^θ in both Laue and Bragg cases is linear in $|F_H|$. This linear rather than quadratic dependence on $|F_H|$ is a distinct and characteristic result of dynamical diffraction.

Standing Waves

As we discussed earlier, near or at a Bragg reflection, the wave field amplitudes, Equation 24, represent standing waves inside the diffracting crystal. In the Bragg reflection geometry, as the incident angle increases through the full Bragg reflection, the selected tie points shift from the α branch to the β branch. Therefore the nodes of the standing wave shift from on the atomic planes ($r = 0$) to in between the atomic planes ($r = d/2$) and the corresponding antinodes shift from in between to on the atomic planes.

For a semi-infinite crystal in the symmetric Bragg case and σ polarization, the standing wave intensity can be written, using Equations 24, 28, and 42, as

$$I = \left| 1 + \sqrt{\frac{P_H}{P_0}} e^{i(\nu + \alpha_H - \mathbf{H}\cdot\mathbf{r})} \right|^2 \tag{46}$$

where ν is the phase of $\eta \pm \sqrt{\eta^2 - 1}$ and α_H is the phase of the structure factor F_H, assuming absorption is negligible. If we define the diffraction plane by choosing an origin such that α_H is zero, then the standing wave intensity as a function of η is determined by the phase factor $\mathbf{H} \cdot \mathbf{r}$ with respect to the origin chosen and the d spacing of the Bragg reflection (Bedzyk and Materlik, 1985). Typical standing wave intensity profiles given by Equation 46 are shown in Figure 6. The phase variable ν and the corresponding reflectivity curve are also shown in Figure 6.

An XSW profile can be observed by measuring the x-ray fluorescence from atoms embedded in the crystal structure since the fluorescence signal is directly proportional to the internal wave field intensity at the atom position (Batterman, 1964). By analyzing the shape of a fluorescence profile, the position of the fluorescing atom with respect to the diffraction plane can be determined. A detailed discussion of nodal plane position shifts of the standing waves in general absorbing crystals has been given by Authier (1986).

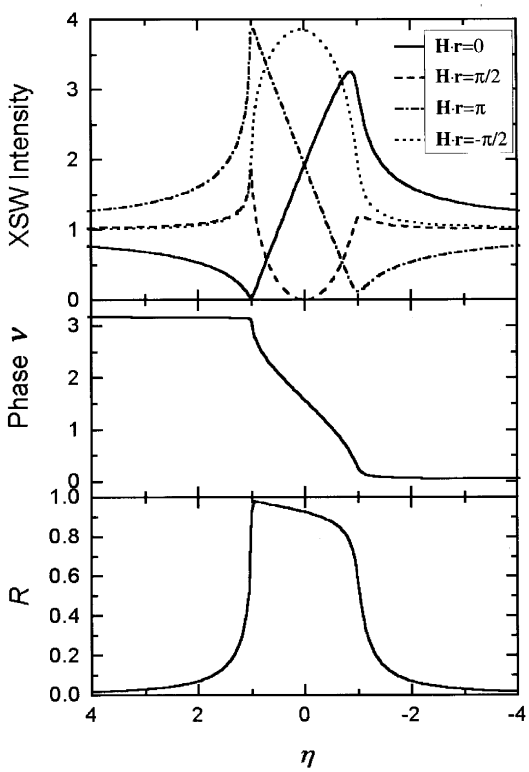


Figure 6. XSW intensity and phase as a function of reduced angular parameter η, along with reflectivity curve, calculated for a semi-infinite GaAs(220) reflection at 1.48 Å.

The standing wave technique has been used to determine foreign atom positions in bulk materials (Batterman, 1969; Golovchenko et al., 1974; Lagomarsino et al., 1984; Kovalchuk and Kohn, 1986). Most recent applications of the XSW technique have been the determination of foreign atom positions, surface relaxations, and disorder at crystal surfaces and interfaces (Durbin et al., 1986; Zegenhagen et al., 1988; Bedzyk et al., 1989; Martines et al., 1992; Fontes et al., 1993; Franklin et al., 1995; Lyman and Bedzyk, 1997). By measuring standing wave patterns for two or more reflections (either separately or simultaneously) along different crystallographic axes, atomic positions can be triangulated in space (Greiser and Materlik, 1986; Berman et al., 1988). More details of the XSW technique can be found in recent reviews by Patel (1996) and Lagomarsino (1996).

The formation of XSWs is not restricted to wide-angle Bragg reflections in perfect crystals. Bedzyk et al. (1988) extended the technique to the regime of specular reflections from mirror surfaces, in which case both the phase and the period of the standing waves vary with the incident angle. Standing waves have also been used to study the spatial distribution of atomic species in mosaic crystals (Durbin, 1998) and quasicrystals (Chung and Durbin, 1995; Jach et al., 1999). Due to a substantial (although imperfect) standing wave formation, anomalous transmission has been observed on the strongest diffraction peaks in nearly perfect quasicrystals (Kycia et al., 1993).

MULTIPLE-BEAM DIFFRACTION

So far, we have restricted our discussion to diffraction cases in which only the incident beam and one Bragg-diffracted beam are present. There are experimental situations, however, in which more than one diffracted beam may be significant and therefore the two-beam approximation is no longer valid. Such situations involving multiple-beam diffraction are dealt with in this section.

Basic Concepts

Multiple-beam diffraction occurs when several sets of atomic planes satisfy Bragg's laws simultaneously. A convenient way to realize this is to excite one Bragg reflection and then rotate the crystal around the diffraction vector. While the **H** reflection is always excited during such a rotation, it is possible to bring another set of atomic planes, **L**, into its diffraction condition and thus to have multiple-beam diffraction. The rotation around the scattering vector **H** is defined by an azimuthal angle, ψ. For x rays, multiple-beam diffraction peaks excited in this geometry were first observed by Renninger (1937); hence, these multiple diffraction peaks are often called the "Renninger peaks." For electrons, multiple-beam diffraction situations exist in almost all cases because of the much stronger interactions between electrons and atoms.

As shown in Figure 7, if atomic planes **H** and **L** are both excited at the same time, then there is always another set of planes, **H–L**, also in diffraction condition. The diffracted beam $\mathbf{k}_L$ by **L** reflection can be scattered again by the **H–L** reflection and this doubly diffracted beam is in the same direction as the **H**-reflected beam $\mathbf{k}_H$. In this sense, the photons (or particles) in the doubly diffracted beam have been through a "detour" route compared to the photons (particles) singly diffracted by the **H** reflection. We usually call **H** the main reflection, **L** the detour reflection, and **H–L** the coupling reflection.

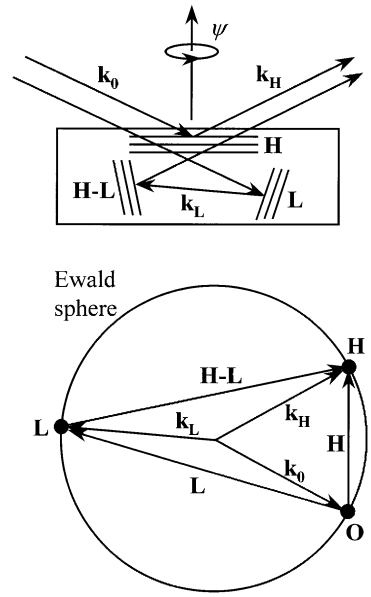


Figure 7. Illustration of a three-beam diffraction case involving **O**, **H**, and **L**, in real space (upper) and reciprocal space (lower).

Depending on the strengths of the structure factors involved, a multiple reflection can cause either an intensity enhancement (peak) or reduction (dip) in the two-beam intensity of **H**. A multiple reflection peak is commonly called the Umweganregung ("detour" in German) and a dip is called the Aufhellung. The former occurs when **H** is relatively weak and both **L** and **H–L** are strong, while the latter occurs when both **H** and **L** are strong and **H–L** is weak. A semiquantitative intensity calculation can be obtained by total energy balancing among the multiple beams, as worked out by Moon and Shull (1964) and Zachariasen (1965).

In most experiments, multiple reflections are simply a nuisance that one tries to avoid since they cause inaccurate intensity measurements. In the last two decades, however, there has been renewed and increasing interest in multiple-beam diffraction because of its promising potential as a physical solution to the well-known "phase problem" in diffraction and crystallography. The phase problem refers to the fact that the data collected in a conventional diffraction experiment are the intensities of the Bragg reflections from a crystal, which are related only to the magnitude of the structure factors, and the phase information is lost. This is a classic problem in diffraction physics and its solution remains the most difficult part of any structure determination of materials, especially for biological macromolecular crystals. Due to an interference effect among the simultaneously excited Bragg beams, multiple-beam diffraction contains the direct phase information on the structure factors involved, and therefore can be used as a way to solve the phase problem.

The basic idea of using multiple-beam diffraction to solve the phase problem was first proposed by Lipcomb (1949), and was first demonstrated by Colella (1974) in theory and by Post (1977) in an experiment on perfect crystals. The method was then further developed by several groups (Chapman et al., 1981; Chang, 1982; Schmidt and Colella, 1985; Shen and Colella, 1987, 1988; Hümmer et al., 1990) to show that it can be applied not only to perfect crystals but also to real, mosaic crystals. Recently, there have been considerable efforts to apply multibeam diffraction to large-unit-cell inorganic and macromolecular crystals (Lee and Colella, 1993; Chang et al., 1991; Hümmer et al., 1991; Weckert et al., 1993). Progress in this area has been amply reviewed by Chang (1984, 1992), Colella (1995, 1996), and Weckert and Hümmer (1997). A recent experimental innovation in reference-beam diffraction (Shen, 1998) allows parallel data collection of three-beam interference profiles using an area detector in a modified oscillation-camera setup, and makes it possible to measure the phases of a large number of Bragg reflections in a relatively short time.

Theoretical treatment of multiple-beam diffraction is considerably more complicated than for the two-beam theory, as evidenced by some of the early works (Ewald and Heno, 1968). This is particularly so in the case of x rays because of mixing of the σ and π polarization states in a multiple-beam diffraction process. Colella (1974), based upon his earlier work for electron diffraction (Colella, 1972), developed a full dynamical theory procedure for multiple-beam diffraction of x rays and a corresponding computer program called NBEAM. With Colella's theory, multiple-beam dynamical calculations have become more practical and more easily performed. On today's powerful computers and software and for not too many beams, running the NBEAM program can be almost trivial, even on personal computers. We will outline the principles of the NBEAM procedure in the NBEAM Theory section.

NBEAM Theory

The fundamental equations for multiple-beam x-ray diffraction are the same as those in the two-beam theory, before the two-beam approximation is made. We can go back to Equation 5, expand the double cross-product, and rewrite it in the following form:

$$\left[\frac{k_0^2}{K_i^2} - (1 + \Gamma F_0)\right]\mathbf{D}_i + \Gamma \sum_{j \neq i} F_{i-j}[\mathbf{u}_i(\mathbf{u}_i \cdot \mathbf{D}_j) - \mathbf{D}_j] = 0 \quad (47)$$

Eigenequation for *D*-field Components. In order to properly express the components of all wave field amplitudes, we define a polarization unit-vector coordinate system for each wave j:

$$\begin{aligned} \mathbf{u}_j &= \mathbf{K}_j/|\mathbf{K}_j| \\ \sigma_j &= \mathbf{u}_j \times \mathbf{n}/|\mathbf{u}_j \times \mathbf{n}| \\ \pi_j &= \mathbf{u}_j \times \sigma_j \end{aligned} \quad (48)$$

where **n** is the surface normal. Multiplying Equation 26 by σ_j and π_j yields

$$\begin{aligned} \left[\frac{k_0^2}{K_i^2} - (1 + \Gamma F_0)\right] D_{i\sigma} &= \Gamma \sum_{j \neq i} F_{i-j}[(\sigma_j \cdot \sigma_i) D_{j\sigma} + (\pi_j \cdot \sigma_i) D_{j\pi}] \\ \left[\frac{k_0^2}{K_i^2} - (1 + \Gamma F_0)\right] D_{i\pi} &= \Gamma \sum_{j \neq i} F_{i-j}[(\sigma_j \cdot \pi_i) D_{j\sigma} + (\pi_j \cdot \pi_i) D_{j\pi}] \end{aligned} \quad (49)$$

Matrix form of the Eigenequation. For an NBEAM diffraction case, Equation 49 can be written in a matrix form if we define a $2N \times 1$ vector $\underline{\mathbf{D}} = (D_{1\sigma}, \ldots, D_{N\sigma}, D_{1\pi}, \ldots, D_{N\pi})$, a $2N \times 2N$ diagonal matrix $\mathbf{T}_{ij}$ with $T_{ii} = k_0^2/K_i^2$ $(i = j)$ and $T_{ij} = 0$ $(i \neq j)$, and a $2N \times 2N$ general matrix $\mathbf{A}_{ij}$ that takes all the other coefficients in front of the wave field amplitudes. Matrix $\underline{\mathbf{A}}$ is Hermitian if absorption is ignored, or symmetric if the crystal is centrosymmetric. Equation 49 then becomes

$$(\underline{\mathbf{T}} + \underline{\mathbf{A}})\underline{\mathbf{D}} = 0 \quad (50)$$

Equation 50 is equivalent to

$$(\underline{\mathbf{T}}^{-1} + \underline{\mathbf{A}}^{-1})\underline{\mathbf{D}} = 0 \quad (51)$$

Strictly speaking the eigenvectors in Equation 51 are actually the **E** fields: $\underline{\mathbf{E}} = \underline{\mathbf{T}} \cdot \underline{\mathbf{D}}$. However, $\underline{\mathbf{D}}$ and $\underline{\mathbf{E}}$ are exchangeable, as discussed in the Basic Principles section.

To find nontrivial solutions of Equation 51, we need to solve the secular eigenvalue equation

$$|\underline{\mathbf{T}}^{-1} + \underline{\mathbf{A}}^{-1}| = 0 \tag{52}$$

with $T_{ii}^{-1} = K_i^2/K_0^2\,(i=j)$ and $T_{ij}^{-1} = 0\,(i \neq j)$. We can write k_j^2 in the form of its normal (n) and tangential (t) components to the entrance surface:

$$K_j^2 = (k_{0\mathrm{n}} + H_{j\mathrm{n}})^2 + k_{j\mathrm{t}}^2 \tag{53}$$

which is essentially Bragg's law together with the boundary condition that $K_{j\mathrm{t}} = k_{j\mathrm{t}}$.

Strategy for Numerical Solutions. If we treat $\mu = K_{0\mathrm{n}/k_0}$ as the only unknown, Equation 52 takes the following matrix form:

$$|\mu^2 - \mu\underline{\mathbf{B}} + \underline{\mathbf{C}}| = 0 \tag{54}$$

where $B_{ij} = -(2H_{j\mathrm{n}}/k_0)\delta_{ij}$ is a diagonal matrix and $C_{ij} = (A-1)_{ij} + \delta_{ij}(H_{j\mathrm{n}}^2 + k_{j\mathrm{t}}^2/k_0^2$. Equation 54 is a quadratic eigenequation that no computer routines are readily available for solving. Colella (1974) employed an ingenious method to show that Equation 51 is equivalent to solving the following linear eigenvalue problem:

$$\begin{pmatrix} \underline{\mathbf{B}} & -\underline{\mathbf{C}} \\ \underline{\mathbf{I}} & 0 \end{pmatrix} \begin{pmatrix} \underline{\mathbf{D}}' \\ \underline{\mathbf{D}} \end{pmatrix} = \mu \begin{pmatrix} \underline{\mathbf{D}}' \\ \underline{\mathbf{D}} \end{pmatrix} \tag{55}$$

where $\underline{\mathbf{I}}$ is a unit matrix, and $\underline{\mathbf{D}}' = \mu\underline{\mathbf{D}}$, which is a redundant $2N$ vector with no physical significance.

Equation 55 can now be solved with standard software routines that deal with linear eigenvalue equations. It is a $4N$th-order equation for $K_{0\mathrm{n}}$, and thus has $4N$ solutions, denoted as $K_{0\mathrm{n}}^l, l = 1, \ldots, 4N$. For each eigenvalue $K_{0\mathrm{n}}$, there is a corresponding $2N$ eigenvector that is stored in $\underline{\mathbf{D}}$, which now is a $2N \times 4N$ matrix and its element labeled $D_{j\sigma}^l$ in its top N rows and $D_{j\pi}^l$ in its bottom N rows. These wave field amplitudes are evaluated at this point only on a relative scale, similar to the amplitude ratio in the two-beam case. For convenience, each $2N$ eigenvector can be normalized to unity:

$$\sum_j^N (|D_{j\sigma}^l|^2 + |D_{j\pi}^l|^2) = 1 \tag{56}$$

In terms of the eigenvalues $K_{0\mathrm{n}}^l$ and the eigenvectors $\mathbf{D}_j^l = (D_{j\sigma}^l, D_{j\pi}^l)$, a general expression for the wave field inside the crystal is given by

$$\mathbf{D}(\mathbf{r}) = \sum_l q_l \sum_j \mathbf{D}_j^l e^{-i\mathbf{K}_j^l \cdot \mathbf{r}} \tag{57}$$

where $\mathbf{K}_j^l = \mathbf{K}_0^l + \mathbf{H}_j$ and q_l's $(l = 1, \ldots, 4N)$ are the coefficients to be determined by the boundary conditions.

Boundary Conditions. In general, it is not suitable to distinguish the Bragg and the Laue geometries in multiple-beam diffraction situations since it is possible to have an internal wave vector parallel to the surface and thus the distinction would be meaningless. The best way to treat the situation, as pointed out by Colella (1974), is to include both the back-diffracted and the forward-diffracted beams in vacuum, associated with each internal beam j. Thus for each beam j, we have two vacuum waves defined by $\mathbf{k}_{\pm j} = \mathbf{k}_{j\mathrm{t}} \pm \mathbf{n}(k_0^2 - k_{j\mathrm{t}}^2)^{1/2}$, where again the subscript t stands for the tangential component. Therefore for an N-beam diffraction from a parallel crystal slab, we have altogether $8N$ unknowns: $4N$ q_l values for the field inside the crystal, $2N$ wave field components of $\mathbf{D}_{-j}^e$ above the entrance surface, and $2N$ components of the wave field $\mathbf{D}_j^e$ below the back surface.

The $8N$ equations needed to solve the above problem are fully provided by the general boundary conditions, Equation 11. Inside the crystal we have

$$\mathbf{E}_j = \sum_l q_l \mathbf{D}_j^l e^{-i\mathbf{K}_j^l \cdot \mathbf{r}} \tag{58}$$

and $\mathbf{H}_j = \mathbf{u}_j \times \mathbf{E}_j$, where the sum is over all eigenvalues l for each jth beam. (We note that in Colella's original formalism converting $\mathbf{D}_j$ to $\mathbf{E}_j$ is not necessary since Equation 51 is already for $\mathbf{E}_j$. This is also consistent with the omissions of all longitudinal components of $\mathbf{E}$ fields, after the eigenvalue equation is obtained, in dynamical theory.) Outside the crystal, we have $\mathbf{D}_j^e$ at the back surface and $\mathbf{D}_{-j}^e$ plus incident beam $\mathbf{D}_0^i$ at the entrance surface. These boundary conditions provide eight scalar equations for each beam j, and thus the $8N$ unkowns can be solved for as a function of $\mathbf{D}_0^i$.

Intensity Computations. Both the reflected and the transmitted intensities, I_{-j} and I_j, for each beam j can be calculated by taking $I_j = |\mathbf{D}_j^e|^2/|\mathbf{D}_0^i|^2$. We should note that the whole computational procedure described above only evaluates the diffracted intensity at <u>one</u> crystal orientation setting with respect to the incident beam. To obtain meaningful information, the computation is usually repeated for a series of settings of the incident angle θ and the azimuthal angle ψ. An example of such two-dimensional (2D) calculations is shown in Figure 8A, which is for a three-beam case, GaAs(335)/(551). In many experimental situations, the intensities in the θ direction are integrated either purposely or because of the divergence in the incident beam. In that case, the integrated intensities versus the azimuthal angle ψ are plotted, as shown in Figure 8B.

Second-Order Born Approximation

From the last segment, we see that the integrated intensity as a function of azimuthal angle usually displays an asymmetric intensity profile, due to the multiple-beam interference. The asymmetry profile contains the phase information about the structure factors involved. Although the NBEAM program provides full account for these multiple-beam interferences, it is rather difficult to gain physical insight into the process and into the structural parameters it depends on.

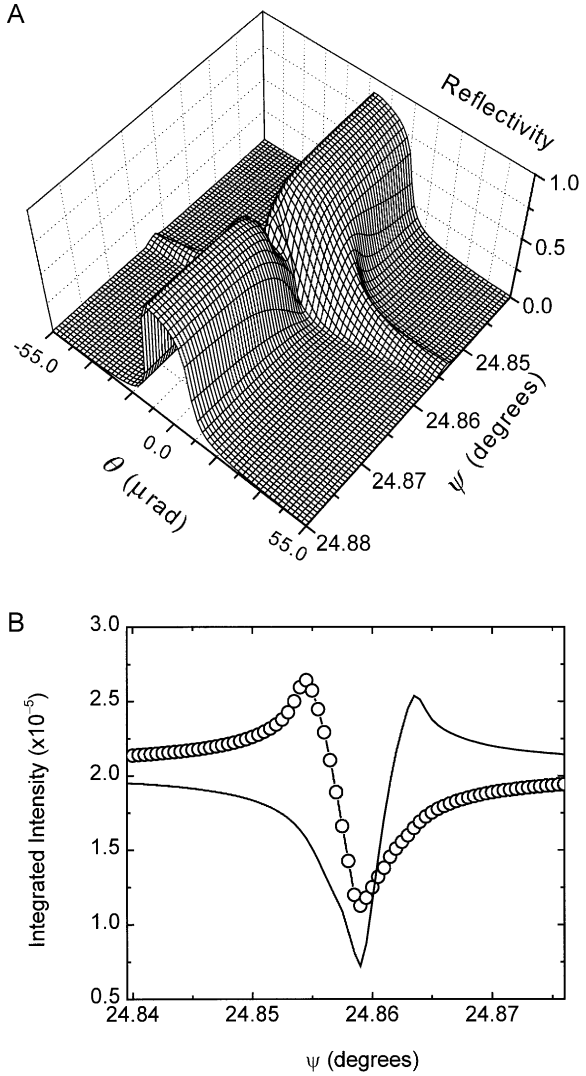


Figure 8. (**A**) Calculated reflectivity using NBEAM for the three-beam case of GaAs(335)/(551), as a function of Bragg angle θ and azimuthal angle ψ. (**B**) Corresponding integrated intensities versus ψ (open circles). The solid-line-only curve corresponds to the profile with an artificial phase of π added in the calculation.

In the past decade or so, there have been several approximate approaches for multiple-beam diffraction intensity calculations based on Bethe approximations (Bethe, 1928; Juretschke, 1982, 1984, 1986; Hoier and Marthinsen, 1983), second-order Born approximation (Shen, 1986), Takagi-Taupin differential equations (Thorkildsen, 1987), and an expanded distorted-wave approximation (Shen, 1999b). In most of these approaches, a modified two-beam structure factor can be defined so that integrated intensities can be obtained through the two-beam equations. In the following section, we will discuss only the second-order Born approximation (for x rays), since it provides the most direct connection to the two-beam kinematic results. The expanded distorted-wave theory is outlined at the end of this unit following the standard distorted-wave theory in surface scattering.

To obtain the Born approximation series, we transform the fundamental Equation 3 into an integral equation by using the Green's function and obtain the following:

$$\mathbf{D}(\mathbf{r}) = \mathbf{D}^{(0)}(\mathbf{r}) + \frac{1}{4\pi}\int d\mathbf{r}' \frac{e^{-ik_0|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\nabla' \times \nabla' \times [\delta\varepsilon(\mathbf{r}')\mathbf{D}(\mathbf{r}')] \tag{59}$$

where $\mathbf{D}^{(0)}(\mathbf{r}) = \mathbf{D}_0 e^{-i\mathbf{k}_0\cdot\mathbf{r}}$ is the incident beam. Since δε is small, we can calculate the scattered wave field $\mathbf{D}(\mathbf{r})$ iteratively using the perturbation theory of scattering (Jackson, 1975). For first-order approximation, we substitute $\mathbf{D}(\mathbf{r}')$ in the integrand by the incident beam $\mathbf{D}^{(0)}(\mathbf{r})$, and obtain a first-order solution $\mathbf{D}^{(1)}(\mathbf{r})$. This solution can then be substituted into the integrand again to provide a second-order approximation, $\mathbf{D}^{(2)}(\mathbf{r})$, and so on. The sum of all these approximate solutions gives rise to the true solution of Equation 59,

$$\mathbf{D}(\mathbf{r}) = \mathbf{D}^{(0)}(\mathbf{r}) + \mathbf{D}^{(1)}(\mathbf{r}) + \mathbf{D}^{(2)}(\mathbf{r}) + \cdots \tag{60}$$

This is essentially the Born series in quantum mechanics.

Assuming that the distance r from the observation point to the crystal is large compared to the size of the crystal (far field approximation), it can be shown (Shen, 1986) that the wave field of the first-order approximation is given by

$$\mathbf{D}^{(1)}(\mathbf{r}) = Nr_e F_{\mathrm{H}} \mathbf{u} \times (\mathbf{u} \times \mathbf{D}_0)(e^{-ik_0 r}/r) \tag{61}$$

where N is the number of unit cells in the crystal, and only one set of atomic planes **H** satisfies the Bragg's condition, $k_0\mathbf{u} = \mathbf{k}_0 + \mathbf{H}$, with **u** being a unit vector. Equation 61 is identical to the scattered wave field expression in kinematic theory, which is what we expect from the first-order Born approximation.

To evaluate the second-order expression, we cannot use Equation 61 as $\mathbf{D}^{(1)}$ since it is valid only in the far field. The original form of $\mathbf{D}^{(1)}$ with Green's function has to be used. For detailed derivations we refer to Shen's (1986). The final second-order wave field $\mathbf{D}^{(2)}$ is expressed by

$$\mathbf{D}^{(2)} = -Nr_e \frac{e^{-ik_0 r}}{r}\mathbf{u} \times \left[\mathbf{u} \times \Gamma \sum_{\mathbf{L}} F_{\mathrm{H-L}} F_{\mathrm{L}} \frac{\mathbf{k}_{\mathrm{L}} \times (\mathbf{k}_{\mathrm{L}} \times \mathbf{D}_0)}{k_0^2 - k_{\mathrm{L}}^2}\right] \tag{62}$$

It can be seen that $\mathbf{D}^{(2)}$ is the detoured wave field involving **L** and **H–L** reflections, and the summation over **L** represents a coherent superposition of all possible three-beam interactions. The relative strength of a given detoured wave is determined by its structure factors and is inversely proportional to the distance $k_0^2 - K_{\mathrm{L}}^2$ of the reciprocal lattice node **L** from the Ewald sphere.

The total diffracted intensity up to second order in Γ is given by a coherent sum of $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$:

$$\begin{aligned} I &= |\mathbf{D}^{(1)} + \mathbf{D}^{(2)}|^2 \\ &= \left| Nr_e \frac{e^{-ik_0 r}}{r}\mathbf{u} \times \left[\mathbf{u} \times F_{\mathrm{H}}\left(\mathbf{D}_0 - \Gamma \sum_{\mathbf{L}} \frac{F_{\mathrm{H-L}} F_{\mathrm{L}}}{F_H} \frac{\mathbf{k}_{\mathrm{L}} \times (\mathbf{k}_{\mathrm{L}} \times \mathbf{D}_0)}{k_0^2 - k_{\mathrm{L}}^2}\right)\right]\right|^2 \end{aligned} \tag{63}$$

Equation 63 provides an approximate analytical expression for multiple-beam diffracted intensities and represents a modified two-beam intensity influenced by multiple-beam interactions. The integrated intensity can be computed by replacing F_H in the kinematic intensity formula by a "modified structure factor" defined by

$$F_H \mathbf{D}_0 \rightarrow F_H \left(\mathbf{D}_0 - \Gamma \sum_{\mathbf{L}} \frac{F_{H-L} F_L}{F_H} \frac{\mathbf{k}_L \times (\mathbf{k}_L \times \mathbf{D}_0)}{k_0^2 - k_L^2} \right) \quad (64)$$

Often, in practice, multiple-beam diffraction intensities are normalized to the corresponding two-beam values. In this case, Equation 63 can be used directly since the prefactors in front of the square brackets will be canceled out. It can be shown (Shen, 1986) that Equation 63 gives essentially the same result as the NBEAM as long as the full three-beam excitation points are excluded, indicating that the second-order Born approximation is indeed a valid approach to multiple-beam diffraction simulations. Equation 63 becomes divergent at the exact three-beam excitation point $k_0 = k_L$. However, the singularity can be avoided numerically if we take absorption into account by introducing an imaginary part in the wave vectors.

Special Multiple-Beam Effects

The second-order Born approximation not only provides an efficient computational technique, but also allows one to gain substantial insight to the physics involved in a multiple-beam diffraction process.

Three-Beam Interactions as the Leading Dynamical Effect. The successive terms in the Born series, Equation 60, represent different levels of multiple-beam interactions. For example, $\mathbf{D}^{(0)}$ is simply the incident beam (**O**), $\mathbf{D}^{(1)}$ consists of two-beam (**O**, **H**) diffraction, $\mathbf{D}^{(2)}$ involves three-beam (**O**, **H**, **L**) interactions, and so on. Equation 62 shows that even when more than three beams are involved, the individual three-beam interactions are the dominant effects compared to higher-order beam interactions. This conclusion is very important to computations of NBEAM effects when N is large. It can greatly simplify even the full dynamical calculations using NBEAM, as shown by Tischler and Batterman (1986). The new multiple-beam interpretation of the Born series also implies that the three-beam effect is the leading term beyond the kinematic first-order Born approximation and thus is the dominant dynamical effect in diffraction. In a sense, the three-beam interactions ($\mathbf{O} \rightarrow \mathbf{L} \rightarrow \mathbf{H}$) are even more important than the multiple scattering in the two-beam case since that involves $\mathbf{O} \rightarrow \mathbf{H} \rightarrow \mathbf{O} \rightarrow \mathbf{H}$ (or higher order) scattering, which is equivalent to a four-beam interaction.

Phase Information. Equation 63 shows explicitly the phase information involved in the multiple-beam diffraction. The interference between the detoured wave $\mathbf{D}^{(2)}$ and the directly scattered wave $\mathbf{D}^{(1)}$ depends on the relative phase difference between the two waves. This phase difference is equal to phase ν of the denominator, plus the phase triplet δ of the structure factor phases α_{H-L}, α_L, and α_H: $\delta = \alpha_{H-L} + \alpha_L - \alpha_H$. It can be shown that although the individual phases α_{H-L}, and α_H depend on the choice of origin in the unit cell, the phase triplet Δ does not; it is therefore called the invariant phase triplet in crystallography. The resonant phase ν depends on whether the reciprocal node **L** is outside ($k_0 < k_L$) or inside ($k_0 > k_L$) the Ewald sphere. As the diffracting crystal is rotated through a three-beam excitation, ν changes by π since **L** is swept through the Ewald sphere. This phase change of π in addition to the constant phase triplet Δ is the cause for the asymmetric three-beam diffraction profiles and allows one to measure the structural phase δ in a diffraction experiment.

Polarization Mixing. For noncoplanar multiple-beam diffraction cases (i.e., **L** not in the plane defined by **H** and $\mathbf{k}_0$), there is in general a mixing of the σ and π polarization states in the detoured wave (Shen, 1991, 1993). This means that if the incident beam is purely σ polarized, the diffracted beam may contain a π-polarized component in the case of multiple-beam diffraction, which does not happen in the case of two-beam diffraction. It can be shown that the polarization properties of the detour-diffracted beam in a three-beam case is governed by the following 2×2 matrix

$$\underline{\mathbf{A}} = \begin{pmatrix} k_L^2 - (\mathbf{L} \cdot \sigma_0)^2 & -(\mathbf{L} \cdot \sigma_0)(\mathbf{L} \cdot \pi_0) \\ -(\mathbf{k}_L \cdot \pi_H)(\mathbf{L} \cdot \sigma_0) & k_L^2 (\pi_H \cdot \pi_0) - (\mathbf{k}_L \cdot \pi_H)(\mathbf{L} \cdot \pi_0) \end{pmatrix} \quad (65)$$

The off-diagonal elements in $\underline{\mathbf{A}}$ indicate the mixing of the polarization states.

This polarization mixing, together with the phase-sensitive multiple-beam interference, provides an unusual coupling to the incident beam polarization state, especially when the incident polarization contains a circularly polarized component. The effect has been used to extract acentric phase information and to determine noncentrosymmetry in quasicrystals (Shen and Finkelstein, 1990; Zhang et al., 1999). If we use a known noncentrosymmetric crystal such as GaAs, the same effect provides a way to measure the degree of circular polarization and can be used to determine all Stokes polarization parameters for an x-ray beam (Shen and Finkelstein, 1992, 1993; Shen et al., 1995).

Multiple-Beam Standing Waves. The internal field in the case of multiple-beam diffraction is a 3D standing wave. This 3D standing wave can be detected, just like in the two-beam case, by observing x-ray fluorescence signals (Greiser and Matrlik, 1986), and can be used to determine the 3D location of the fluorescing atom—similar to the method of triangulation by using multiple separate two-beam cases. Multiple-beam standing waves are also responsible for the so-called super-Borrmann effect because of additional lowering of the wave field intensity around the atomic planes (Borrmann and Hartwig, 1965).

Polarization Density Matrix

If the incident beam is partially polarized—that is, it includes an unpolarized component—calculations in the case of multiple-beam diffraction can be rather complicated. One can simplify the algorithm a great deal by using a polarization density matrix as in the case of magnetic x-ray scattering (Blume and Gibbs, 1988). A polarization matrix is defined by

$$\underline{\rho} = \frac{1}{2}\begin{pmatrix} 1+P_1 & P_2 - iP_3 \\ P_2 + iP_3 & 1-P_1 \end{pmatrix} \tag{66}$$

where (P_1, P_2, P_3) are the normalized Stokes-Poincaré polarization parameters (Born and Wolf, 1983) that characterize the σ and π linear polarization, $\pm 45^\circ$ tilted linear polarization, and left- and right-handed circular polarization, respectively.

A polarization-dependent scattering process, where the incident beam ($D_{0\sigma}, D_{0\pi}$) is scattered into ($D_{\mathrm{H}\sigma}, D_{\mathrm{H}\pi}$), can be described by a 2×2 matrix $\underline{\mathbf{M}}$ whose elements $M_{\sigma\sigma}, M_{\sigma\pi}$, and $M_{\pi\pi}$ represent the respective $\sigma \rightarrow \sigma$, $\sigma \rightarrow \pi, \pi \rightarrow \sigma$, and $\pi \rightarrow \pi$ scattering amplitudes:

$$\begin{pmatrix} D_{\mathrm{H}\sigma} \\ D_{\mathrm{H}\pi} \end{pmatrix} = \begin{pmatrix} M_{\sigma\sigma} & M_{\pi\sigma} \\ M_{\sigma\pi} & M_{\pi\pi} \end{pmatrix} \begin{pmatrix} D_{0\sigma} \\ D_{0\pi} \end{pmatrix} \tag{67}$$

It can be shown that with the density matrix $\underline{\rho}$ and scattering matrix $\underline{\mathbf{M}}$, the scattered new density matrix $\underline{\rho}_{\mathrm{H}}$ is given by $\underline{\rho}_{\mathrm{H}} = \underline{\mathbf{M}}\underline{\rho}\underline{\mathbf{M}}^{\dagger}$, where $\underline{\mathbf{M}}^{\dagger}$ is the Hermitian conjugate of $\underline{\mathbf{M}}$. The scattered intensity I_{H} is obtained by calculating the trace of the new density matrix

$$I_{\mathrm{H}} = \mathrm{Tr}(\underline{\rho}_{\mathrm{H}}) \tag{68}$$

This equation is valid for any incident beam polarization, including when the beam is partially polarized. We should note that the method is not restricted to dynamical theory and is widely used in other physics fields such as quantum mechanics. In the case of multiple-beam diffraction, the matrix $\underline{\mathbf{M}}$ can be evaluated using either the NBEAM program or one of the perturbation approaches.

GRAZING-ANGLE DIFFRACTION

Grazing-incidence diffraction (GID) of x rays or neutrons refers to situations where either the incident or the diffracted beam forms a small angle less than or in the vicinity of the critical angle of a well-defined crystal surface. In these cases, both a Bragg-diffracted beam and a specular-reflected beam can occur simultaneously. Although there are only two beams, **O** and **H**, inside the crystal, the usual two-beam dynamical diffraction theory cannot be applied to this situation without some modifications (Afanasev and Melkonyan, 1983; Cowan et al., 1986; Hoche et al., 1986). These special considerations, however, can be automatically taken into account in the NBEAM theory discussed in the Multiple-Beam Diffraction section, as shown by Durbin and Gog (1989).

A GID geometry may include the following situations: (1) specular reflection, (2) coplanar GID involving highly asymmetric Bragg reflections, and (3) GID in an inclined geometry. Because of the substantial decrease in the penetration depths of the incident beam in these geometries, there have been widespread applications of GID using synchrotron radiation in recent years in materials studies of surface structures (Marra et al., 1979), depth-sensitive disorder and phase transitions (Dosch, 1992; Rhan et al., 1993; Krimmel et al., 1997; Rose et al., 1997), and long-period multilayers and superlattices (Barbee and Warburton, 1984; Salditt et al., 1994). We devote this section first to the basic concepts in GID geometries. A recent review on these topics has been given by Holy (1996); see also SURFACE X-RAY DIFFRACTION. In the Distorted-wave Born Approximation section, we present the principle of this approximation (Vineyard, 1982; Dietrich and Wagner, 1983, 1984; Sinha et al., 1988), which provides a bridge between the dynamical Fresnel formula and the kinematic theory of surface scattering of x rays and neutrons.

Specular Reflectivity

It is straightforward to show that the Fresnel's optical reflectivity, which is widely used in studies of mirrors (e.g., Bilderback, 1981), can be recovered in the dynamical theory for x-ray diffraction. We recall that in the case of one beam, the solution to the dispersion equation is given by Equation 7. Assuming a semi-infinite crystal and using the general boundary condition, Equation 11, we have the following equations across the interface (see Appendix for definition of terms):

$$\begin{aligned} D_0^{\mathrm{i}} + D_0^{e} &= D_0 \\ k_0 \sin\theta (D_0^{\mathrm{i}} - D_0^{\mathrm{e}}) &= k_0 \sin\theta' D_0 \end{aligned} \tag{69}$$

where θ and θ' are the incident angles of the external and the internal incident beams (Fig. 1B). By using Equation 7 and the fact that $\mathbf{K}_0$ and $\mathbf{k}_0$ can differ only by a component normal to the surface, we arrive at the following wave field ratios (for small angles):

$$r_0 \equiv \frac{D_0^{\mathrm{e}}}{D_0^{\mathrm{i}}} = \frac{\theta - \sqrt{\theta^2 - \theta_{\mathrm{c}}^2}}{\theta + \sqrt{\theta^2 - \theta_{\mathrm{c}}^2}} \tag{70a}$$

$$t_0 \equiv \frac{D_0}{D_0^{\mathrm{i}}} = \frac{2\theta}{\theta + \sqrt{\theta^2 - \theta_{\mathrm{c}}^2}} \tag{70b}$$

with the critical angle defined as $\theta_{\mathrm{c}} = (\Gamma F_0)^{1/2}$. For most materials, θ_{c} is on the order of a few milliradians. In general, θ_{c} can be complex in order to take into account absorption. Obviously, Equations 70 gives the same reflection and transmission coefficients as the Fresnel theory in visible optics (see, e.g., Jackson, 1974). The specular reflectivity R is given by the square of the magnitude of Equation 70a: $R = |r_0|^2$, while $|t_0|^2$ of Equation 70b is the internal wave field intensity at the surface. An example of $|r_0||^2$ and $|t_0|^2$ is shown in Figure 9A,B for a GaAs surface.

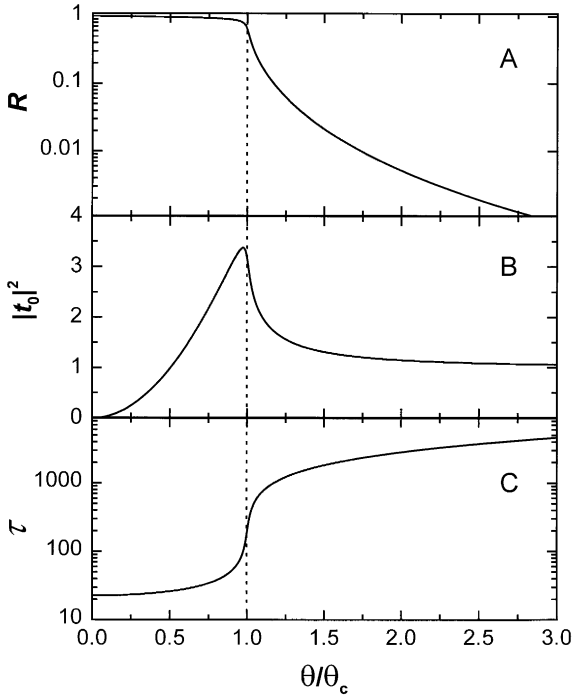


Figure 9. (**A**) Fresnel's reflectivity curve for a GaAs surface at 1.48 Å. (**B**) Intensity of the internal field at the surface. (**C**) Penetration depth.

At $\theta \gg \theta_c$, θ in Equations 70a,b should be replaced by the original sin θ and the Fresnel reflectivity $|r_0|^2$ varies as $1/(2\sin\theta)^4$, or as $1/q^4$ with q being the momentum transfer normal to the surface. This inverse fourth power law is the same as that derived in kinematic theory (Sinha et al., 1988) and in the theory of small-angle scattering (Porod, 1952, 1982). At first glance, the $1/q^4$ asymptotic law is drastically different from the crystal truncation rod $1/q^2$ behavior for the Bragg reflection tails. A more careful inspection shows that the difference is due to the integral nature of the reflectivity over a more fundamental physical quantity called differential cross-section, $d\sigma/d\Omega$, which is defined as the incident flux scattered into a detector area that forms a solid angle $d\Omega$ with respect to the scattering source. In both Fresnel reflectivity and Bragg reflection cases, $d\sigma/d\Omega \sim 1/q^2$ in reciprocal space units. Reflectivity calculations in both cases involve integrating over the solid angle and converting the incident flux into an incident intensity; each would give rise to a factor of 1/sin θ (Sinha et al., 1988). The only difference now is that in the case of Bragg reflections, this factor is simply 1/sin θ_B, which is a constant for a given Bragg reflection, whereas for Fresnel reflectivity cases, sin $\theta \sim q$ results in an additional factor of $1/q^2$.

Evanescent Wave

When $\theta < \theta_c$, the normal component K_{0n} of the internal wave vector $\mathbf{K}_0$ is imaginary so that the x-ray wave field inside the material diminishes exponentially as a function of depth, as given by

$$\mathbf{D}_0(\mathbf{r}) = t_0 e^{-\mathrm{Im}(K_{0n})r_n} e^{-i\mathbf{k}_t\cdot\mathbf{r}_t} \qquad (71)$$

where t_0 is given by Equation 70b and r_n and $\mathbf{r}_t$ are the respective coordinates normal and parallel to the surface.

The characteristic penetration depth $\tau(1/e)$ value of the intensity) is given by $\tau = 1/[2\,\mathrm{Im}(K_{0n})]$, where $\mathrm{Im}\,(K_{0n})$ is the imaginary part of K_{0n}. A plot of τ as a function of the incident angle θ is shown in Figure 9C. In general, a penetration depth (known as skin depth) as short as 10 to 30 Å can be achieved with Fresnel's specular reflection when $\theta < \theta_c$. The limit at $\theta = 0$ is simply given by $\tau = \lambda/(4\pi\theta_c)$ with λ being the x-ray wavelength. This makes the x-ray reflectivity–related measurement a very useful tool for studying surfaces of various materials.

At $\theta > \theta_c$, τ becomes quickly dominated by true photoelectric absorption and the variation is simply geometrical. The large variation of τ around $\theta \sim \theta_c$ forms the basis for such depth-controlled techniques as x-ray fluorescence under total external reflection (de Boer, 1991; Hoogenhof and de Boer, 1994), grazing-incidence scattering and diffraction (Dosch, 1992; Lied et al., 1994; Dietrich and Hasse, 1995; Gunther et al., 1997), and grazing-incidence x-ray standing waves (Hashizume and Sakata, 1989; Jach et al., 1989; Jach and Bedzyk, 1993).

Multilayers and Superlattices

Synthetic multilayers and superlattices usually have long periods of 20 to 50 Å. Since the Bragg angles corresponding to these periods are necessarily small in an ordinary x-ray diffraction experiment, the superlattice diffraction peaks are usually observed in the vicinity of specular reflections. Thus dynamical theory is often needed to describe the diffraction patterns from multilayers of amorphous materials and superlattices of nearly perfect crystals.

A computational method to calculate the reflectivity from a multilayer system was first developed by Parratt (1954). In this method, a series of recursive equations on the wave field amplitudes is set up, based on the boundary conditions at each interface. Assuming that the last layer is a substrate that is sufficiently thick, one can find the solution of each layer backward and finally obtain the reflectivity from the top layer. For details of this method, we refer the readers to Parratt's original paper (1954) and to a more recent matrix formalism reviewed by Holy (1996).

It should be pointed out that near the specular region, the internal crystalline structures of the superlattice layers can be neglected, and only the average density of each layer would contribute. Thus the reflectivity calculations for multilayers and for superlattices are identical near the specular reflections. The crystalline nature of a superlattice needs to be taken into account near or at Bragg reflections. With the help of Takagi-Taupin equations, lattice mismatch and variations along the growth direction can also be taken into account, as shown by Bartels et al. (1986). By treating a semi-infinite single crystal as an extreme case of a superlattice or multilayer, one can calculate the reflectivity for the entire range from specular to all of the Bragg reflections along a given crystallographic axis (Caticha, 1994).

X-ray diffraction studies of laterally structured superlattices with periods of ~0.1 to 1 μm, such as surface

gratings and quantum wire and dot arrays, have been of much interest in materials science in recent years (Bauer et al., 1996; Shen, 1996b). Most of these studies can be dealt with using kinematic diffraction theory (Aristov et al., 1988), and a rich amount of information can be obtained such as feature profiles (Shen et al., 1993; Darhuber et al., 1994), roughness on side wall surfaces (Darhuber et al., 1994), imperfections in grating arrays (Shen et al., 1996b), size-dependent strain fields (Shen et al., 1996a), and strain gradients near interfaces (Shen and Kycia, 1997). Only in the regimes of total external reflection and GID are dynamical treatments necessary as demonstrated by Tolan et al. (1992, 1995) and by Darowski et al. (1997).

Grazing-Incidence Diffraction

Since most GID experiments are performed in the inclined geometry, we will focus only on this geometry and refer the highly asymmetric cases to the literature (Hoche et al., 1988; Kimura and Harada, 1994; Holy, 1996).

In an inclined GID arrangement, both the incident beam and the diffracted beam form a small angle with respect to the surface, as shown in Figure 10A, with the scattering vector parallel to the surface. This geometry involves two internal waves, **O** and **H**, and three external waves, incident **O**, specular reflected $\overline{\mathbf{O}}$ and diffracted **H** beams. With proper boundary conditions, the diffraction problem can be solved analytically as shown by several authors (Afanasev and Melkonyan, 1983; Cowan et al., 1986; Hoche et al., 1986; Hung and Chang, 1989; Jach et al., 1989). Durbin and Gog (1989) applied the NBEAM program to GID geometry.

A characteristic dynamical effect in GID geometry is a double-critical-angle phenomenon due to the diameter gap of the dispersion surface for **H** reflection. This can be seen intuitively from simple geometric considerations. Inside the crystal, only two beams, **O** and **H**, are excited and thus the usual two-beam theory described in the Two-Beam Diffraction section applies. The dispersion surface inside the crystal is exactly the same as shown in Figure 2A. The only difference is the boundary condition. In the GID case, the surface normal is perpendicular to the page in Figure 2, and therefore the circular curvature out of the page needs to be taken into account. For simplicity, we consider only the diameter points on the dispersion surface for one polarization state. A cut through the diameter points L and Q in Figure 2 is shown schematically in Figure 10B; this consists of three concentric circles representing the hyperboloids of revolution α and β branches, and the vacuum sphere at point L. At very small incident angles, we see that no tie points can be excited and only total specular reflection can exist. As the incident angle increases so that $\phi > \phi_c^{\alpha}$, α tie points are excited but the β branch remains extinguished. Thus specular reflectivity would maintain a lower plateau, until $\phi > \phi_c^{\beta}$ when both α and β modes can exist inside the crystal. Meanwhile, the Bragg reflected beam should have been fully excited when $\phi_c^{\alpha} < \phi < \phi_c^{\beta}$, but because of the partial specular reflection its diffracted intensity is much reduced. These effects can be clearly seen in the example shown in Figure 11, which is for a Ge(220) reflection with a (1-11) surface orientation.

If the Bragg's condition is not satisfied exactly, then the circle labeled L in Figure 10B will be split into two concentric ones representing the two spheres centered at **O** and **H**, respectively. We then see that the exit take-off angles can be different for the reflected **O** beam and the diffracted **H** beam. With a position-sensitive linear detector and a range of incident angles, angular profiles (or rod profiles) of diffracted beams can be observed directly, which can provide depth-sensitive structural information near a crystal surface (Dosch et al., 1986; Bernhard et al., 1987).

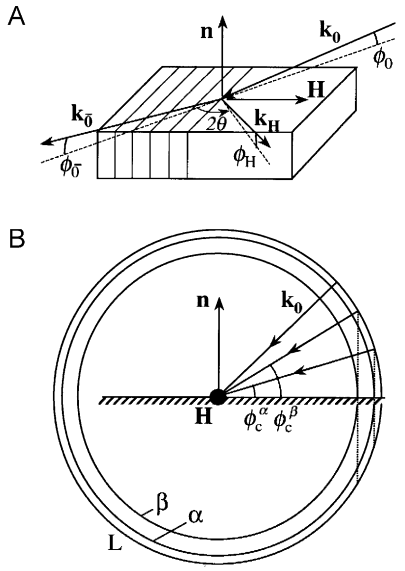


Figure 10. (**A**) Schematic of the grazing-incidence diffraction geometry. (**B**) A cut through the diameter points of the dispersion surface.

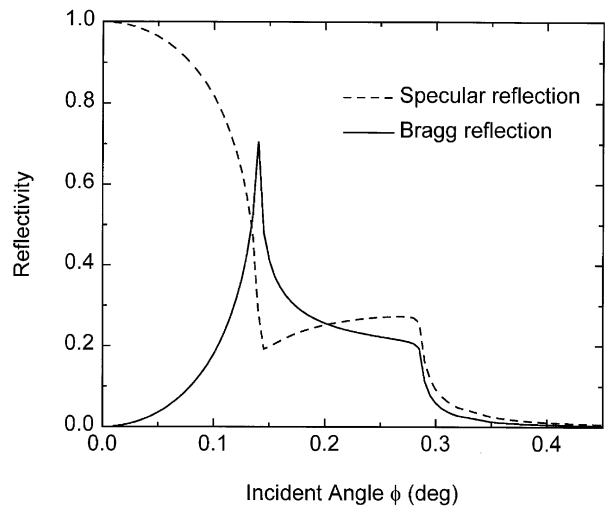


Figure 11. Specular and Bragg reflectivity at the center of the rocking curve for the Ge(220) reflection with a (1-11) surface orientation.

Distorted-Wave Born Approximation

GID, which was discussed in the last section, can be viewed as the dynamical diffraction of the internal evanescant wave, Equation 71, generated by specular reflection under grazing-angle conditions. If the rescattering mechanism is relatively weak, as in the case of a surface layer, then dynamical diffraction theory may not be necessary and the Born approximation can be substituted to evaluate the scattering of the evanescant wave. This approach is called the distorted-wave Born approximation (DWBA) in quantum mechanics (see, e.g., Schiff, 1955), and was first applied to x-ray scattering from surfaces by Vineyard (1982) and by Dietrich and Wagner (1983, 1984). It was noted by Dosch et al. (1986) that Vineyard's original treatment did not handle the exit-angle dependence properly because of a missing factor in its reciprocity arrangement.

The DWBA has been applied to several different scattering situations, including specular diffuse scattering from a rough surface, crystal truncation rod scattering near a surface, diffuse scattering in multilayers, and near-surface diffuse scattering in binary alloys (X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS). The underlying principle is the same for all these cases and we will only discuss specular diffuse scattering to illustrate these principles.

From a dynamical theory point of view, the DWBA is schematically shown in Figure 12A. An incident beam $\mathbf{k}_0$ creates an internal incident beam $\mathbf{K}_0$ and a specular reflected beam $\mathbf{k}_{\bar{0}}$. We then assume that the internal beam $\mathbf{K}_0$ is scattered by a weak "Bragg reflection" at a lateral momentum transfer $\mathbf{q}_t$. Similar to the two-beam case in dynamical theory, we draw two spheres centered at $\mathbf{q}_t$ shown as the dashed circles in Figure 12A. However, the internal diffracted wave vector is determined by kinematic scattering as $\mathbf{K}_{\bar{q}} = \mathbf{K}_0 + \mathbf{k}$, where $\mathbf{q}$ includes both the lateral component $\mathbf{q}_t$ and a component $\mathbf{q}_n$ normal to the surface, defined by the usual 2θ angle. Therefore only one of the tie points on the internal sphere is excited, giving rise to $\mathbf{K}_{\bar{q}}$. Outside the surface, we have two tie points that yield $\mathbf{k}_q$ and $\mathbf{k}_{\bar{q}}$, respectively, as defined in dynamical theory. Altogether we have six beams, three associated with $\mathbf{O}$ and three associated with $\mathbf{q}$. The connection between the $\mathbf{O}$ and the $\mathbf{q}$ beams is through the internal kinematic scattering

$$\mathbf{D}_{\bar{q}} = S(\mathbf{q}_t)\mathbf{D}_0 \tag{72}$$

where $S(\mathbf{q}_t)$ is the surface scattering form factor. As will be seen later, $|S(\mathbf{q}_t)|^2$ represents the scattering cross-section per unit surface area defined by Sinha et al. (1988) and equals the Fourier transform of the height-height correlation function $C(\mathbf{r}_t)$ in the case of not-too-rough surfaces.

To find the diffuse-scattered exit wave field $\mathbf{D}^e_{\bar{q}}$, we use the optical reciprocity theorem of Helmhotz (Born and Wolf, 1983) and reverse the directions of all three wave vectors of the $\mathbf{q}$ beams. We see immediately that the situation is identical to that discussed at the beginning of this section for Fresnel reflections. Thus, we should have

$$\mathbf{D}^e_{\bar{q}} = t_q \mathbf{D}_{\bar{q}} \qquad t_q \equiv \frac{2\theta_q}{\theta_q + \sqrt{\theta_q^2 - \theta_c^2}} \tag{73}$$

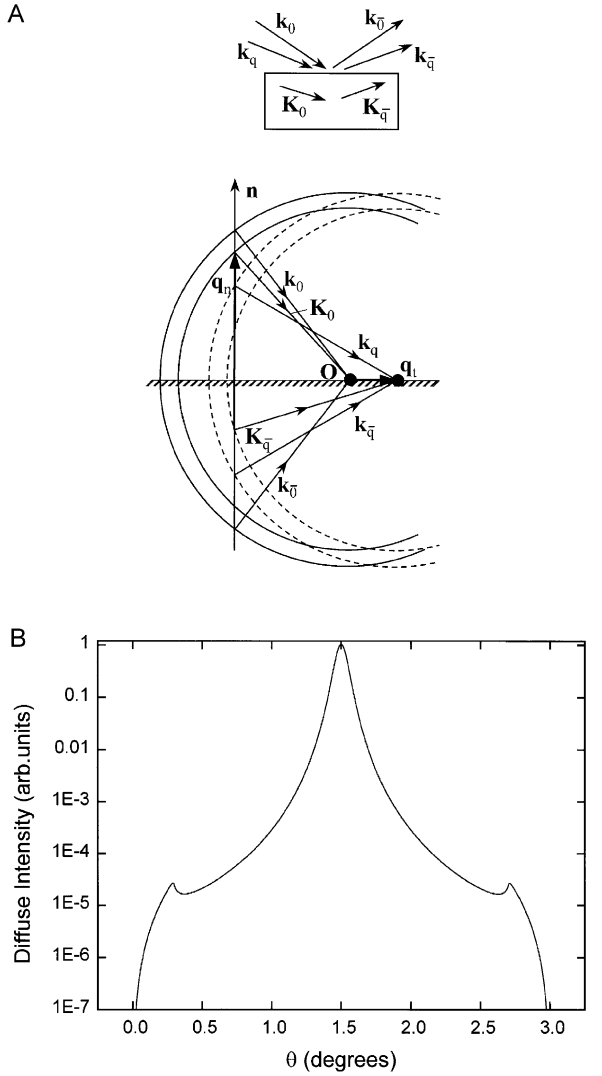


Figure 12. (**A**) Dynamical theory illustration of the distorted-wave Born approximation. (**B**) Typical diffuse scattering profile in specular reflectivity with Yoneda wings.

Using Equations 70b and 72, we obtain that

$$\mathbf{D}^e_{\bar{q}} = t_0 t_q S(\mathbf{q}_t)\mathbf{D}^i_0 \tag{74}$$

and the diffuse scattering intensity is simply given by

$$I_{\text{diff}} = |\mathbf{D}^e_{\bar{q}}/\mathbf{D}^i_0|^2 = |t_0|^2|t_q|^2|S(\mathbf{q}_t)|^2 \tag{75}$$

Apart from a proper normalization factor, Equation 75 is the same as that given by Sinha et al. (1988). Of course, here the scattering strength $|S(\mathbf{q}_t)|^2$ is only a symbolic quantity. For the physical meaning of various surface roughness correlation functions and its scattering forms, we refer to the article by Sinha et al. (1988) for more a detailed discussion.

In a specular reflectivity measurement, one usually uses so-called rocking scans to record a diffuse scattering profile. The amount of diffuse scattering is determined by

the overall surface roughness and the shape of the profile is determined by the lateral roughness correlations. An example of computer-simulated rocking scan is shown in Figure 12B for a GaAs surface at 1.48 Å with the detector $2\theta = 3^\circ$. The parameter $|S(\mathbf{q}_t)|^2$ is assumed to be a Lorentzian with a correlation length of 4000 Å. The two peaks at $\theta \sim 0.3^\circ$ and 2.7° correspond to the situation where the incident or the exit beam makes the critical angle with respect to the surface. These peaks are essentially due to the enhancement of the evanescent wave (standing wave) at the critical angle (Fig. 9B) and are often called the Yoneda wings, as they were first observed by Yoneda (1963).

Diffuse scattering of x rays, neutrons, and electrons is widely used in materials science to characterize surface morphology and roughness. The measurements can be performed not only near specular reflection but also around nonspecular crystal truncation rods in grazing-incidence inclined geometry (Shen et al., 1989; Stepanov et al., 1996). Spatially correlated roughness and morphologies in multilayer systems have also been studied using diffuse x-ray scattering (Headrick and Baribeau, 1993; Baumbach et al., 1994; Kaganer et al., 1996; Paniago et al., 1996; Darhuber et al., 1997). Some of these topics are discussed in detail in KINEMATIC DIFFRACTION OF X RAYS and in the units on x-ray surface scattering (see X-RAY TECHNIQUES).

Expanded Distorted-Wave Approximation

The scheme of the distorted-wave approximation can be extended to calculate nonspecular scattering that includes multilayer diffraction peaks from a multilayer system where a recursive Fresnel's theory is usually used to evaluate the distorted-wave (Kortright and Fischer-Colbrie, 1987; Holy and Baumbach, 1994). Recently, Shen (1999b,c) has further developed an expanded distorted-wave approximation (EDWA) to include multiple-beam diffraction from bulk crystals where a two-beam dynamical theory is applied to obtain the distorted internal waves. In Shen's EDWA theory, a sinusoidal Fourier component **G** is added to the distorting susceptibility component, which represents a charge-density modulation of the **G** reflection. Instead of the Fresnel theory, a two-beam dynamical theory is employed to evaluate the distorted-wave, while the subsequent scattering of the distorted-wave is again handled by the first-order Born approximation. We now briefly outline this EDWA approach.

Following the formal distorted-wave description given in Vineyard (1982), $\delta\varepsilon(\mathbf{r})$ in the fundamental equation 3 is separated into a distorting component $\delta\varepsilon_1(\mathbf{r})$ and the remaining part $\delta\varepsilon_2(\mathbf{r}) : \delta\varepsilon(\mathbf{r}) = \delta\varepsilon_1(\mathbf{r}) + \delta\varepsilon_2(\mathbf{r})$, where $\delta\varepsilon_1(\mathbf{r})$ contains the homogeneous average susceptibility, plus a single predominant Fourier component **G**:

$$\delta\varepsilon_1(\mathbf{r}) = -\Gamma(F_0 + F_G e^{-i\mathbf{G}\cdot\mathbf{r}} + F_{\mathrm{G}}^{-} e^{i\mathbf{G}\cdot\mathbf{r}}) \tag{76}$$

and the remaining $\delta\varepsilon_2(\mathbf{r})$ is

$$\delta\varepsilon_2(\mathbf{r}) = -\Gamma \sum_{L \neq 0, \pm \mathrm{G}} F_{\mathrm{L}} e^{-i\mathbf{L}\cdot\mathbf{r}} \tag{77}$$

Since $F_{\mathrm{G}} = |F_G| \exp(i\alpha_{\mathrm{C}})$ and $F_{\mathrm{G}}^{-} = F_{\mathrm{G}}^{*} = |F_{\mathrm{G}}| \exp(-i\alpha_{\mathrm{G}})$ if absorption is negligible, it can be seen that the additional component in Equation 76 represents a sinusoidal distortion, $-2\Gamma|F_{\mathrm{G}}| \cos(\alpha_{\mathrm{G}} - \mathbf{G}\cdot\mathbf{r})$

The distorted wave $\mathbf{D}_1(\mathbf{r})$, due only to $\delta\varepsilon_1(\mathbf{r})$, satisfies the following equation:

$$(\nabla^2 + k_0^2)\mathbf{D}_1 = -\nabla \times \nabla \times (\delta\varepsilon_1 \mathbf{D}_1) \tag{78}$$

which is a standard two-beam case since only **O** and **G** Fourier components exist in $\delta\varepsilon_1(\mathbf{r})$, and can therefore be solved by the two-beam dynamical theory (Batterman and Cole, 1964; Pinsker, 1978). It can be shown that the total distorted wave $\mathbf{D}_1(\mathbf{r})$ can be expressed as follows:

$$\mathbf{D}_1(\mathbf{r}) = \mathbf{D}_0(r_0 e^{-i\mathbf{K}_0\cdot\mathbf{r}} + r_{\mathrm{G}} e^{i\alpha_{\mathrm{G}}} e^{-i\mathbf{K}_{\mathrm{G}}\cdot\mathbf{r}}) \tag{79}$$

where

$$\begin{cases} r_0 = 1 \\ r_{\mathrm{G}} = -\sqrt{|b|}(\eta_{\mathrm{G}} \pm \sqrt{\eta_{\mathrm{G}}^2 - 1}) \end{cases} \tag{80}$$

in the semi-infinite Bragg case and

$$\begin{cases} r_0 = \cos(A\eta_{\mathrm{G}}) + i \sin(A\eta) \\ r_{\mathrm{G}} = i\sqrt{|b|} \sin(A\eta_{\mathrm{G}})/\eta_{\mathrm{G}} \end{cases} \tag{81}$$

in the thin transparent Laue case. Here standard notations in the Two-Beam Dynamical Theory section are used. It should be noted that the amplitudes of these distorted waves, given by Equations 80 and 81, are slow varying functions of depth z through parameter A, since A is much smaller than $K_0 \cdot \mathbf{r}$ or $\mathbf{K}_{\mathrm{G}} \cdot \mathbf{r}$ by a factor of $\sim \Gamma|F_{\mathrm{G}}|$, which ranges from 10^{-5} to 10^{-6} for inorganic to 10^{-7} to 10^{-8} for protein crystals.

We now consider the rescattering of the distorted-wave $\mathbf{D}_1(\mathbf{r})$, Equation 79, by the remaining part of the susceptibility $\delta\varepsilon_2(\mathbf{r})$ defined in Equation 77. Using the first-order Born approximation, the scattered wave field $\mathbf{D}(\mathbf{r})$ is given by

$$\mathbf{D}(\mathbf{r}) = \frac{e^{-ik_0 r}}{4\pi r} \int d\mathbf{r}'\, e^{ik_0 \mathbf{u}\cdot\mathbf{r}'} \nabla' \times \nabla' \times [\delta\varepsilon_2(\mathbf{r}')\mathbf{D}_1(\mathbf{r}')] \tag{82}$$

where **u** is a unit vector and r is the distance from the sample to the observation point, and the integral is evaluated over the sample volume. The amplitudes r_0 and r_{G} can be factored out of the integral because of their much weaker spatial dependence than $\mathbf{K}_0 \cdot \mathbf{r}$ $\mathbf{K}_{\mathrm{G}} \cdot \mathrm{r}$ as mentioned above. The primary extinction effects in Bragg cases and the Pendellösung effects in Laue cases are taken into account by first evaluating intensity $I_{\mathrm{H}}(z)$ scattered by a volume element at a certain depth z, and then taking an average over z to obtain the final diffracted intensity.

It is worth noting that the distorted wave, Equation 79, can be viewed as the new incident wave for the Born approximation, Equation 59, and it consists of two beams, $\mathbf{K}_0$ and $\mathbf{K}_{\mathrm{G}}$. These two incident beams can each produce its own diffraction pattern. If reflection **H** satisfies Bragg's

law, $k_0\mathbf{u} = \mathbf{K}_0 + \mathbf{H} \equiv \mathbf{K}_H$, and is excited by $\mathbf{K}_0$, then there always exists a reflection **H–G**, excited by $\mathbf{K}_G$, such that the doubly scattered wave travels along the same direction as $\mathbf{K}_H$, since $\mathbf{K}_G + \mathbf{H} - G = \mathbf{K}_H$. With this in mind and using the algebra given in the Second Order Born Approximation section, it is easy to show that Equation 82 gives rise to the following scattered wave:

$$\mathbf{D}_H = Nr_e\mathbf{u} \times (\mathbf{u} \times \mathbf{D}_0)\frac{e^{-ik_0r}}{r}(F_H r_0 + F_{H-G} r_G e^{i\alpha_G}) \qquad (83)$$

Normalizing to the conventional first-order Born wave field $\mathbf{D}_H^{(1)}$ defined by Equation 61, Equation 83 can be rewritten as

$$\mathbf{D}_H = \mathbf{D}_H^{(1)}(r_0 + |F_{H-G}/F_H| r_G e^{i\delta}) \qquad (84)$$

where $\delta = \alpha_{H-G} + \alpha_G - \alpha_H$ is the invariant triplet phase widely used in crystallography. Finally, the scattered intensity into the $\mathbf{k}_H = \mathbf{K}_H = k_0\mathbf{u}$ direction is given by $I_H = (1/t)\int_0^t |\mathbf{D}_H|^2 dz$, which is averaged over thickness t of the crystal as discussed in the last paragraph. Numerical results show that the EDWA theory outlined here provides excellent agreement with the full NBEAM dynamical calculations even at the center of a multiple reflection peak. For further information, refer to Shen (1999b,c).

SUMMARY

In this unit, we have reviewed the basic elements of dynamical diffraction theory for perfect or nearly perfect crystals. Although the eventual goal of obtaining structural information is the same, the dynamical approach is considerably different from that in kinematic theory. A key distinction is the inclusion of multiple scattering processes in the dynamical theory whereas the kinematic theory is based on a single scattering event.

We have mainly focused on the Ewald–von Laue approach of the dynamical theory. There are four essential ingredients in this approach: (1) dispersion surfaces that determine the possible wave fields inside the material; (2) boundary conditions that relate the internal fields to outside incident and diffracted beams; (3) intensities of diffracted, reflected, and transmitted beams that can be directly measured; and (4) internal wave field intensities that can be measured indirectly from signals of secondary excitations.

Because of the interconnections of different beams due to multiple scattering, experimental techniques based on dynamical diffraction can often offer unique structural information. Such techniques include determination of impurity locations with x-ray standing waves, depth profiling with grazing-incidence diffraction and fluorescence, and direct measurements of phases of structure factors with multiple-beam diffraction. These new and developing techniques have benefited substantially from the rapid growth of synchrotron radiation facilities around the world. With more and newer-generation facilities becoming available, we believe that dynamical diffraction study of various materials will continue to expand in application and become more common and routine to materials scientists and engineers.

ACKNOWLEDGMENTS

The author would like to thank Boris Batterman, Ernie Fontes, Ken Finkelstein, and Stefan Kycia for critical reading of this manuscript. This work is supported by the National Science Foundation through CHESS under grant number DMR-9311772.

LITERATURE CITED

Afanasev, A. M. and Melkonyan, M. K. 1983. X-ray diffraction under specular reflection conditions. Ideal crystals. *Acta Crystallogr. Sec. A* 39:207–210.

Aleksandrov, P. A., Afanasev, A. M., and Stepanov, S. A. 1984. Bragg-Laue diffraction in inclined geometry. *Phys. Status Solidi A* 86:143–154.

Anderson, S. K., Golovchenko, J. A., and Mair, G. 1976. New application of x-ray standing wave fields to solid state physics. *Phys. Rev. Lett.* 37:1141–1144.

Andrews, S. R. and Cowley, R. A. 1985. Scattering of X-rays from crystal surfaces. *J. Phys. C* 18:6427–6439.

Aristov, V. V., Winter, U., Nikilin, A. Y., Redkin, S. V., Snigirev, A.A., Zaumseil, P., and Yunkin, V.A. 1988. Interference thickness oscillations of an x-ray wave on periodically profiled silicon. *Phys. Status Solidi A* 108:651–655.

Authier, A. 1970. Ewald waves in theory and experiment. *In* Advances in Structure Research by Diffraction Methods Vol. 3 (R. Brill and R. Mason, eds.) pp. 1–51. Pergamon Press, Oxford.

Authier, A. 1986. Angular dependence of the absorption-induced nodal plane shifts of x-ray stationary waves. *Acta Crystallogr. Sec. A* 42:414–425.

Authier, A. 1992. Dynamical theory of x-ray diffraction *In* International Tables for Crystallography, Vol. B (U. Shmueli, ed.) pp. 464–480. Academic, Dordrecht.

Authier, A. 1996. Dynamical theory of x-ray diffraction—I. Perfect crystals; II. Deformed crystals. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Barbee, T. W. and Warburton, W. K. 1984. X-ray evanescent and standing-wave fluorescence studies using a layered synthetic microstructure. *Mater. Lett.* 3:17–23.

Bartels, W. J., Hornstra, J., and Lobeek, D. J. W. 1986. X-ray diffraction of multilayers and superlattices. *Acta Crystallogr. Sec. A* 42:539–545.

Batterman, B. W. 1964. Effect of dynamical diffraction in x-ray fluorescence scattering. *Phys. Rev.* 133:759–764.

Batterman, B. W. 1969. Detection of foreign atom sites by their x-ray fluorescence scattering. *Phys. Rev. Lett.* 22:703–705.

Batterman, B. W. 1992. X-ray phase plate. *Phys. Rev. B* 45:12677–12681.

Batterman, B. W. and Bilderback, D. H. 1991. X-ray monochromators and mirrors. *In* Handbook on Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.) pp. 105–153. North-Holland, New York.

Batterman, B. W. and Cole, H. 1964. Dynamical diffraction of x-rays by perfect crystals. *Rev. Mod. Phys.* 36:681–717.

Bauer, G., Darhuber, A. A., and Holy, V. 1996. Structural characterization of reactive ion etched semiconductor nanostructures using x-ray reciprocal space mapping. *Mater. Res. Soc. Symp. Proc.* 405:359–370.

Baumbach, G. T., Holy, V., Pietsch, U., and Gailhanou, M. 1994. The influence of specular interface reflection on grazing incidence X-ray diffraction and diffuse scattering from superlattices. *Physica B* 198:249–252.

Bedzyk, M. J., Bilderback, D. H., Bommarito, G. M., Caffrey, M., and Schildkraut, J. S. 1988. Long-period standing waves as molecular yardstick. *Science* 241:1788–1791.

Bedzyk, M. J., Bilderback, D., White, J., Abruna, H. D., and Bommarito, M.G. 1986. Probing electrochemical interfaces with x-ray standing waves. *J. Phys. Chem.* 90:4926–4928.

Bedzyk, M. J. and Materlik, G. 1985. Two-beam dynamical solution of the phase problem: A determination with x-ray standing-wave fields. *Phys. Rev. B* 32:6456–6463.

Bedzyk, M. J., Shen, Q., Keeffe, M., Navrotski, G., and Berman, L. E. 1989. X-ray standing wave surface structural determination for iodine on Ge (111). *Surf. Sci.* 220:419–427.

Belyakov, V. and Dmitrienko, V. 1989. Polarization phenomena in x-ray optics. *Sov. Phys. Usp.* 32:697–719.

Berman, L. E., Batterman, B. W., and Blakely, J. M. 1988. Structure of submonolayer gold on silicon (111) from x-ray standing-wave triangulation. *Phys. Rev.* B 38:5397–5405.

Bernhard, N., Burkel, E., Gompper, G., Metzger, H., Peisl, J., Wagner, H., and Wallner, G. 1987. Grazing incidence diffraction of X-rays at a Si single crystal surface: Comparison of theory and experiment. *Z. Physik B* 69:303–311.

Bethe, H. A. 1928. *Ann. Phys. (Leipzig)* 87:55.

Bilderback, D. H. 1981. Reflectance of x-ray mirrors from 3.8 to 50 keV (3.3 to 0.25 A). *SPIE Proc.* 315:90–102.

Bilderback, D. H., Hoffman, S. A., and Thiel, D. J. 1994. Nanometer spatial resolution achieved in hand x-ray imaging and Laue diffraction experiments. *Science* 263:201–203.

Blume, M. and Gibbs, D. 1988. Polarization dependence of magnetic x-ray scattering. *Phys. Rev. B* 37:1779–1789.

Born, M. and Wolf, E. 1983. Principles of Optics, 6th ed. Pergamon, New York.

Borrmann, G. 1950. Die Absorption von Rontgenstrahlen in Fall der Interferenz. *Z. Phys.* 127:297–323.

Borrmann, G. and Hartwig, Z. 1965. *Z. Kristallogr. Kristallgeom. Krystallphys. Kristallchem.* 121:401.

Brummer, O., Eisenschmidt, C., and Hoche, H. 1984. Polarization phenomena of x-rays in the Bragg case. *Acta Crystallogr. Sec. A* 40:394–398.

Caticha, A. 1993. Diffraction of x-rays at the far tails of the Bragg peaks. *Phys. Rev. B* 47:76–83.

Caticha, A. 1994. Diffraction of x-rays at the far tails of the Bragg peaks. II. Darwin dynamical theory. *Phys. Rev. B* 49:33–38.

Chang, S. L. 1982. Direct determination of x-ray reflection phases. *Phys. Rev. Lett.* 48:163–166.

Chang, S. L. 1984. Multiple Diffraction of X-Rays in Crystals. Springer-Verlag, Heidelberg.

Chang, S.-L. 1998. Determination of X-ray Reflection Phases Using N-Beam Diffraction. *Acta Crystallogr. A* 54:886–894.

Chang, S. L. 1992. X-ray phase problem and multi-beam interference. *Int. J. Mod. Phys.* 6:2987–3020.

Chang, S.-L., Chao, C.-H., Huang, Y.-S., Jean, Y.-C., Sheu, H.-S., Liang, F.-J., Chien, H.-C., Chen, C.-K., and Yuan, H. S. 1999. Quantitative phase determination for macromolecular crystals using stereoscopic multibeam imaging. *Acta Crystallogr.* A 55:933–938.

Chang, S. L., King, H. E., Jr., Huang, M.-T., and Gao, Y. 1991. Direct phase determination of large macromolecular crystals using three-beam x-ray interference. *Phys. Rev. Lett.* 67:3113–3116.

Chapman, L. D., Yoder, D. R., and Colella, R. 1981. Virtual Bragg scattering: A practical solution to the phase problem. *Phys. Rev. Lett.* 46:1578–1581.

Chikawa, J.-I. and Kuriyama, M. 1991. Topography. *In* Handbook on Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.) pp. 337–378. North-Holland, New York.

Chung, J.-S. and Durbin, S. M. 1995. Dynamical diffraction in quasicrystals. *Phys. Rev. B* 51:14976–14979.

Cole, H., Chambers, F. W., and Dunn, H. M. 1962. Simultaneous diffraction: Indexing Umweganregung peaks in simple cases. *Acta Crystallogr.* 15:138–144.

Colella, R. 1972. N-beam dynamical diffraction of high-energy electrons at glancing incidence. General theory and computational methods. *Acta Crystallogr. Sec. A* 28:11–15.

Colella, R. 1974. Multiple diffraction of x-rays and the phase problem. computational procedures and comparison with experiment. *Acta Crystallogr. Sec. A* 30:413–423.

Colella, R. 1991. Truncation rod scattering: Analysis by dynamical theory of x-ray diffraction. *Phys. Rev. B* 43:13827–13832.

Colella, R. 1995. Multiple Bragg scattering and the phase problem in x-ray diffraction. I. Perfect crystals. *Comments Cond. Mater. Phys.* 17:175–215.

Colella, R. 1996. Multiple Bragg scattering and the phase problem in x-ray diffraction. II. Perfect crystals; Mosaic crystals. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Cowan, P. L., Brennan, S., Jach, T., Bedzyk, M. J., and Materlik, G. 1986. Observations of the diffraction of evanescent x rays at a crystal surface. *Phys. Rev. Lett.* 57:2399–2402.

Cowley, J. M. 1975. Diffraction Physics. North-Holland Publishing, New York.

Darhuber, A. A., Koppensteiner, E., Straub, H., Brunthaler, G., Faschinger, W., and Bauer, G. 1994. Triple axis x-ray investigations of semiconductor surface corrugations. *J. Appl. Phys.* 76:7816–7823.

Darhuber, A. A., Schittenhelm, P., Holy, V., Stangl, J., Bauer, G., and Abstreiter, G. 1997. High-resolution x-ray diffraction from multilayered self-assembled Ge dots. *Phys. Rev. B* 55:15652–15663.

Darowski, N., Paschke, K., Pietsch, U., Wang, K. H., Forchel, A., Baumbach, T., and Zeimer, U. 1997. Identification of a buried single quantum well within surface structured semiconductors using depth resolved x-ray grazing incidence diffraction. *J. Phys. D* 30:L55–L59.

Darwin, C. G. 1914. The theory of x-ray reflexion. *Philos. Mag.* 27:315–333; 27:675–690.

Darwin, C. G. 1922. The reflection of x-rays from imperfect crystals. *Philos. Mag.* 43:800–829.

de Boer, D. K. G. 1991. Glancing-incidence X-ray fluorescence of layered materials. *Phys. Rev. B* 44:498–511.

Dietrich, S. and Haase, A. 1995. Scattering of X-rays and neutrons at interfaces. *Phys. Rep.* 260: 1–138.

Dietrich, S. and Wagner, H. 1983. Critical surface scattering of x-rays and neutrons at grazing angles. *Phys. Rev. Lett.* 51: 1469–1472.

Dietrich, S. and Wagner, H. 1984. Critical surace scattering of x-rays at Grazing Angles. *Z. Phys. B* 56:207–215.

Dosch, H. 1992. Evanescent X-rays probing surface-dominated phase transitions. *Int. J. Mod. Phys. B* 6:2773–2808.

Dosch, H., Batterman, B. W., and Wack., D. C. 1986. Depth-controlled grazing-incidence diffraction of synchrotron x-radiation. *Phys. Rev. Lett.* 56:1144–1147.

Durbin, S. M. 1987. Dynamical diffraction of x-rays by perfect magnetic crystals. *Phys. Rev. B* 36:639–643.

Durbin, S. M. 1988. X-ray standing wave determination of Mn sublattice occupancy in a Cd_{1x} Mn_x Te mosaic crystal. *J. Appl. Phys*. 64:2312–2315.

Durbin, S. M. 1995. Darwin spherical-wave theory of kinematic surface diffraction. *Acta Crystallogr. Sec. A* 51:258–268.

Durbin, S. M., Berman, L. E., Batterman, B. W., and Blakely, J. M. 1986. Measurement of the silicon (111) surface contraction. *Phys. Rev. Lett.* 56:236–239.

Durbin, S. M. and Follis, G. C. 1995. Darwin theory of heterostructure diffraction. *Phys. Rev. B* 51:10127–10133.

Durbin, S. M. and Gog, T. 1989. Bragg-Laue diffraction at glancing incidence. *Acta Crystallogr. Sec. A* 45:132–141.

Ewald, P. P. 1917. Zur Begrundung der Kristalloptik. III. Die Kristalloptic der Rontgenstrahlen. *Ann. Physik (Leipzig)* 54:519–597.

Ewald, P. P. and Heno, Y. 1968. X-ray diffraction in the case of three strong rays. I. Crystal composed of non-absorbing point atoms. *Acta Crystallogr. Sec. A* 24:5–15.

Feng, Y. P., Sinha, S. K., Fullerton, E. E., Grubel, G., Abernathy, D., Siddons, D. P., and Hastings, J. B. 1995. X-ray Fraunhofer diffraction patterns from a thin-film waveguide. *Appl. Phys. Lett*. 67:3647–3649.

Fewster, P. F. 1996. Superlattices. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Fontes, E., Patel, J. R., and Comin, F. 1993. Direct measurement of the asymmetric diner buckling of Ge on Si(001). *Phys. Rev. Lett.* 70:2790–2793.

Franklin, G. E., Bedzyk, M. J., Woicik, J. C., Chien L., Patel, J. R., and Golovchenko, J.A. 1995. Order-to-disorder phase-transition study of Pb on Ge(111). *Phys. Rev. B* 51:2440–2445.

Funke, P. and Materlik, G. 1985. X-ray standing wave fluorescence measurements in ultra-high vacuum adsorption of Br on Si(111)-(1X1). *Solid State Commun.* 54:921.

Giles, C., Malgange, C., Goulon, J., de Bergivin, F., Vettier, C., Dartyge, E., Fontaine, A., Giorgetti, C., and Pizzini, S. 1994. Energy-dispersive phase plate for magnetic circular dichroism experiments in the x-ray range. *J. Appl. Crystallogr.* 27:232–240.

Golovchenko, J. A., Batterman, B. W., and Brown, W. L. 1974. Observation of internal x-ray wave field during Bragg diffraction with an application to impurity lattice location. *Phys. Rev. B* 10:4239–4243.

Golovchenko, J. A., Kincaid, B. M., Levesque, R. A., Meixner, A. E., and Kaplan, D. R. 1986. Polarization Pendellosung and the generation of circularly polarized x-rays with a quarter-wave plate. *Phys. Rev. Lett*. 57:202–205.

Golovchenko, J. A., Patel, J. R., Kaplan, D. R., Cowan, P. L., and Bedzyk, M. J. 1982. Solution to the surface registration problem using x-ray standing waves. *Phys. Rev. Lett.* 49:560.

Greiser, N. and Matrlik, G. 1986. Three-beam x-ray standing wave analysis: A two-dimensional determination of atomic positions. *Z. Phys. B* 66:83–89.

Gunther, R., Odenbach, S., Scharpf, O., and Dosch, H. 1997. Reflectivity and evanescent diffraction of polarized neutrons from Ni(110). *Physica B* 234-236:508–509.

Hart, M. 1978. X-ray polarization phenomena. *Philos. Mag. B* 38:41–56.

Hart, M. 1991. Polarizing x-ray optics for synchrotron radiation. *SPIE Proc.* 1548:46–55.

Hart, M. 1996. X-ray optical beamline design principles. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Hashizume, H. and Sakata, O. 1989. Dynamical diffraction of X-rays from crystals under grazing-incidence conditions. *J. Crystallogr. Soc. Jpn.* 31:249–255; *Coll. Phys. C* 7:225–229.

Headrick, R. L. and Baribeau, J. M. 1993. Correlated roughness in Ge/Si superlattices on Si(100). *Phys. Rev. B* 48:9174–9177.

Hirano, K., Ishikawa, T., and Kikuta, S. 1995. Development and application of x-ray phase retarders. *Rev. Sci. Instrum.* 66:1604–1609.

Hirano, K., Izumi, K., Ishikawa, T., Annaka, S., and Kikuta, S. 1991. An x-ray phase plate using Bragg case diffraction. *Jpn. J. Appl. Phys*. 30:L407–L410.

Hoche, H. R., Brummer, O., and Nieber, J. 1986. Extremely skew x-ray diffraction. *Acta Crystallogr. Sec. A* 42:585–586.

Hoche, H.R., Nieber, J., Clausnitzer, M., and Materlik, G. 1988. Modification of specularly reflected x-ray intensity by grazing incidence coplanar Bragg-case diffraction. *Phys. Status Solidi A* 105:53–60.

Hoier, R. and Marthinsen, K. 1983. Effective structure factors in many-beam x-ray diffraction—use of the second Bethe approximation. *Acta Crystallogr. Sec. A* 39:854–860.

Holy, V. 1996. Dynamical theory of highly asymmetric x-ray diffraction. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B.K. Tanner, eds.). Plenum, New York.

Holy, V. and Baumbach, T. 1994. Nonspecular x-ray reflection from rough multilayers. *Phys. Rev. B* 49:10668–10676.

Hoogenhof, W. W. V. D. and de Boer, D. K. G. 1994. GIXA (glancing incidence X-ray analysis), a novel technique in near-surface analysis. *Mater. Sci. Forum (Switzerland)* 143–147:1331–1335.

Hümmer, K. and Billy, H. 1986. Experimental determination of triplet phases and enantiomorphs of non-centrosymmetric structures. I. Theoretical considerations. *Acta Crystallogr. Sec. A* 42:127–133.

Hümmer, K., Schwegle, W., and Weckert, E. 1991. A feasibility study of experimental triplet-phase determination in small proteins. *Acta Crystallogr. Sec. A* 47:60–62.

Hümmer, K., Weckert, E., and Bondza, H. 1990. Direct measurements of triplet phases and enantiomorphs of non-centrosymmetric structures. Experimental results. *Acta Crystallogr. Sec. A* 45:182–187.

Hung, H. H. and Chang, S. L. 1989. Theoretical considerations on two-beam and multi-beam grazing-incidence x-ray diffraction: Nonabsorbing cases. *Acta Crystallogr. Sec. A* 45:823–833.

Jach, T. and Bedzyk, M.J. 1993. X-ray standing waves at grazing angles. *Acta Crystallogr. Sec. A* 49:346–350.

Jach, T., Cowan, P. L., Shen, Q., and Bedzyk, M. J. 1989. Dynamical diffraction of x-rays at grazing angle. *Phys. Rev. B* 39:5739–5747.

Jach, T., Zhang, Y., Colella, R., de Boissieu, M., Boudard, M., Goldman, A. I., Lograsso, T. A., Delaney, D. W., and Kycia, S. 1999. Dynamical diffraction and x-ray standing waves from

2-fold reflections of the quasicrystal AlPdMn. *Phys. Rev. Lett.* 82:2904–2907.

Jackson, J. D. 1975. Classical Electrodynamics, 2nd ed. John Wiley & Sons, New York.

James, R. W. 1950. The Optical Principles of the Diffraction of X-rays. G. Bell and Sons, London.

Juretschke, J. J. 1982. Invariant-phase information of x-ray structure factors in the two-beam Bragg intensity near a three-beam point. *Phys. Rev. Lett.* 48:1487–1489.

Juretschke, J. J. 1984. Modified two-beam description of x-ray fields and intensities near a three-beam diffraction point. General formulation and first-order solution. *Acta Crystallogr. Sec. A* 40:379–389.

Juretschke, J. J. 1986. Modified two-beam description of x-ray fields and intensities near a three-beam diffraction point. Second-order solution. *Acta Crystallogr. Sec. A* 42:449–456.

Kaganer, V. M., Stepanov, S. A., and Koehler, R. 1996. Effect of roughness correlations in multilayers on Bragg peaks in X-ray diffuse scattering. *Physica B* 221:34–43.

Kato, N. 1952. Dynamical theory of electron diffraction for a finite polyhedral crystal. *J. Phys. Soc. Jpn.* 7:397–414.

Kato, N. 1960. The energy flow of x-rays in an ideally perfect crystal: Comparison between theory and experiments. *Acta Crystallogr.* 13:349–356.

Kato, N. 1974. X-ray diffraction. *In* X-ray Diffraction (L. V. Azaroff, R. Kaplow, N. Kato, R. J. Weiss, A. J. C. Wilson, and R. A. Young, eds.). pp. 176–438. McGraw-Hill, New York.

Kikuchi, S. 1928. *Proc. Jpn. Acad. Sci.* 4:271.

Kimura, S., and Harada, J. 1994. Comparison between experimental and theoretical rocking curves in extremely asymmetric Bragg cases of x-ray diffraction. *Acta Crystallogr. Sec. A* 50:337.

Klapper, H. 1996. X-ray diffraction topography: Application to crystal growth and plastic deformation. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Kortright, J. B. and Fischer-Colbrie, A. 1987. Standing wave enhanced scattering in multilayer structures. *J. Appl. Phys.* 61:1130–1133.

Kossel, W., Loeck, V., and Voges, H. 1935. *Z. Phys.* 94:139.

Kovalchuk, M. V. and Kohn, V. G. 1986. X-ray standing wave—a new method of studying the structure of crystals. *Sov. Phys. Usp.* 29:426–446.

Krimmel, S., Donner, W., Nickel, B., Dosch, H., Sutter, C., and Grubel, G. 1997. Surface segregation-induced critical phenomena at FeCo(001) surfaces. *Phys. Rev. Lett.* 78:3880–3883.

Kycia, S. W., Goldman, A. I., Lograsso, T. A., Delaney, D. W., Black, D., Sutton, M., Dufresne, E., Bruning, R., and Rodricks, B. 1993. Dynamical x-ray diffraction from an icosahedral quasicrystal. *Phys. Rev.* B 48:3544–3547.

Lagomarsino, S. 1996. X-ray standing wave studies of bulk crystals, thin films and interfaces. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Lagomarsino, S., Scarinci, F., and Tucciarone, A. 1984. X-ray stading waves in garnet crystals. *Phys. Rev. B* 29:4859–4863.

Lang, J. C., Srajer, G., Detlefs, C., Goldman, A. I., Konig, H., Wang, X., Harmon, B. N., and McCallum, R. W. 1995. Confirmation of quadrupolar transitions in circular magnetic X-ray dichroism at the dysprosium L_{III} edge. *Phys. Rev. Lett.* 74:4935–4938.

Lee, H., Colella, R., and Chapman, L. D. 1993. Phase determination of x-ray reflections in a quasicrystal. *Acta Crystallogr. Sec. A* 49:600–605.

Lied, A., Dosch, H., and Bilgram, J. H. 1994: Glancing angle X-ray scattering from single crystal ice surfaces. *Physica B* 198:92–96.

Lipcomb, W. N. 1949. Relative phases of diffraction maxima by multiple reflection. *Acta Crystallogr.* 2:193–194.

Lyman, P. F. and Bedzyk, M. J. 1997. Local structure of Sn/Si(001) surface phases. *Surf. Sci.* 371:307–315.

Malgrange, C. 1996. X-ray polarization and applications. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Marra, W. L., Eisenberger, P., and Cho, A. Y. 1979. X-ray total-external-relfection Bragg diffraction: A structural study of the GaAs-Al interface. *J. Appl. Phys.* 50:6927–6933.

Martines, R. E., Fontes, E., Golovchenko, J. A., and Patel, J. R. 1992. Giant vibrations of impurity atoms on a crystal surface. *Phys. Rev. Lett.* 69:1061–1064.

Mills, D. M. 1988. Phase-plate performance for the production of circularly polarized x-rays. *Nucl. Instrum. Methods A* 266:531–537.

Moodie, A.F., Crowley, J. M., and Goodman, P. 1997. Dynamical theory of electron diffraction. In International Tables for Crystallography, Vol. B (U. Shmueki, ed.). p. 481. Academic Dordrecht, The Netherlands.

Moon, R. M. and Shull, C. G. 1964. The effects of simultaneous reflection on single-crystal neutron diffraction intensities. *Acta Crystallogr. Sec. A* 17:805–812.

Paniago, R., Homma, H., Chow, P. C., Reichert, H., Moss, S. C., Barnea, Z., Parkin, S. S. P., and Cookson, D. 1996. Interfacial roughness of partially correlated metallic multilayers studied by nonspecular X-ray reflectivity. *Physica B* 221:10–12.

Parrat, P. G. 1954. Surface studies of solids by total reflection of x-rays. *Phys. Rev.* 95:359–369.

Patel, J. R. 1996. X-ray standing waves. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, B. K. and Tanner, eds.). Plenum, New York.

Patel, J. R., Golovchenko, J. A., Freeland, P. E., and Gossmann, H.-J. 1987. Arsenic atom location on passivated silicon (111) surfaces. *Phys. Rev. B* 36:7715–7717.

Pinsker, Z. G. 1978. Dynamical Scattering of X-rays in Crystals. Springer Series in Solid-State Sciences, Springer-Verlag, Heidelberg.

Porod, G. 1952. Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden system en. *Kolloid. Z.* 125:51–57; 108–122.

Porod, G. 1982. *In* Small Angle X-ray Scattering (O. Glatter and O. Kratky, eds.). Academic Press, San Diego.

Post, B. 1977. Solution of the x-ray phase problem. *Phys. Rev. Lett.* 39:760–763.

Prins, J. A. 1930. Die Reflexion von Rontgenstrahlen an absorbierenden idealen Kristallen. *Z. Phys.* 63:477–493.

Renninger, M. 1937. Umweganregung, eine bisher unbeachtete Wechselwirkungserscheinung bei Raumgitter-interferenzen. *Z. Phys.* 106:141–176.

Rhan, H., Pietsch, U., Rugel, S., Metzger, H., and Peisl, J. 1993. Investigations of semiconductor superlattices by depth-sensitive X-ray methods. *J. Appl. Phys.* 74:146–152.

Robinson, I. K. 1986. Crystal truncation rods and surface roughness. *Phys. Rev. B* 33:3830–3836.

Rose, D., Pietsch, U., and Zeimer, U. 1997. Characterization of $In_xGa_{1-x}As$ single quantum wells, buried in GaAs[001], by grazing incidence diffraction. *J. Appl. Phys.* 81:2601–2606.

Salditt, T., Metzger, T. H., and Peisl, J. 1994. Kinetic roughness of amorphous multilayers studied by diffuse x-ray scattering. *Phys. Rev. Lett.* 73:2228–2231.

Schiff, L. I. 1955. Quantum Mechanics, 2nd ed. McGraw-Hill, New York.

Schlenker, M. and Guigay, J.-P. 1996. Dynamical theory of neutron scattering. *In* X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Schmidt, M. C. and Colella, R. 1985. Phase determination of forbidden x-ray reflections in V_3Si by virtual Bragg scattering. *Phys. Rev. Lett.* 55:715–718.

Shastri, S. D., Finkelstein, K. D., Shen, Q., Batterman, B. W., and Walko, D. A. 1995. Undulator test of a Bragg-reflection elliptical polarizer at 7.1 keV. *Rev. Sci. Instrum.* 66:1581.

Shen, Q. 1986. A new approach to multi-beam x-ray diffraction using perturbation theory of scattering. *Acta Crystallogr. Sec. A* 42:525–533.

Shen, Q. 1991. Polarization state mixing in multiple beam diffraction and its application to solving the phase problem. *SPIE Proc.* 1550:27–33.

Shen, Q. 1993. Effects of a general x-ray polarization in multiple-beam Bragg diffraction. *Acta Crystallogr. Sec. A* 49:605–613.

Shen, Q. 1996a. Polarization optics for high-brightness synchrotron x-rays. *SPIE Proc.* 2856:82.

Shen, Q. 1996b. Study of periodic surface nanosctructures using coherent grating x-ray diffraction (CGXD). *Mater. Res. Soc. Symp. Proc.* 405:371–379.

Shen, Q. 1998. Solving the phase problem using reference-beam x-ray diffraction. *Phys. Rev. Lett.* 80:3268–3271.

Shen, Q. 1999a. Direct measurements of Bragg-reflection phases in x-ray crystallography. *Phys. Rev. B* 59:11109–11112.

Shen, Q. 1999b. Expanded distorted-wave theory for phase-sensitive x-ray diffraction in single crystals. *Phys. Rev. Lett.* 83:4764–4787.

Shen, Q. 1999c. A distorted-wave approach to reference-beam x-ray diffraction in transmission cases. *Phys. Rev. B.* 61:8593–8597.

Shen, Q., Blakely, J. M., Bedzyk, M. J., and Finkelstein, K. D. 1989. Surface roughness and correlation length determined from x-ray-diffraction line-shape analysis on Ge(111). *Phys. Rev. B* 40:3480–3482.

Shen, Q. and Colella, R. 1987. Solution of phase problem for crystallography at a wavelength of 3.5 Å. *Nature* (London) 329: 232–233.

Shen, Q. and Colella, R. 1988. Phase observation in organic crystal benzil using 3.5 Å x-rays. *Acta Crystallogr. Sec. A* 44:17–21.

Shen, Q. and Finkelstein, K. D. 1990. Solving the phase problem with multiple-beam diffraction and elliptically polarized x rays. *Phys. Rev. Lett.* 65:3337–3340.

Shen, Q. and Finkelstein, K. D. 1992. Complete determination of x-ray polarization using multiple-beam Bragg diffraction. *Phys. Rev. B* 45:5075–5078.

Shen, Q. and Finkelstein, K. D. 1993. A complete characterization of x-ray polarization state by combination of single and multiple Bragg reflections. *Rev. Sci. Instrum.* 64:3451–3456.

Shen, Q. and Kycia, S. 1997. Determination of interfacial strain distribution in quantum-wire structures by synchrotron x-ray scattering. *Phys. Rev. B* 55:15791–15797.

Shen, Q., Kycia, S. W., Schaff, W. J., Tentarelli, E. S., and Eastman, L. F. 1996a. X-ray diffraction study of size-dependent strain in quantum wire structures. *Phys. Rev. B* 54:16381–16384.

Shen, Q., Shastri, S., and Finkelstein, K. D. 1995. Stokes polarimetry for x-rays using multiple-beam diffraction. *Rev. Sci. Instrum.* 66:1610–1613.

Shen, Q., Umbach, C. C., Weselak, B., and Blakely, J. M. 1993. X-ray diffraction from a coherently illuminated Si(001) grating surface, *Phys. Rev. B* 48:17967–17971.

Shen, Q., Umbach, C. C., Weselak, B., and Blakely, J. M. 1996b. Lateral correlation in mesoscopic on silicon (001) surface determined by grating x-ray diffuse scattering. *Phys. Rev. B* 53: R4237–4240.

Sinha, S. K., Sirota, E. B., Garoff, S., and Stanley, H. B. 1988. X-ray and neutron scattering from rough surfaces. *Phys. Rev. B* 38:2297–2311.

Stepanov, S. A., Kondrashkina, E. A., Schmidbauer, M., Kohler, R., Pfeiffer, J.-U., Jach, T., and Souvorov, A. Y. 1996. Diffuse scattering from interface roughness in grazing-incidence X-ray diffraction. *Phys. Rev. B* 54:8150–8162.

Takagi, S. 1962. Dynamical theory of diffraction applicable to crystals with any kind of small distortion. *Acta Crystallogr.* 15:1311–1312.

Takagi, S. 1969. A dynamical theory of diffraction for a distorted crystal. *J. Phys. Soc. Jpn.* 26:1239–1253.

Tanner, B. K. 1996. Contrast of defects in x-ray diffraction topographs. In X-ray and Neutron Dynamical Diffraction: Theory and Applications (A. Authier, S. Lagomarsino, and B. K. Tanner, eds.). Plenum, New York.

Taupin, D. 1964. Theorie dynamique de la diffraction des rayons x par les cristaux deformes. *Bull. Soc. Fr. Miner. Crist.* 87:69.

Thorkildsen, G. 1987. Three-beam diffraction in a finite perfect crystal. *Acta Crystallogr. Sec. A* 43:361–369.

Tischler, J. Z. and Batterman, B. W. 1986. Determination of phase using multiple-beam effects. *Acta Crystallogr. Sec. A* 42:510–514.

Tolan, M., Konig, G., Brugemann, L., Press, W., Brinkop, F., and Kotthaus, J. P. 1992. X-ray diffraction from laterally structured surfaces: Total external reflection and grating truncation rods. *Eur. Phys. Lett.* 20:223–228.

Tolan, M., Press, W., Brinkop, F., and Kotthaus, J. P. 1995. X-ray diffraction from laterally structured surfaces: Total external reflection. *Phys. Rev. B* 51:2239–2251.

Vineyard, G. H. 1982. Grazing-incidence diffraction and the distorted-wave approximation for the study of surfaces. *Phys. Rev. B* 26:4146–4159.

von Laue, M. 1931. Die dynamische Theorie der Rontgenstrahlinterferenzen in neuer Form. *Ergeb. Exakt. Naturwiss.* 10:133–158.

Wang, J., Bedzyk, M. J., and Caffrey, M. 1992. Resonance-enhanced x-rays in thin films: A structure probe for membranes and surface layers. *Science* 258:775–778.

Warren, B. E. 1969. X-Ray Diffraction. Addison Wesley, Reading, Mass.

Weckert, E. and Hümmer, K. 1997. Multiple-beam x-ray diffraction for physical determination of reflection phases and its applications. *Acta Crystallogr. Sec. A* 53:108–143.

Weckert, E., Schwegle, W., and Hummer, K. 1993. Direct phasing of macromolecular structures by three-beam diffraction. *Proc. R. Soc. London A* 442:33–46.

Yahnke, C. J., Srajer, G., Haeffner, D. R., Mills, D. M, and Assoufid. L. 1994. Germanium x-ray phase plates for the production of circularly polarized x-rays. *Nucl. Instrum. Methods A* 347:128–133.

Yoneda, Y. 1963. Anomalous Surface Reflection of X-rays. *Phys. Rev.* 131:2010–2013.

Zachariasen, W. H. 1945. Theory of X-ray Diffraction in Crystals. John Wiley & Sons, New York.

Zachariasen, W. H. 1965. Multiple diffraction in imperfect crystals. *Acta Crystallogr. Sec. A* 18:705–710.

Zegenhagen, J., Hybertsen, M. S., Freeland, P. E., and Patel, J. R. 1988. Monolayer growth and structure of Ga on Si(111). *Phys. Rev. B* 38:7885–7892.

Zhang, Y., Colella, R., Shen, Q., and Kycia, S. W. 1999. Dynamical three-beam diffraction in a quasicrystal. *Acta Crystallogr. A* 54:411–415.

KEY REFERENCES

Authier et al., 1996. See above.

Contains an excellent selection of review papers on modern dynamical theory topics.

Batterman and Cole, 1964. See above.

One of the most cited articles on x-ray dynamical theory.

Colella, 1974. See above.

Provides a fundamental formulation for NBEAM dynamical theory.

Zachariasen, 1945. See above.

A classic textbook on x-ray diffraction theories, both kinematic and dynamical.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

A	Effective crystal thickness parameter
b	Ratio of direction cosines of incident and diffracted waves
$\mathbf{D}$	Electric displacement vector
$\mathbf{D}_0$	Incident electric displacement vector
$\mathbf{D}_{\bar{0}}$	Specular reflected wave field
$\mathbf{D}_\mathrm{H}$	Fourier component $\mathbf{H}$ of electric displacement vector
$\mathbf{D}_0^\mathrm{i}, \mathbf{D}_\mathrm{H}^\mathrm{i}$	Internal wave fields
$\mathbf{D}_0^\mathrm{e}, \mathbf{D}_\mathrm{H}^\mathrm{i}$	External wave fields
$\mathbf{E}$	Electric field vector
F_0'	Real part of F_0
F_0''	Imaginary part of F_0
F_H	Structure factor of reflection $\mathbf{H}$
$\mathbf{G}$	Reciprocal lattice vector for reference reflection
$\mathbf{H}$	Reciprocal lattice vector
$\overline{\mathbf{H}}$	Negative of $\mathbf{H}$
$\mathbf{H}$–$\mathbf{G}$	Difference between two reciprocal lattice vectors
I_H	Intensity of reflection $\mathbf{H}$
$K_{0\mathrm{n}}$	Component of internal incident wave vector normal to surface
$\mathbf{K}_0$	Incident wave vector inside crystal
$\mathbf{k}_0$	Incident wave vector outside crystal
$\mathbf{K}_\mathrm{H}$	Wave vector inside crystal
$\mathbf{k}_\mathrm{H}$	Wave vector outside crystal
$\mathbf{L}$	reciprocal lattice vector for a detour reflection
$\underline{\mathbf{M}}, \underline{\mathbf{A}}$	Matrices used with polarization density matrix
n	Index of refraction
N	Number of unit cells participating in diffraction
$\mathbf{n}$	Unit vector along surface normal
$\mathbf{O}$	Reciprocal lattice origin
P	Polarization factor
P_1, P_2, P_3	Stokes-Poincare polarization parameters
P_H/P_0	Total diffracted power normalized to the incident power
$\mathbf{r}$	Real-space position vector
R	Reflectivity
r_0, r_G	Distorted-wave amplitudes in expanded distorted-wave theory
r_0, t_0	Fresnel reflection and transmission coefficient at an interface
r_e	Classical radius of an electron, 2.818×10^{-5} angstroms
$S, \mathbf{S}_\mathrm{T}$	Poynting vector
$\underline{\mathbf{T}}, \underline{\mathbf{A}}, \underline{\mathbf{B}}, \underline{\mathbf{C}}$	Matrices used in NBEAM theory
$\mathbf{u}$	Unit vector along wave propagation direction
w	Intrinsic diffraction width, Darwin width
Γ	$= r_e\lambda^2(\pi V_\mathrm{c})$
α_H	Phase of FH, structure factor of reflection $\mathbf{H}$
α, β	Branches of dispersion surface
$\delta\varepsilon(\mathbf{r})$	Susceptibility function of a crystal
$\varepsilon(\mathbf{r})$	Dielectric function of a crystal
ε_0	Dielectric constant in vacuum
γ_0	Direction cosine of incident wave vector
γ_H	Direction cosine of diffracted wave vector
η, η_G	Angular deviation from θ_B normalized to Darwin width
λ	Wavelength
μ_0	Linear absorption coefficient
ν	Intrinsic dynamical phase shift
π	Polarization unit vector within scattering plane
θ	Incident angle
θ_B	Bragg angle
θ_c	Critical angle
$\underline{\rho}$	Polarization density matrix
ρ_0	Average charge density
$\rho(\mathbf{r})$	Charge density
σ	Polarization unit vector perpendicular to scattering plane
τ	Penetration depth of an evanescent wave
ξ_0	Correction to dispersion surface $\mathbf{O}$ due to two-beam diffraction
ξ_H	Correction to dispersion surface $\mathbf{H}$ due to two-beam diffraction
ψ	Azimuthal angle around the scattering vector

QUN SHEN
Cornell University
Ithaca, New York

COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS

INTRODUCTION

Diffuse intensities in alloys are measured by a variety of techniques, such as x ray, electron, and neutron scattering. Above a structural phase-transformation boundary, typically in the solid-solution phase where most materials processing takes place, the diffuse intensities yield valuable information regarding an alloy's tendency to order. This has been a mainstay characterization technique for binary alloys for over half a century. Although multicomponent metallic alloys are the most technologically important, they also pose a great experimental and theoretical challenge. For this reason, a vast majority of experimental and theoretical effort has been made on binary systems, and most investigated "ternary" systems are either limited to a small percentage of ternary solute (say, to investigate electron-per-atom effects) or they are pseudo-binary systems. Thus, for multicomponent alloys the questions are: how can you interpret diffuse scattering experiments on such systems and how does one theoretically predict the ordering behavior?

This unit discusses an electronic-based theoretical method for calculating the structural ordering in multicomponent alloys and understanding the electronic origin for this chemical-ordering behavior. This theory is based on the ideas of concentration waves using a modern electronic-structure method. Thus, we give examples (see Data Analysis and Initial Interpretation) that show how we determined the electronic origin behind the unusual ordering behavior in a few binary and ternary alloy systems that were not understood prior to our work. From the start, the theoretical approach is compared and contrasted to other complimentary techniques for completeness. In addition, some details are given about the theory and its underpinnings. Please do not let this deter you from jumping ahead and reading Data Analysis and Initial Interpretation and Principles of the Method. For those not familiar with electronic properties and how they manifest themselves in the ordering properties, the discussion following Equation 27 may prove useful for understanding Data Analysis and Initial Interpretation.

Importantly, for the more general multicomponent case, we describe in the context of concentration waves how to extract more information from diffuse-scattering experimental data (see Concentration Waves in Multicomponent Alloys). Although developed to understand the calculated diffuse-scattering intensities, this analysis technique allows one to determine completely the type of ordering described in the numerous chemical pair correlations that must be measured. In fact, what is required (in addition to the ordering wavevector) is an ordering "polarization" of the concentration wave that is contained in the diffuse intensities. The example case of face-centered cubic (fcc) Cu_2NiZn is given.

For definitions of the symbols used throughout the unit, see Table 1.

For binary or multicomponent alloys, the atomic short-range order (ASRO) in the disordered solid-solution phase is related to the thermally induced concentration fluctuations in the alloy. Such fluctuations in the chemical site occupations are the (infinitesimal) deviations from a homogeneously random state, and are directly related to the chemical pair correlations in the alloy (Krivoglaz, 1969). Thus, the ASRO provides valuable information on the atomic structure to which the disordered alloy is tending—i.e., it reveals the chemical ordering tendencies in the high-temperature phase (as shown by Krivoglaz, 1969; Clapp and Moss, 1966; de Fontaine, 1979; Khachaturyan, 1983; Ducastelle, 1991). Importantly, the ASRO can be determined experimentally from the diffuse scattering intensities measured in reciprocal space either by x rays (X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS), neutrons (NEUTRON TECHNIQUES), or electrons (LOW-ENERGY ELECTRON DIFFRACTION; Sato and Toth, 1962; Moss, 1969; Reinhard et al., 1990). However, the underlying microscopic or electronic origin for the ASRO cannot be determined from such experiments, only their observed indirect effect on the order. Therefore, the calculation of diffuse intensities in high-temperature, disordered alloys based on electronic density-functional theory (DFT; SUMMARY OF ELECTRONIC STRUCTURE METHODS) and the subsequent connection of those intensities to its microscopic origin(s) provides a fundamental understanding of the experimental data and phase instabilities. These are the principal themes that we will emphasize in this unit.

The chemical pair correlations determined from the diffuse intensities are written usually as normalized probabilities, which are then the familiar Warren-Cowley parameters (defined later). In reciprocal space, where scattering data is collected, the Warren-Cowley parameters are denoted by $\alpha_{\mu\nu}(\mathbf{k})$, where μ and ν label the species (1 to N in an N-component alloy) and where $\mathbf{k}$ is the scattering wave vector. In the solid-solution phase, the sharp Bragg diffraction peaks (in contrast to the diffuse peaks) identify the underlying Bravais lattice symmetry, such as, fcc and body-centered cubic (bcc), and determine the possible set of available wave vectors. (We will assume heretofore that there is no change in the Bravais lattice.) The diffuse maximal peaks in $\alpha_{\mu\nu}(\mathbf{k})$ at wave vector $\mathbf{k}_0$ indicate that the disordered phase has low-energy ordering fluctuations with that periodicity, and $\mathbf{k}_0$ is not where the Bragg reflection sits. These fluctuations are not stable but may be long-lived, and they indicate the nascent ordering tendencies of the disordered alloy. At the so-called spinodal temperature, T_{sp}, elements of the $\alpha_{\mu\nu}(\mathbf{k}=\mathbf{k}_0)$ diverge, indicating the absolute instability of the alloy to the formation of a long-range ordered state with wavevector $\mathbf{k}_0$. Hence, it is clear that the fluctuations are related to the disordered alloy's stability matrix. Of course, there may be more than one (symmetry unrelated) wavevector prominent, giving a more complex ordering tendency.

Because the concentrations of the alloy's constituents are then modulated with a wave-like periodicity, such orderings are often referred to as "concentration waves" (Khachaturyan, 1972, 1983; de Fontaine, 1979). Thus, *any* ordered state can be thought of as a modulation of the disordered state by a thermodynamically stable concentration wave. Keep in mind that any arrangement of atoms on a Bravais lattice (sites labeled by i) may be

Table 1. Table of Symbols

Symbol	Meaning
AuFe	Standard alloy nomenclature such that underlined element is majority species, here Au-rich
$L1_0$, $L1_2$, $L1_1$, etc.	Throughout we use the Strukturbericht notation (http://dave.nrl,navy.mil) lattice where A are monotonic (e.g., A_1 = fcc; A_2 = bcc), B_2 [e.g., CsCl with (111) wavevector ordering], and $L1_0$ (e.g., CuAu with ⟨100⟩ wavevector ordering), and so on
⟨...⟩	Thermal/configurational average
Bold symbols	Vectors
k and **q**	Wavevectors in reciprocal space
$\mathbf{k}_0$	Specific set of symmetry-related, ordering wavevector
Star of **k**	The star of a wavevector is the set of symmetry equivalent **k** values, e.g., in fcc, the star of **k** = (100) is {(100), (010), (001)}
N	Number of elements in a multicomponent alloy, giving $N-1$ independent degrees of freedom because composition is conserved
(h, k, l)	General k-space (reciprocal lattice) point in the first Brillouin zone
i, j, k, etc.	Refer to enumeration of real-space lattice site
μ, ν, etc.	Greek symbols refer to elements in alloy, i.e., species labels
$\mathbf{R}_i$	Real-space lattice position for ith site
$\xi_{\mu,i}$	Site occupation variable (1, if μ-type atoms at ith site, 0 otherwise)
σ	Branch index for possible multicomponent ordering polarizations, i.e., sublattice occupations relative to "host" element (see text).
$c_{\mu,i}$	Concentration of μ type atoms at ith site, which is the thermal average of $\xi_{\mu,i}$. As this is between 0 and 1, this can also be thought of as a site-occupancy probability.
$\delta_{\mu,\nu}$	Kronecker delta (1 if subscripts are same, 0 otherwise). Einstein summation is not used in this text.
$q_{\mu\nu,ij}$	Real-space atomic pair-correlation function (not normalized). Generally, it has two species labels, and two site indices.
$\alpha_{\mu\nu,ij}$	Normalized real-space atomic pair-correlation function, traditionally referred to as the Warren-Cowley short-range-order parameter. Generally, it has two species labels, and two site indices. $\alpha_{ii} = 1$ by definition (see text).
$q_{\mu\nu}(\mathbf{k})$	Fourier transform of atomic pair-correlation function
$\alpha_{\mu\nu}(\mathbf{k})$	Experimentally measured Fourier transform of normalized pair-correlation function, traditionally referred to as the Warren-Cowley short-range-order parameter. Generally, it has two species labels. For binary alloy, no labels are required, which is more familiar to most people.
$\mathbf{e}^{\sigma}_{\mu}(\mathbf{k})$	For N-component alloys, element of eigenvector (or eigenmode) for concentration-wave composed of $N-1$ branches σ and $N-1$ independent species μ. This is 1 for binary alloy, but between 0 and 1 for an N-component alloy. As we report, this can be measured experimentally to determine the sublattice ordering in a multicomponent alloy, as done recently by ALCHEMI measurements.
$\eta^{\sigma}_{S}(T)$	Temperature-dependent long-range-order parameter for branch index σ, which is between 0 (disordered phase) and 1 (fully ordered phase)
T	Temperature (units are given in text)
F	Free energy
Ω	Grand potential of alloy. With subscript "*e*," it is the electronic grand potential of the alloy, where the electronic degrees of freedom have not been integrated out
$N(E)$	The electronic integrated density of states at an energy E
$n(E)$	The electronic density of states at an energy E
$t_{\alpha,i}$	Single-site scattering matrix, which determines how an electron will scatter off a single atom
$\tau_{\alpha,ii}$	Electronic scattering-path operator, which completely details how an electron scatters through an array of atoms

Fourier-wave decomposed, i.e., considered a "concentration wave." For a binary ($A_{1-c}B_c$), the concentration wave (or site occupancy) is simply $c_i = c + \Sigma_k[Q(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{R}_i} + \text{c.c.}]$, with the wavevectors limited to the Brillouin zone−associated $Q(\mathbf{k})$ with the underlying Bravais lattice of the disordered alloy, and where the amplitudes dictate strength of ordering (c.c. stands for complex conjugate). For example, a peak in $\alpha_{\mu\nu}(\mathbf{k}_0 = \{001\})$ within a 50-50 binary fcc solid solution indicates an instability toward alternating layers along the z direction in real space, such as in the Cu-Au structure [designated as $L1_0$ in Strukturbericht notation (see Table 1) and having alternating Cu/Au layers along (001)]. Of course, at high temperatures, all wavevectors related by the symmetry operations of the disordered lattice (referred to as a *star*) are degenerate, such as the ⟨100⟩ star comprised of (100), (010), and (001). In contrast, a $\mathbf{k}_0 = (000)$ peak indicates clustering because the associated wavelength of the concentration modulation is very long range. Interpretation of the results of our first-principles calculations is greatly facilitated by the concentration wave concept, especially for multicomponent alloys, and we will explain results in that context.

In the high-temperature disordered phase, where most materials processing takes place, this local atomic ordering governs many materials properties. In addition, these incipient ordering tendencies are *often* indicative of the long-range order (LRO) found at lower temperatures, even if the transition is first order; that is, the ASRO is a precusor of the LRO phase. For these two additional reasons, it is important to predict and to understand fundamentally this ubiquitous alloying behavior. To be precise for the experts and nonexperts alike, strictly speaking,

the fluctuations in the disordered state reveal the low-temperature, long-range ordering behavior for a second-order transition, with critical temperature $T_c = T_{sp}$. On the other hand, for first-order transitions (with $T_c > T_{sp}$), symmetry arguments indicate that this can be, but does not have to be, the case (Landau, 1937a,b; Lifshitz, 1941, 1942; Landau and Lifshitz, 1980; Khachaturyan, 1972, 1983). It is then possible that the system undergoes a first-order transition to an ordering that preempts those indicated by the ASRO and leads to LRO of a different periodicity unrelated to $\mathbf{k}_0$. Keep in mind, while not every alloy has an experimentally realizable solid-solution phase, the ASRO of the hypothetical solid-solution phase is still interesting because it is indicative of the ordering interactions in the alloy, and, is typically indicative of the long-ranged ordered phases.

Most metals of technological importance are alloys of more than two constituents. For example, the easy-forming, metallic glasses are composed of four and five elements (Inoue et al., 1990; Peker and Johnson, 1993), and traditional steels have more than five active elements (Lankford et al., 1985). The enormous number of possible combinations of elements makes the search for improved or novel metallic properties a daunting proposition for both theory and experiment. Except for understanding the "electron-per-atom" (e/a) effects due to small ternary additions, measurement of ASRO and interpretation of diffuse scattering experiments in multicomponent alloys is, in fact, a largely uncharted area. In a binary alloy, the theory of concentration waves permits one to determine the structure indicated by the ASRO given *only* the ordering wavevector (Khachaturyan, 1972, 1983; de Fontaine, 1975, 1979). In multicomponent alloys, however, the concentration waves have additional degrees of freedom corresponding to polarizations in "composition space," similar to "branches" in the case of phonons in alloys (Badalayan et al., 1969; de Fontaine, 1973; Althoff et al., 1996); thus, more information is required. These polarizations are determined by the electronic interactions and they determine the sublattice occupations in partially ordered states (Althoff et al., 1996).

From the point of view of alloy design, and at the root of alloy theory, identifying and understanding the electronic origins of the ordering tendencies at high temperatures and the reason why an alloy adopts a specific low-temperature state gives valuable guidance in the search for new and improved alloys via "tuning" an alloy's properties at the most fundamental level. In metallic alloys, for example, the electrons cannot be allocated to specific atomic sites, nor can their effects be interpreted in terms of pairwise interactions. For addressing ASRO in specific alloys, it is generally necessary to solve the many-electron problem as realistically and as accurately as possible, and then to connect this solution to the appropriate compositional, magnetic, or displacive correlation functions measured experimentally.

To date, most studies from first-principle approaches have focused on binary alloy phase diagrams, because even for these systems the thermodynamic problem is extremely nontrivial, and there is a wealth of experimental data for comparison. This unit will concentrate on the techniques employed for calculating the ASRO in binary and multicomponent alloys using DFT methods. We will not include, for example, simple parametric phase stability methods, such as CALPHAD (Butler et al., 1997; Saunders, 1996; Oates et al., 1996), because they fail to give any fundamental insight and cannot be used to predict ASRO. In what follows, we give details of the chemical pair correlations, including connecting what is measured experimentally to that developed mathematically. Because we use an electronic DFT based, mean-field approach, some care will be taken throughout the text to indicate innate problems, their solutions, quantitative and qualitative errors, and resolution accomplished within mean-field means (but would agree in great detail with more accurate, if not intractable, means). We will also discuss at some length the interesting means of interpreting the type of ASRO in multicomponent alloys from the diffuse intensities, important for both experiment and theory. Little of this has been detailed elsewhere, and, with our applications occurring only recently, this important information is not widely known.

Before presenting the electronic basis of the method, it is helpful to develop a fairly unique approach based on classical density-functional theory that not only can result in the well-known, mean field equations for chemical potential and pair correlation but may equally allow a DFT-based method to be developed for such quantities. Because the electronic DFT underpinnings for the ASRO calculations are based on a rather mathematical derivation, we try to discuss the important physical content of the DFT-based equations through truncated versions of them, which give the essence of the approach. In the Data Analysis and Initial Interpretation section, we discuss the role of several electronic mechanisms that produce strong CuAu [$L1_0$ with (001) wavevector] order in NiPt, Ni_4Mo [or $(1\frac{1}{2}0)$ wavevector] ordering in AuFe alloys, both commensurate and incommensurate order in fcc Cu-Ni-Zn alloys, and the novel CuPt [or $L1_1$, with $(\frac{1}{2}\frac{1}{2}\frac{1}{2})$ wave vector] order in fcc CuPt. Prior to these results, and very relevant for NiPt, we discuss how charge within a homogeneously random alloy is actually correlated through the local chemical environment, even though there are no chemical correlations.

At minimum, a DFT-based theory of ASRO, whose specific advantage is the ability to connect features in the ASRO in multicomponent alloys with features of the electronic structure of the disordered alloy, would be very advantageous for establishing trends, much the way Hume-Rothery established empirical relationships to trends in alloy phase formation. Some care will be given to list briefly where such calculations are relevent, in evolution or in defect, as well as those that complement other techniques. It is clear then that this is not an exhaustive review of the field, but an introduction to a specific approach.

Competitive and Related Techniques

Traditionally, DFT-based band structure calculations focus on the possible ground-state structures. While it is clearly valuable (and by no means trivial) to predict the

ground-state crystal structure from first principles, it is equally important to expand this to partially ordered and disordered phases at high temperatures. One reason for this is that ASRO measurements and materials processing take place at relatively high temperatures, typically in a disordered phase. Basically, today, this calculation can be done in two distinct (and usually complementary) ways.

First, methods based on effective chemical interactions obtained from DFT methods have had successes in determining phase diagrams and ASRO (Asta and Johnson, 1997; Wolverton and Zunger, 1995a; Rubin and Finel, 1995). This is, e.g., the idea behind the cluster-expansion method proposed by Connolly and Williams (1983), also referred to as the structural inversion method (SIM). Briefly, in the cluster-expansion method a fit is made to the formation energies of a few (up to several tens of) ordered lattice configurations using a generalized Ising model (which includes 2-body, 3-body, up to N-body clusters, whatever is required [in principle] to produce effective-chemical interactions (ECIs). These ECIs approximate the formation energetics of all other phases, including homogeneously random, and are used as input to some classical statistical mechanics approach, like Monte Carlo or the cluster-variation method (CVM), to produce ASRO or phase-boundary information.

While this is an extremely important first-principles method, and is highlighted elsewhere in this chapter (PREDICTION OF PHASE DIAGRAMS), it is difficult from this approach to discern any electronic origin because all the underlying electronic information has been integrated out, obscuring the quantum mechanical origins of the ordering tendencies. Furthermore, careful (reiterative) checks have to be made to validate the convergence of the fit with a number of structures, of stoichiometries, and the range and multiplet structure of interactions. The inclusion of magnetic effects or multicomponent additions begins to add such complexity that the cluster expansion becomes more and more difficult (and delicate), and the size of the electronic-structure unit cells begins to grow very large (depending on the DFT method, growing as N to N^3, where N is the number of atoms in the unit cell). The use of the CVM, e.g., quickly becomes uninviting for multicomponent alloys, and it then becomes necessary to rely on Monte Carlo methods for thermodynamics, where interpretation sometimes can be problematic. Nevertheless, this approach can provide ASRO and LRO information, including phase-boundary (global stability) information. If, however, you are interested in calculating the ASRO for just one multicomponent alloy composition, it is more reliable and efficient to perform a *fixed-composition* SIM using DFT methods to get the effective ECIs, because fewer structures are required and the subtleties of composition do not have to be reproduced (McCormack et al., 1997). In this mode, the fitted interactions are more stable and multiplets are suppressed; however, global stability information is lost.

A second approach (the concentration-wave approach), which we shall present below, involves use of the (possible) high-temperature disordered phase (at fixed composition) as a reference and looks for the types of local concentration fluctuations and ordering instabilities that are energetically allowed as the temperature is lowered. Such an approach can be viewed as a linear-response method for thermodynamic degrees of freedom, in much the same way that a phonon dynamical matrix may be calculated within DFT by expanding the vibrational (infinitesimal) displacements about the ideal Bravais lattice (i.e., the high-symmetry reference state; Gonze, 1997; Quong and Lui, 1997; Pavone et al., 1996; Yu and Kraukauer, 1994). Such methods have been used for three decades in classical DFT descriptions of liquids (Evans, 1979), and, in fact, there is a 1:1 mapping from the classical to electronic DFT (Györffy and Stocks, 1983). These methods may therefore be somewhat familiar in mathematical foundation. Generally speaking, a theory that is based on the high-temperature, disordered state is not biased by any *a priori* choice of chemical structures, which may be a problem with more traditional total-energy or cluster-expansion methods. The major disadvantage of this approach is that no global stability information is obtained, because only the local stability at one concentration is addressed. Therefore, the fact that the ASRO for a specific concentration can be directly addressed is both a strength and shortcoming, depending upon one's needs. For example, if the composition dependence of the ASRO at five specific compositions is required, only five calculations are necessary, whereas in the first method described above, depending on the complexity, a great many alloy compositions and structural arrangements at those compositions are still required for the fitting (until the essential physics is somehow, maybe not transparently, included). Again, as emphasized in the introduction, a great strength of the first-principles concentration-wave method is that the electronic mechanisms responsible for the ordering instabilities may be obtained.

Thus, in a great many senses, the two methods above are very complementary, rather than competing. Recently, the two methods have been used simultaneously on binary (Asta and Johnson, 1997) and ternary alloys (Wolverton and de Fontaine, 1994; McCormack et al., 1997). Certain results from both methods agree very well, but each method provides additional (complementary) information and viewpoints, which is very helpful from a computer alloy design perspective.

Effective Interactions from High-Temperature Experiments

While not really a first-principles method, it is worth mentioning a third method with a long-standing history in the study of alloys and diffuse-scattering data—using inverse Monte Carlo techniques based upon a generalized Ising model to extract ECIs from experimental diffuse-scattering data (Masanskii et al., 1991; Finel et al., 1994; Barrachin et al., 1994; Pierron-Bohnes et al., 1995; Le Bolloc'h et al., 1997). Importantly, such techniques have been used typically to extract the Warren-Cowley parameters in real space from the **k**-space data because it is traditional to interpret the experiment in this fashion. Such ECIs have been used to perform Monte Carlo calculations of phase boundaries, and so on. While it may be useful to extract the Warren-Cowley parameters via this route, it is important to understand some fundamental points

that have not been appreciated until recently: the ECIs so obtained (1) are **not** related to any fundamental alloy Hamiltonian; (2) are **parameters** that achieve a best fit to the measured ASRO; and (3) should not be trusted, in general, for calculating phase boundaries.

The origin and consequences of these three remarks are as follows. It should be fairly obvious that, given enough ECIs (i.e., fitting degrees of freedom), a fit of the ASRO is possible. For example, one may use many pairs of ECIs, or fewer pairs if some multiplet interactions are included, and so on (Finel et al., 1994; Barrachin et al., 1994). Therefore, it is clear that the fit is not unique and does not represent anything fundamental; hence, points 1 and 2 above. The only important matter for the fitting of the ASRO is the **k**-space location of the maximal intensities and their heights, which reveal both the type and strength of the ASRO, at least for binaries where such a method has been used countless times. Recently, a very thorough study was performed on a simple model alloy Hamiltonian to exemplify some of these points (Wolverton et al., 1997). In fact, while different sets of ECIs may satisfy the fitting procedure and lead to a good reproduction of the experimental ASRO, there is no *a priori* guarantee that all sets of ECIs will lead to equivalent predictions of other physical properties, such as grain-boundary energies (Finel et al., 1994; Barrachin et al., 1994).

Point 3 is a little less obvious. If both the type and strength of the ASRO are reproduced, then the ECIs are accurately reproducing the energetics associated with the infinitesimal-amplitude concentration fluctuations in the high-temperature disordered state. They may not, however, reflect the strength of the finite-amplitude concentration variations that are associated with a (possibly strong) first-order transition from the disordered to a long-range ordered state. In general, the energy gained by a first-order transformation is larger than suggested by the ASRO, which is why $T_c > T_{sp}$. In the extreme case, it is quite possible that the ASRO produces a set of ECIs that produce ordering type phase boundaries (with a negative formation energy), whereas the low-temperature state is phase separating (with a positive formation energy). An example of this can be found in the Ni-Au system (Wolverton and Zunger, 1997).

Keep in mind, however, that this is a generic comment and much understanding can certainly be obtained from such studies. Nevertheless, this should emphasize the need (1) to determine the underlying origins for the fundamental thermodynamic behavior, (2) to connect high and low temperature properties and calculations, and (3) to have complementary techniques for a more thorough understanding.

PRINCIPLES OF THE METHOD

After establishing general definitions and the connection of the ASRO to the alloy's free energy, we show, a simple standard Ising model, the well-known Krivoglaz-Clapp-Moss form (Krivoglaz, 1969; Clapp and Moss, 1966), connecting the so-called "effective chemical interactions" and the ASRO. We then generalize these concepts to the more accurate formulation involving the electronic grand potential of the disordered alloy, which we base on a DFT Hamiltonian.

Since we wish to derive the pair correlations from the electronic interactions inherent in the high-temperature state, it is most straightforward to employ a simple two-state, Ising-like variable for each alloy component and to enforce a single-occupancy constraint on each site in the alloy. This approach generates a model which straightforwardly deals with an arbitrary number of species, in contrast to an approach based on an N-state spin model (Ceder et al., 1994), which produces a mapping between the spin and concentration variables that is nonlinear.

With this Ising-like representation, any atomic configuration of an alloy (whether ordered, partially ordered, or disordered) is described by a set of occupation variables, $\{\xi_{\mu,i}\}$, where μ is the species label and i labels the lattice site. The variable $\xi_{\mu,i}$ is equal to 1 if an atom of species μ occupies the site i; otherwise it is 0. Because there can be only one atom per lattice site (i.e., a single-occupancy constraint: $\Sigma_\mu \xi_{\mu,i} = 1$) there are $(N - 1)$ independent occupation variables at each site for an N-component alloy. This single-occupancy constraint is implemented by designating one species as the "host" species (say, the Nth one) and treating the host variables as dependent. The site probability (or sublattice concentration) is just the thermodynamic average (denoted by $\langle \ldots \rangle$) of the site occupations, i.e., $c_{\mu,i} = \langle \xi_{\mu,i} \rangle$ which is between 0 and 1. For the disordered state, with no long-range order, $c_{\mu,i} = \bar{c}_\mu$ for all sites i. (Obviously, the presence of LRO is reflected by a nonzero value of $c_{\mu,i} = \bar{c}_\mu$, which is one possible definition of a LRO parameter.) In all that follows, because the meaning without a site index is obvious, we will forego the overbar on the *average* concentration.

General Background on Pair Correlations

The atomic pair-correlation functions, that is, the correlated fluctuations about the average probabilities, are then properly defined as:

$$\begin{aligned} q_{\mu\nu,ij} &= \langle (\xi_{\mu,i} - c_{\mu,i})(\xi_{\nu,j} - c_{\nu,j}) \rangle \\ &= \langle \xi_{\mu,i} \xi_{\nu,j} \rangle - \langle \xi_{\mu,i} \rangle \langle \xi_{\nu j} \rangle \end{aligned} \tag{1}$$

which reflects the presence of ASRO. Note that pair correlation is of rank $(N - 1)$ for an N-component alloy because of our choice of independent variables (the "host" is dependent). Once the portion of rank $(N - 1)$ has been determined, the "dependent" part of the full N-dimensional correlation function may be found by the single-occupancy constraint. Because of the dependencies introduced by this constraint, the N-dimensional pair-correlation function is a singular matrix, whereas, the "independent" portion of rank $(N - 1)$ is nonsingular (it has an inverse) everywhere above the spinodal temperature.

It is important to notice that, by definition, the site-diagonal part of the pair correlations, i.e., $\langle \xi_{\mu,i}\, \xi_{\mu,j} \rangle$, obeys a sum rule because $(\xi_{\mu,i})^2 = \xi_{\mu,i}$,

$$q_{\mu\nu,ii} = c_\mu (\delta_{\mu\nu} - c_\nu) \tag{2}$$

where $\delta_{\mu\nu}$ is a Kronecker delta (and there is no summation over repeated indices). For a binary alloy, with $c_A + c_B = 1$, there is only one independent composition, say, c_A and c_B is the "host" so that there is only one pair correlation and $q_{AA,ii} = c_A(1 - c_A)$. It is best to define the pair correlations in terms of the so-called Warren-Cowley parameters as:

$$\alpha_{\mu\nu,ij} = \frac{q_{\mu\nu,ij}}{c_\mu(\delta_{\mu\nu} - c_\nu)} \tag{3}$$

Note that for a binary alloy, the single pair correlation is $\alpha_{AA,ij} = q_{AA,ij}/[c_A(1 - c_A)]$ and the AA subscripts are not needed. Clearly, the Warren-Cowley parameters are normalized to range between ± 1, and, hence, they are the joint probabilities of finding two particular types of atoms at two particular sites.

The pair correlations defined in Equation 3 are, of course, the same pair correlations that are measured in diffuse-scattering experiments. This is seen by calculating the scattering intensity by averaging thermodynamically the square of the scattering amplitude, $A(\mathbf{k})$. For example, the $A(\mathbf{k})$ for a binary alloy with N_s atoms is given by $(1/N_s)\Sigma_i[f_A\xi_{A,i} + f_B(1 - \xi_{A,i}))]e^{i\mathbf{k}\cdot\mathbf{R}_i}$, for on site i you are either scattering off an "A" atom or "not an A" atom. Here f_μ is the scattering factor for x rays; use b_μ for neutrons. The scattering intensity, $I(\mathbf{k})$, is then

$$\begin{aligned} I(\mathbf{k}) &= \langle |A(\mathbf{k})|^2 \rangle \\ &= \delta_{\mathbf{k},0}[c_A f_A + (1 - c_A)f_B]^2 \qquad \text{(Bragg term)} \\ &\quad + (f_A - f_B)^2 \frac{1}{N_s} \Sigma_{ij} q_{\mu\nu,ij} e^{i\mathbf{k}\cdot(\mathbf{R}_i - \mathbf{R}_j)} \qquad \text{(diffuse term)} \end{aligned} \tag{4}$$

The first term in the scattering intensity is the Bragg scattering found from the average lattice, with an intensity given by the compositionally averaged scattering factor. The second term is the so-called diffuse-scattering term, and it is the Fourier transform of Equation 1. Generally, the diffuse scattering intensity for an N-component alloy (relevant to experiment) is then

$$I_{\text{diff}}(\mathbf{k}) = \sum_{\mu \neq \nu = 1}^{N} (f_\mu - f_\nu)^2 q_{\mu\nu}(\mathbf{k}) \tag{5}$$

or the sum may also go from 1 to $(N - 1)$ if $(f_\mu - f_\nu)^2$ is replaced by $f_\mu f_\nu$. The various ways to write this arise due to the single-occupancy constraint. For direct comparison to scattering experiments, theory needs to calculate $q_{\mu\nu}(\mathbf{k})$. Similarly, the experiment can only measure the $\mu \neq \nu$ portion of the pair correlations because it is the only part that has scattering contrast (i.e., $f_\mu - f_\nu \neq 0$) between the various species. The remaining portion is obtained by the constraints. In terms of experimental Laue units [i.e., $I_{\text{laue}} = (f_\mu - f_\nu)^2 c_\mu(\delta_{\mu\nu} - c_\nu)$], $I_{\text{diff}}(\mathbf{k})$ may also be easily given in terms of Warren-Cowley parameters.

When the free-energy curvature, i.e., $q^{-1}(\mathbf{k})$, goes through zero, the alloy is unstable to chemical ordering, and $q(\mathbf{k})$ and $\alpha(\mathbf{k})$ diverge at T_{sp}. So measurements or calculations of $q(\mathbf{k})$ and $(\mathbf{k})$ are a direct means to probe the free energy associated with concentration fluctuations. Thus, it is clear that the chemical fluctuations leading to the observed ASRO arise from the curvature of the alloy free energy, just as phonons or positional fluctuations arise from the curvature of a free energy (the dynamical matrix).

It should be realized that the above comments could just as well have been made for magnetization, e.g., using the mapping $(2\xi - 1) \rightarrow \sigma$ for the spin variables. Instead of chemical fields, there are magnetic fields, so that $q(\mathbf{k})$ becomes the magnetic susceptibility, $\chi(\mathbf{k})$. For a disordered alloy with magnetic fluctuations present, one will also have a cross-term that represents the magnetochemical correlations, which determine how the magnetization on an atomic site varies with local chemical fluctuations, or vice versa (Staunton et al., 1990; Ling et al., 1995a). This is relevant to magnetism in alloys covered elsewhere in this unit (see Coupling of Magnetic Effects and Chemical Order).

Sum Rules and Mean-Field Errors

By Equations 2 and 3, $\alpha_{\mu\nu,ii}$ should always be 1; that is, due to the (discrete) translational invariance of the disordered state, the Fourier transform is well defined and

$$\alpha_{\mu\nu,ii} = \alpha_{\mu\nu}(\mathbf{R} = 0) = \int d\mathbf{k}\alpha_{\mu\nu}(\mathbf{k}) = 1 \tag{6}$$

This intensity sum rule is used to check the experimental errors associated with the measured intensities (see SYMMETRY IN CRYSTALLOGRAPHY, KINEMATIC DIFFRACTION OF X RAYS and DYNAMICAL DIFFRACTION). Within most mean-field theories using model Hamiltonians, unless care is taken, Equations 2 and 6 are violated. It is in fact this violation that accounts for the major errors found in mean-field estimates of transition temperatures, because the diagonal (or intrasite) elements of the pair correlations are the largest. Lars Onsager first recognized this in the 1930s for interacting electric dipoles (Onsager, 1936), where he found that a mean-field solution produced the wrong physical sign for the electrostatic energy. Onsager found that by enforcing the equivalents of Equations 4 or 6 (by subtracting an approximate field arising from self-correlations), a more correct physical behavior is found. Hence, we shall refer to the mathematical entities that enforce these sum rules as Onsager corrections (Staunton et al., 1994). In the 1960s, mean-field, magnetic-susceptibility models that implemented this correction were referred to as mean-spherical models (Berlin and Kac, 1952), and the connection to Onsager corrections themselves were referred to as reaction or cavity fields (Brout and Thomas, 1967). Even today this correction is periodically rediscovered and implemented in a variety of problems. As this has profound effects on results, we shall return to how to implement the sum rules within mean-field approaches later, in particular, within our first-principles technique, which incorporates the corrections self-consistently.

Concentration Waves in Multicomponent Alloys

While the concept of concentration waves in binary alloys has a long history, only recently have efforts returned to

the multicomponent alloy case. We briefly introduce the simple ideas of ordering waves, but take this as an opportunity to explain how to interpret ASRO in a multicomponent alloy system where the wavevector alone is not enough to specify the ordering tendency (de Fontaine, 1973; Althoff et al., 1996).

As indicated in the introduction, any arrangement of atoms on a Bravais lattice may be thought of as a modulation of the disordered state by a thermodynamically stable concentration wave. That is, one may Fourier decompose the ordering wave for each site and species on the lattice:

$$c_{\alpha i} = c_{\alpha}^{0} + \sum_{j} [Q_{\alpha}(\mathbf{k}_j) e^{i\mathbf{k}_j \cdot \mathbf{R}_i} + c.c] \tag{7}$$

A binary $A_c B_{(1-c)}$ alloy has a special symmetry: on each site, if the atom is not an A type atom, then it is definitely a B type atom. One consequence of this A-B symmetry is that there is only *one* independent local composition, $\{c_i\}$ (for all sites i), and this greatly simplifies the calculation and the interpretation of the theoretical and experimental results. Because of this, the structure (or concentration wave) indicated by the ASRO is determined *only* by the ordering wavevector (Khachaturyan, 1972, 1983; de Fontaine, 1975, 1979); in this sense, the binary alloys are a special case. For example, for CuAu, the low-temperature state is a layered $L1_0$ state with alternating layers of Cu and Au. Clearly, with $c_{\mathrm{Cu}} = 1/2$, and $c_{\mathrm{Au}} = 1 - c_{\mathrm{Cu}}$, the "concentration wave" is fully described by

$$c_{\mathrm{Cu},i}(\mathbf{R}_i) = \frac{1}{2} + \frac{1}{2}\eta(T) e^{i(2\pi/a)(001)\cdot\mathbf{R}_i} \tag{8}$$

where a single wavevector, $\mathbf{k}$=(001), in units of $2\pi/a$, where a is the lattice parameter, indicates the type of modulation. Here, $\eta(T)$ is the long-range order parameter. So, knowing the composition of the alloy and the energetically favorable ordering wavevector, you fully define the type of ordering. Both bits of information are known from the experiment: the ASRO of CuAu indicates (Moss, 1969) the star of $\mathbf{k} = (001)$ is the most energetically favorable fluctuation. The amplitude of the concentration wave is related to the energy gain due to ordering, as can be seen from a simple chemical, pairwise-interaction model with interactions $V(R_i - R_j)$. The energy difference between the disordered and short-range ordered state is $\frac{1}{2}\sum_{\mathbf{k}} |Q(\mathbf{k})|^2 V(\mathbf{k})$ for infinitesimal ordering fluctuations.

Multicomponent alloys (like an A-B-C alloy) do not possess the binary A-B symmetry and the ordering analysis is therefore more complicated. Because the concentration waves have additional degrees of freedom, more information is needed from experiment or theory. For a bcc ABC_2 alloy, for example, the particular ordering also requires the relative polarizations in the Gibbs "composition space," which are the concentrations of "A relative to C" and "B relative to C" on each sublattice being formed. The polarizations are similar to "branches" for the case of phonons in alloys (Badalayan et al., 1969; de Fontaine, 1973; Althoff et al., 1996). The polarizations of the ordering wave thus determine the sublattice occupations in partially ordered states (Althoff et al., 1996).

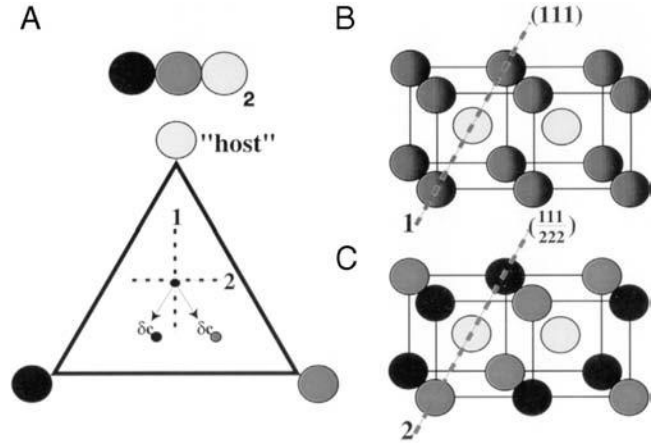


Figure 1. (**A**) the Gibbs triangle with an example of two possible polarization paths that ultimately lead to a Heusler or $L2_1$ type ordering at fixed composition in a bcc ABC_2 alloy. Note that unit vectors that describe the change in the A (black) and B (dark gray) atomic concentrations are marked. First, a $B2$-type order is formed from a $\mathbf{k}_0 = (111)$ ordering wave; see (**B**); given polarization 1 (upper dashed line), A and B atoms randomly populate the cube corners, with C (light gray) atoms solely on the body centers. Next, the polarization 2 (lower dashed line) must occur, which separates the A and B onto separate cube corners, creating the Huesler structure (via a $\mathbf{k}_1 = \mathbf{k}_0/2$ symmetry allowed periodicity); see (**C**). Of course, other polarizations for $B2$ state are possible as determined from the ASRO.

An example of two possible polarizations is given in Figure 1, part A, for the case of $B2$-type ordering in a ABC_2 bcc alloy. At high temperatures, $\mathbf{k} = (111)$ is the unstable wave vector and produces a $B2$ partially ordered state. However, the amount of A and B on the two sublattices is dictated by the polarization: with polarization 1, for example, Figure 1, part B, is appropriate. At a lower temperature, $\mathbf{k} = (\frac{1}{2}\frac{1}{2}\frac{1}{2})$ ordering is symmetry allowed, and then the alloy forms, in this example, a Heusler-type $L2_1$ alloy because only polarization 2 is possible (see Figure 1, part C; in a binary alloy, the Heusler would be the DO_3 or Fe_3Al prototype because there are two distinct "Fe" sites). However, for $B2$ ordering, keep in mind that there are an infinite number of polarizations (types of partial order) that can occur, which must be determined from the electronic interactions on a system-by-system basis.

In general, then, the concentration wave relevant for specifying the type of ordering tendencies in a ternary alloy, as given by the wavevectors in the ASRO, can be written as (Althoff et al., 1996)

$$\begin{bmatrix} c_A(\mathbf{R}_i) \\ c_B(\mathbf{R}_i) \end{bmatrix} = \begin{bmatrix} c_A \\ c_B \end{bmatrix} + \sum_{s,\sigma} \eta_s^{\sigma}(T) \begin{bmatrix} e_A^{\sigma}(\mathbf{k}_s) \\ e_B^{\sigma}(\mathbf{k}_s) \end{bmatrix} \left\{ \sum_{j_s} \gamma^{\sigma}(\mathbf{k}_{j_s};\, \{e_{\alpha}^{\sigma}\}) e^{i\mathbf{k}j_s \cdot \mathbf{R}_i} \right\} \tag{9}$$

The generalization to N-component alloys follows easily in this vector notation. The same results can be obtained by mapping the problem of "molecules on a lattice" investigated by Badalayan et al. (1969). Here, the amplitude of the ordering wave has been broken up into a product of a temperature-dependent factor and two others:

$Q_\alpha(\mathbf{k}_{j_s}) = \eta_s(T)e_\alpha(\mathbf{k}_s)\gamma(\mathbf{k}_{j_s})$, in a spirit similar to that done for the binary alloys (Khachaturyan, 1972, 1983). Here, the η are the temperature-dependent, long-range-order parameters that are normalized to 1 at zero temperature in the fully ordered state (if it exists); the e_α are the eigenvectors specifying the relative polarizations of the species in the proper thermodynamic Gibbs space (see below; also see de Fontaine, 1973; Althoff et al., 1996); the values of γ are geometric coefficients that are linear combinations of the eigenvectors at finite temperature, hence the k dependence, but must be simple ratios of numbers at zero temperatures in a fully ordered state (like the $\frac{1}{2}$ in the former case of CuAu). In regard to the summation labels: s refers to the contributing stars [e.g., (100) or $(1\frac{1}{2}0)$; σ refers to branches, or, the number of chemical degrees of freedom (2 for a ternary); j_s refers the number of wavevectors contained in the star [for fcc, (100) has 3 in the star] (Khachaturyan, 1972, 1983; Althoff et al., 1996).

Notice that only $\eta_s^\sigma(T)$ are quantities not determined by the ASRO, for they depend on thermodynamic averages in a partially or fully ordered phase with those specific probability distributions. For $B2$-type (two sublattices, I and II) ordering in an ABC_2 bcc alloy, there are two order parameters, which in the partially ordered state can be, e.g., $\eta_1 = c_A(\mathrm{I}) - c_A(\mathrm{II})$ and $\eta_2 = c_B(\mathrm{I}) - c_B(\mathrm{II})$. Scattering measurements in the partially ordered state can determine these by relative weights under the superlattice spots that form, or they can be obtained by performing thermodynamic calculations with Monte Carlo or CVM. On a stoichiometric composition, the values of γ are simple geometric numbers, although, from the notation, it is clear they can be different for each member of a star, hence, the different ordering of Cu-Au at $L1_2$ and $L1_0$ stoichiometries (Khachaturyan, 1983). Thus, the eigenvectors $e_\alpha(\mathbf{k}_s)$ at the unstable wavevector(s) give the ordering of A (or B) relative to C. These eigenvectors are completely determined by the *electronic interactions*. What are these eigenvectors and how does one get them from any calculation or measurement? This is a bit tricky.

First, let us note what the $e_\alpha(\mathbf{k}_s)$ are not, so as to avoid confusion between high-T and low-T approaches which use concentration-wave ideas. In the high-temperature state, each component of the eigenvector is degenerate among a given star. By the symmetry of the disordered state, this must be the case and it may be removed from the "j_s" sum (as done in Equation 9). However, below a first-order transition, it is possible that the $e_\alpha(\mathbf{k}_s)$ is temperature and star dependent, for instance, but this cannot be ascertained from the ASRO. Thus, from the point of view of determining the ordering tendency from the ASRO, the $e_\alpha(\mathbf{k}_s)$ do not vary among the members of the star, and their temperature dependence is held fixed after it is determined just above the transition. This does not mean, as assumed in pair-potential models, that the interactions (and, therefore, polarizations) are given *a priori* and do not change as a function of temperature; it only means that averages in the disordered state cannot necessarily give you averages in the partially ordered state. Thus, in general, the $e_\alpha(\mathbf{k})$ may also have a dependence on members of the star, because $e_\alpha(\mathbf{k}_{j_s})\gamma(\mathbf{k}_{j_s})$ has to reflect the symmetry operations of the ordered distribution when writing a concentration wave. We do not address this possibility here.

Now, what is $e_\alpha(\mathbf{k})$ and how do you get it? In Figure 1, part A; the unit vectors for the fluctuations of A and B compositions are shown within the ternary Gibbs triangle: only within this triangle are the values of c_A, c_B, and c_C allowed (because $c_A + c_B + c_C = 1$). Notice that the unit vectors for δc_A and δc_B fluctuations are at an oblique angle, because the Gibbs triangle is an oblique coordinate system. The free energy associated with concentration fluctuations is $\Delta F = \delta\mathbf{c}^T\mathbf{q}^{-1}\delta\mathbf{c}$, using matrix notation with species labels suppressed (note that superscript T is a transpose operation). The matrix $\mathbf{q}^{-1}(\mathbf{k})$ is symmetric and square in $(N-1)$ species (let us take species C as the "host"). As such, it seems "obvious" that the eigenvectors of $\mathbf{q}^{-1}$ are required because they reflect the "principal directions" in free energy space which reveal the true order. However, its eigenvectors, $\mathbf{e}_C$, produce a host-dependent, unphysical ordering! That is, Equation 9 would produce negative concentrations in some cases.

Immediately, you see the problem. The Gibbs triangle is an oblique coordinate system and, therefore, the eigenvectors must be obtained in a properly orthogonal Cartesian coordinate system (de Fontaine, 1973). By an oblique coordinate transform, defined by $\delta\mathbf{c} = \mathbf{T}\mathbf{x}$, $\Delta F_x = \mathbf{x}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{q}^{-1}\mathbf{T})\mathbf{x}$, but still $\Delta F_x = \Delta F$. From $\mathbf{T}^{\mathrm{T}}\mathbf{q}^{-1}\mathbf{T}$, we find a set of host-independent eigenvectors, $\mathbf{e}_X$; in other words, regardless of which species you take as the host, you always get the same eigenvectors! Finally, the physical eigenvectors we seek in the Gibbs space are then $\mathbf{e}_G = \mathbf{T}\mathbf{e}_X$ (since $\delta\mathbf{c} = \mathbf{T}\mathbf{x}$). It is important to note that $\mathbf{e}_C$ is not the same as $\mathbf{e}_G$ because $\mathbf{T}^{\mathrm{T}} \neq \mathbf{T}^{-1}$ in an oblique coordinate system like the Gibbs triangle, and, therefore, $\mathrm{T}\mathbf{T}^{\mathrm{T}}$ is not 1.

It is the $\mathbf{e}_G$ that reveal the true principal directions in free-energy space, and these parameters are related to linear combinations of elements of $\mathbf{q}^{-1}(\mathbf{k} = \mathbf{k}_0)$ at the pertinent unstable wavevector(s). If nothing else, the reader should take away that these quantities can be determined theoretically or experimentally via the diffuse intensities. Of course, any error in the theory or experiment, such as not maintaining the sum rules on $\mathbf{q}$ or α, will create a subsequent error in the eigenvectors and hence the polarization. Nevertheless, it is possible to obtain from the ASRO both wavevector and "wave polarization" information which determines the ordering tendencies (also see the appendix in Althoff et al., 1996).

To make this a little more concrete, let us reexamine the previous bcc ABC_2 alloy. In the bcc alloys, the first transformation from disordered $A2$ to the partially order $B2$ phase is second order, with $\mathbf{k} = (111)$ and no other wavevectors in the star. The modulation (111) indicates that the bcc lattice is being separated into two distinct sublattices. If the polarization 1 in Figure 1, part A, was found, it indicates that species C is going to be separated on its own sublattice; whereas, if polarization 2 was found initially, species C would be equally placed on the two sublattices. Thus, the polarization already gives a great deal of information about the ordering in the $B2$ partially ordered phase and, in fact, is just the slope of the line in the Gibbs triangle. This is the basis for the recent graphical representation of ALCHEMI (atom location by channeling

electron microscopy) results in $B2$-ordering ternary intermetallic compounds (Hou et al., 1997). There are, in principle, two order parameters because of the two branches in a ternary alloy case. The order-parameter η_2, say, can be set to zero to obtain the $B2$-type ordering, and, because the eigenvalue, λ_2, of eigenmode e_2 is higher in energy than that of e_1, i.e., $\lambda_1 < \lambda_2$, only e_1 is the initially unstable mode. See Johnson et al. (1999) for calculations in Ti-Al-Nb bcc-based alloys, which are directly compared to experiment (Hou, 1997).

We close this description of interpreting ASRO in ternary alloys by mentioning that the above analysis generalizes completely for quaternaries and more complex alloys. The important chemical space progresses: binary is a line (no angles needed), ternary is a triangle (one angle), quaternaries are pyramids (two angles, as with Euler rotations), and so on. So the oblique transforms become increasingly complex for multidimensional spaces, but the additional information, along with the unstable wavevector, is contained within the ASRO.

Concentration Waves from a Density-Functional Approach

The present first-principles theory leads naturally to a description of ordering instabilities in the homogeneously random state in terms of static concentration waves. As discussed by Khachaturyan (1983), the concentration-wave approach has several advantages, which are even more relevant when used in conjunction with an electronic-structure approach (Staunton et al., 1994; Althoff et al., 1995, 1996). Namely, the method (1) allows for interatomic interaction at arbitrary distances, (2) accounts for correlation effects in a long-range interaction model, (3) establishes a connection with the Landau-Lifshitz thermodynamic theory of second-order phase transformations, and (4) does not require *a priori* assumptions about the atomic superstructure of the ordered phases involved in the order-disorder transformations, allowing the possible ordered-phase structures to be predicted from the underlying correlations.

As a consequence of the electronic-structure basis to be discussed later, realistic contributions to the effective chemical interactions in metals arise, e.g., from electrostatic interactions, Fermi-surface effects, and strain fields, all of which are inherently long range. Analysis within the method is performed entirely in reciprocal space, allowing for a description of atomic clustering and ordering, or of strain-induced ordering, none of which can be included within many conventional ordering theories. In the present work, we neglect all elastic effects, which are the subject of ongoing work. As with the experiment, the electronic theory leads naturally to a description of the ASRO in terms of the temperature-dependent, two-body compositional-correlation function in reciprocal space. As the temperature is lowered, (usually) one wavevector becomes prominent in the ASRO, and the correlation function ultimately diverges there.

It is probably best to derive some standard relations that are applicable to both simple models and DFT-based approaches. The idea is simply to show that certain simplifications lead to well-known and venerable results, such as the Krivoglaz-Clapp-Moss formula (Krivoglaz, 1969; Clapp and Moss, 1966), where, by making fewer simplifications, an electronic DFT-based theory can be formulated, which nevertheless, is a mean-field theory of configurational degrees of freedom. While it is certainly much easier to derive approximations for pair correlations using very standard mean-field treatments based on effective interactions, as has been done traditionally, an electronic-DFT-based approach would require much more development along those lines. Consequently, we shall proceed in a much less common way, which can deal with all possibilities. In particular, we shall give a derivation for binary alloys and state the result for the N-component generalization, with some clarifying remarks added.

As shown for an A-B binary system (Györffy and Stocks, 1983), it is straightforward to adapt the density-functional ideas of classical liquids (Evans, 1979) to a "lattice-gas" model of alloy configurations (de Fontaine, 1979). The fundamental DFT theorem states that, in the presence of an external field, $H_{\text{ext}} = \Sigma_n \varpi_n \xi_n$ with external chemical potential, ϖ_n, there is a grand potential (not yet the thermodynamic one),

$$\mathbf{\Omega}[T, V, N, \nu, \{c_n\}] = F[T, V, N, \{c_n\}] - \Sigma_n(\varpi_n - \nu)c_n \quad (10)$$

such that the internal Helmholtz free energy $F[\{c_n\}]$ is a unique functional of the local concentrations c_n, that is, $\langle \xi_n \rangle$, meaning F is independent of ϖ_n. Here, T, V, N, and ν are, respectively, the temperature, volume, number of unit cells, and chemical potential difference $(\nu_A - \nu_B)$. The equilibrium configuration is specified by the stationarity condition;

$$\left.\frac{\partial \Omega}{\partial c_n}\right|_{\{c_n^0\}} = 0 \quad (11)$$

which determines the Euler-Lagrange equations for the alloy problem. Most importantly, from arguments given by Evans (Evans, 1979), it can be proven that Ω is a minimum at $\{c_n^0\}$ and equal to the proper thermodynamic grand potential $\Omega[T, V, N, \nu]$ (Györffy et al., 1989). In terms of $\nu_i = (\varpi_i - \nu)$, an effective chemical potential difference, Ω is a generating function for a hierarchy of correlation functions (Evans, 1979). The first two are

$$-\frac{\partial \Omega}{\partial \nu_i} = c_i \qquad \text{and} \qquad -\frac{\partial^2 \Omega}{\partial \nu_i \partial \nu_j} = \beta q_{ij} \quad (12)$$

This second generator is the correlation function that we require for stability analysis.

Some standard DFT tricks are useful at this point and these also happen to be the equivalent tricks originally used to derive the electronic-DFT Kohn-Sham equations (also known as the single-particle Schrödinger equations; Kohn and Sham, 1965). Although F is not yet known, we can break this complex functional up into a known non-interacting part (given by point entropy for the alloy problem):

$$F_0 = \beta^{-1} \sum_n [c_n \ln c_n + (1 - c_n) \ln(1 - c_n)] \quad (13)$$

and an interacting part Φ, defined by $F = F_0 - \Phi$. Here, β^{-1} is the temperature, $k_B T$ where k_B is the Boltzmann constant. In the DFT for electrons, the noninteracting part was taken as the single-particle kinetic energy (Kohn and Sham, 1965), which is again known exactly. It then follows from Equation 11 that the Euler-Lagrange equations that the $\{c_n^0\}$ satisfy are:

$$\beta^{-1}\ln\left[\frac{c_n^0}{(1-c_n^0)}\right] - S_n^{(1)} - \nu_n = 0 \tag{14}$$

which determines the contribution to the local chemical potential differences in the alloy due to all the interactions, if $S^{(1)}$ can be calculated (in the physics literature, $S^{(1)}$ would be considered a self-energy functional). Here it has been helpful to define a new set of correlation functions generated from the functional derivatives of Φ with respect to concentration variable; the first two correlation functions being:

$$S_i^{(1)} \equiv -\frac{\partial\Phi}{\partial c_i} \qquad \text{and} \qquad S_{ij}^{2} \equiv -\frac{\partial^2\Phi}{\partial c_i \partial c_j} \tag{15}$$

In the classical theory of liquids, the $S^{(2)}$ is an Ornstein-Zernike (Ornstein, 1912; Ornstein and Zernike, 1914 and 1918; Zernike, 1940) direct-correlation function (with density instead of concentration fluctuations; Stell, 1969).

Note that there are as many coupled equations in Equation 14 as there are atomic sites. If we are interested in, for example, the concentration profile around an antiphase boundary, Equation 14 would in principle provide that information, depending upon the complexity of Φ and whether we can calculate its functional derivatives, which we shall address momentarily. Also, recognize that $S^{(2)}$ is, by its very nature, the stability matrix (with respect to concentration fluctuations) of the interacting part of the free energy. Keep in mind that Φ, in principle, must contain all many-body-type interactions, including all entropy contributions beyond the point entropy that was used as the noninteracting portion. If it was based on a fully electronic description, it must also contain "particle-hole" entropy associated with the electronic density of states at finite temperature (Staunton et al., 1994).

The significance of $S^{(2)}$ can immediately be found by performing a stability analysis of the Euler-Lagrange equations; that is, take the derivatives of Equation 14 w.r.t. c_i, or, equivalently, expand the equation to find the fluctuations about c_i^0 (i.e., $c_i = c_i^0 + \delta c_i$), to find out how fluctuations affect the local chemical potential difference. The result is the stability equations for a general inhomogeneous alloy system:

$$\frac{\delta_{ij}}{\beta c_i(1-c_i)} - S_{ij}^{(2)} - \frac{\partial\nu_i}{\partial c_j} = 0 \tag{16}$$

Through the DFT theorem and generating functionals, the response function $\partial\nu_i/\partial c_j$, which tells how the concentrations vary with changes in the applied field, has a simple relationship to the true pair-correlation function:

$$\frac{\partial\nu_i}{\partial c_j} = \left(\frac{\partial c_j}{\partial\nu_i}\right)^{-1} \equiv \left(\frac{\delta^2\Omega}{\delta\nu_i\delta\nu_j}\right)^{-1} \equiv \beta^{-1}(q^{-1})_{ij} \equiv [\beta c(1-c)]^{-1}(\alpha^{-1})_{ij} \tag{17}$$

where the last equality arises through the definition of Warren-Cowley parameters. If Equation 16 is now evaluated in the random state where (discrete) translational invariance holds, and the connection between the two types of correlation functions is used (i.e., Equation 17), we find:

$$\alpha(\mathbf{k}) = \frac{1}{1 - \beta c(1-c)S^{(2)}(\mathbf{k})} \tag{18}$$

Note that here we have evaluated the exact functional in the homogeneously random state with $c_i^0 = c\,\forall\, i$, which is an approximation because in reality there are some changes to function induced by the developed ASRO. In principle, we should incorporate this ASRO in the evaluation to more properly describe the situation. For peaks at finite wavevector $\mathbf{k}_0$, it is easy to see that absolute instability of the binary alloy to ordering occurs when $\beta c(1-c)S^{(2)}(\mathbf{k}=\mathbf{k}_0) = 1$ and the correlations diverge. The alloy would be unstable to ordering with that particular wavevector. The temperature, T_{sp}, where this may occur, is the so-called "spinodal temperature." For peaks at $\mathbf{k}_0 = 0$, i.e., long-wavelength fluctuations, the alloy would be unstable to clustering.

For the N-component case, a similar derivation is applicable (Althoff et al., 1996; Johnson, 2001) with multiple chemical fields, $\varpi_{\alpha n}$, chemical potential differences, ν_α (relative to the ν_N), and effective chemical potential differences $\nu_{\alpha n} = (\varpi_{\alpha n} - \nu_\alpha)$. One cannot use the simple c and $(1-c)$ relationship in general and must keep all the labels relative to the Nth component. Most importantly, when taking compositional derivatives, the single-occupancy constraint must be handled properly, i.e., $\partial c_{\alpha i}/\partial c_{\beta j} = \delta_{ij}[(\delta_{\alpha\beta} - \delta_{\alpha N})(1-\delta_{\beta N})]$. The generalized equations for the pair correlations are, when evaluated in the homogeneously random state:

$$\left[q^{-1}(\mathbf{k})\right]_{\alpha\beta} = \left[\frac{\delta_{\alpha\beta}}{c_\alpha} + \frac{1}{c_N}\right] - \beta S_{\alpha\beta}^{(2)}(\mathbf{k}) \tag{19}$$

where neither α nor β can be the Nth component. This may again be normalized to produce the Warren-Cowley pairs. With the constraint implemented by designating the Nth species as the "host," the (nonsingular) portion of the correlation function matrices are rank $(N-1)$. For an A-B binary, $c_\alpha = c_A = c$ and $c_N = c_B = 1 - c$ because only the $\alpha = \beta = A$ term is valid ($N = 2$ and matrices are rank 1), and we recover the familiar result, Equation 18.

Equation 19 is, in fact, a most remarkable result. It is completely general and exact! However, it is based on some still unknown functional $S^{(2)}(\mathbf{k})$, which is not a pairwise interaction but a pair-correlation function arising

from the interacting part of the free energy. Also, Equations 18 and 19 properly conserve spectral intensity, $\alpha_{\alpha\beta}(\mathbf{R}=0)=1$, as required in Equation 6. Notice that $S^{(2)}(\mathbf{k})$ has been defined without referring to pair potentials or any larger sets of ECIs. In fact, we shall discuss how to take advantage of this to make a connection to first-principles electronic-DFT calculations of $S^{(2)}(\mathbf{k})$. First, however, let us discuss some familiar mean-field results in the theory of pair correlations in binary alloys by picking approximate $S^{(2)}(\mathbf{k})$ functionals. In such a context, Equation 18 may be cautiously thought of as a generalization of the Krivoglaz-Clapp-Moss formula, where $S^{(2)}$ plays the role of a concentration- and (weakly) temperature-dependent effective pairwise interaction.

Connection to Well-Known Mean-Field Results

In the concentration-functional approach, one mean-field theory is to take the interaction part of the free energy as the configurational average of the alloy Hamiltonian, i.e., $\Phi_{\mathrm{MF}}=-\langle H[\{\xi_n\}]\rangle$, where the averaging is performed with an inhomogeneous product probability distribution function, $P[\{\xi_n\}]=\prod_n P_n(\xi_n)$, with $P_n(1)=c_n$ and $P_n(0)=1-c_n$. Such a product distribution yields the mean-field results $\langle\xi_i\xi_j\rangle=c_ic_j$, e.g., usually called the random-phase approximation in the physics community. For an effective chemical interaction model based on pair potentials and using $\langle\xi_i\xi_j\rangle=c_ic_j$, then

$$\Phi\rightarrow\Phi_{\mathrm{MF}}=\frac{1}{2}\sum_{ij}c_iV_{ij}c_j \tag{20}$$

and therefore, $S^{(2)}(\mathbf{k})=-V(\mathbf{k})$, which no longer has a direct electronic connection for the pairwise correlations. As a result, we recover the Krivoglaz-Clapp-Moss result (Krivoglaz, 1969; Clapp and Moss, 1966), namely:

$$\alpha(\mathbf{k})=\frac{1}{[1+\beta c(1-c)V(\mathbf{k})]} \tag{21}$$

and the Gorsky-Bragg-Williams equation of state would be reproduced by Equation 14. To go beyond such a mean-field result, fluctuation corrections would have to be added to Φ_{MF}. That is, the probability distribution would have to be more than a separable product. One consequence of the uncorrelated configurational averaging (i.e., $\langle\xi_i\xi_j\rangle=c_ic_j$) is a substantial violation of the spectral intensity sum rule $\alpha(\mathbf{R}=0)=1$. This was recognized early on and various scenarios for normalizing the spectral intensity have been used (Clapp and Moss, 1966; Vaks et al., 1966; Reinhard and Moss, 1993).

A related effect of such a mean-field averaging is that the system is "overcorrelated" through the mean fields. This occurs because the effective chemical fields are produced by averaging over all sites. As such, the local composition on the ith site interacts with all the remaining sites through that average field, which already contains effects from the ith site; so the ith site has a large self correlation. The "mean" field produces a correlation because it contains field information from all sites, which is the reason that although assuming $\langle\xi_i\xi_j\rangle=c_ic_j$, which says that no pairs are correlated, we managed to obtain atomic short-range order, or a pair correlation. So, the mean-field result properly has a correlation, although it is too large a self-correlation, and there is a slight lack of consistency due to the use of "mean" fields. In previous comparisons of Ising models (e.g., to various mean-field results), this excessive self-correlation gave rise to the often quoted 20% error in transition temperatures (Brout and Thomas, 1965).

Improvements to Mean-Field Theories

While this could be a chapter unto itself, we will just mention a few key points. First, just because one uses a mean-field theory does not necessarily make the results bad. That is, there are many different breeds of mean-field approximations. For example, the CVM is a mean-field approximation for cluster entropy, being much better than the Gorsky-Bragg-Williams approximation, which uses only point entropy. In fact, the CVM is remarkably robust, giving in many cases results similar to "exact" Monte Carlo simulations (Sanchez and de Fontaine, 1978, 1980; Ducastelle, 1991). However, it too does have limitations, such as a practical restriction to small interaction ranges or multiplet sizes.

Second, when addressing an alloy problem, the complexity of Φ (the underlying Hamiltonian) matters, not only how it is averaged. The overcorrelation in the mean-field approximation, e.g., while often giving a large error in transition temperatures in simple alloy models, is not a general principle. If the correct physics giving rise to the ordering phenomena in a particular alloy is well described by the Hamiltonian, very good temperatures can result. If entropy was the entire driving force for the ordering, and Φ did not have any entropy included, we would get quite poor results. On the other hand, if electronic band filling was the overwhelming contribution to the structural transformation, then a Φ that included that information in a reasonable way, but that threw out higher-order entropy, would give very good results; much better, in fact, than the often quoted "20% too high in transition temperature." We shall indeed encounter this in the results below.

Third, even simple improvements to mean-field methods can be very useful, as we have already intimated when discussing the Onsager cavity-field corrections. Let us see what the effect is of just ensuring the sum rule required in Equation 6. The Onsager corrections (Brout and Thomas, 1967) for the above mean-field average amounts to the following coupled equations in the alloy problem (Staunton et al., 1994; Tokar, 1997; Boriçi-Kuqo et al., 1997), depending on the mean-field Φ used

$$\alpha(\mathbf{k},T)=\frac{1}{1-\beta c(1-c)[S^{(2)}_{\mathrm{MF}}(\mathbf{k},T)-\Lambda(T)]} \tag{22}$$

and, using Equation 6,

$$\nabla(T)=\frac{1}{\Omega_{\mathrm{BZ}}}\int d\mathbf{k}S^{(2)}_{\mathrm{MF}}(\mathbf{k})\alpha(\mathbf{k},T;\Lambda) \tag{23}$$

where Λ is the temperature-dependent Onsager correction and Ω_{BZ} is the Brillouin zone volume of the random alloy Bravais lattice. This coupled set of equations may be solved by standard Newton-Raphson techniques. For the N-component alloy case, these become a coupled set of matrix equations where all the matrices (including Λ) have two subscripts identifying the pairs, as given in Equation 19 and appropriate internal sums over species are made (Althoff et al., 1995, 1996). An even more improved and general approach has been proposed by Chepulskii and Bugaev (1998a,b).

The effect of the Onsager correction is to renormalize the mean-field $S^{(2)}(\mathbf{k})$ producing an effective $S^{(2)}_{\text{eff}}(\mathbf{k})$ which properly conserves the intensity. For an exact $S^{(2)}(\mathbf{k})$, Λ is zero, by definition. So, the closer an approximate $S^{(2)}(\mathbf{k})$ satisfies the sum rule, the less important are the Onsager corrections. At high temperatures, where $\alpha(\mathbf{k})$ is 1, it is clear from Equation 23 that $\Lambda(T)$ becomes the average of $S^{(2)}(\mathbf{k})$ over the Brillouin zone, which turns out to be a good "seed" value for a Newton-Raphson solution. It is important to emphasize that Equations 22 and 23 may be derived in numerous ways. However, for the current discussion, we note that Staunton et al. (1994) derived these relations from Equation 14 by adding the Onsager cavity-field corrections while doing a linear-response analysis that can add additional complexity, (i.e., more q dependence than evidenced by Equation 22 and Equation 23). Such an approach can also yield the equivalent to a high-T expansion to second order in β—as used to explain the temperature-dependent shifts in ASRO (Le Bulloc'h et al., 1998). Now, for an Onsager-corrected mean-field theory, as one gets closer to the spinodal temperature, $\Lambda(T \cong T_{\text{sp}})$ becomes larger and larger because $\alpha(\mathbf{k})$ is diverging and more error has to be corrected. Improved entropy mean-field approaches, such as the CVM, still suffer from errors associated with the intensity sum rule, which are mostly manifest around the transition temperatures (Mohri et al., 1985).

For a pairwise Hamiltonian, it is assumed that $V_{ii} = 0$, otherwise it would just be an arbitrary shift of the energy zero, which does not matter. However, an interesting effect in the high-T limit for the (mean-field) pair-potential model is that $\Lambda = \frac{1}{\Omega_{BZ}} \int d\mathbf{k} V(\mathbf{k}) = V_{ii}$, which is not generally zero, because V_{ii} is not an interaction but a self-energy correction, (i.e., Φ_{ii}, which must be finite in mean-field theory just to have a properly normalized correlation function). As evidenced from Equation 16, the correlation function can be written as $\underline{\underline{\alpha}}^{-1} = \underline{\underline{\Sigma}} - \underline{\underline{V}}$ in terms of a self-energy Σ, as can be shown more properly from field-theory (Tokar, 1985, 1997). However, Σ in this case is the exact self-energy, rather than just $[\beta c(1-c)]^{-1}$ for the Krivoglaz-Clapp-Moss mean-field case. Moreover, the zeroth-order result for the self-energy yields the Onsager correction (Masanskii et al., 1991; Tokar, 1985, 1997), i.e., $\Sigma = [\beta c(1-c)]^{-1} + \Lambda(T)$. Therefore, V_{ii}, or more properly $\Lambda(T)$, is manifestly not arbitrary. These techniques have little to say regarding complicated many-body Hamiltonians, however.

It would be remiss not to note that for short-range order in strongly correlated situations the mean-field results, even using Onsager corrections, can be topologically incorrect. An example is the interesting toy model of a disordered (electrostatically screened) binary Madelung lattice (Wolverton and Zunger, 1995b; Boriçi-Kuqo et al., 1997), in which there are two types of point charges screened by rules depending on the nearest-neighbor occupations. In such a pairwise case, including intrasite self-correlations, the intensity sums are properly maintained. However, self-correlations beyond the intrasite (at least out to second neighbors) are needed in order to correct $\underline{\underline{\alpha}}^{-1} = \underline{\underline{\Sigma}} - \underline{\underline{V}}$ and its topology (Wolverton and Zunger, 1995a,b; Boriçi-Kuqo et al., 1997).

In less problematic cases, such as backing out ECIs from experimental data on real alloys, it is found that the zeroth-order (Onsager) correction plus additional first-order corrections agrees very well with those ECIs obtained using inverse Monte Carlo methods (Reinhard and Moss, 1993; Masanskii et al., 1991). When a second-order correction was included, no difference was found between the ECIs from mean-field theory and inverse Monte Carlo, suggesting that lengthy simulations involved with the inverse Monte Carlo techniques may be avoided (Le Bolloc'h et al., 1997). However, as warned before, the inverse mapping is not unique, so care must be taken when using such information.

Notice that even in problem cases, improvements made to mean-field theories properly reflect most of the important physics, and can usually be handled more easily than more exacting approaches. What is important is that a mean-field treatment is not in itself patently inappropriate or wrong. It is, however, important to have included the correct physics for a given system. Including the correct physics for a given alloy is a system-specific requirement, which usually cannot be known *a priori*. Hence, our choice is to try and handle chemical and electronic effects, all on an equal footing, and represented from a highly accurate, density-functional basis.

Concentration Waves from First-Principles, Electronic-Structure Calculations

What remains to be done is to connect the formal derivation given above to the system-dependent, electronic structure of the random substitutional alloy. In other words, we must choose a Φ, which we shall do in a mean-field approach based on local density approximation (LDA) to electronic DFT (Kohn and Sham, 1965). In the adiabatic approximation, $\Phi_{\text{MF}} = -\langle \Omega_e \rangle$, where Ω_e is the electronic grand potential of the electrons for a specific configuration (where we have also lumped in the ion-ion contribution). To complete the formulation, a mean-field configurational averaging of Ω_e is required in analytic form, and must be dependent on all sites in order to evaluate the functional derivatives analytically. Note that using a local density approximation to electronic DFT is also, in effect, a mean-field theory of the electronic degrees of freedom. So, even though they will be integrated out, the electronic degrees of freedom are all handled on a par with the configurational degrees of freedom contained in the noninteracting contribution to the chemical free energy.

For binaries, Györffy and Stocks (1983) originally discussed the full adaptation of the above ideas and its

implementation including only electronic band-energy contributions based on the Korringa-Kohn-Rostocker (KKR) coherent potential approximation (CPA) electronic-structure calculations. The KKR electronic-structure method (Korringa, 1947; Kohn and Rostoker, 1954) in conjunction with the CPA (Soven, 1967; Taylor, 1968) is now a well-proven, mean-field theory for calculating electronic states and energetics in random alloys (e.g., Johnson et al., 1986, 1990). In particular, the ideas of Ducastelle and Gautier (1976) in the context of tight-binding theory were used to obtain $\langle \Omega_e \rangle$ within an inhomogeneous version of the KKR-CPA, where all sites are distinct so that variational derivatives could be made.

As shown by Johnson et al. (1986, 1990), the electronic grand potential for any alloy configuration may be written as:

$$\Omega_e = -\int_{-\infty}^{\infty} d\varepsilon N(\varepsilon;\mu) f(\varepsilon-\mu) + \int_{-\infty}^{\mu} d\mu' \int_{-\infty}^{\infty} d\varepsilon \frac{dN(\varepsilon;\mu')}{d\mu'} f(\varepsilon-\mu') \qquad (24)$$

where the first term is the single-particle, or band-energy, contribution, which produces the local (per site) electronic density of states, $n_i(\varepsilon;\mu)$, and the second term properly gives the "double-counting" corrections. Here $f(\varepsilon - \mu)$ is the Fermi occupation factor from finite-temperature effects on the electronic chemical potential, μ (or Fermi energy at $T = 0$ K). Hence, the band-energy term contains all electron-hole effects due to electronic entropy, which may be very important in some high-T alloys. The $N(\varepsilon;\mu) = \Sigma_i \int_{-\infty}^{\varepsilon} d\varepsilon' n_i(\varepsilon;\mu)$, and is the integrated density of states as typically discussed in band-structure methods. We may obtain an analytic expression for Ω_e as long as an analytic expression for $N(\varepsilon;\mu)$ exists (Johnson et al., 1986, 1990). Within the CPA, an analytic expression for $N(\varepsilon;\mu)$ in either a homogeneously or inhomogeneously disordered state is given by the generalized Lloyd formula (Faulkner and Stocks, 1980). Hence, we can determine Ω^{CPA} for a inhomogeneously random state. As with any DFT, besides the extrinsic variables T, V, and μ (temperature, volume, and chemical potential, respectively), Ω_e^{CPA} is only a functional of the CPA charge density, $\{\rho_{\alpha,i}\}$ for all species and sites.

In terms of KKR multiple-scattering theory, the inhomogeneous CPA is pictorially understood by "replacing" the individual atomic scatterers at the ith site (i.e., $t_{\alpha,i}$) by a CPA effective scatterer per site (i.e., $t_{c,i}$) (Györffy and Stocks, 1983; Staunton et al., 1994). It is more appropriate to average scattering properties rather than potentials to determine a random system's properties (Soven, 1967; Taylor, 1968). For an array of CPA scatterers, $\tau_{c,ii}$ is a (site-diagonal) KKR scattering-path operator that describes the scattering of an electron from all sites given that it starts and ends at the ith site. The t_c values are determined from the requirement that replacing the effective scatterer by any of the constituent atomic scatterers (i.e., $t_{\alpha,i}$) does not on average change the scattering properties of the entire system as given by $\tau_{c,ii}$ (Fig. 2). This

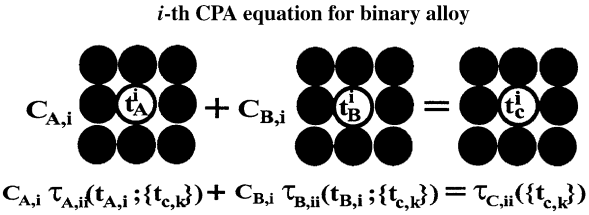


Figure 2. Schematic of the required average scattering condition, which determines the inhomogeneous CPA self-consistent equations. t_c and t_α are the site-dependent, single-site CPA and atomic scattering matrices, respectively, and τ_c is the KKR scattering path operator describing the entire electronic scattering in the system.

requirement is expressed by a set of CPA conditions, $\Sigma_\alpha c_{\alpha,i}\tau_{\alpha,ii} = \tau_{c,ii}$, one for each lattice site. Here, $\tau_{\alpha,ii}$ is the site-diagonal, scattering-path operator for an array of CPA scatterers with a single impurity of type α at the ith site (see Fig. 2) and yields the required set of $\{\rho_{\alpha,i}\}$. Notice that each of the CPA single-site scatterers can in principle be different (Györffy et al., 1989). Hence, the random state is inhomogeneous and scattering properties vary from site to site. As a consequence, we may relate any type of ordering (relative to the homogeneously disordered state) directly to electronic interactions or properties that lower the energy of a particular ordering wave.

While these inhomogeneous equations are generally intractable for solution, the inhomogeneous CPA must be considered to calculate analytically the response to variations of the local concentrations that determine $S_{ij}^{(2)}$. This allows all possible configurations (or different possible site occupations) to be described. By using the homogeneous CPA as the reference, all possible orderings (wave vectors) may be compared simultaneously, just as in with phonons in elemental systems (Pavone et al., 1996; Quong and Lui, 1997). The conventional single-site homogeneous CPA (used for total-energy calculations in random alloys; Johnson et al., 1990) provides a soluble highest symmetry reference state to perform a linear-response description of the inhomogeneous CPA theory.

Those ideas have been recently extended and implemented to address multicomponent alloys (Althoff et al., 1995, 1996), although the initial calculations still just include terms involving the band-energy only (BEO). For binaries, Staunton et al. (1994) have worked out the details and implemented calculations of atomic-short-range order that include all electronic contributions, e.g., electrostatic and exchange correlation. The coupling of magnetic and chemical degrees of freedom have been addressed within this framework by Ling et al. (1995a, b), and references cited therein. The full DFT theory has thus far been applied mostly to binary alloys, with several successes (Staunton et al., 1990; Pinski et al., 1991, 1998; Johnson et al., 1994; Ling et al., 1995; Clark et al., 1995). The extension of the theory to incorporate atomic displacements from the average lattice is also ongoing, as is the inclusion of all terms beyond the band energy for multicomponent systems.

Due to unique properties of the CPA, the variational nature of DFT within the KKR-CPA is preserved, i.e., $\delta\Omega_e^{\mathrm{CPA}}/\delta\rho_{\alpha,i} = 0$ (Johnson et al., 1986, 1990). As a result, only the explicit concentration variations are required to obtain equations for $S^{(1)}$, the change in (local) chemical potentials, i.e.:

$$\frac{\delta\Omega^{\mathrm{CPA}}}{\partial c_{\alpha,i}} = -(1-\delta_{\alpha N})Im\int_{-\infty}^{\infty} d\varepsilon f(\varepsilon-\mu_e) \times [N_{\alpha,i}(\varepsilon) - N_{N,i}(\varepsilon)] + \cdots \quad (25)$$

Here, $f(\varepsilon-\mu)$ is the Fermi filling factor and μ_e is the electronic chemical potential (Fermi energy at $T = 0$ K); $N_\alpha(\varepsilon)$ is the CPA site integrated density of states for the α species, and the Nth species has been designated the "host." The ellipses refer to the remaining direct concentration variations of the Coulomb and exchange-correlation contributions to Ω^{CPA}, and the term shown is the band-energy-only contribution (Staunton et al., 1994). This band-energy term is completely determined for each site by the change in band-energy found by replacing the "host" species by an α species. Clearly, $S^{(1)}$ is zero if the α species is the "host" because this cannot change the chemical potential.

For the second variation for $S^{(2)}$, it is not so nice, because there are implicit changes to the charge densities (related to $\tau_{c,ii}$ and $\tau_{\alpha,ii}$) and the electronic chemical potentials, μ. Furthermore, these variations must be limited by global charge neutrality requirements. These restricted variations, as well as other considerations, lead to dielectric screening effects and charge "rearrangement"–type terms, as well as standard Madelung-type energy changes (Staunton et al., 1994).

At this point, we have (in principle) kept all terms contributing to the electronic grand potential, except for static displacements. However, in many instances, only the terms directly involving the underlying electronic structure predominantly determine the ordering tendencies (Ducastelle, 1991), as it is argued that screening and near-local charge neutrality make the double-counting terms negligible. This is in general not so, however, as discussed recently (Staunton et al., 1994; Johnson et al., 1994). For simplicity's sake, we consider here only the important details within the band-energy-only (BEO) approach, and only state important differences for the more general case when necessary. Nonetheless, it is remarkable that the BEO contributions actually address a great many alloying effects that determine phase stability in many alloy systems, such as, band filling (or electron-per-atom, e/a; Hume-Rothery, 1963), hybridization (arising from diagonal and off-diagonal disorder), and so-called electronic topological transitions (Lifshitz, 1960), which encompass Fermi-surface nesting (Moss, 1969) and van Hove singularity (van Hove, 1953) effects. We shall give some real examples of such effects (how they physically come about) and how these effects determine the ASRO, including in disordered fcc Cu-Ni-Zn (see Data Analysis and Initial Interpretation).

Within a BEO approach, the expression for the band-energy part of $S^{(2)}_{\alpha\beta}(\mathbf{q};\varepsilon)$ is

$$S^{(2)}_{\alpha\beta}(\mathbf{q};\varepsilon) = -\frac{1}{\pi}(1-\delta_{\alpha N})(1-\delta_{\beta N})\,Im\int_{-\infty}^{\infty} d\varepsilon f(\varepsilon-\mu) \times \left\{\sum_{L_1L_2L_3L_4} M_{\alpha,L_1L_2}(\varepsilon) X_{L_2L_1;L_3L_4}(\mathbf{q};\varepsilon) M_{\beta,L_3L_4}(\varepsilon)\right\} \quad (26)$$

where L refers to angular momentum indices of the spherical harmonic basis set (i.e., contributions from s, p, d, etc. type electrons in the alloy), and the matrix elements may be found in the referenced papers (multicomponent: Johnson, 2001; binaries: Györffy and Stocks, 1983; Staunton et al., 1994). This chemical "ordering energy," arising only from changes in the electronic structure from the homogeneously random, is associated with perturbing the concentrations on two sites. There are $N(N-1)/2$ independent terms, as we expected. There is some additional q dependence resulting from the response of the CPA medium, which has been ignored for simplicity's sake to present this expression. Ignoring such q dependence is the same as what is done for the generalized perturbation methods (Duscastelle and Gautier, 1976; Duscastelle, 1991).

As the key result, the main q dependence of the ordering typically arises mainly from the convolution of the electronic structure given by

$$X_{L_2L_1;L_3L_4}(\mathbf{q};\varepsilon) = \frac{1}{\Omega_{\mathrm{BZ}}}\int d\mathbf{k}\left[\tau_{c;L_2L_3}(\mathbf{k}+\mathbf{q};\varepsilon)\tau_{c;L_4L_1}(\mathbf{k};\varepsilon)\right] - \tau_{c;ii;L_2L_3}(\varepsilon)\tau_{c;ii;L_4L_1}(\varepsilon) \quad (27)$$

which involves only changes to the CPA medium due to off-diagonal scattering terms. This is the difficult term to calculate. It is determined by the underlying electronic structure of the random alloy and must be calculated using electronic density functional theory. How various types of chemical ordering are revealed from $S^{(2)}_{\alpha\beta}(\mathbf{q};\varepsilon)$ is discussed later (see Data Analysis and Initial Interpretation). However, it is sometimes helpful to relate the ordering directly to the electronic dispersion through the Bloch spectral functions $A_{\mathrm{B}}(\mathbf{k};\varepsilon) \propto \mathrm{Im}\ \tau_c(\mathbf{k};\varepsilon)$ (Györffy and Stocks, 1983), where τ_c and the configurationally averaged Green's functions and charge densities are also related (Faulkner and Stocks, 1980). The Bloch spectral function defines the average dispersion in the alloy system. For ordered alloys, $A_{\mathrm{B}}(\mathbf{k};\varepsilon)$ consists of delta functions in $\mathbf{k}$-space whenever the dispersion relationship is satisfied, i.e., $\delta(\varepsilon-\varepsilon_k)$, which are the electronic "bands." In a disordered alloy, these "bands" broaden and shift (in energy) due to disorder and alloying effects. The loci of peak positions at ε_{F}, if the widths of the peaks are small on the scale of the Brillouin zone dimension, defines a "Fermi surface" in a disordered alloy. The widths reflect, for example, the inverse lifetimes of electrons, determining such quantities as resistivity (Nicholson and Brown, 1993). Thus, if only electronic states near the Fermi surface play the dominant role in determining the ordering tendency from the convolution integral, the reader can already imagine how

Fermi-surface nesting gives a large convolution from flat and parallel portions of electronic states, as detailed later.

Notably, the species- and energy-dependent matrix elements in Equation 26 can be very important, as discussed later for the case of NiPt. To appreciate how band-filling effects (as opposed to Fermi-surface-related effects) are typically expected to affect the ordering in an alloy, it is useful to summarize as follows. In general, the band-energy-only part of $S^{(2)}_{\alpha\beta}(\mathbf{q};\varepsilon)$ is derived from the filling of the electronic states and harbors the Hume-Rothery electron-per-atom rules (Hume-Rothery, 1963), for example. From an analysis using tight-binding theory, Ducastelle and others (e.g., Ducastelle, 1992) have shown what ordering is to be expected in various limiting cases where the transition metal alloys can be characterized by diagonal disorder (i.e., difference between on site energies is large) and off-diagonal disorder (i.e., the constituent metals have different d band widths). The standard lore in alloy theory is then as follows: if the d band is either nearly full or empty, then $S^{(2)}_{\text{Band}}(\mathbf{q})$ peaks at $|\mathbf{q}| = 0$ and the system clusters. On the other hand, if the bands are roughly half-filled, then $S^{(2)}_{\text{Band}}(\mathbf{q})$ peaks at finite $|\mathbf{q}|$ values and the system orders. For systems with the d band nearly filled, the system is filling antibonding type states unfavorable to order, whereas, the half-filled band would have the bonding-type states filled and the antibonding-type states empty favoring order (this is very similar to the ideas learned from molecular bonding applied to a continuum of states). Many alloys can have their ordering explained on this basis. However, this simple lore is inapplicable for alloys with substantial off-diagonal disorder, as recently discussed by Pinski et al. (1991, 1998), and as explained below (see Data Analysis and Initial Interpretation) sections.

While the "charge effects" are important to include as well (Mott, 1937), let us mention the overall gist of what is found (Staunton et al., 1994). There is a "charge-rearrangement" term that follows from implicit variations of the charge on site i and the concentration on site j, which represents a dielectric response of the CPA medium. In addition, charge density-charge density variations lead to Madelung-type energies. Thus, $S^{(2)}_{\text{total}}(\mathbf{q}) = S^{(2)}_{c,c}(\mathbf{q}) + S^{(2)}_{c,\rho}(\mathbf{q}) + S^{(2)}_{\rho,\rho}(\mathbf{q})$. The additional terms also affect the Onsager corrections discussed above (Staunton et al., 1994). Importantly, the density of states at the Fermi energy reflects the number of electrons available in the metal to screen excess charges coming from the solute atoms, as well as local fluctuations in the atomic densities due to the local environments (see Data Analysis and Initial Interpretation). In a binary alloy case, e.g., where there is a large density of states at the Fermi energy (ε_F), $S^{(2)}$ reduces mainly to a screened Coulomb term (Staunton et al., 1994), which determine the Madelung-like effects. In addition, the major q dependence arises from the excess charge at the ion positions via the Fourier transform (FT) of the Coulomb potential, $C(\mathbf{q}) = \text{FT}|R_i - R_j|^{-1}$,

$$S^{(2)}(\mathbf{q}) \approx S^{(2)}_{c,c}(\mathbf{q}) - \frac{e^2 \Delta Q^2 [C(\mathbf{q}) - R_{nn}^{-1}]}{1 + l_{\text{scr}}^2 [C(\mathbf{q}) - R_{nn}^{-1}]} \qquad (28)$$

where $\Delta Q = q_A - q_B$ is the difference in average excess charge (in units of e, electron charge) on a site in the homogeneous alloy, as determined by the self-consistent KKR-CPA. The average excess charge $q_{\alpha,i} = Z_{\alpha i} - \int_{\text{cell}} d\mathbf{r}\, \rho_{\alpha i}(\mathbf{r})$ (with Z_i the atomic number on a site). Here, l_{scr} is the system-dependent, metallic screening length. The nearest-neighbor distance, R_{nn}, arises due to charge correlations from the local environment (Pinski et al., 1998), a possibly important intersite electrostatic energy within metallic systems previously absent in CPA-based calculations—essentially reflecting that the disordered alloy already contains a large amount of electrostatic (Madelung) energy (Cole, 1997).

The proper (or approximate) physical description and importance of "charge correlations" for the formation energetics of random alloys have been investigated by numerous approaches, including simple models (Magri et al., 1990; Wolverton and Zunger, 1995b), in CPA-based, electronic-structure calculations (Abrikosov et al., 1992; Johnson and Pinski, 1993; Korzhavyi et al., 1995), and large supercell calculations (Faulkner et al., 1997), to name but a few. The sum of the above investigations reveal that for disordered and partially ordered metallic alloys, these atomic (local) charge correlations may be reasonably represented by a single-site theory, such as the coherent potential approximation. Including only the average effect of the charges on the nearest-neighbors shell (as found in Equation 28) has been shown to be sufficient to determine the energy of formation in metallic systems (Johnson and Pinski, 1993; Korzhavyi et al., 1995; Ruban et al., 1995), with only minor difference between various approaches that are not of concern here. Below (see Data Analysis and Initial Interpretation) we discuss the effect of incorporating such charge correlations into the concentration-wave approach for calculating the ASRO in random substitutional alloys (specifically fcc NiPt; Pinski et al., 1998).

DATA ANALYSIS AND INITIAL INTERPRETATION

Hybridization and Charge Correlation Effects in NiPt

The alloy NiPt, with its d band almost filled, is an interesting case because it stands as a glaring exception to traditional band-filling arguments from tight-binding theory (Treglia and Ducastelle, 1987): a transition-metal alloy will cluster, i.e., phase separate, if the Fermi energy lies near either d band edge. In fact, NiPt strongly orders in the CuAu (or ⟨100⟩-based) structure, with its phase diagram more like an fcc prototype (Massalski et al., 1990). Because Ni and Pt are in the same column of the periodic table, it is reasonable to assume that upon alloying there should be little effect from electrostatics and only the change in the band energy should really be governing the ordering. Under such an assumption, a tight-binding calculation based on average off-diagonal matrix elements reveals that no ordering is possible (Treglia and Ducastelle, 1987).

Such a band-energy-only calculation of the ASRO in NiPt was, in fact, one of the first applications of our thermodynamic linear-response approach based on the CPA (Pinski et al., 1991, 1992), and it gave almost quantitative

Table 2. The Calculated $\mathbf{k_0}$ (in Units of $2\pi/a$, Where a is the Lattice Constant) and T_{sp} (in K) for fcc Disordered NiPt (Including Scalar-relativistic Effects) at Various Levels of Approximation Using the Standard KKR-CPA (Johnson et al., 1986) and a Mean-field, Charge-correlated KKR-CPA (Johnson and Pinski, 1993), Labeled scr-KKR-CPA (Pinski et al., 1998)

Method	BEO[a]		BEO + Onsager[b]		BEO + Coulomb[c]		BEO + Coulomb + Onsager[d]	
KKR-CPA	100	1080	100	780	100	6780	$\frac{1}{2}\frac{1}{2}\frac{1}{2}$	1045
scr-KKR-CPA	100	1110	100	810	100	3980	100	905

[a]Band-energy-only (BEO) results.
[b]BEO plus Onsager corrections.
[c]Results including the charge-rearrangement effects associated with short-range ordering.
[d]Results of the full theory. Experimentally, NiPt has a T_c of 918 K (Massalski, et al., 1990).

agreement with experiment. However, our more complete theory of ASRO (Staunton et al., 1994), which includes Madelung-type electrostatic effects, dielectric effects due to rearrangement of charge, and Onsager corrections, yielded results for the transition temperature and unstable wavevector in NiPt that were simply wrong, whereas for many other systems we obtained very good results (Johnson et al., 1994). By incorporating the previously described screening contributions to the calculation of ASRO in NiPt (Pinski et al., 1998), the wave vector and transition temperature were found to be in exceptional agreement with experiment, as evidenced in Table 2.

In the experimental diffuse scattering on NiPt, a (100) ordering wave vector was found, which is indicative of CuAu ($L1_0$)-type short-range order (Dahmani et al., 1985), with a first-order transition temperature of 918 K. From Table 2, we see that using the improved screened (scr)-KKR-CPA yields a (100) ordering wave vector with a spinodal temperature of 905 K. If only the band-energy contributions are considered, for either the KKR-CPA or its screened version, the wave vector is the same and the spinodal temperature is about 1100 K (without the Onsager corrections). Essentially, the BEO approximation is reflecting most of the physics, as was anticipated based on their being in the same column of the periodic table. What is also clear is that the KKR-CPA, which contains a much larger Coulomb contribution, has necessarily much larger spinodal temperature (T_{sp}) before the Onsager correction is included. While the Onsager correction therefore must be very large to conserve spectral intensity, it is, in fact, the dielectric effects incorporated into the Onsager corrections that are trying to reduce such a large electrostatic contribution and change the wave vector into disagreement, i.e., $\mathbf{q} = (\frac{1}{2}\frac{1}{2}\frac{1}{2})$, even though the T_{sp} remains fairly good at 1045 K. The effect of the screening contributions to the electrostatic energy (as found in Equation 28) is to reduce significantly the effect of the Madelung energy (T_{sp} is reduced 40% before Onsager corrections); therefore, the dielectric effects are not as significant and do not change the wave vector dependence.

Ultimately, the origin for the ASRO reduces back to what is happening in the band-energy-only situation, for it is predominantly describing the ordering and temperature dependence and most of the electrostatic effects are canceling one another. The large electronic density of states at the Fermi level (Fig. 3) is also important for it is those electrons that contribute to screening and the dielectric response.

What remains to tell is why fcc NiPt wants to order with $\mathbf{q} = (100)$ periodicity. Lu et al. (1993) stated that relativistic effects induce the chemical ordering in NiPt. Their work showed that relativistic effects (specifically, the Darwin and mass-velocity terms) lead to a contraction of the s states, which stabilized both the disordered and ordered phases relative to phase separation, but their work did not explain the origin of the chemical ordering. As marked in electronic density of states (DOS) for disordered NiPt in Figure 3 (heavy, hatched lines), there is a large number of low-energy states below the Ni-based d band that arise due to hybridization with the Pt sites. These d states are of t_{2g} symmetry whose lobes point to the nearest-neighbor sites in an fcc lattice. Therefore, the system can lower its energy by modulating itself with a (100) periodicity to create lots of such favorable (low-energy, d-type) bonds between nearest-neighbor Ni and Pt sites. This basic explanation was originally given by Pinski et al. (1991, 1992).

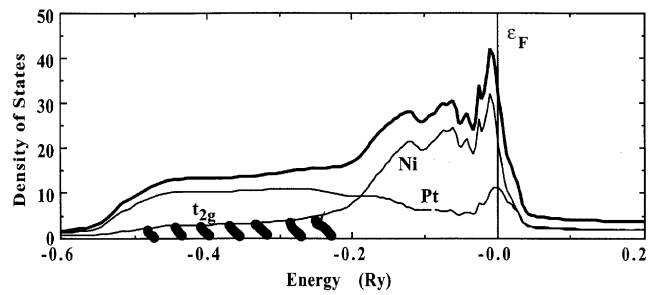


Figure 3. The calculated scr-KKR-CPA electronic density of states (states/Ry-atom) versus energy (Ry) for scalar-relativistic, disordered $Ni_{50}Pt_{50}$. The hybridized d states of t_{2g}-symmetry created due to an electronic size effect related to the difference in electronic bandwidths between Ni and Pt are marked by thick, hatched lines. The apparent pinning of the density of states at the Fermi level for Ni and Pt reflect the fact that the two elements fall in the same column of the periodic table, and there is effectively no "charge transfer" from electronegativity effects.

Pinski et al. (1991, 1992) pointed out that this hybridization effect arises due to what amounts to an electronic "size effect" related to the difference in bandwidths between Ni (little atom, small width) and Pt (big atom, large width), which is related to off-diagonal disorder in tight-binding theory. The lattice constant of the alloy plays a role in that it is smaller (or larger) than that of Pt (or Ni) which further increases (decreases) the bandwidths, thereby further improving the hybridization. Because Ni and Pt are in the same column of the periodic table, the Fermi level of the Ni and Pt d bands is effectively pinned, which greatly affects this hybridization phenomenon. See Pinski et al. (1991, 1992) for a more complete treatment.

It is noteworthy that in metallic NiPt ordering originates from effects that are well below the Fermi level. Therefore, usual ideas regarding reasons for ordering used in substitutional metallic alloys about e/a effects, Fermi-surface nesting, or filling of (anti-) bonding states, that is, all effects are due to the electrons around the Fermi level, should not be considered "cast in stone." The real world is much more interesting! This in hindsight turns out also to explain the failure of tight binding for NiPt: because off-diagonal disorder is important for Ni-Pt, it must be well described, that is, not to approximate those matrix elements by usual procedures. In effect, some system-dependent information of the alloying and hydridization must be included when establishing the tight-binding parameters.

Coupling of Magnetic Effects and Chemical Order

This hybridization (electronic "size") effect that gives rise to (100) ordering in NiPt is actually a more ubiquitous effect than one may at first imagine. For example, the observed $\mathbf{q} = (1\frac{1}{2}0)$, or Ni_4Mo-type, short-range order in paramagnetic, disordered AuFe alloys that have been fast quenched from high-temperature, results partially from such an effect (Ling, 1995b). In paramagnetic, disordered AuFe, two types of disorder (chemical and magnetic) must be described simultaneously [this interplay is predicted to allow changes to the ASRO through magnetic annealing (Ling et al., 1995b)].

For paramagnetic disordered AuFe alloys, the important point in the present context is that a competion arises between an electronic band-filling (or e/a) effect, which gives a clustering, or $\mathbf{q} = (000)$ type ASRO, and the stronger hybridization effect, which gives a $\mathbf{q} = (100)$ ASRO. The competition between clustering and ordering arises due to the effects from the magnetism (Ling et al., 1995b). Essentially, the large exchange splitting between the Fe majority and minority d band density of states results in the majority states being fully populated (i.e., they lie below the Fermi level), whereas the Fermi level ends up in a peak in the minority d band DOS (Fig. 4). Recall from usual band-filling-type arguments that filling bonding-type states favor chemical ordering, while filling antibonding-type states oppose chemical ordering (i.e., favor clustering). Hence, the hybridization "bonding states" that are created below the Fe d band due to interaction with the wider band Au (just as in NiPt) promotes ordering (Fig. 4), whereas the band filling of the minority

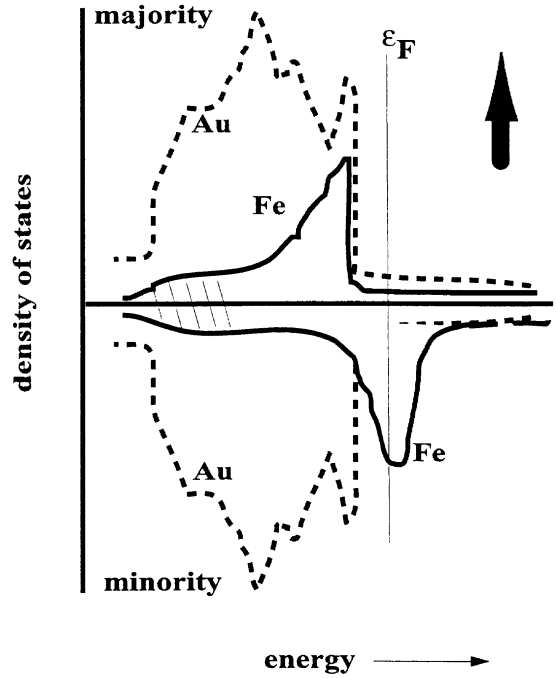


Figure 4. A schematic drawing of the electronic density of states (states/Ry-atom) versus energy (Ry) for scalar-relativistic, chemically disordered, and magnetically disordered (i.e., paramagnetic) $Au_{75}Fe_{25}$ using the CPA to configurationally average over both chemical and magnetic degrees of freedom. This represents the "local" density of states (DOS) for a site with its magnetization along the local z axis (indicated by the heavy vertical arrow). Due to magnetic disorder, there are equivalent DOS contributions from $-z$ direction, obtained by reflecting the DOS about the horizontal axis, as well as in the remaining 4π orientations. As with NiPt, the hybridized d states of t_{2g} symmetry are marked by hatched lines for both majority ($\uparrow$) and minority ($\downarrow$) electron states.

d band (which behave as "antibonding" states because of the exchange splitting) promotes clustering, with a compromise to $(1\frac{1}{2}0)$ ordering.

In the calculation, this interpretation is easily verified by altering the band filling, or e/a, in a rigid-band sense. As the Fermi level is lowered below the exchange-split minority Fe peak in Figure 4, the calculated ASRO rapidly becomes (100)-type, simply because the unfavorable antibonding states are being depopulated. Charge-correlation effects that were important for Ni-Pt are irrelevant for Au-Fe. By "magnetic annealing" the high-T AuFe in a magnetic field, we can utilize this electronic interplay to alter the ASRO to $\langle 100 \rangle$.

Multicomponent Alloys: Fermi-Surface Nesting, van Hove Singularities, and e/a in fcc Cu-Ni-Zn

Broadly speaking, the ordering in the related fcc binaries of Cu-Ni-Zn might be classified according to their phase diagrams (Massalski et al., 1990) as strongly ordering in NiZn, weakly ordering in CuZn, and clustering in CuNi. Perhaps then, it is no surprise that the phase diagram of Cu-Ni-Zn alloys (Thomas, 1972) reflects this, with clustering in Zn-poor regions, K-state effects (e.g., reduced

resistance with cold working), ⟨100⟩ short- (Hashimoto et al., 1985) and long-range order (van der Wegen et al., 1981), as well as $(1\frac{1}{4}0)$ (or DO_{23}-type) ASRO (Reinhard et al., 1990), and incommensurate-type ordering in the Ni-poor region.

Hashimoto et al. (1985) has shown that the three Warren-Cowley pair parameters for Cu_2NiZn reflect the above ordering tendencies of the binaries with strong ⟨100⟩-type ASRO in the Ni-Zn channel, and no fourfold diffuse scattering patterns, as is common in noble-metal-based alloys. Along with the transmission electron microscopy results of van der Wegen et al. (1981), which also suggest ⟨100⟩-type long-range order, it was assumed that Fermi-surface nesting, which is directly related to the geometry of the Fermi surface and has long been known to produce fourfold diffuse patterns in the ASRO, is not operative in this system. However, the absence of fourfold diffuse patterns in the ASRO, while necessary, is not sufficient to establish the nonexistence of Fermi-surface nesting (Althoff et al., 1995, 1996).

Briefly stated, and most remarkably, Fermi-surface effects (due to nesting and van Hove states) are found to be responsible for all the commensurate and incommensurate ASRO found in the Cu-rich, fcc ternary phase field. However, a simple interpretation based solely in terms of e/a ratio (Hume-Rothery, 1963) is not possible because of the added complexity of disorder broadening of the electronic states and because both composition and e/a may be independently varied in a ternary system, unlike in binary systems. Even though Fermi-surface nesting is operative, which is traditionally said to produce a four-fold incommensurate peak in the ASRO, a [100]-type of ASRO is found over an extensive composition range for the ternary, which indicates an important dependence of the nesting wavevector on e/a and disorder. In the random state, the broadening of the alloy's Fermi surface from the disorder results in certain types of ASRO being stronger or persisting over wider ranges of e/a than one determines from sharp Fermi surfaces.

For the fcc Cu-Ni-Zn, the electron states near the Fermi energy, ε_F, play the dominant role in determining the ordering tendency found from $S^{(2)}(\mathbf{q})$ (Althoff et al., 1995, 1996). In such a case, it is instructive to interpret (not calculate) $S^{(2)}(\mathbf{q})$ in terms of the convolution of Bloch spectral functions $A_B(\mathbf{k}; \varepsilon)$ (Györffy and Stocks, 1983). The Bloch spectral function defines the average dispersion in the system and $A_B(\mathbf{k}; \varepsilon) \propto \mathrm{Im}\,\tau_c(\mathbf{k}; \varepsilon)$. As mentioned earlier, for ordered alloys $A_B(\mathbf{k}; \varepsilon)$ consists of delta functions in $\mathbf{k}$ space whenever the dispersion relationship is satisfied, i.e., $\delta(\varepsilon - \varepsilon_k)$, which are the electronic "bands." In a disordered alloy, these "bands" broaden and shift (in energy) due to disorder and alloying effects. The loci of peak positions at ε_F, if the widths of the peaks are small on the scale of the Brillouin zone dimension, defines a "Fermi surface" in a disordered alloy.

Provided that k-, energy-, and species-dependent matrix elements can be roughly neglected in $S^{(2)}(\mathbf{q})$, and that only the energies near ε_F are pertinent because of the Fermi factor (NiPt was a counterexample to all this), then the q-dependent portion of $S^{(2)}(\mathbf{q})$ is proportional to a convolution of the spectral density of states at

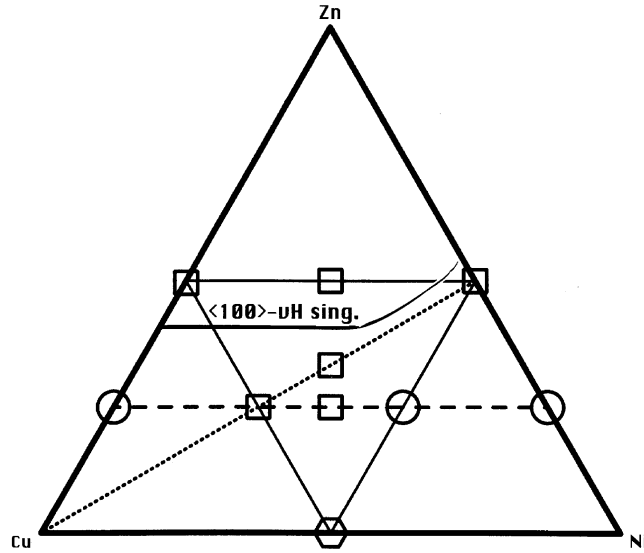


Figure 5. The Cu-Ni-Zn Gibbs triangle in atomic percent. The dotted line is the Cu isoelectronic line. The ASRO is designated as: squares, ⟨100⟩ ASRO; circles, incommensurate ASRO; hexagon, clustering, or (000) ASRO. The additional line marked ⟨100⟩-vH establishes roughly where the fcc Fermi surface of the alloys has spectral weight (due to van Hove singularities) at the ⟨100⟩ zone boundaries, suggesting bcc is nearing in energy to fcc. For fcc CuZn, this occurs at 40% Zn, close to the maximum solubility limit of 38% Zn before transformation to bcc CuZn. Beyond this line a more careful determination of the electronic free energy is required to determined fcc or bcc stability.

ε_F (the Fermi surface; Györffy and Stocks, 1983; Györffy et al., 1989), i.e.:

$$S^{(2)}_{\alpha\beta}(\mathbf{q}) \propto \int d\mathbf{k} A_B(\mathbf{k}; \varepsilon_F) A_B(\mathbf{k} + \mathbf{q}; \varepsilon_F) \qquad (29)$$

With the Fermi-surface topology playing the dominate role, ordering peaks in $S^{(2)}(\mathbf{q})$ can arise from states around ε_F in two ways: (1) due to a spanning vector that connects parallel, flat sheets of the Fermi surface to give a large convolution (so-called **Fermi-surface nesting**; Györffy and Stocks, 1983), or, (2) due to a spanning vector that promotes a large joint density of states via convolving points where van Hove singularities (van Hove, 1953) occur in the band structure at or near ε_F (Clark et al., 1995). For fcc Cu-Ni-Zn, both of these Fermi-surface-related phenomena are operative, and are an ordering analog of a Peierls transition.

A synopsis of the calculated ASRO is given in Figure 5 for the Gibbs triangle of fcc Cu-Ni-Zn in atomic percent. All the trends observed experimentally are completely reproduced: Zn-poor Cu-Ni-Zn alloys and Cu-Ni binary alloys show clustering-type ASRO; along the line $Cu_{0.50+x}Ni_{0.25-n}Zn_{0.25}$ (the dashed line in the figure), $Cu_{75}Zn$ shows $(1\frac{1}{4}0)$-type ASRO, which changes to commensurate (100)-type at Cu_2NiZn, and then to fully incommensurate around $CuNi_2Zn$, where the K-state effects are observed. K-state effects have been tied to the short-range order (Nicholson and Brown, 1993). Most interestingly, a large

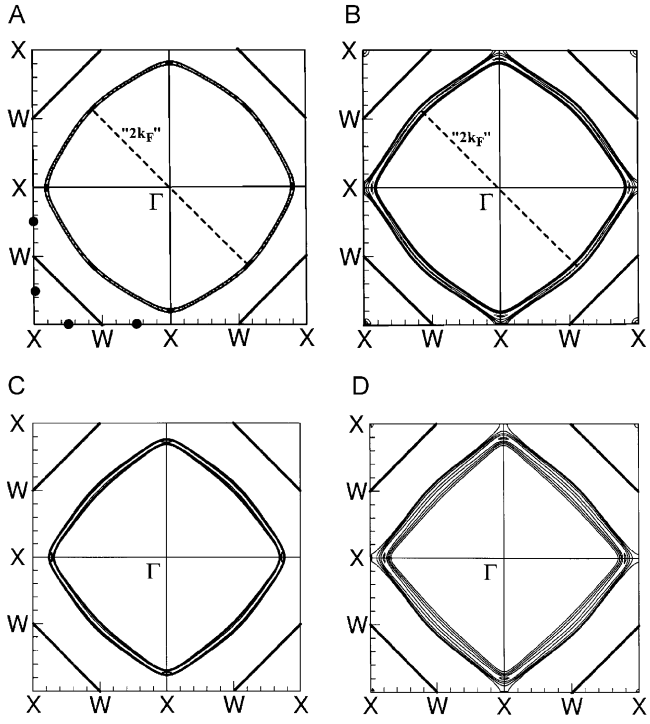


Figure 6. The Fermi surface, or $A_B(\mathbf{k};\varepsilon_F)$, in the {100} plane of the first Brillouin zone for fcc alloys with a lattice constant of 6.80 a.u.: (**A**) $Cu_{75}Zn_{25}$, (**B**) $Cu_{25}Ni_{25}Zn_{50}$, (**C**) $Cu_{50}Ni_{25}Zn_{25}$, and (**D**) $Ni_{50}Zn_{50}$. Note that (**A**) and (**B**) have $e/a = 1.25$ and (**C**) and (**D**) have $e/a = 1.00$. As such, the caliper dimensions of the Fermi surface, as measured from peak to peak (and typically referred to as "$2k_F$"), are identical for the two pairs. The widths change due to increased disorder: NiZn has the greatest difference between scattering properties and therefore the largest widths. In the lower left quadrant of (**A**) are the fourfold diffuse spots that occur due to nesting. The fourfold diffuse spots may be obtained graphically by drawing circles (actually spheres) of radius "$2k_F$" from all Γ points and finding the common intersection of such circles along the *X-W-X* high symmetry lines.

region of (100)-type ordering is calculated around the Cu isoelectronic line (the dotted line in the figure), as is observed (Thomas, 1972).

The Fermi surface in the ⟨100⟩ plane of $Cu_{75}Zn$ is shown in Figure 6, part A, and is reminiscient of the Cu-like "belly" in this plane. The caliper dimensions, or so-called "$2k_F$," of the Fermi surface in the [110] direction is marked; it is measured peak to peak and determines the nesting wavevector. It should be noted that perpendicular to this plane ([001] direction) this rather flat portion of Fermi surface continues to be rather planar, which additionally contributes to the convolution in Equation 29 (Althoff et al., 1996). In the lower left quadrant of Figure 6, part A, are the fourfold diffuse spots that occur due to the nesting. As shown in Figure 6, parts C and D, the caliper dimensions of the Fermi surface in the ⟨100⟩ plane are the same along the Cu isoelectronic line (i.e., constant e/a = 1.00). For NiZn and Cu_2NiZn, this "$2k_F$" gives a (100)-type ASRO because its magnitude matches the $\Delta k = |(000) - (110)|$, or $\Gamma - X$, distance perfectly. The spectral widths change due to increased disorder. NiZn has the greatest difference between scattering properties and therefore the largest widths (see Fig. 6). The increasing disorder with decreasing Cu actually helps improve the convolution of the spectral density of states, Equation 29, and strengthens the ordering, as is evidenced experimentally through the phase-transformation temperatures (Massalski et al., 1990).

As one moves off this isoelectronic line, the caliper dimensions change and an incommensurate ASRO is found, as with $Cu_{75}Zn$ and $CuNi_2Zn$ (see Fig. 6, parts A and B). As Zn is added, eventually van Hove states (van Hove, 1953) appear at (100) points or X-points (see Fig. 6, part D) due to symmetry requirements of the electronic states at the Brillouin zone boundaries. These van Hove states create a larger convolution integral favoring (100) order over incommensurate order. For $Cu_{50}Zn_{50}$, one of the weaker ordering cases, a competition with temperature is found between spanning vectors arising from Fermi-surface-nesting and van Hove states (Althoff et al., 1996). For compositions such as $CuNiZn_2$, the larger disorder broadening and increase in van Hove states make the (100) ASRO dominant.

It is interesting to note that the appearance of van Hove states at (100) points, such as for $Cu_{60}Zn_{40}$, where Zn has a maximum solubility of 38.5% experimentally (Thomas, 1972; Massalski et al., 1990) occurs like precursors to the observed fcc-to-bcc transformations (see rough sketch in the Gibbs triangle; Fig. 5). A detailed discussion that clarifies this correlation has been given recently about the effect of Brillouin zone boundaries in the energy difference between fcc and bcc Cu-Zn (Paxton et al., 1997).

Thus, all the incommensurate and commensurate ordering can be explained in terms of Fermi-surface mechanisms that were dismissed experimentally as a possibility due to the absence of fourfold diffuse scattering spots. Also, disorder broadening in the random alloy plays a role, in that it actual helps the ordering tendency by improving the (100) nesting features. The calculated T_{sp} and other details may be found in Althoff et al. (1996). This highlights one of the important roles for theory: to determine the underlying electronic mechanism(s) responsible for order and make predictions that can be verified from experiment.

Polarization of the Ordering Wave in Cu_2NiZn

As we have already discussed, a ternary alloy like fcc $ZnNiCu_2$ does not possess the A-B symmetry of a binary; the analysis is therefore more complicated due to the concentration waves having "polarization" degrees of freedom, requiring more information from experiment or theory. In this case, the extra degree of freedom introduced by the third component leads also to additional ordering transitions at lower temperatures. These polarizations (as well as the unstable wavevector) are determined by the electronic interactions; also they determine the sublattice occupations that are (potentially) made inequivalent in the partially ordered state (Althoff et al., 1996).

The relevent star of $\mathbf{k}_0 = \langle 100 \rangle$ ASRO—comprised of (100), (010), (001) vectors—found for $ZnNiCu_2$ is a precursor to the partially ordered state that may be determined

approximately from the eigenvectors of $q^{-1}(\mathbf{k}_0)$, as discussed previously (see Principles of the Method). We have written the alloy in this way because Cu has been arbitrarily taken as the "host." If the eigenvectors are normalized, then there is but one parameter that describes the eigenvectors of Δ_F in the Cartesian or Gibbsian coordinates, which can be written:

$$\begin{bmatrix} e_1^{\mathrm{zn}}(\mathbf{k}_0) \\ e_1^{\mathrm{Ni}}(\mathbf{k}_0) \end{bmatrix} = \begin{bmatrix} \sin\theta_{\mathbf{k}_0} \\ -\cos\theta_{\mathbf{k}_0} \end{bmatrix} ; \begin{bmatrix} e_2^{\mathrm{zn}}(\mathbf{k}_0) \\ e_2^{\mathrm{Ni}}(\mathbf{k}_0) \end{bmatrix} = \begin{bmatrix} \cos\theta_{\mathbf{k}_0} \\ \sin\theta_{\mathbf{k}_0} \end{bmatrix} \qquad (30)$$

If $\theta_{\mathbf{k}}$ is taken as the parameter in the Cartesian space, then in the Gibb's space the eigenvectors are appropriate linear combinations of $\theta_{\mathbf{k}}$. The "angle" $\theta_{\mathbf{k}}$ is fully determined by the electronic interactions and plays the role of "polarization angle" for the concentration wave with the $\mathbf{k}_0$ wavevector. Details are fully presented in the appendix of Althoff et al. (1996). However, the lowest energy concentration mode in Gibbs space at $T = 1000\,\mathrm{K}$ for the $\mathbf{k}_0 = \langle 100 \rangle$ is given by $e^{\mathrm{Zn}} = 1.0517$ and $e^{\mathrm{Ni}} = -0.9387$, where one calculates $T_{\mathrm{sp}} = 985\,\mathrm{K}$, including Onsager corrections (experimental $T_{\mathrm{c}} = 774$ K; Massalski et al., 1990).

For a ternary alloy, the matrices are of rank 2 due to the independent degrees of freedom. Therefore, there are two possible order parameters, and hence, two possible transitions as the temperature is lowered (based on our knowledge from high-T information). For the partially ordered state, the long-range order parameter associated with the higher energy mode can be set to zero. Using this information in Equation 9, as discussed by Althoff et al. (1996), produces an atomic distribution in real space for the partially ordered state as in Table 3. Clearly, there is already a trend to a tetragonal, $L1_0$-like state with Zn-enhanced on cube corners, as observed (van der Wegen et al., 1981) in the low-temperature, fully ordered state (where Zn is on the fcc cube corners, Cu occupies the faces in the central plane, and Ni occupies the faces in the Zn planes). However, there is still disorder on all the sublattices. The $\langle 100 \rangle$ wave has broken the disordered fcc cube into a four-sublattice structure, with two sublattices degenerate by symmetry.

Unfortunately, the partially ordered state assessed from TEM measurements (van der Wegen et al, 1981) suggests that it is $L1_2$-like, with Cu/Ni disordered on all the cube faces and predominately Zn on the fcc corners. Interestingly, if domains of the calculated $L1_0$ state occur with an equal distribution of tetragonal axes, then a state with $L1_2$ symmetry is produced, similar to that supposed by TEM. Also, because the discussion is based on the stability of the high-temperature disordered state, the temperature for the second transition cannot be gleaned from the eigenvalues directly. However, a simple estimate can be made. Before any sublattice has a negative occupation value, which occurs for $\eta = 0.49$ (see Ni in Table 3), the second long-range order parameter must become finite and the second mode becomes accessible. As the transition temperature is roughly proportional to $\Delta \mathrm{E}_{\mathrm{order}}$, or η^2, then $T^{\mathrm{II}} = (1 - \eta^2) T^{\mathrm{I}}$ (assuming that the Landau coefficients are the same). Therefore, $T^{\mathrm{II}}/T^{\mathrm{I}} = 0.76$, which is close to the experimental value of 0.80 (i.e., 623 K/774 K) (Massalski et al., 1990). Further discussion and comparison with experiment may be found elsewhere (Althoff et al., 1996), along with allowed ordering due to symmetry restrictions.

Table 3. Atomic Distributions in Real Space of the Partially Ordered State to which the Disordered State with Stochiometry Cu_2NiZn is Calculated to be Unstable at T_c

Sublattice	Alloy Component	Site-Occupation Probability[a]
	Zn	$0.25 + 0.570\eta(T)$
1: Zn rich	Ni	$0.25 - 0.510\eta(T)$
	Cu	$0.50 - 0.060\eta(T)$
	Zn	$0.25 - 0.480\eta(T)$
2: Ni rich	**Ni**	$0.25 + 0.430\eta(T)$
	Cu	$0.50 + 0.050\eta(T)$
	Zn	$0.25 - 0.045\eta(T)$
3 and ≤ 4: Random	Ni	$0.25 + 0.040\eta(T)$
	Cu	$0.50 + 0.005\eta(T)$

[a] η is the long-range-order parameter, where $0 \leq \eta \leq 1$. Values were obtained from Althoff et al. (1996).

Electronic Topological Transitions: van Hove Singularities in CuPt

The calculated ASRO for $Cu_{50}Pt_{50}$ (Clark et al., 1995) indicates an instability to concentration fluctuations with a $\mathbf{q} = (\frac{1}{2}\frac{1}{2}\frac{1}{2})$, consistent with the observed $L1_1$ or CuPt ordering (Massalski et al., 1990). The $L1_1$ structure consists of alternating fcc (111) layers of Cu and Pt, in contrast with the more common $L1_0$ structure, which has alternating (100) planes of atoms. Because CuPt is the only substitutional metallic alloy that forms in the $L1_1$ structure (Massalski et al., 1990), it is appropriate to ask: what is so novel about the CuPt system and what is the electronic origin for the structural ordering?

The answers follow directly from the electronic properties of disordered CuPt near its Fermi surface, and arise due to what Lifshitz (1960) termed an "electronic topological transition." That is, due to the topology of the electronic structure, electronic states, which are possibly unfavorable, may be filled (or unfilled) due to small changes in lattice or chemical structure, as arising from Peierls instabilities. Such electronic topological transitions may affect a plethora of observables, causing discontinuities in, e.g., lattice constants and specific heats (Bruno et al., 1995). States due to van Hove singularities, as discussed in fcc Cu-Ni-Zn, are one manifestation of such topological effects, and such states are found in CuPt. In Figure 7, the Fermi surface of disordered CuPt around the L point has a distinctive "neck" feature similar to elemental Cu. Furthermore, because ε_F cuts the density of states near the top of a feature that is mainly Pt-d in character (see Fig. 8, part A) pockets of d holes exist at the X points (Fig. 7). As a result, the ASRO has peaks at $(\frac{1}{2}\frac{1}{2}\frac{1}{2})$ due to the spanning vector $X - L = (0, 0, 1) - (\frac{1}{2}\frac{1}{2}\frac{1}{2})$ (giving a large joint electron density of states in Equation 29), which is a member of the star of L. Thus, the $L1_1$ structure is stabilized by a Peierls-like mechanism arising from the

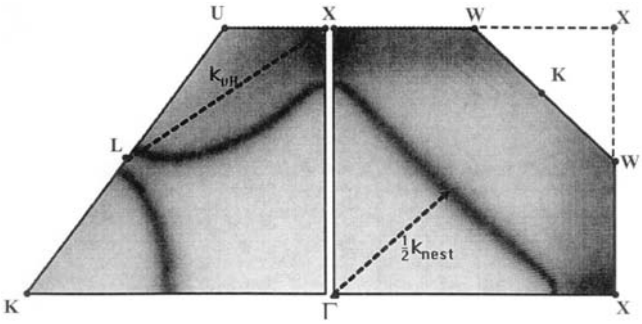


Figure 7. $A_B(\mathbf{k}; \varepsilon_F)$ for disordered fcc CuPt, i.e., the Fermi surface, for portions of the ⟨110⟩ (Γ-*X*-*U*-*L*-*K*), and ⟨100⟩ (Γ-*X*-*W*-*K*-*W*-*X*-*L*) planes. Spectral weight is given by relative gray scale, with black as largest and white as background. Note the neck at *L*, and the smeared pockets at *X*. The widths of the peaks are due to the chemical disorder experienced by the electrons as they scatter through the random alloy. The spanning vector, k_{vH}, associated with states near van Hove singularities, as well as typical "$2k_F$" Fermi-surface nesting are clearly labeled. The more Cu in the alloy the fewer *d* holes, which makes the "$2k_F$" mechanism more energetically favorable (if the dielectric effects are accounted for fully; Clark et al., 1995).

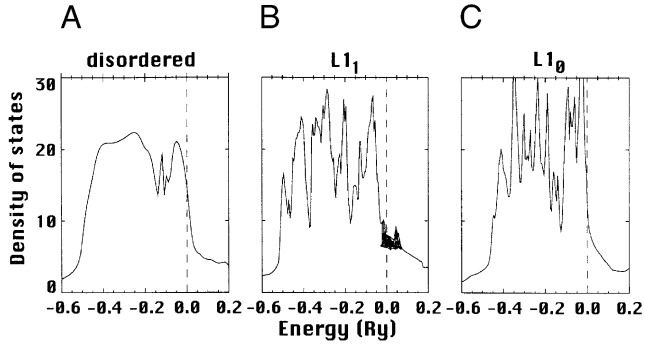


Figure 8. Scalar-relativistic total densities of states for (**A**) disordered CuPt, using the KKR-CPA method; ordered CuPt in the (**B**) $L1_1$ and (**C**) $L1_0$ structures, using the LMTO method. The dashed line indicates the Fermi energy. Note the change of scale in partial Pt state densities. The bonding (antibonding) states created by the $L1_1$ concentration wave just below (above) the Fermi energy are shaded in black.

hybridization between van Hove singularities at the high-symmetry points. This hybridization is the only means the system has to fill up the few remaining (antibonding) Pt *d* states, which is why this $L1_1$ ordering is rather unique to CuPt. That is, by ordering along the (111) direction, all the states at the *X* points—(100), (010), and (001)—may be equally populated, whereas only the states around (100) and (010) are fully populated with an (001) ordering wave consistent with $L1_0$ type order. See Clark et al. (1995) for more details.

This can be easily confirmed as follows. By increasing the number of *d* holes at the *X* points, $L1_1$ ordering should not be favored because it becomes increasingly more difficult for a $(\frac{1}{2}\frac{1}{2}\frac{1}{2})$ concentration wave to occupy all the *d* holes at *X*. Indeed, calculations repeated with the Fermi level lowered by 30 mRy (in a rigid-band way) into the Pt *d*-electron peak near ε_F results in a large clustering tendency (Clark et al., 1995). By filling the Pt *d* holes of the disordered alloy (raise ε_F by 30 mRy, see Fig. 8), thereby removing the van Hove singularities at ε_F, there is no great advantage to ordering into $L1_1$ and $S^{(2)}(\mathbf{q})$ now peaks at all *X* points, indicating $L1_0$-type ordering (Clark et al., 1995).

This can be confirmed from ordered band-structure calculations using the linear muffin tin orbital method (LMTO) within the atomic sphere approximation (ASA). In Figure 8, we show the calculated LMTO electronic densities of states for the $L1_0$ and $L1_1$ configurations for comparison to the density of states for the CPA disordered state, as given by Clark et al. (1995). In the disordered case, Figure 8, part A, ε_F cuts the top of the Pt *d* band, which is consistent with the *X* pockets in the Fermi surface. In the $L1_1$ structure, the density of states at ε_F is reduced, since the modulation in concentration introduces couplings between states at ε_F. The $L1_0$ density of states in Figure 8, part C demonstrates that not all ordered structures will produce this effect. Notice the small Peierls-type set of bonding and antibonding peaks that exist in the $L1_1$ Pt *d*-state density in Figure 8, part B (darkened area). Furthermore, the $L1_0 - L1_1$ energy difference is 2.3 mRy per atom with LMTO (2.1 mRy with full-potential method; Lu et al., 1991) in favor of the $L1_1$ structure, which confirms the associated lowering of energy with $L1_1$-type ordering. We also note that without the complete description of bonding (particularly *s* contributions) in the alloy, the system would not be globally stable, as discussed by (Lu et al., 1991).

The ordering mechanism described here is similar to the conventional Fermi surface nesting mechanism. However, conventional Fermi surface nesting takes place over extended regions of *k* space with spanning vectors between almost parallel sheets. The resulting structures tend to be long-period superstructures (LPS), which are observed in Cu-, Ag-, and Au-rich alloys (Massalski et al., 1990). In contrast, in the mechanism proposed for CuPt, the spanning vector couples only the regions around the *X* and *L* points in the fcc Brillouin zone, and the large joint density of states results from van Hove singularities that exist near ε_F. The van Hove mechanism will naturally lead to high-symmetry structures with short periodicities, since the spanning vectors tend to connect high-symmetry points (Clark et al., 1995).

What is particularly interesting in $Cu_{1-c}Pt_c$ is that the $L1_1$ ordering (at $c \approx 0.5$) and the one-dimensional LPS associated with Fermi-surface nesting (at $c \approx 0.73$) are both found experimentally (Massalski et al., 1990). Indeed, there are nested regions of Fermi surface in the (100) plane (see Fig. 7) associated with the *s*-*p* electrons, as found in Cu-rich Cu-Pd alloys (Györffy and Stocks, 1983). The Fermi-surface nesting dimension is concentration dependent, and, $\alpha(\mathbf{q})$ peaks at $\mathbf{q} = (1,0.2,0)$ at 73% Cu, provided both band-energy and double-counting terms are included

(Clark et al., 1995). Thus, a cross-over is found from a conventional Fermi-surface ordering mechanism around 75% Cu to ordering dominated by the novel van Hove-singularity mechanism at 50%. At higher Pt concentrations, $c \approx 0.25$, the ASRO peaks at L with subsidiary peaks at X, which is consistent with the ordered tetragonal fcc-based superstructure of $CuPt_3$ (Khachaturyan, 1983). Thus, just as in Cu-Ni-Zn alloys, nesting from s-p states and van Hove singularities (in this case arising from d states) both play a role, only here the effects from van Hove singularities cause a novel, and observed, ordering in CuPt.

On the Origin of Temperature-Dependent Shifts of ASRO Peaks

The ASRO peaks in Cu_3Au and Pt_8V at particular (h, k, l) positions in reciprocal-space have been observed to shift with temperature. In Cu_3Au, the four-fold split diffuse peaks at $(1k0)$ positions about the (100) points in reciprocal-space coalesce to one peak at T_c, i.e., $k \rightarrow 0$; whereas, the splitting in k increases with increasing temperature (Reichert et al., 1996). In Pt_8V, however, there are twofold split diffuse peaks at $(1 \pm \Delta h,0,0)$ and the splitting, Δh, decreases with increasing temperature, distinctly opposite to Cu_3Au (Le Bulloc'h et al., 1998).

Following the Cu_3Au observations, several explanations have been offered for the increased splitting in Cu_3Au, all of which cite entropy as being responsible for increasing the fourfold splitting (Reichert et al., 1996, 1997; Wolverton and Zunger, 1997). It was emphasized that the behavior of the diffuse scattering peaks shows that its features are not easily related to the energetics of the alloy, i.e., the usual Fermi-surface nesting explanation of fourfold diffuse spots (Reichert et al., 1997). However, entropy is not an entirely satisfactory explanation for two reasons. First, it does not explain the opposite behavior found for Pt_8V. Second, entropy by its very nature is dimensionless, having no q dependence that can vary peak positions.

A relatively simple explanation has been recently offered by Le Bulloc'h et al. (1998), although it is not quantitative. They detail how the temperature dependence of peak splitting of the ASRO is affected differently depending on whether the splitting occurs along $(1k0)$, as in Cu_3Au and Cu_3Pd, or whether it occurs along $(h00)$ as in Pt_8V. However, the origin of the splitting is always just related to the underlying chemical interactions and energetics of the alloy. While the electronic origin for the splitting would be obtained directly from our DFT approach, this subtle temperature and entropy effect would not be properly described by the method under its current implementation.

CONCLUSION

For multicomponent alloys, we have described how the "polarization of the ordering waves" may be obtained from the ASRO. Besides the unstable wavevector(s), the polarizations are the additional information required to define the ordering tendency of the alloy. This can also be obtained from the measured diffuse scattering intensities, which, heretofore, have not been appreciated. Furthermore, it has been the main purpose of this unit to give an overview of an electronic-structure-based method for calculating atomic short-range order in alloys from first principles. The method uses a linear-response approach to obtain the thermodynamically induced ordering fluctuations about the random solid solution as described via the coherent-potential approximation. Importantly, this density functional-based concentration-wave theory is generalized in a straightforward manner to multicomponent alloys, which is extremely difficult for most other techniques.

While the approach is clearly mean-field (as thoroughly outlined), it incorporates on an equal footing many of the electronic and entropic mechanisms that may compete on a system-by-system basis. This is especially notable when in metals the important energy differences are the order of a few mRy, where 1 mRy is 158 K; thus, even a 10% error in calculated temperature scales at this point is amazingly good, although for some systems we have done much better. When the ASRO indicates the low-temperature, ordered state (as is usually the case), then it is possible to determine the underlying electronic mechanism responsible for the phase transformation. In any case, it does determine the origin of the atomic short-range order. Nevertheless, the first-principles concentration wave theory does not include possibly important effects, such as statistical fluctuations beyond mean field or cluster entropy, which may give rise to first-order transformations entirely distinct from the premonitory fluctuations or temperature effects in the ASRO. In such cases, more accurate statistical methods, such as Monte Carlo, may be employed if the energetics that are relevant (as determined by some means) can also be obtained with sufficient relative accuracy.

A few electronic mechanisms were showcased which gave rise to ordering in various binary and ternary alloys. Fermi-surface effects explain all the commensurate and incommensurate ordering tendencies in fcc Cu-Ni-Zn alloys, in contrast to interpretations made from experimental data. A hybridization effect that occurs well below the Fermi level produces the strong $L1_0$-type order in NiPt. The hybridization and well-known band-filling (or, e/a) effects explain the ASRO in AuFe alloys, if magnetic exchange splitting is incorporated. A novel van Hove singularity mechanism, which arises due to the topology of the electronic structure of the disordered alloy, explains the unique $L1_1$-type order found in CuPt. Without the ability to connect the ASRO to electronic effects, many of these effects would have been impossible to identify via traditional band-structure applications, even for the cases of long-range order, which speaks to the usefulness of the technique. The first-principles, concentration-wave technique may be even more useful in multicomponent alloys. Currently, several ternary alloy systems are under investigation to determine the site-occupancy preferences for partially ordered $B2$ Nb-Al-Ti alloys, as recently measured by Hou et al. (1997) and Johnson et al. (1999). Nevertheless, contributions from size and electrostatic effects

must still be included in the multicomponent case. At this point, only more applications and comparisons with experiment on a system-by-system basis will reveal important new insights into origins of alloy phase stability.

ACKNOWLEDGMENTS

This work was supported by the Division of Materials Science, Office of Basic Energy Sciences, U.S. Department of Energy when Dr. Johnson was at Sandia National Laboratories (contract DE-AC04-94AL85000) and recently at the Frederick Seitz Materials Research Laboratory (contract DEFG02-96ER45439). Dr. Johnson would like to thank Alphonse Finel for informative discussions regarding his recent work, and, the NATO AGARD program for making the discussion possible with a visit to O.N.E.R.A., France. It would be remiss not to acknowledge also the many collaborators throughout developments and applications: Jeff Althoff, Mark Asta, John Clark, Beniamino Ginatempo, Balazs Györffy, Jeff Hoyt, Michael Ling, Bill Shelton, Phil Sterne, and G. Malcolm Stocks.

LITERATURE CITED

Althoff, J. D., Johnson, D. D., and Pinski, F. J., 1995. Commensurate and incommensurate ordering tendencies in the ternary fcc Cu-Ni-Zn system. *Phys. Rev. Lett*. 74:138.

Althoff, J. D., Johnson, D. D., Pinski, F. J., and Staunton, J. B. 1996. Electronic origins of ordering in multicomponent metallic alloys: Application to the Cu-Ni-Zn system. *Phys. Rev. B* 53:10610.

Asta, M. and Johnson, D. D. 1997. Thermodynamic properties of fcc-based Al-Ag alloys. *Comp. Mater. Sci*. 8:64.

Badalayan, D. A., Khachaturyan, A. G., and Kitaigorodskii, A. I. 1969. Theory of order-disorder phase transformations in molecular crystals. II. *Kristallografiya* 14:404.

Barrachin, M., Finel, A., Caudron, R., Pasturel, A., and Francois, A. 1994. Order and disorder in Ni_3V effective pair interactions and the role of electronic excitations. *Phys. Rev. B* 50:12980.

Berlin, T. H. and Kac, M. 1952. Spherical model of a ferromagnet. *Phys. Rev*. 86:821.

Boriçi135>i-Kuqo, M. and Monnier, R. 1997. Short-range order in the random binary Madelung lattice. *Comp. Mater. Sci*. 8:16.

Brout, R. and Thomas, H. 1965. Phase Transitions. W. A. Benjamin, Inc., Menlo Park, Calif.

Brout, R. and Thomas, H. 1967. Molecular field theory, the Onsager reaction field and the spherical model. *Physics* 3:317.

Bruno, E., Ginatempo, B., and Guiliano, E. S. 1995. Fermi-surface and electronic topological transition in metallic random alloys (I): Influence on equilibrium properties. *Phys. Rev*. 52:14544.

Butler, C. J., McCartney, D. G., Small, C. J., Horrocks, F. J., and Saunders, N. 1997. Solidification microstructure and calculated phase equilibria in the Ti-Al-Mn system. *Acta Mater. (USA)*. 45:2931.

Ceder, G., Garbulsky, G. D., Avis, D., and Fukuda, K. 1994. Ground states of a ternary fcc lattice model with nearest-neighbor and next-nearest-neighbor interactions. *Phys. Rev. B* 49:1.

Chepulskii, R. V. and Bugaev, V. N. 1998. Analytical methods for calculations of the short-range order in alloys (a) I General theory, *J. Phys.: Conds. Matter* 10:7309. (b) II Numerical accuracy studies, *J. Phys.: Conds. Matter* 10:7327.

Clapp, P. C. and Moss, S. C. 1966. Correlation functions of disordered binary alloys I. *Phys. Rev*. 142:418.

Clark, J., Pinski, F. J., Johnson, D. D., Sterne, P. A., Staunton, J. B., and Ginatempo, B. 1995. van Hove singularity induced $L1_1$ ordering in CuPt. *Phys. Rev. Lett*. 74:3225.

Cole, R. J., Brooks, N. J., and Weightman, P. 1997. Madelung potentials and disorder broadening of core photoemission spectra in random alloys. *Phys. Rev. Lett*. 78P:3777.

Connolly, J. W. D. and Williams, A. R. 1983. Density-functional theory applied to phase transformations in transition-metal alloys. *Phys. Rev. B* 27:RC5169.

Dahmani, C. E., Cadeville, M. C., Sanchez, J. M., and Moran-Lopez, J. L. 1985. Ni-Pt phase diagram: experiment and theory. *Phys. Rev. Lett*. 55:1208.

Duscastelle, F. 1991. Order and phase stability in alloys. (F. de Boer and D. Pettifor, eds.) p. 303–313. North-Holland, Amsterdam.

Duscastelle, F. and Gautier, F. 1976. Generalized perturbation theory in disordered transitional alloys: Application to the calculation of ordering energies. *J. Phys.* F6:2039.

de Fontaine, D. 1973. An analysis of clustering and ordering in multicomponent solid solutions. I. Fluctuations and kinetics. *J. Phys. Chem. Solids* 34:1285.

de Fontaine, D. 1975. k-space symmetry rules for order-disorder reactions. *Acta Metall*. 23:553.

de Fontaine, D. 1979. Configurational thermodynamics of solid solutions. *Solid State Phys*. 34:73.

Evans, R. 1979. Density-functional theory for liquids. *Adv. Phys*. 28:143.

Faulkner, J. S. and Stocks, G. M. 1980. Calculating properties with the coherent-potential approximation. *Phys. Rev. B* 21:3222.

Faulkner, J. S., Wang, Yang, and Stocks, G. M. 1997. Coulomb energies in alloys. *Phys. Rev. B* 55:7492.

Finel, A., Barrachin, M., Caudron, R., and Francois, A. 1994. Effective pairwise interactions in Ni3V, in metallic alloys: Experimental and theoretical perspectives. (J.S. Faulkner and R. Jordon, eds.). NATO-ASI Series Vol. 256, page 215. Kluwer Academic Publishers, Boston.

Gonze, X. 1997. First-principles responses of solids to atomic displacements and homogeneous electric fields: Implementation of a conjugate-gradient algorithm. *Phys. Rev. B* 55:10337.

Györffy, B. L., Johnson, D. D., Pinski, F. J., Nicholson, D. M., and Stocks, G. M. 1989. The electronic structure and state of compositional order in metallic alloys (G. M. Stocks and A. Gonis, eds.) NATO-ASI Series Vol. 163: Alloy Phase Stability. Kluwer Academic Publishers, Boston.

Györffy, B. L. and Stocks, G. M. 1983. Concentration waves and fermi surfaces in random metallic alloys. *Phys. Rev. Lett*. 50:374.

Hashimoto, S., Iwasaki, H., Ohshima, K., Harada, J., Sakata, M., and Terauchi, H. 1985. Study of local atomic order in a ternary $Cu_{47}Ni_{29}Zn_{24}$ alloy using anomalous scattering of synchrotron radiation. *J. Phys. Soc. Jpn*. 54:3796.

Hou, D. H., Jones, I. P., and Fraser, H. L. 1997. The ordering tie-line method for sublattice occupancy in intermetallic compounds. *Philos. Mag. A* 74:741.

Hume-Rothery, W., 1963. Electrons, Atoms, Metals, and Alloys, 3rd ed. Dover, New York.

Inoue, A., Zhang, T., and Matsumoto, T. 1990. Zr-Al-Ni amorphous alloys with high glass transition temperatures and significant supercooled liquid region. *Mater. Trans. JIM* 31:177.

Johnson, D. D. 1999. Atomic short-range order precursors to Heusler phase in disordered BCC ternary alloys. To be submitted for publication.

Johnson, D. D. 2001. An electronic-structure-based theory of atomic short-range order and phase stability in multicomponent alloys: With application to $CuAuZn_2$. *Phys. Rev. B*. Submitted.

Johnson, D. D. and Asta, M. 1997. Energetics of homogeneously-random fcc Al-Ag alloys: A detailed comparison of computational methods. *Comput. Mater. Sci*. 8:54.

Johnson, D. D., Asta, M. D., and Althoff, J. D. 1999. Temperature-dependent chemical ordering in bcc-based ternary alloys: a theoretical study in Ti-Al-Nb. *Philos. Mag. Lett*. 79:551.

Johnson, D. D., Nicholson, D. M., Pinski, F. J., Stocks, G. M., and Györffy, B.L. 1986. Density-functional theory for random alloys: Total energy within the coherent-potential approximation. *Phys. Rev. Lett*. 56:2088.

Johnson, D. D., Nicholson, D. M., Pinski, F. J., Stocks, G. M., and Györffy, B.L. 1990. Total energy and pressure calculations for random substitutional alloys. *Phys. Rev. B* 41:9701.

Johnson, D. D. and Pinski, F. J. 1993. Inclusion of charge correlations in the calculation of the energetics and electronic structure for random substitutional alloys. *Phys. Rev. B* 48:11553.

Johnson, D. D., Staunton, J. B., and Pinski, F. J. 1994. First-principles all-electron theory of atomic short-range ordering in metallic alloys: DO_{22}- versus $L1_2$-like correlations. *Phys. Rev. B* 50:1473.

Khachaturyan, A. G. 1972. Atomic structure of ordered phases: Stability with respect to formation of antiphase domains. *Zh. Eksp. Teor. Fiz*. 63:1421.

Khachaturyan, A. G. 1983. Theory of structural transformations in solids. John Wiley & Sons, New York.

Kohn, W. and Rostaker, N, 1954. *Phys. Rev*. 94:1111.

Kohn, W. and Sham, L. J. 1965. Self-consistent equations including exchange and correlation effects. *Phys. Rev*. 140:A1133.

Korringa, J. 1947. On the calculation of the energy of a Bloch wave in a metal. *Physica* 13:392.

Korzhavyi, P. A., Ruban, A. V., Abrikosov, I. A., and Skriver, H. L. 1995. Madelung energy for random metallic alloys in the coherent potential approximation. *Phys. Rev*. 51:5773.

Krivoglaz, M., 1969. Theory of x-ray and thermal-neutron scattering by real crystals. Plenum, New York.

Landau, L. D. 1937a. Theory of phase transformations I. *Phys. Zeits. d. Sowjetunion* 11:26.

Landau, L. D. 1937b. Theory of phase transformations II, *Phys. Zeits. d. Sowjetunion*, 11:545.

Landau, L. D. and Lifshitz, E. M. 1980. Statistical Physics, 3rd ed. Pergamon Press, New York.

Lankford, W. T. Jr., Samways, N., Craven, R., and McGannon, H. (eds.), 1985. The Making, Shaping and Forming of Steels. AISE, Herbick and Hild, Pittsburgh.

Le Bolloc'h, D., Caudron, R., and Finel, A. 1998. Experimental and theoretical study of the temperature and concentration dependence of the short-range order in Pt-V alloys. *Phys. Rev. B* 57:2801.

Le Bolloc'h, D., Cren, T., Caudron, R., and Finel, A. 1997. Concentration variation of the effective pair interactions measured on the Pt-V system. Evaluation of the gamma-expansion method. *Comp. Mater. Sci*. 8:24.

Lifshitz, E. M. 1941. A theory of phase transition of 2nd kind. I. Measurement of the elementary crystal unit cells during the phase transition of the 2nd kind. *Zh. Eksp. Teor. Fiz*. 11:255.

Lifshitz, E. M. 1942. (Title unavailable). Akad. Nauk SSSR Izvestiia Seriia Fiz. 7:251.

Lifshitz, I. M. 1960. High-pressure anomalies of electron properties of a metal. *Zh. Eksp. Teor. Fiz*. 38:1569.

Ling, M. F., Stuanton, J. B., and Johnson, D. D. 1995a. All-electron, linear-response theory of local environment effects in magnetic, metallic alloys and multilayers. *J. Phys*. (*Condensed Matter*) 7:1863.

Ling, M. F., Staunton, J. B., Johnson, D. D., and Pinski, F. J. 1995b. Origin of the (1 1/2 0) atomic short-range order in Au-rich Au-Fe alloys. *Phys. Rev. B* 52:R3816.

Lu, Z. W., Wei, S.-H., and Zunger, A. 1991. Long-range order in binary late-transition-metal alloys. *Phys. Rev. Lett*. 66:1753.

Lu, Z. W., Wei, S.-H., and Zunger, A. 1993. Relativistically-induced ordering and phase separation in intermetallic compounds. *Europhys. Lett*. 21:221.

Magri, R., Wei, S.-H., and Zunger, A. 1990. Ground-state structures and the random-state energy of the Madelung lattice. *Phys. Rev. B* 42:11388.

Masanskii, I. V., Tokar, V. I., and Grishchenko, T. A. 1991. Pair interactions in alloys evaluated from diffuse-scattering data. *Phys. Rev. B* 44:4647.

Massalski, T. B., Okamoto, H., Subramanian, P. R., and Kacprzak, L. 1990. Binary Alloy Phase Diagrams, 2nd ed. ASM International, Materials Park, Ohio.

McCormack, R., Asta, M., Hoyt, J. J., Chakoumakos, B. C., Misture, S. T., Althoff, J. D., and Johnson, D. D. 1997. Experimental and theoretical investigation of order-disorder in Cu_2AlMn. *Comp. Mater. Sci*. 8:39.

Mohri, T., Sanchez, J. M., and de Fontaine, D. 1985. Short-range order diffuse intensity calculations in the cluster variation method. *Acta Metall*. 33:1463.

Moss, S. C. 1969. Imaging the Fermi surface through diffraction scattering from a concentrated disordered alloy. *Phys. Rev. Lett*. 22:1108.

Mott, N. 1937. The energy of the superlatiice in β brass. *Proc. Phys. Soc. Lond*. 49:258.

Nicholson, D. M. and Brown, R. H. 1993. Electrical resistivity of $Ni_{0.8}Mo_{0.2}$: Explanation of anomalous behavior in short-range ordered alloys. *Phys. Rev. Lett*. 21:3311.

Oates, W. A., Wenzl, H., and Mohri, T. 1996. On putting more physics into CALPHAD solution models. *CALPHAD, Comput. Coupling Phase Diagr. Thermochem. (UK). 20:37*.

Onsager, L. 1936. Electric moments of molecules in liquids. *J. Am. Chem. Soc*. 58:1486.

Ornstein, L. S. 1912. Accidental deviations of density in mixtures. *K. Akad. Amsterdam* 15:54.

Ornstein, L. S. and Zernike, F. 1914. Accidental deviation of density and opalescence at the critical point. *K. Akad. Amsterdam* 17:793.

Ornstein, L. S. and Zernike, F. 1918. The linear dimensions of density variations. *Phys. Z*. 19:134.

Pavone, P., Bauer, R., Karch, K., Schuett, O., Vent, S., Windl, W., Strauch, D., Baroni, S., and De Gironcoli, S. 1996. *Ab initio* phonon calculations in solids. *Physica B* 219–220:439.

Paxton, A. T., Methfessel, M., and Pettifor, D. G. 1997. A band-structure view of the Hume-Rothery electron phases. *Proc. R. Soc. Lond. A* 453:1493.

Peker, A. and Johnson, W. L. 1993. A highly processable metallic glass: $Zr_{41.2}Ti_{13.8}Cu_{12.5}$ $Ni_{10.0}Be_{22.5}$. *Appl. Phys. Lett*. 63:2342.

Pierron-Bohnes, V., Kentzinger, E., Cadeville, M. C., Sanchez, J. M., Caudron, R., Solal, F., and Kozubski, R. 1995. Experimental determination of pair interactions in a $Fe_{0.804}V_{0.196}$ single crystal. *Phys. Rev. B* 51:5760.

Pinski, F. J., Ginatempo, B., Johnson, D. D., Staunton, J. B., and Stocks, G. M., and Györffy, B. L. 1991. Origins of compositional order in NiPt. *Phys. Rev. Lett*. 66:766.

Pinski, F. J., Ginatempo, B., Johnson, D. D., Staunton, J. B., Stocks, G. M. 1992. Reply to comment. *Phys. Rev. Lett*. 68: 1962.

Pinski, F. J., Staunton, J. B., and Johnson, D. D. 1998. Charge-correlation effects in calculations of atomic short-range order in metallic alloys. *Phys. Rev. B* 57:15177.

Quong, A. A. and Liu. A. Y. 1997. First-principles calculations of the thermal expansion of metals. *Phys. Rev. B* 56:7767.

Reichert, H., Moss, S. C., and Liang, K. S. 1996. Anomalous temperature dependence of x-ray diffuse scattering in Cu_3Au. *Phys. Rev. Lett.* 77:4382.

Reichert, H., Tsatskis, I., and Moss, S. C. 1997. Temperature dependent microstructure of Cu_3Au in the disordered phase. *Comp. Mater. Sci.* 8:46.

Reinhard, L. and Moss, S. C. 1993. Recent studies of short-range order in alloys: The Cowley theory revisited. *Ultramicroscopy* 52:223.

Reinhard, L., Schönfeld, B., Kostorz, G., and Bührer, 1990. Short-range order in α-brass. *Phys. Rev. B* 41:1727.

Ruban, A. V., Abrikosov, I. A., and Skriver, H. L. 1995. Ground-state properties of ordered, partially ordered, and random Cu-Au and Ni-Pt alloys. *Phys. Rev. B* 51:12958.

Rubin, G. and Finel, A. 1995. Application of first-principles methods to binary and ternary alloy phase diagram predictions. *J. Phys. (Condensed) Matter* 7:3139.

Sanchez, J. M. and de Fontaine, D. 1978. The fcc Ising model in the cluster variation method. *Phys. Rev. B* 17:2926.

Sanchez, J. M. and de Fontaine, D. 1980. Ordering in fcc lattices with first- and second-neighbor interactions. *Phys. Rev. B* 21:216.

Sato, H. and Toth, R. S. 1962. Long period superlattices in alloys. II. *Phys. Rev.* 127:469.

Saunders, N. 1996. When is a compound energy not a compound energy? A critique of the 2-sublattice order/disorder model, CALPHAD. *Comput. Coupling Phase Diagr. Thermochem.* 20:491.

Soven, P. 1967. Coherent-potential model of substitutional disordered alloys. *Phys. Rev. B* 156:809.

Staunton, J. B., Johnson, D. D., and Pinski, F. J. 1990. Theory of compositional and magnetic correlations in alloys: Interpretation of a diffuse neutron-scattering experiment on an iron-vanadium single crystal. *Phys. Rev. Lett*. 65:1259.

Staunton, J. B., Johnson, D. D., and Pinski, F. J. 1994. Compositional short-range ordering in metallic alloys: Band-filling, charge-transfer, and size effects from a first-principles, all-electron, Landau-type theory. *Phys. Rev. B* 50:1450.

Stell, G. 1969. Some critical properties of the Ornstein-Zernike system. *Phys. Rev*. 184:135.

Taylor, D. W. 1968. Vibrational properties of imperfect crystals with large defect concentrations. *Phys. Rev*. 156:1017.

Thomas, H. 1972. Über den elektrischen Widerstand von Kupfer-Nickel-Zink-Legierungen und den Einfluß einer Tieftemperatur-Verformung. *Z. Metallk*. 63:106.

Tokar, V. I. 1985. A new series expansion for lattice statistics. *Phys. Lett*. 110A:453.

Tokar, V. I. 1997. A new cluster method in lattice statistics. *Comp. Mat. Sci*. 8:8.

Treglia, G. and Ducastelle, F. 1987. Is ordering in PtNi alloys induced by spin-orbit interactions. *J. Phys. F* 17:1935.

Vaks, V. G., Larkin, A. I., and Likin, S. A. 1966. Self-consistent method for the description of phase transitions. *Zh. Eksp. Teor. Fiz*. 51:361.

van der Wegen, G. J. L., De Rooy, A., Bronsveld, P. M., and De Hosson, J. Th. M. 1981. The order/disorder transition in the quasibinary cross section $Cu_{50}Ni_{50-x}Zn_x$. *Scr. Met*. 15:1359.

van Hove, L. 1953. The occurrence of singularities in the elastic frequency distribution of a crystal. *Phys. Rev*. 89:1189.

Wolverton, C. and de Fontaine, D. 1994. Cluster expansions of alloy energetics in ternary intermetallics. *Phys. Rev. B* 49: 8627.

Wolverton, C. and Zunger, A. 1995a. Ising-like description of structurally relaxed ordered and disordered alloys. *Phys. Rev. Lett*. 75:3162–3165.

Wolverton, C. and Zunger, A. 1995b. Short-range order in a binary Madelung Lattice. *Phys. Rev. B* 51:6876.

Wolverton, C. and Zunger, A. 1997. Ni-Au: A testing ground for theories of phase stability. *Comp. Mater. Sci*. 8:107.

Wolverton, C., Zunger, A., and Schonfeld, B. 1997. Invertible and non-invertible alloy Ising problems. *Solid State Commun*. 101:519.

Yu, R. and Krakauer, H. 1994. Linear-response calculations within the linearized augmented plane-wave method. *Phys. Rev. B* 49:4467.

Zernike, F. 1940. The propogation of order in co-operative phenomena. Part I: The AB case. *Physica* 7:565.

KEY REFERENCES

Althoff et al., 1995. See above.

Althoff et al., 1996. See above.

Provides a detailed explanation for the electronic origins of ASRO in fcc Cu-Ni-Zn. The 1996 paper details all contributing binaries, along with several ternaries, along with an appendix that describes how to determine the polarization of the ternary concentration waves from the diffuse scattering intensities, as applied to Cu_2NiZn.

Asta and Johnson, 1997. See above.

Johnson and Asta, 1997. See above.

These references provide a proper comparison of complimentary methods briefly discussed in the text, and show that when done carefully the methods agree. Moreover, the first paper shows how this information may be used to calculate the equilibrium (or metastable) phase diagram of an alloy with very good agreement to the assessed phase diagram, as well as how calculations help interpretation when there is contradictory experimental data.

Brout and Thomas, 1967. See above.

Some original details on connection of Onsager corrections and mean-spherical models in simple model Hamiltonians.

Clark et al., 1995. See above.

Application of the present approach, which was the first theory to explain the $L1_1$ *ordering in CuPt. Furthermore, it detailed how electronic van Hove singularities play the key role in producing such a unique ordering in CuPt.*

Evans, 1979. See above.

Provides a very complete reference for classical density-functional theory as applied to liquids, but which is the basis for the connection between electronic and classical DFT as performed here.

Györffy and Stocks, 1983. See above.

The first paper to detail the conceptual framework for the first-principles, electronic DFT for calculating ASRO within a classical DFT, and also to show how the concentration effects on the random alloy Fermi surface explains the concentration-dependent shifts in diffuse scattering peaks, with application to Cu-Pd alloys.

Ling et al., 1995b. See above.

Details how magnetism and chemical correlations are intimately connected, and how the quenching rate, which determined the magnetic state, affects the ASRO observed, which may be altered via magnetic annealing.

Staunton et al., 1990. See above.

First application of the first-principles concentration wave technique applied to a magnetic binary system and details how useful such calculations may be in explaining the origins of the ASRO features in the experimental data. Also details how quenching a sample can play an important role in establishing what type of exchange-split electronic structure (e.g., ferromagnetic and disordered paramagetic) gives rise to the chemical fluctuations.

Staunton et al., 1994. See above.

Johnson et al., 1994. See above.

First complete details, and applications in several alloy systems, of the fully self-consistent, all-electron, mean-field, density-functional-based theory for calculating ASRO for binary metallic alloys, which includes band, charge, and dielectric effects, along with Onsager corrections.

DUANE D. JOHNSON
FRANK J. PINSKI
JULIE B. STAUNTON
University of Illionis
Urbana, Illinois

This page intentionally left blank

MECHANICAL TESTING

INTRODUCTION

The mechanical behavior of materials is concerned primarily with the response of materials to forces or loads. This behavior ultimately governs the usefulness of materials in a variety of applications from automotive and jet engines to skyscrapers, as well as to more common products such as scissors, electric shavers, coffee mugs, and cooking utensils. The forces or loads that these materials experience in their respective applications make it necessary to identify the limiting values that can be withstood without failure or permanent deformation. Indeed, in many cases it is necessary to know not only the response of the materials to an applied load, but also to be able to predict their behavior under repeated loading and unloading. In other applications, it is also necessary to determine the time-dependent behavior and routine wear of materials under the applied loads and under operating conditions.

Knowledge of the mechanical behavior of materials is also necessary during manufacturing processes. For example, it is often necessary to know the values of temperature and loading rates that minimize the forces necessary during mechanical forming and shaping of components. Determination of the resulting microstructures during shaping and forming is an integral part of these operations. This is an area that clearly combines the determination of mechanical properties with that of the microstructure. Of course, the atomistic concept of flow and materials failure is an integral part of the determination of mechanical properties.

Basic mechanical property and materials strength measurements are obtained by standardized mechanical tests, many of which are described in this chapter. Each unit in this chapter not only covers details, principles, and practical aspects of the various tests, but also provides comprehensive reference to the standard procedures and sample geometries dictated by the ASTM or other standards agencies. This chapter also covers additional mechanical testing techniques such as high-strain-rate testing, measurements under pressure, and tribological and wear testing. As such, the chapter covers a broad spectrum of techniques used to assess materials behavior in a variety of engineering applications.

Whereas mechanical testing has a long history and is largely a mature field, methods continue to take advantage of advances in instrumentation and in fundamental understanding of mechanical behavior, particularly in more complex systems. Indeed, the first category of materials one thinks of that benefits from knowledge and control of mechanical properties is that of structural materials, i.e., substantial components of structures and machines. However, these properties are also crucial to the most technologically sophisticated applications in, for example, the highest-density integrated electronics and atom-by-atom deposited multilayers. In this respect, mechanical testing remains an active field at the forefront of modern technology.

REZA ABBASCHIAN

TENSION TESTING

INTRODUCTION

Of all mechanical properties, perhaps the most fundamental are related to what happens when a material is subjected to simple uniaxial tension. In its essence, a tensile test is carried out by attaching a specimen to a load-measuring device, applying a load (or imposing a given deformation), and measuring the load and corresponding deformation. A schematic of a specimen and test machine is provided in Figure 1 (Schaffer et al., 1999).

The result obtained from a tensile test is a so-called stress/strain curve, a plot of stress (force/unit area) versus strain (change in length/original length), illustrated schematically in Figures 1 and 2 (Dieter, 1985). The results of such a test (along with the results of other tests, to be sure) are basic to determination of the suitability of a given material for a particular load-bearing application. In this regard the results obtained from such a test are of great engineering significance. Tensile-test results also provide a great deal of information relative to the fundamental mechanisms of deformation that occur in the specimen. Coupled with microscopic examination, tensile test results are used to develop theories of hardening and to develop new alloys with improved properties. Tensile tests can be used to obtain information on the following types of property:

Elastic Properties. These are essentially those constants that relate stresses to strains in the (usually) small strain regime where deformation is reversible. This is the linear region in Figures 1B and 2. Deformation is said to be reversible when a specimen or component that has been subjected to tension returns to its original dimensions after the load is removed. Young's modulus (the slope of the linear, reversible portion of Fig. 1B) and Poisson's ratio (the ratio of strain in the loading direction to strain in the transverse direction) are typical examples of elastic properties.

Plastic Properties. Plastic properties are those that describe the relationships between stresses and strains when the deformation is large enough to be irreversible. Typical plastic properties are the yield stress (the stress at which deformation become permanent, denoted by the symbol σ_{ys}), the extent of hardening with deformation (referred to as strain hardening), the maximum in the

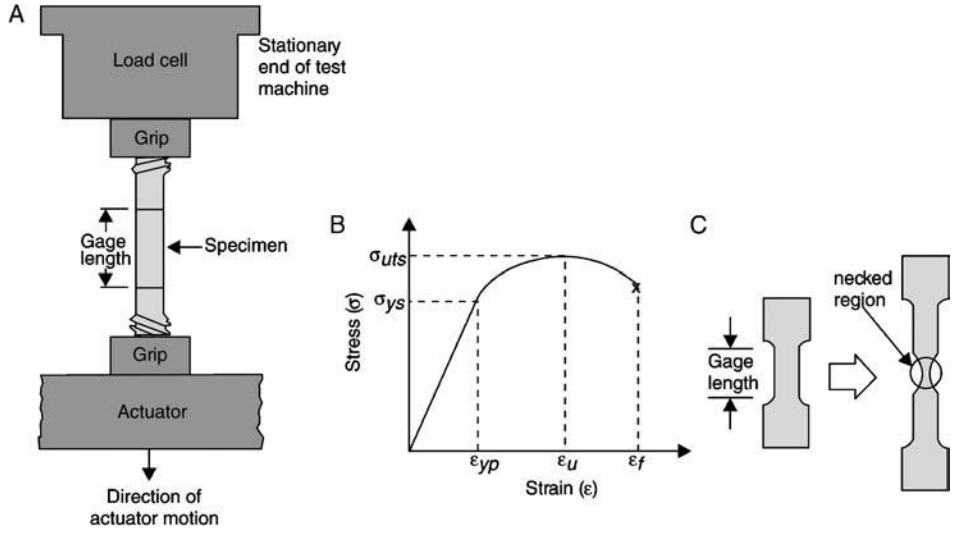


Figure 1. Schematic of specimen attached to testing machine (Schaffer et al., 1999).

stress/strain plot (the ultimate tensile strength, denoted σ_{uts}), the total elongation, and the percent reduction in area. These various quantities are illustrated in Figure 2 and discussed in more detail below.

Indication of the Material's Toughness. In a simplified view, toughness is the ability of a material to absorb energy in being brought to fracture. Intuitively, toughness is manifested by the absorption of mechanical work and is related to the area under the curve of stress versus strain, as shown in Figure 3 (Schaffer et al., 1999).

Tensile testing is carried out at different temperatures, loading rates, and environments; the results of these tests are widely used in both engineering applications and scientific studies.

In the following sections, basic principles are developed, fundamentals of tensile testing are pointed out, and references are provided to the detailed techniques employed in tensile testing.

Competitive and Related Techniques

In addition to the mechanical tests described in the following sections, information can be obtained about elastic properties through vibrational analysis, and information about plastic properties (e.g., tensile strength) may be obtained from microhardness testing (HARDNESS TESTING). While such information is limited, it can be obtained quickly and inexpensively.

PRINCIPLES OF THE METHOD

Analysis of Stress/Strain Curves

The stress/strain curve will typically contain the following distinct regions.

1. An initial linear portion in which the deformation is uniform and reversible, meaning that the specimen

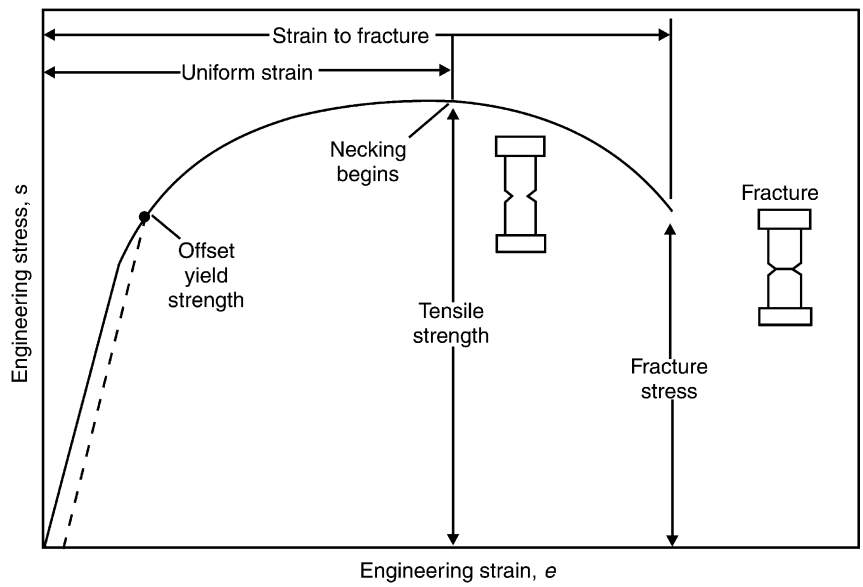


Figure 2. Schematic stress/strain curve for metallic material (Dieter, 1985).

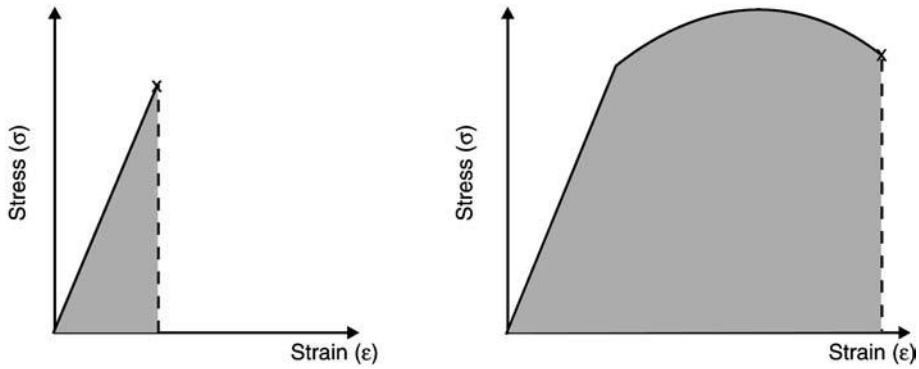


Figure 3. Area under stress/strain curve for (left) brittle and (right) ductile materials (Schaffer et al., 1999).

comes back to its original shape when the load is released. Such behavior is referred to as elastic and is seen in the initial straight-line regions of Figures 1 and 2.

2. A region of rising load in which the deformation is uniform and permanent (except, of course, for the elastic component), as illustrated in Figure 1.
3. A region in which the deformation is permanent and concentrated in a small localized region or "neck." The region of nonuniform deformation is indicated in Figure 1, and necking is illustrated schematically in Figure 4 (Dieter, 1985).

These regions are discussed in the sections that follow.

Elastic Deformation

Under the application of a specified external load, atoms (or molecules) are displaced by a small amount from their equilibrium positions. This displacement results in an increase in the internal energy and a corresponding force, which usually varies linearly for small displacements from equilibrium. This initial linear variation (which is reversible upon release of the load) obeys what is known as Hooke's law when expressed in terms of force and displacement:

$$\mathbf{F} = k\mathbf{x} \tag{1}$$

where k is the Hooke's law constant, $\mathbf{x}$ = displacement, and $\mathbf{F}$ is the applied force.

In materials testing, Hooke's law is more frequently expressed as a relationship between stress and strain:

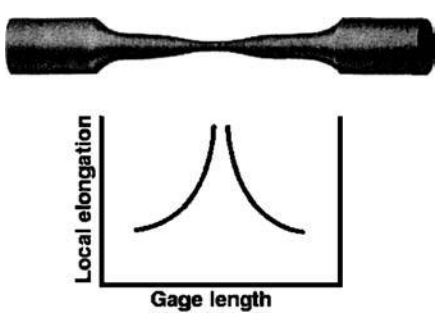


Figure 4. Illustration of necking in a metallic specimen and the local variation in the strain (Dieter, 1985).

$$\sigma = E \cdot \varepsilon \tag{2}$$

where σ is stress (in units of force/area normal to force), ε = strain (displacement/initial length), and E = Young's modulus.

Definitions of Stress and Strain

While stress always has units of force/area, there are two ways in which stress may be calculated. The true stress, usually represented by the symbol σ, is the force divided by the *instantaneous* area. The engineering stress is usually represented by the symbol S and is the force divided by the *original* area. Since the cross-sectional area decreases as the specimen elongates, the true stress is always larger than the engineering stress. The relationship between the true stress and the engineering stress is easily shown to be:

$$\sigma = S(1 + e) \tag{3}$$

where e is the *engineering* strain.

The engineering strain is defined as displacement (Δl) divided by the original length (l_0) and is denoted by e. That is:

$$e = \frac{\Delta l}{l_0} \tag{4}$$

The true strain is given by

$$\varepsilon = \int_{l_0}^{l} \frac{dl}{l} = \ln\left(\frac{i}{l_0}\right) \tag{5}$$

where l is the instantaneous length. This term is simply the sum of all of the instantaneous strains. The true strain and engineering strain are related by the equation:

$$\varepsilon = \ln(1 + e) \tag{6}$$

There is not a significant difference in the true strain and engineering strain until the engineering strain reaches ~10%. The difference between the conventional engineering stress/strain curve and the true stress/strain curve is illustrated in Figure 5 (from Schaffer et al., 1999).

Young's modulus is an indicator of the strength of the interatomic (or intermolecular) bonding and is related to the curvature and depth of the energy-versus-position

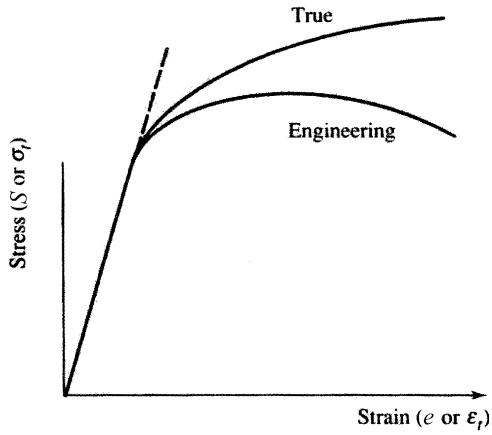


Figure 5. Illustration of the difference between the engineering and true stress/strain curves (Schaffer et al., 1999).

curve. Young's modulus may be obtained from the initial portion of the stress/strain curves shown in Figures 1 and 2.

Uniform Plastic Deformation

After the initial linear portion of the stress-strain curve (obtained from the load deflection curve), metallic and polymeric materials will begin to exhibit permanent deformation (i.e., when the load is released the object being loaded does not return to its initial no-load position). For most metals, the load required to continue deformation will rise continually up to some strain level. Throughout this regime, all deformation is uniform, as illustrated in Figures 1 and 2.

In some materials, such as low-carbon steels that have been annealed, there is an initial post-yield increment of nonuniform deformation, which is termed Luder's strain. Luder's strain is caused by bands of plastic activity passing along the gage length of the specimen, and is associated with dynamic interactions between dislocations (the defect responsible for plastic deformation) and carbon atoms. This is shown in Figure 6 (Dieter, 1985). In Figure 6, the stress at which there is a dropoff is referred to as the upper yield point and the plateau stress immediately following this drop is called the lower yield point. Such behavior is

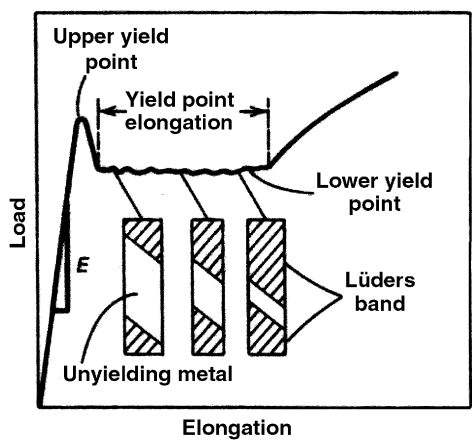


Figure 6. Illustration of the upper and lower yield points and Lüders strain in a mild steel tensile curve (Dieter, 1985).

predominantly limited to annealed steels with low carbon content.

Stress-strain curves for many metals in the region of uniform strain are well described by the following equation:

$$\sigma = K\varepsilon^n \tag{7}$$

where K is strength coefficient and n is strain-hardening exponent.

The strength coefficient and strain-hardening coefficients can be obtained by representing the stress/strain curve on a log/log basis:

$$\log\sigma = n\log\varepsilon + \log K \tag{8}$$

The slope of Equation 8 is the strain-hardening exponent. The value of K is simply obtained by noting that it is the value of the stress at a strain of 1.

Nonuniform Plastic Deformation

At some point, the increased strength of the material due to strain hardening is no longer able to balance the decreased cross-sectional area due to deformation. Thus a maximum in the load/displacement or stress/strain curve is reached. At this point $dP = 0$ (i.e., zero slope, which implies that elongation occurs with no increment in the load P) and if Equation 7 is obeyed, the strain value is given by:

$$\varepsilon_n = n \tag{9}$$

In Equation 9 the subscript n refers to "necking," by which it is meant that at this point all subsequent deformation is concentrated in a local region called a "neck" (Fig. 4) and is nonuniform. This nonuniformity is usually referred to as "plastic instability." Independent of the form of the stress/strain curve, the onset of plastic instability (assuming that deformation takes place at constant volume) occurs when the following condition is achieved:

$$\frac{d\sigma}{e} = \frac{\sigma}{1+e} \tag{10}$$

Equation 10 is the basis of the so-called Considere construction, which may be used to find the onset of plastic instability. This is done by projecting a tangent to the stress/strain curve from the point $(-1,0)$. If this is done, Equation 10 is obviously satisfied and the point of tangency represents the necking strain. It is important to realize that in this case the term σ is the true stress and e is the engineering strain. The Considere construction is illustrated in Figure 7 (Dieter, 1985).

Factors Affecting the Form of the Stress/Strain Curve

The stress/strain curve that is obtained depends upon the following variables:

The material and its microstructure;

The test temperature;

The testing rate;

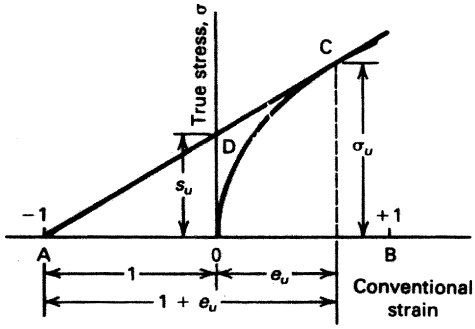


Figure 7. Considere's construction for determining the onset of necking (Dieter, 1985). Here, "u" refers to the point of necking.

The geometry of the test specimen;

The characteristics of the testing machine (as discussed above);

The mode in which the test is carried out.

By the "mode" is meant whether the test is carried out by controlling the rate of load application (load control), the rate of machine displacement (displacement control), or the rate of specimen strain (strain control).

Of these six factors, the first three may be considered intrinsic factors. The effects of the last three (discussed in Practical Aspects and Method Automation) tend to be less appreciated, but they are no less important. Indeed, some results that are considered to be fundamental to the material are largely influenced by the latter three variables, which may be considered to be extrinsic to the actual material. The first three factors are discussed below.

Material and Microstructure

The material and microstructure plays a critical role on the form of the stress/strain curve. This is illustrated in Figure 8 (Schaffer, 1999) for three different materials. The curve labeled Material I would be typical of a very brittle material such as a ceramic, a white cast iron, or a high-carbon martensitic steel. The curve labeled Material II is representative of structural steels, low-strength aluminum alloys, and copper alloys, for example. The curve labeled Material III is representative of polymeric materials.

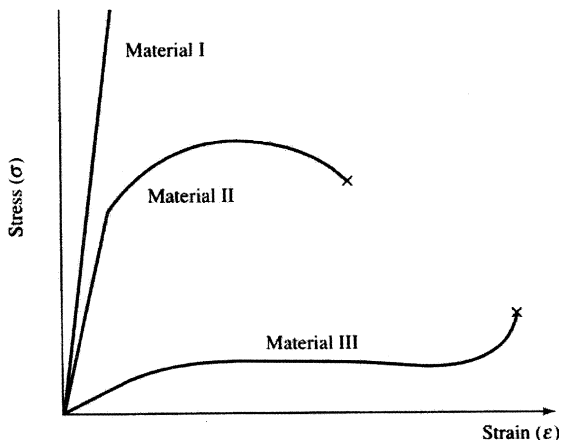


Figure 8. Schematic illustration of stress/strain curves for a brittle material (I), a ductile metal (II), and a polymer (III) (Schaffer et al., 1999).

The details of the microstructure are also very important. For example, if a given type of metal contains precipitates it is likely to be harder and to show less extension than if it does not contain precipitates. Precipitates are typically formed as a result of a specific heat treatment, which in effect is the imposition of a time/temperature cycle on a metal. In the heat treatment of steel, the formation of martensite is frequently a necessary intermediate phase. Martensite is very strong but is intrinsically brittle, meaning it has little or no ductility. It will have a stress-strain curve similar to that of Material I. If the martensitic material is heated to a temperature in the range of 200°C, precipitates will form and the crystal structure will change to one that is intrinsically more ductile; a stress/strain curve similar to Material II will then be obtained. Grain size also has a significant effect. Fine-grained materials both are stronger and exhibit more ductility than coarse-grained materials of the same chemistry. The degree of prior cold working is also important. Materials that are cold worked will show higher yield strengths and less ductility, since cold working produces defects in the crystalline structure that impede the basic mechanisms of plastic deformation.

Effects of Temperature and Strain Rate

Since deformation is usually assisted by thermal energy, the stress/strain curve at high temperature will usually lie below that at low temperature. Related to the effects of thermal energy, stress/strain curves at high rate will usually lie above those obtained at low rates, since at high rate more of the energy must be supplied athermally. The stress as a junction of strain rate for a given strain and temperature is usually expressed through the simplified equation:

$$\sigma_{\varepsilon,T} = C\dot{\varepsilon}^m \tag{11}$$

where C and m are material constants and $\dot{\varepsilon}$ is the strain rate.

Yield-Point Phenomena

In some materials, such as mild steels, a load drop is observed at the onset of plastic deformation, as shown in Figures 6 and 9B. This is easily understood in terms of some elementary concepts in the physics of deformation.

There is a fundamental relationship between the strain rate, the dislocation density, and the velocity of dislocation movement. (Dislocations are defects in the material that are responsible for the observed deformation behavior.) The strain rate is given by:

$$\dot{\varepsilon} = b\rho v \tag{12}$$

where b = Burger's vector (presumed to be constant), v = velocity of dislocation movement, and ρ is density of dislocations.

At the yield point, there is a rapid increase in the density of mobile dislocations. Since b is constant, this means

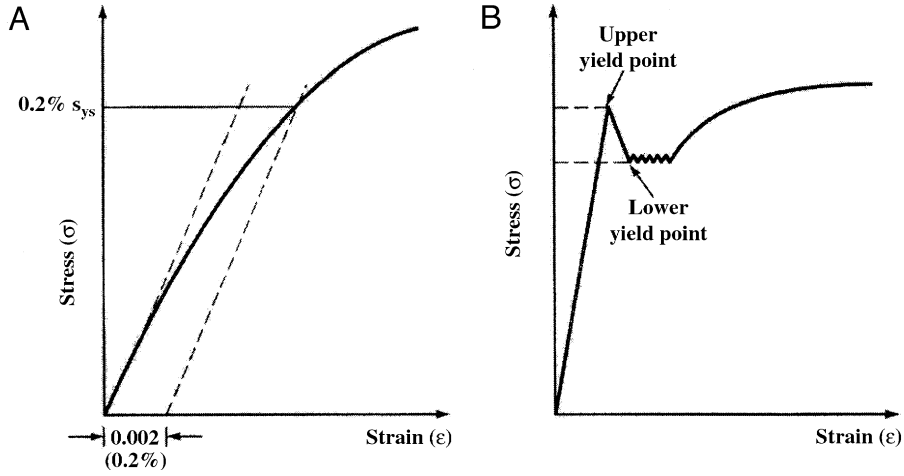


Figure 9. Illustration of stress/strain curves (**A**) for typical metals and (**B**) for mild steel in which there is an upper and lower yield point. Determination of the 0.2% offset yield is also shown in (**A**).

that v must decrease if a constant strain rate is being maintained. However, there is a relationship between dislocation velocity and stress that is expressed as:

$$\frac{v}{v_0} = \left(\frac{\sigma}{\sigma_0}\right)^n \tag{13}$$

where $\mathbf{V}_0$ and σ_0 are constants and n is an exponent on the order of 30 or 40 depending on the material. Now if **v** decreases, then Equation 13 shows that the stress σ will decrease, resulting in a load drop. However, if the control mode is a given rate of load application (i.e., load control), then no load drop will be observed since when the density of mobile dislocations increases, the machine will simply move faster in order to maintain the constant rate of loading.

PRACTICAL ASPECTS OF THE METHOD

The Basic Tensile Test

The basic tensile test consists of placing a specimen in a test frame, loading the specimen under specified conditions, and measuring the loads and corresponding displacements. The fundamental result of the tensile test is a curve relating the simultaneous loads and displacements. This curve is called the load displacement record. The load displacement record is converted to a stress/strain curve by dividing the load by the cross-sectional area and the elongation by the gage-length of the specimen as discussed in the following sections. To carry out a tensile test, four essential components are required. They are:

- A specimen of appropriate geometry;
- A machine to apply and measure the load;
- A device to measure the extension of the specimen;
- An instrument to record the simultaneous values of load and displacement.

Testing may be done at high or low temperatures and in a variety of environments other than air. Accessories are required to carry out such tests, which would include a high-temperature furnace, an extensometer adapted for use with such a furnace, and possibly a chamber for testing in inert or aggressive environments or under vacuum.

Typically, the following information can be obtained from the load displacement record of a tensile test or from the specimen:

- Young's modulus, E;
- The yield strength, σ_{ys}—or, more conventionally, the 0.2% offset yield as defined in Figure 9A (Schaffer et al., 1999);
- The strength coefficient and strain hardening exponent (Eq. 7);
- The ultimate tensile strength ($\sigma_{uts} = P_{max}/A_0$) as defined in Figures 1B and 2(P_{max} is maximum load);
- The total strain to failure, $\varepsilon_f = (l_f - l_0)/l_0$;

The percent reduction in area from the specimen, $\%RA=(A_0 - A_f)/A_0$ (where A_0 and A_f are the initial and final areas, respectively).

The ASTM has developed a standard test procedure for tensile testing (ASTM, 1987, E 8 & E 8M) that provides details on all aspects of this test procedure. The reader is strongly encouraged to consult the following documents related to tensile testing, calibration, and analysis of test results (all of which can be found in ASTM, 1987):

- Designation E 8: Standard Methods of Tension Testing of Metallic Materials (pp. 176–198).
- Designation E 8M: Standard Methods of Tension Testing of Metallic Materials [Metric] (pp. 199–222).
- Designation E 4: Standard Practices for Load Verification of Testing Machines (pp. 119–126).
- Designation E 83: Standard Practice for Verification and Classification of Extensometers (pp. 368–375).
- Designation E 1012: Standard Practice for Verification of Specimen Alignment Under Tensile Loading (pp. 1070–1078).

Designation E 21: Standard Recommended Practice for Elevated Temperature Tests of Metallic Materials (pp. 272–281).

Designation E 74: Standard Practices of Calibration of Force-Measuring Instruments for Verifying the Load Indication of Testing Machines (pp. 332–340).

Designation E 1 1 1: Standard Test Method for Young's Modulus, Tangent Modulus, and Chord Modulus (pp. 344–402).

The specimens and apparatus used in testing are considered in the sections that follow.

Specimen Geometry

The effect of geometry is also significant. As discussed above, the total elongation is made up of a uniform and a nonuniform component. Since the nonuniform component is localized, its effect on the overall elongation will be less for specimens having a long gage section. This means that the total strain to failure for a given material will be less for a long specimen of given diameter than it will be for a short specimen of the same diameter. Thus care must be exercised when comparing test results to ensure that geometrically similar specimens were used. It is standard in the US to use specimens in which the ratio of the length to diameter is 4:1. Since the total strain depends upon this ratio, comparisons can be made between different specimen sizes provided that this ratio is maintained. A more fundamental measure of the ductility is the percent reduction in area (%RA), which is independent of the specimen diameter. Typical specimen geometries are illustrated in ASTM (1987), E 83.

Test Machines, Extensometers, and Test Machine Characteristics

Test Machine. Testing is carried out using machines of varying degrees of sophistication. In its most basic form, a machine consists of:

a. A load cell attached to an upper support and a linkage to connect the load cell to the specimen;
b. A lower crosshead (or piston) that connects to the specimen via another linkage; and
c. A means to put the lower crosshead or piston into motion and thereby apply a force to the specimen.

The applied forces are transmitted through a load frame, which may consist of two or more columns, and the lower crosshead or piston may be actuated either mechanically or hydraulically. The load cell consists of a heavy metal block onto which strain gages are attached, usually in the configuration of a Wheatstone bridge. As force is applied, the load cell suffers displacement,and this displacement is calibrated to the applied load (ASTM, 1987, E 4). Most modern test machines have sophisticated electronic controls to aid in applying precise load-time or displacement-time profiles to the specimen.

Extensometers. The extension of the specimen must be measured with an extensometer in order to obtain the displacement corresponding to a given load. There are various ways to measure small displacements with great accuracy. One technique that has become very popular is the so-called clip-on gage. This gage consists of two spring arms attached to a small block. Strain gages are attached to both spring arms and are connected to form a Wheatstone bridge similar to the load cell. However, the calibration is done in terms of displacement (ASTM, 1987, E 83), and it is possible to measure very small displacements with great accuracy using this instrument.

While some indication of the specimen deformation can be obtained by monitoring the displacement of the crosshead (or hydraulic ram, depending on the nature of the test machine), it is, of course, preferable to attach an extensometer directly to the specimen. In this way the extension of the specimen is measured unambiguously. Two problems arise if the extension of the specimen is equated to the displacement of the lower crosshead or piston: (1) the gage length is assumed to be the region between the shoulders of the specimen and, more importantly, (2) deflection in the load train occurs to a rather significant degree. Thus, unless the machine stiffness is known and factored into account, the extension of the specimen will be overestimated.

In addition, even if the machine stiffness is accounted for, the rate of straining of the specimen will be variable throughout the plastic region since the proportion of actual machine and specimen deflection changes in a nonlinear way. Given these sources of error it is highly desirable to measure the machine deflection directly using an extensometer. Of course this is not always possible, especially when testing in severe environments and/or at high temperatures. In such cases other techniques must be used to obtain reasonable estimates of the strain in the specimen. Typical extensometers are shown in Martin (1985).

Testing Machine Characteristics and Testing Mode. The characteristics of the testing machine are also very important. If a machine is "soft" (i.e., there is considerable deflection in the load train during testing), then events that would tend to give rise to load drops (such as initial yielding in low-carbon steels, as discussed previously) can be masked by soft machines. In essence, the machine will spring back during the event that would otherwise cause a load drop and tend to maintain, at least to some degree, a constant load. Similarly, testing under conditions of load control would fully mask load drops, while testing under strain control would, by the same token, maximize the observability of load-drop phenomena.

The preceding brief discussion illustrates that due care must be exercised when carrying out a test in order to obtain the maximum amount of information.

Testing at Extreme Temperatures and in Controlled Environments

While it is clearly not possible to consider all possible combinations of test temperature and environment, a few general comments on the subject of testing at low and high

temperatures and in environments other than ambient are in order. These are important in various advanced technology applications. For example, jet and rocket engines operate at temperatures that approach 1500 K, power generation systems operate at temperatures only slightly lower, and cryogenic applications such as refrigeration systems operate well below room temperature. Furthermore, a wide range of temperatures is encountered in the operation of ships and planes.

High-Temperature Testing. Care must be exercised to assure a uniform temperature over the gage length of the specimen. For practical purposes, the temperature should not vary by more than a few degrees over the entire gage length. If the specimen is heated using a resistance furnace, a so-called "chimney effect" can occur if the top of the furnace is open. This occurs because hot air rises, and as it rises in the tube of the furnace, cold air is drawn in, which tends to cool the bottom of the specimen and create an excessive temperature gradient. This effect can be reduced by simply providing shutters at the top of the furnace that just allow the force rods to enter into the heating chamber but that block the remaining open area.

Testing in a resistance furnace is also complicated by the way in which the extensometer is attached to the specimen and by the need to avoid temperature gradients. A small access hole is provided in the furnace and probes of alumina or quartz are attached to the specimen to define the gage length. These arms are then attached to the clip-on gage mentioned previously to measure specimen displacement. In some cases, "divots" are put in the specimen to assure positive positioning of the probe arms. However, this practice is not recommended since the divots themselves may affect the test results, especially the measured ductility. It is possible to adjust the radius of curvature of the probe arms and the tension holding the extensometer to the specimen so that slippage is avoided. A very popular way of heating the specimen is to use an induction generator and coil. By careful design of the coil, a very constant temperature profile can be established and the extensometer can be easily secured to the specimen by putting the probe arms through the coil.

Low-Temperature Testing. Problems similar to those discussed above apply to low-temperature testing. Refrigeration units with circulating gases or fluids can be used, but constant mixing is required to avoid large temperature gradients. Uniform temperatures can be achieved with various mixtures of liquids and dry ice. For example, dry ice and acetone will produce a constant-temperature bath, but the temperature is limited. Extensometry becomes even more difficult than at high temperatures if a fluid bath is used. In such instances, an arrangement in which a cylindrical linearly variable differential transformer (i.e., LVDT-based extensometer) is attached to the specimen may be useful. The body of the LVDT is attached to the bottom of the gage section and the core is attached to the top. As the specimen elongates, a signal is generated that is proportional to the relative displacement of the core and LVDT body.

Environmental Testing. Tensile testing is also carried out in various environments or in vacuum. Since environmental attack is accelerated at high temperatures, such testing is often done at high temperatures. In such instances, an environmental chamber is used that can be evacuated and then, if appropriate, back-filled with the desired environment (e.g., oxygen in a neutral carrier gas at a prescribed partial pressure). All of the comments that were made relative to extensometry and temperature control apply here as well, with the added complication of the apparatus used to provide environmental control. In addition to being able to measure the temperature, it is highly desirable to measure the gaseous species present. These may arise from so-called "internal" leaks (e.g., welds that entrap gasses but that have a very small hole allowing such gasses to escape) or from impure carrier gasses. Gaseous species can be measured by incorporating a mass spectrometer into the environmental chamber.

METHOD AUTOMATION

As discussed above, the basic needs are a machine to apply a force, an instrument to measure the extension of the specimen, and a readout device to record the experimental results. However, it is frequently desirable to apply the load to the specimen in a well-defined manner. For example, it is well known that materials are sensitive to the rate at which a strain is applied and it is thus important to be able to load the specimen in such a way as to maintain a constant strain rate. In essence, this requirement imposes two conditions on the test:

a. The control mode must be of the displacement type; and
b. The displacement that is measured and controlled must be specimen displacement, as opposed to displacement of the crosshead or hydraulic ram.

In this situation the following are required:

a. An extensometer attached directly to the specimen, and
b. A machine with the ability to compare the desired strain/time profile to the actual strain/time profile and make instantaneous adjustments in the strain rate so as to minimize, as far as possible, the differences between the command signal (i.e., the desired strain rate) and the resultant signal (i.e., the actual strain rate).

Clearly these conditions cannot be met if the test machine can only move the lower crosshead at a predetermined rate, since this will not take deflection in the load train into account as discussed above. The control mode just described is referred to as "closed-loop control," and is the way in which most modern testing is carried out. Modern machines are instrumented in such as way as to be able to operate in strain control, load control, and displacement control modes. In addition, manufacturers now supply controllers in which a combined signal may

be used as the control mode. This is called combinatorial control. An example of combinatorial control would be to load a specimen such that the rate of stress × strain (i.e., power) was constant. A schematic of the closed-loop control concept is provided in Martin (1985).

It is important to realize that the results that are obtained can depend quite sensitively on the control mode. Essentially, the results of a tensile test reflect not only the material's intrinsic properties but also, more fundamentally, the interaction between the material being tested and the test machine. This is perhaps best illustrated for mild steel. If tested under strain control (or under displacement control, if the test machine is very stiff relative to the specimen), this material will exhibit an upper and lower yield point. On the other hand, the upper and lower yield points are completely eliminated if the specimen is tested under load control and the machine is very compliant relative to the specimen. The yield-point phenomenon is illustrated in Figure 5. Reasons for this behavior are discussed above (see Principles of the Method).

SAMPLE PREPARATION

What sort of test specimen should be used will depend upon the form of the material that is being tested and its ductility. The ASTM (1987, E 8 & E 8M) has defined a variety of specimens that can be used for various product forms; typical specimens are shown in ASTM (1987) E 8. It is important to ensure that the tolerances shown in these figures are respected, since excessive nonparallelism and variations in thickness will adversely affect the interpretation and reliability of results.

Generally speaking, specimen preparation for metallic materials requires tools found in most machine shops, such as lathes and surface grinding machines. Within reasonable limits, the tensile and yield strengths are not affected by surface finish except in the case of materials of very limited ductility, for which low-stress grinding is used to obtain honed surfaces with no residual stresses or disturbed surface layers. The parameters for steels and high-temperature Ni-base alloys can be found in ASM (1985). Safety considerations for metallic materials specimen preparation involves only those considerations which are normal in good machining practice. For composites with a polymer matrix, specimens are generally made by infiltration of the polymer (usually an epoxy) around the fibers. In this case, vacuum equipment may be used and care must be taken to ensure proper venting of hazardous fumes.

An additional requirement for obtaining reliable results is that the alignment of the load cell of the test machine and the specimen be coaxial. This will eliminate bending moments and produce a result in which only tensile forces are measured. Techniques for alignment are discussed in detail in ASTM (1987), E 1012.

DATA ANALYSIS AND INITIAL INTERPRETATION

As previously mentioned, the fundamental result of a tensile test is a load-displacement curve. Since the load required to bring a specimen to a given state (e.g., to fracture) depends upon the cross-sectional area, load is not a fundamental measure of material behavior. Therefore, the load must be converted to stress by dividing by the cross-sectional area. Similarly, the extension depends upon the actual length of the specimen and is also not a fundamental quantity. Elongation is put on a more fundamental basis by dividing by the length of the specimen, and the resulting quantity is called strain. The result of a tensile test is then a stress/strain curve, as shown in Figures 1B and 2. There are typically three distinct regions to a stress/strain curve, which have been discussed above (see Principles of the Method).

When comparing test results, it is generally found that the yield and tensile strength are independent of specimen geometry. However, this is not the case for the percent elongation which is often used as a measure of ductility. If specimens of circular cross-section are considered, then the percent elongation will be higher the smaller the ratio of the gage length to diameter. This can be understood in part by noting that the elongation associated with necking will be constant for a given diameter. Thus, the contribution of the nonuniform deformation to the total percent elongation will increase as the gage length decreases. Since different length/diameter specimens are used in different countries, care must be exercised in making comparisons. However, the percent reduction in area (%RA) is independent of diameter. Since this quantity is a good measure of ductility and is independent of diameter, it is recommended for making ductility comparisons.

The scatter associated with such tests is very small (<1%) for well-machined and well-aligned specimens.

Table 1. Typical Tensile Properties of Engineering Materials

Material	Modulus (MPa)	Tensile Strength (MPa)	Yield Strength (MPa)	Elongation (%)	Reduction in Area (%)
1020 steel (As-rolled)	206,850	448.2	330.9	36	59
4340 steel (normalized)	206,850	1279	861.8	12	36
4341 steel (quenched and tempered at 205°C)	206,850	1875	1675	10	38
2024-T3 Al	70,329	393	290	12	—
7075-T6 Al	70,329	538	483	7	—
Ti-6Al-4V (titanium alloy, solution-treated and aged)	119,973	1180	1085	6	11

Differences in yield and ultimate strengths are then indications of differences in material properties. It is customary to use three to five tests to characterize tensile properties. To have a statistically valid test, it is usually considered that the specimen diameter should be at least 25 times the average grain size. For larger grains, specimen-to-specimen variation becomes more pronounced.

Some typical tensile properties are listed in Table 1.

SUMMARY

Very important engineering and scientific information can be obtained from a tensile test. While the test is relatively straightforward to perform, care must be exercised in all aspects of the test, including:

a. selection of the appropriate specimen geometry;
b. proper specimen fabrication;
c. careful alignment of the specimen within the testing machine;
d. appropriate control of the testing conditions including temperature, loading rate, and control mode;
e. analysis and presentation of the test results.

Well-established standards are available that, if followed, will ensure the validity of the results that are obtained.

LITERATURE CITED

American Society for Testing and Materials (ASTM). 1987. ASTM Annual Book of Standards, 1987, Vol. 03.01. ASTM, Philadelphia.

Provides detailed procedures for testing and data reduction.

ASM. 1985. Metals Handbook: Desk Edition. pp. 24–27. American Society for Metals, Metals Park, Ohio.

Provides information on all aspects of production, properties, and use of metallic materials.

Dieter, G. E. 1985. Metals Handbook, 9th ed. Mechanical Testing, Vol. 8, pp. 20–27. American Society for Metals (ASM), Metals Park, Ohio.

Provides an overview of tensile testing.

Martin, J. J. 1985. Metals Handbook, 9th ed. Mechanical Testing, Vol. 8, pp. 47–51. ASM, Metals Park, Ohio.

Provides information on test equipment.

Schaffer, J. P., Saxena, A., Antolovich, S. D., Sanders, T., Jr., and Warner, S. B. 1999. The Science and Design of Engineering Materials. McGraw-Hill, New York.

Provides a tutorial on mechanical testing.

KEY REFERENCES

American Society for Testing and Materials (ASTM), 1987. See above.

Provides detailed procedures for testing and data reduction.

ASM, 1985. See above.

Provides information on all aspects of production, properties, and use of metallic materials.

Dieter, 1985. See above.

Provides an overview of tensile testing.

Martin, 1985. See above.

Provides information on test equipment.

Schaffer et al., 1999. See above.

Provides a tutorial on mechanical testing.

STEPHEN D. ANTOLOVICH
Washington State University
Pullman, Washington

HIGH-STRAIN-RATE TESTING OF MATERIALS

INTRODUCTION

Experimental methods of probing material behavior at high rates of strain are concerned with measuring the variation of mechanical properties, such as strength and ductility, that can vary with strain rate. Strain rate, $\dot{\varepsilon}$, is the rate of change of strain (defined as the relative change in the length of a test sample) with respect to time (t). Little scientific or engineering attention was historically paid to the effects of high strain rates until increased manufacturing production techniques (Lindholm, 1971, 1974; Follansbee, 1985; Field et al., 1994)—i.e., high-speed wire drawing and cold rolling—as well as studies supporting military technologies concerned with ballistics (Hopkinson, 1914; Carrington and Gayler, 1948), armor, and detonation physics necessitated further knowledge. Interest in the high-rate mechanical behavior of materials has continued to expand during the last four decades, driven by demands for increased understanding of material response in impact events. High-rate tests also provide the data critically needed to create predictive constitutive model descriptions of materials. These descriptions strive to capture the fundamental relationships between the ways that each of the independent variables—stress, strain rate, strain, and temperature—independently affect the constitutive behavior of materials (Follansbee and Kocks, 1988; Klepaczko, 1988; Chen and Gray, 1996). Robust material models capturing the physics of high-rate material response are required for large-scale finite-element simulations of (1) automotive crashworthiness; (2) aerospace impacts, including foreign-object damage such as that during bird ingestion in jet engines and meteorite impact on satellites; (3) dynamic structural loadings such as that occurring during earthquakes; (4) high-rate manufacturing processes, including high-rate forging and machining; and (5) projectile/armor interactions.

Measurements of the mechanical properties of materials are normally conducted via loading test samples in compression, tension, or torsion. Mechanical testing frames can be used to achieve nominally constant loading rates for limited plastic strains and thereby a constant engineering strain rate. Typical screw-driven or servohydraulic testing machines are routinely used to measure

the stress-strain response of materials up to strain rates as high as $1\,s^{-1}$. Specially designed testing machines, typically equipped with high-capacity servohydraulic valves and high-speed control and data acquisition instrumentation, can be used during compression testing to achieve strain rates as high as $200\,s^{-1}$. Above this strain rate regime, $\dot{\varepsilon} > 200\,s^{-1}$, alternate techniques using projectile-driven impacts to induce stress-wave propagation have been developed.

Chief among these techniques is the split-Hopkinson pressure bar, which is capable of achieving the highest uniform uniaxial stress loading of a specimen in compression at nominally constant strain rates of the order of $10^3\,s^{-1}$. Stress is measured by using an elastic element in series with the specimen of interest. Stress waves are generated via an impact event and the elastic elements utilized are long bars such that the duration of the loading pulse is less than the wave transit time in the bar. Utilizing this technique, the dynamic stress-strain response of materials at strain rates up to $2\times10^4\,s^{-1}$ and true strains of 0.3 can be readily achieved in a single test.

Historical Background

The Hopkinson bar technique is named after Bertram Hopkinson (Hopkinson, 1914) who, in 1914, used the induced wave propagation in a long elastic bar to measure the pressures produced during dynamic events. Through the use of momentum traps of different lengths, Hopkinson studied the shape and evolution of stress pulses as they propagated down long rods as a function of time. Based on this pioneering work, the experimental apparatus utilizing elastic stress-wave propagation in long rods to study dynamic processes was named the Hopkinson pressure bar. Later work by Davies (1948a,b) and Kolsky (1949) utilized two Hopkinson pressure bars in series, with the sample sandwiched in between, to measure the dynamic stress-strain response of materials. (A note on nomenclature: The terms input/incident bar and output/transmitted bar will be used interchangeably.) This technique thereafter has been referred to as either the split-Hopkinson pressure bar (Lindholm and Yeakly, 1968; Follansbee, 1985), Davies bar (Kolsky, 1964), or Kolsky bar (Kolsky, 1949; Follansbee, 1985).

This unit describes the techniques involved in measuring the high-strain-rate stress-strain response of materials in compression utilizing a split-Hopkinson pressure bar, hereafter abbreviated as SHPB. Emphasis will be given to the method for collecting and analyzing compressive high-rate mechanical property data and to discussion of the critical experimental variables that must be controlled to yield valid and reproducible high-strain-rate stress-strain data.

Competitive and Related Techniques

In addition to the original split-Hopkinson pressure bar developed to measure the compressive response of a material, the Hopkinson technique has been modified for loading samples in uniaxial tension (Harding et al., 1960; Lindolm and Yeakley, 1968), torsion (Duffy et al., 1971), and simultaneous torsion-compression (Lewis and Goldsmith, 1973). The basic theory of bar data reduction based upon one-dimensional stress wave analysis, as presented below (see Principles of the Method), is common to all three loading states. Of the different Hopkinson bar techniques—compression, tension, and torsion—the compression bar remains the most readily analyzed and least complex method to achieve a uniform high-rate stress state.

The additional complications encountered in the tensile and torsional Hopkinson techniques are related to (1) the modification of the pressure bar ends to accommodate gripping of complex samples, which can alter wave propagation in the sample and bars; (2) the potential need for additional diagnostics to calculate true stress; (3) an increased need to accurately incorporate inertial effects into data reduction to extract quantitative material constitutive behavior; and (4) the more complicated stress-pulse generation systems required for tensile and torsion bars. Alteration of the bar ends to accommodate threaded or clamped samples leads to complex boundary conditions at the bar/specimen interface and therefore introduces uncertainties in the wave mechanics description of the test (Lindholm and Yeakley, 1968). When complex sample geometries are used, signals measured in the pressure bars record the structural response of the entire sample, not just the gauge section, where plastic deformation is assumed to be occurring. When plastic strain occurs in the sections adjacent to the sample's uniform gauge area, accurate determination of the stress-strain response of the material is more complicated. In these cases, additional diagnostics, such as high-speed photography, are mandatory to quantify the loaded section of the deforming sample. In the tensile bar case, an additional requirement is that exact quantification of the sample cross-sectional area as a function of strain is necessary to achieve true-stress data.

An additional complexity inherent to both the tension and torsion Hopkinson loading configurations has to do with the increased sample dimensions required. Valid dynamic characterization of many material product forms, such as thin sheet materials and small-section bar stock, may be significantly complicated or completely impractical using either tensile or torsion Hopkinson bars due to an inability to fabricate test samples.

High-rate tensile loading of a material may also be conducted utilizing an expanding ring test (Hoggatt and Recht, 1969; Gourdin et al., 1989). This technique, which requires very specialized equipment, employs the sudden radial acceleration of a ring due to detonation of an explosive charge or electromagnetic loading. Once loaded via the initial impulse, the ring expands radially and thereafter decelerates due to its own internal circumferential stresses. While this technique has been utilized to determine the high-rate stress-strain behavior of a material, it is complicated by the fact that the strain rate changes throughout the test. This variable rate is determined first by the initial loading history, related to the initial shock and/or magnetic pulse, and then by the rapid-strain-rate decelerating gradient during the test following the initial

impulse. The varied difficulties of the expanding ring technique, along with the expense, sample size, and shape, have limited its use as a standard technique for quantitative measurement of dynamic tensile constitutive behavior.

The remaining alternate method of probing the mechanical behavior of materials at high strain rates, of the order of $10^3\,s^{-1}$, is the Taylor rod impact test. This technique, named after its developer G.I. Taylor (1948), entails firing a solid cylinder of the material of interest against a massive and rigid target as shown schematically in Figure 1.

The deformation induced in the rod by the impact shortens the rod as radial flow occurs at the impact surface. The fractional change in the rod length can then, by assuming one-dimensional rigid-plastic analysis, be related to the dynamic yield strength. By measuring the overall length of the impacted cylinder and the length of the undeformed (rear) section of the projectile, the dynamic yield stress of the material can be calculated according to the formula (Taylor, 1948):

$$\sigma = \frac{\rho V^2 (L - X)}{2(L - L_1)\ln(L/X)} \tag{1}$$

where σ is the dynamic yield stress of the material, ρ is the material's density, V is the impact velocity, and L, X, and L_1 are the dimensional quantities of the bar length and deformed length as defined in Figure 1. The Taylor test technique offers an apparently simplistic method to ascertain some information concerning the dynamic strength properties of a material. However, this test represents an integrated test, rather than a unique experiment with a uniform stress state or strain rate like the split-Hopkinson pressure bar test. Accordingly, the Taylor test has been most widely used as a validation experiment in concert with two-dimensional finite-element calculations. In this approach, the final length and cylinder profile of the Taylor sample is compared with code simulations to validate the material constitutive model implemented in the finite-element code (Johnson and Holmquist, 1988; Maudlin et al., 1995, 1997). Comparisons with the recovered Taylor sample provide a check on how accurately the code can calculate the gradient in deformation stresses and strain rates leading to the final strains imparted to the cylinder during the impact event.

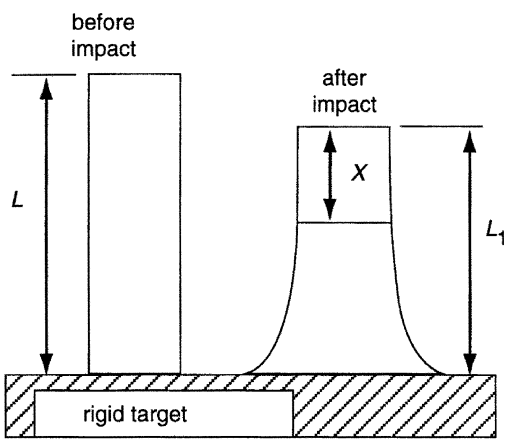


Figure 1. Schematic of Taylor impact test showing the initial and final states of the cylindrical sample (Taylor, 1948).

New Developments

The split-Hopkinson pressure bar technique, as a tool for quantitative measurement of the high-rate stress-strain behavior of materials, is far from static, with many new improvements still evolving. One- and two-dimensional finite-element models of the split-Hopkinson pressure bar have proven their ability to simulate test parameters and allow pretest setup validation checks as an aid to planning. Novel methods of characterizing sample diametrical strains are being developed using optical diagnostic techniques (Valle et al., 1994; Ramesh and Narasimhan, 1996). Careful attention to controlling wave reflections in the SHPB has also opened new opportunities to study defect/damage evolution in brittle materials during high-rate loading histories (Nemat-Nasser et al., 1991). Finally, researchers are exploring exciting new methods for in situ dispersion measurements (Wu and Gorham, 1997) on pressure bars, which offer opportunities for increased signal resolution in the future.

PRINCIPLES OF THE METHOD

The determination of the stress-strain behavior of a material being tested in a Hopkinson bar, whether it is loaded in compression as in the present instance or in a tensile bar configuration, is based on the same principles of one-dimensional elastic-wave propagation within the pressure loading bars (Lindholm, 1971; Follansbee, 1985).

As identified originally by Hopkinson (1914) and later refined by Kolsky (1949), the use of a long, elastic bar to study high-rate phenomena in materials is feasible using remote elastic bar measures of sample response because the wave propagation behavior in such a geometry is well understood and mathematically predictable. Accordingly, the displacements or stresses generated at any point can be deduced by measuring the elastic wave at any point, x, as it propagates along the bar.

In what follows, the subscripts 1 and 2 will be used to denote the incident and transmitted side of the specimen, respectively. The strains in the bars will be designated $\varepsilon_i, \varepsilon_r, \varepsilon_t$ and the displacements of the ends of the specimen u_1, u_2 at the input bar/specimen and specimen/output bar interfaces as given schematically in the magnified view of the specimen in Figure 2.

From elementary wave theory, it is known that the solution to the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c_b^2}\frac{\partial^2 u}{\partial t^2} \tag{2}$$

can be written

$$u = f(x - c_b t) + g(x + c_b t) = u_i + u_r \tag{3}$$

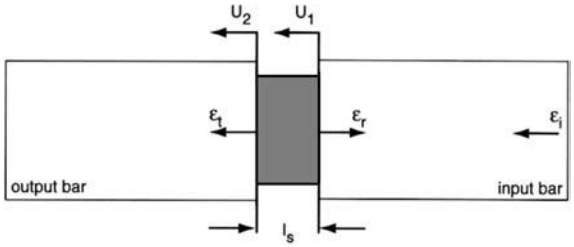


Figure 2. Expanded view of input bar/specimen/output bar region.

for the input bar, where f and g are functions describing the incident and reflected wave shapes and c_b is the wave speed in the rod.

By definition, the one-dimensional strain is given by

$$\varepsilon = \frac{\partial u}{\partial x} \tag{4}$$

So differentiating Equation 3 with respect to X, the strain in the incident rod is given by

$$\varepsilon = f' + g' = \varepsilon_i + \varepsilon_r \tag{5}$$

Differentiating Equation 3 with respect to time and using Equation 5 gives

$$\dot{u} = c_b(-f' + g') = c_b(-\varepsilon_i + \varepsilon_r) \tag{6}$$

for the input bar.

Since the output bar has only the transmitted wave, $u = h(x - c_b t)$, propagating in it,

$$\dot{u} = -c_b \varepsilon_t \tag{7}$$

in the output bar.

Equations 6 and 7 are true everywhere, including at the ends of the pressure bars.

The strain rate in the specimen is, by definition, given by

$$\dot{\varepsilon} = \frac{(\dot{u}_1 - \dot{u}_2)}{l_s} \tag{8}$$

where l_s is the instantaneous length of the specimen and $\dot{u}_1$ and $\dot{u}_2$ are the velocities at the incident bar–specimen and specimen–output bar interfaces, respectively.

Substituting Equations 6 and 7 into Equation 8 gives

$$\dot{\varepsilon} = \frac{c_b}{l_s}(-\varepsilon_i + \varepsilon_r + \varepsilon_t) \tag{9}$$

By definition, the forces in the two bars are

$$F_1 = AE(\varepsilon_i + \varepsilon_r) \tag{10}$$

$$F_2 = AE\varepsilon_t \tag{11}$$

where A is the cross-sectional area of the pressure bar and E is the Young's modulus of the bars (normally equal, given identical material is used for the input and output bars).

Assuming that after an initial ringing up period (the period during which the forces on the ends of the specimen become equal), where the exact time depends on the sample sound speed and sample geometry, the specimen is in force equilibrium; and assuming that the specimen is deforming uniformly, a simplification can be made equating the forces on each side of the specimen, i.e., $F_1 = F_2$. Comparing Equations 10 and 11 therefore means that

$$\varepsilon_t = \varepsilon_i + \varepsilon_r \tag{12}$$

Substituting this criterion into Equation 9 yields

$$\dot{\varepsilon} = \frac{2c_b\varepsilon_r}{l_s} \tag{13}$$

The stress is calculated from the strain gauge signal measure of the transmitted force divided by the instantaneous cross-sectional area of the specimen, A_s:

$$\sigma(t) = \frac{AE\varepsilon_t}{A_s} \tag{14}$$

Utilizing Equations 13 and 14 to determine the dynamic stress-strain curve of a material is termed a "one-wave" analysis because it uses only the reflected wave for strain in the sample and only the transmitted wave for stress in the sample. Hence, it assumes that stress equilibrium is assured in the sample. Conversely, the stress in the sample at the incident bar–sample interface can be calculated using a momentum balance of the incident and reflected wave pulses, termed a "two-wave" stress analysis since it is a summation of the two waves at this interface. However, it is known that such a condition cannot be correct at the early stages of the test because of the transient effect that occurs when loading starts at the input bar–specimen interface while the other face remains at rest. Time is required for stress-state equilibrium to be achieved.

Numerous researchers have adopted a "three-wave" stress analysis that averages the forces on both ends of the specimen to track the ring-up of the specimen to a state of stable stress. The term "three-wave" indicates that all three waves are used to calculate an average stress in the sample; the transmitted wave to calculate the stress at the specimen–transmitted interface (back stress) and the combined incident and reflected pulses to calculate the stress at the incident bar–specimen interface (front stress). In the three-wave case, the specimen stress is then simply the average of the two forces divided by the combined interface areas:

$$\sigma(t) = \frac{F_1(t) + F_2(t)}{2A_s} \tag{15}$$

Substituting Equations 10 and 11 into Equation 15 then gives

$$\sigma(t) = \frac{AE}{2A_s}(\varepsilon_i + \varepsilon_r + \varepsilon_t) \tag{16}$$

From these equations, the stress-strain curve of the specimen can be computed from the measured reflected and transmitted strain pulses as long as the volume of the specimen remains constant, i.e., $A_0 l_0 = A_s l_s$ (where l_0 is the original length of the specimen and A_0 its original cross-sectional area) and the sample is free of barreling (i.e., friction effects are minimized). (*Note*: The stipulation of constant volume precludes, by definition, the testing of foams or porous materials.)

PRACTICAL ASPECTS OF THE METHOD

While there is no worldwide standard design for a split-Hopkinson pressure bar test apparatus, the various designs share common features. A compression bar test apparatus consists of (1) two long, symmetric, high-strength bars, (2) bearing and alignment fixtures to allow the bars and striking projectile to move freely but in precise axial alignment, (3) a gas gun or alternate device for accelerating a projectile to produce a controlled compressive wave in the input bar, (4) strain gauges mounted on both bars to measure the stress wave propagation, and (5) the associated instrumentation and data acquisition system to control, record, and analyze the wave data.

A short sample is sandwiched between the input and output bar as shown schematically in Figure 3. The use of a bar on each side of the material sample to be tested allows measurement of the displacement, velocity, and/or stress conditions, and therefore provides an indication of the conditions on each end of the sample.

The bars used in a Hopkinson bar setup are most commonly constructed from a high-strength material: AISI-SAE 4340 steel or maraging steel or a nickel alloy such as Inconel. This is because the yield strength of the pressure-bar material determines the maximum stress attainable within the deforming specimen. Bars made of Inconel, whose elastic properties are essentially invariant up to 600°C, are often utilized for elevated-temperature Hopkinson bar testing. Because a lower-modulus material increases the signal-to-noise level, the selection of a material with lower strength and lower elastic modulus for the bars is sometimes desirable to facilitate high-resolution dynamic testing of low-strength materials such as polymers or foams. Researchers have utilized bar materials ranging from maraging steel (210 GPa) to titanium (110 GPa) to aluminum (90 GPa) to magnesium (45 GPa) and finally to polymer bars (< 20 GPa) (Gary et al., 1995, 1996; Gray et al., 1997).

The length, l, and diameter, d, of the pressure bars are chosen to meet a number of criteria for test validity as well as the maximum strain rate and strain level desired. First, the length of the pressure bars must assure one-dimensional wave propagation for a given pulse length; for experimental measurements on most engineering materials this requires ~10-bar diameters. To allow wave reflection, each bar should exceed a l/d ratio of ~20. Second, the maximum strain rate desired must be considered in selecting the bar diameter, where the highest-strain-rate tests require the smallest diameter bar. The third consideration affecting the selection of the bar length is the amount of total strain desired to be imparted into the specimen; the absolute magnitude of this strain is related to the length of the incident wave. The pressure bar must be at least twice as long as the incident wave if the incident and reflected waves are to be recorded without interference. In addition, since the bars remain elastic during the test, the displacement and velocity of the bar interface between the sample and the bar can be accurately determined. Depending on the sample size, for strains >30% it may be necessary for the split-Hopkinson bars to have a l/d ratio of 100 or more.

For proper operation, split-Hopkinson bars must be physically straight, free to move without binding, and carefully mounted to assure optimum axial alignment. Precision bar alignment is required for both uniform and one-dimensional wave propagation within the pressure bars as well as for uniaxial compression within the specimen during loading. Bar alignment cannot be forced by overconstraining or forceful clamping of curved pressure bars in an attempt to "straighten" them, as this clamping violates the boundary conditions for one-dimensional wave propagation in an infinite cylindrical solid. Bar motion must not be impeded by the mounting bushings utilized; the bar must remain free to move readily along its axis. Accordingly, it is essential to apply precise dimensional specifications during construction and assembly. In typical bar installations, as illustrated schematically in Figure 3, the pressure bars are mounted to a common rigid base to provide a rigid, straight mounting platform. Individual mounting brackets with slip bearings through which the bars pass are typically spaced every 100 to 200 mm,

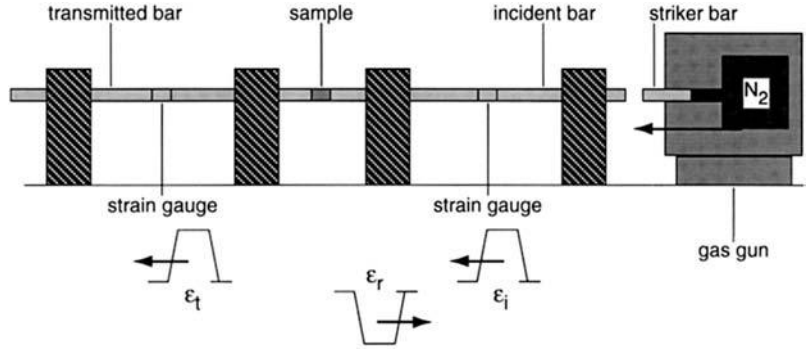


Figure 3. Schematic of a split-Hopkinson bar.

depending on the bar diameter and stiffness. Mounting brackets are generally designed so that they can be individually translated to facilitate bar alignment.

The most common method of generating an incident wave in the input bar is to propel a striker bar to impact the incident bar. The striker bar is normally fabricated from the same material and is of the same diameter as the pressure bars. The length and velocity of the striker bar are chosen to produce the desired total strain and strain rate within the specimen. While elastic waves can also be generated in an incident bar through the adjacent detonation of explosives at the free end of the incident bar, as Hopkinson did, it is more difficult to ensure a one-dimensional excitation within the incident bar by this means.

The impact of a striker bar with the free end of the incident bar develops a longitudinal compressive incident wave in this bar, designated ε_i, as denoted in Figure 3. Once this wave reaches the bar–specimen interface, a part of the pulse, designated ε_r, is reflected while the remainder of the stress pulse passes through the specimen and, upon entering the output bar, is termed the transmitted wave, ε_t. The time of passage and the magnitude of these three elastic pulses through the incident and transmitted bars are recorded by strain gauges normally cemented at the midpoint positions along the length of the two bars. Figure 4 is an illustration of the data measured as a function of time for the three wave signals. The incident and transmitted wave signals represent compressive loading pulses, while the reflected wave is a tensile wave.

If we use the wave signals from the gauges on the incident and transmitted bars as a function of time, the forces and velocities at the two interfaces of the specimen can be determined. When the specimen is deforming uniformly, the strain rate within the specimen is directly proportional to the amplitude of the reflected wave. Similarly, the stress within the sample is directly proportional to the amplitude of the transmitted wave. The reflected wave is also integrated to obtain strain and is plotted against stress to give the dynamic stress-strain curve for the specimen.

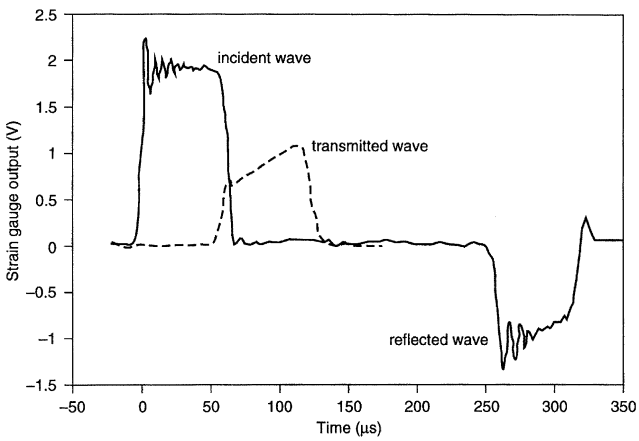


Figure 4. Strain gauge data, after signal conditioning and amplification, from a Hopkinson bar test of a nickel alloy sample showing the three waves measured as a function of time. (Note that the transmitted wave position in time is arbitrarily superimposed on the other waveforms.)

To analyze the data from a Hopkinson bar test, the system must be calibrated prior to testing. Calibration of the entire Hopkinson bar setup is obtained in situ by comparing the constant amplitude of a wave pulse with the impact velocity of the striker bar for each bar separately. Operationally, this is accomplished by applying a known-velocity pulse to the input bar, and then to the transmitted bar, with no sample present. Thereafter, the impact of the striker with the input bar in direct contact with the transmitted bar, with no specimen, gives the coefficient of transmission. Accurate measurement of the velocity, V, of the striker bar impact into a pressure bar can be obtained using the linear relationship

$$\varepsilon_j = \frac{V}{2c_b} \tag{17}$$

where ε_j is the strain in the incident or transmitted bar, depending on which is being calibrated, and c_b is the longitudinal wave speed in the bar. Equation 17 applies if the impacting striker bar and the pressure bar are the same material and have the same cross-sectional area. Careful measurement of the striker velocity—using a laser interruption scheme, for example, in comparison to the elastic strain signal in a pressure bar—can then be used to calculate a calibration factor for the pressure bar being calibrated.

Optimum data resolution requires careful design of the sample size for a given material followed by selection of an appropriate striker bar length and velocity to achieve test goals. Determination of the optimal sample length first requires consideration of the sample rise time, t, required for a uniform uniaxial stress state to equilibrate within the sample. It has been estimated (Davies and Hunter, 1963) that this requires three (actually π) reverberations of the stress pulse within the specimen to achieve equilibrium. For a plastically deforming solid obeying the Taylor–von Karman theory, time follows the relationship

$$t^2 > \frac{\pi^2 \rho_s l_s^2}{\partial\sigma/\partial\varepsilon} \tag{18}$$

where ρ_s is the density of the specimen, l_s is the specimen length, and $\partial\sigma/\partial\varepsilon$ is the stage II work hardening rate of the true-stress/true-strain curve for the material to be tested. For rise times less than that given by Equation 18, the sample should not be assumed to be deforming uniformly and stress-strain data will accordingly be in error.

One approach for achieving a uniform stress state during split-Hopkinson pressure-bar testing is to decrease the sample length, such that the rise time, t, from Equation 18 is as small as possible. Because other considerations of scale (see Sample Preparation) limit the range of l/d ratios appropriate for a specimen dependent on material, the specimen length may not be decreased without a concomitant decrease in the specimen and bar diameters. The use of small-diameter bars (<6 mm) to achieve higher strain

rates is a common practice in split-Hopkinson pressure-bar testing.

Because the value of t from Equation 18 has a practical minimum, an alternate method to achieve stress-state equilibrium at low strains is to increase the rise time of the incident wave. Utilizing self-similar impedance materials for the striker and incident bar (i.e., a symmetric impact) yields a short-rise-time pulse that approximates a square wave. The rise time of such a square-wave pulse is likely to be less than t in Equation 18 in most instances. Contrarily, if the rise time of the incident wave pulse is increased to a value more comparable with the time required to ring up the specimen, then the data will be valid at a lower strain. Furthermore, because the highly dispersive short-wavelength components arise from the leading and trailing edges of the waves, a longer-rise-time pulse will contain fewer of these components than will a sharply rising pulse (Frantz et al., 1984; Follansbee, 1985). The tradeoff to this solution is a lowering of the applied strain rate.

Experimentally, the rise time of the incident wave can be increased by placing a soft, deformable metal shim between the striker and the incident bar during impact. The choice of material and thickness for this shim, or "tip material" (Frantz et al., 1984), depends on the desired strain rate and the strength of the specimen. Typically, the tip material is selected to have the same strength as the specimen and is 0.1 to 2 mm in thickness. An additional benefit of this layer is that it can result in a more uniform strain rate throughout the experiment. However, for thick shims the strain rate will not be constant and will ramp up during the test. The exact selection of the optimal tip material and thickness for a given test sample is not readily calculated and remains a matter of experience via trial and error.

Test Setup

Once the system has been calibrated and the optimal specimen length, l_s, has been selected, preparations for testing can be considered. At a constant strain rate, the maximum strain that can be achieved in a specimen is directly proportional to the length of the striker bar utilized, L:

$$\varepsilon = 2\,\dot{\varepsilon}\frac{L}{C_b} \tag{19}$$

The nominal strain rate in the specimen may be very roughly approximated, erring on the low side, by considering momentum conservation between the striker bar and incident bar. It can similarly be shown (Follansbee, 1985) that

$$\dot{\varepsilon} = \frac{V}{l_s} \tag{20}$$

where V is the striker bar velocity; this will overestimate the strain rate. This relationship has been found to be a good first approximation for soft metals, such as annealed copper, at high striker bar velocities. Equations 19 and 20 can be used to approximate the striker bar length and striker bar velocity required to achieve a desired strain and strain rate for a given sample size. These formulas provide a good starting point for selecting test parameters to achieve a valid SHPB test. However, for samples possessing high initial yield strengths, the yield drops (this is often seen in low-carbon steels or other refractory metals) and/or very high strain-hardening responses will necessitate increased striker bar lengths and impact velocities to achieve the desired strain value.

Stress-State Equilibrium

The classic split-Hopkinson pressure bar equations relating strain gauge measurement to stress-strain behavior in the deforming specimen require that the specimen must deform uniformly; this is opposed by both radial and longitudinal inertia and by frictional constraint at the specimen/pressure bar interfaces. To understand the procedure for validating attainment of a uniform stress state in the sample in a SHPB test, it is instructive to examine the different analyses used to calculate sample stress from the pressure bar strains. In the one-wave analysis, the sample stress is directly proportional to the bar strain measured in the output bar as calculated using Equation 14. This waveform characteristically exhibits low oscillation amplitude because the deforming sample effectively damps much of the high-frequency oscillations inherent in the incident pulse as it propagates through the sample. More importantly, the one-wave stress analysis reflects the conditions at the sample-transmitted bar interface and is often referred to as the sample "back stress."

Alternatively, in a two-wave analysis, the sum of the synchronized incident and reflected bar waveforms (which are opposite in sign) is proportional to the sample front stress and represents the conditions at the interface between the incident/reflected bar and the sample. Unfortunately, both the incident and reflected waveforms contain substantial inherent oscillations that, compared to the transmitted waveform, cause uncertainty in the interpretation of stress, especially near the yield point. In addition, these harmonic oscillations are subject to dispersion due to the wave speed dependence of different frequencies that causes asynchronization of the raw overlapped waveforms and, therefore, inaccuracy in the calculation of the front stress.

A dispersion correction analysis has been developed (Follansbee and Frantz, 1983; Frantz et al., 1984; Follansbee, 1985) to account for these changes in phase angle of the primary mode harmonic oscillation of all three strain signals. This analysis results in more accurate and smoother stress-strain curves, especially near the yield point. Finally, a third stress-calculation variation that considers the complete set of three measured bar waveforms, the three-wave analysis, is simply the average of the front and back stress. The three-wave average is calculated as described by Equation 16.

Sample equilibrium can be checked by comparing the one-wave and three-wave (or two-wave) stress-strain response (Follansbee and Frantz, 1983; Gray et al., 1997; Wu and Gorham, 1997). Recent studies have shown that both inertia and wave propagation effects can significantly affect the stress differences across the length of a specimen deformed in a compression SHPB (Gray et al., 1997; Wu and Gorham, 1997).

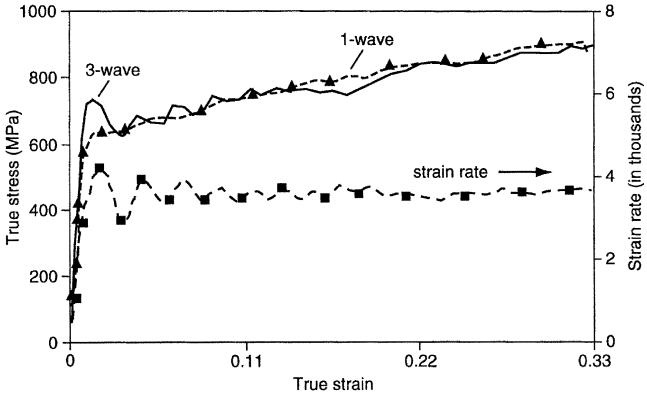


Figure 5. Comparison of stress-strain response of a drawing-sheet steel (DQSK) showing the one- and three-wave stress curves and the strain rate.

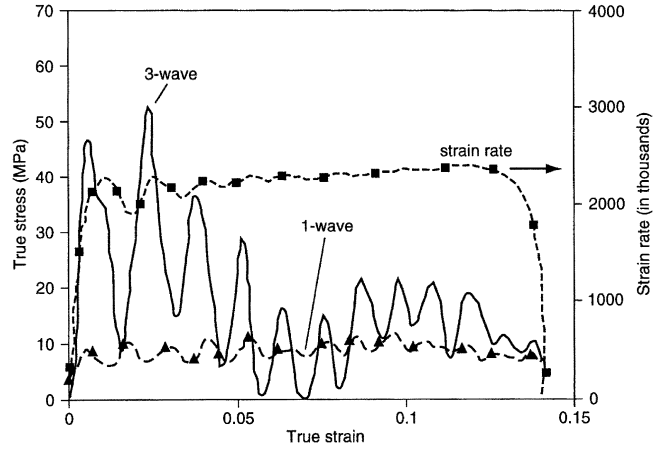


Figure 6. Comparison of room temperature stress-strain response of the polymer Adiprene-L100 (Gray et al., 1997) for a 6.35-mm long sample showing the one- and three-wave stress curves in addition to strain rate.

When the stress state is uniform throughout the sample, the three-wave stress oscillates equally above and below the one-wave stress. Figure 5 shows the one-wave, three-wave, and strain rate data as a function of strain for a SHPB test conducted on a drawing-quality sheet steel. In this illustration, the front and back stress data reductions exhibit very similar response beyond $\sim$0.02 strain, verifying that the sample attained a uniform stress state. This check on the stress equilibrium, when taken with the verification of attaining an essentially constant strain rate throughout the test through a careful balance of striker-bar length, striker-bar velocity, and tip material, demonstrates a high-precision, valid material characterization measurement.

Contrarily, when the stress state is *not* uniform throughout the SHPB sample, the three-wave stress diverges and exceeds the one-wave stress values. Previous Hopkinson bar studies of ceramic materials using this one-wave versus three-wave comparison have shown quite dramatically that a sample is not in stress equilibrium when divergence is observed (Blumenthal and Gray, 1989). In ceramic and cermet materials, this divergence correlates very well with the onset of nonuniform plastic flow and/or premature fracture events.

Similar results have revealed that the slow longitudinal sound speeds typical for some polymeric materials make stress equilibrium during SHPB testing difficult to achieve (Walley et al., 1991; Walley and Field, 1994; Gary et al., 1996). The pronounced difference in the initial one- and three-wave signals for an Adiprene L-100 sample, as shown in Figure 6, can accordingly be viewed as an indication of a sluggish sample ring-up to stress-state equilibrium, compared to the incident wave risetime, and a marginally valid Hopkinson bar test at strains $>5\%$, even though a stable strain rate is indicated throughout the entire test. The data in Figure 6 therefore substantiate the need to examine the technique of using thinner sample aspect ratios when studying the high strain rate constitutive response of low-sound-speed, dispersive materials (Gray et al., 1997; Wu and Gorham, 1997).

Because of the transient effects that are dominant during the ring-up until stress equilibrium is achieved (well over 1% plastic strain in the sample), it is impossible to accurately measure the compressive Young's modulus of materials at high strain rates using the SHPB. The compressive Young's modulus of a material is best measured using ultrasonic techniques.

Increased resolution of the ring-up during SHPB testing of materials with high sound speeds and/or low fracture toughness values can be achieved by directly measuring the strain in the sample via strain gauges bonded directly on the sample (Blumenthal, 1992). Testing of ceramics, cermets, thermoset epoxies, and geological materials requires accurate measurement of the local strains. The difficulty with this technique is that (1) reproducible gauge application on small samples is challenging and labor intensive, and (2) the specimens often deform to strains greater than the gauges can survive (nominally 5% strain) and so can only be used once.

Testing as a Function of Temperature

The derivation of a robust model description of mechanical behavior often requires quantitative knowledge of the coincident influence of temperature variations. Accurate measurement of the high-rate response utilizing a SHPB at temperatures other than ambient, however, presents several technical and scientific challenges. Because the stress and strain as a function of time within the deforming specimen are determined by strain gauge measurements made on the elastic pressure bars, the longitudinal sound speed and elastic modulus of the pressure bars, both of which vary with temperature, are important parameters. The pronounced effect of temperature on the elastic properties of viscoelastic materials, which have been proposed as alternate bar materials to achieve increased stress resolution, has been a chief barrier to their adoption as a means to measure the high-rate response of polymers over a range of temperatures (Gray et al., 1997). In this case, both a rate and temperature-dependent constitutive model for the elastomer pressure bar material itself would be required to reduce SHPB data for the sample of interest.

The combined length of both pressure bars (which can easily exceed 2 m) makes it operationally impossible to heat or cool the entire bar assembly with any uniformity of temperature. Even if this were feasible, it would require a bar material capable of withstanding the high or low temperatures desired as well as the development of new temperature-calibrated strain gauges and robust epoxies to attach them to the bars. Operationally, therefore, the most common techniques for elevated-temperature testing include heating only the sample and perhaps a short section of the pressure bars and then correcting for the temperature-gradient effects on the properties of the pressure bars if they are significant. Researchers have shown that based on an assessment of the temperature gradient, either estimated or measured with a thermocouple, the strain gauge signals can be corrected for the temperature-dependent sound velocity and elastic modulus in the bars (Lindholm and Yeakley, 1968). This procedure has been utilized at temperatures up to 613°C (Lindholm and Yeakley, 1968). The selection of a bar material, such as Inconel, that exhibits only a small variability in its elastic properties up to ~600°C is an alternative that avoids the need for corrections.

An alternate method of performing elevated-temperature SHPB tests that eliminates the need for corrections to the strain gauge data is to heat only the specimen and no part of the bars. This is accomplished by starting with the bars separated from the sample. The sample is heated independently in a specially designed furnace. Just prior to firing the striker bar the pressure bars are automatically brought into contact with the heated sample (Frantz et al., 1984). If the contact time between the sample and the bars is minimized to <200 ms, the specimen temperature and the temperature at the ends of the bars remain approximately constant. This approach, while requiring careful attention to the timing between the movement of the pressure bars and the firing of the striker bar, has been used successfully at temperatures up to 1200°C (Sizek and Gray, 1993).

METHOD AUTOMATION

Acquisition of split-Hopkinson-bar data requires a high degree of automation due to the high-speed nature of the stress loading technique and the need for signal measurement and storage equipment capable of corresponding acquisition speeds. The measurement instrumentation required for a split-Hopkinson pressure bar test includes two strain gauge signal conditioners and a means of recording these signals. The straingauge signal conditioners must have a frequency response of ≥1 MHz to achieve adequate data resolution; units with sampling frequencies as high as 1 GHz are now available. Until recently, oscilloscopes were used almost exclusively to capture and record the strain gauge signal pulses. When using one oscilloscope to capture the strain gauge data, the transmitted wave and integrated reflected wave can be fed to an oscilloscope with *x-y* capability to directly yield the dynamic stress-strain curve for the specimen; the reflected wave may be fed through an operational amplifier to yield a signal directly proportional to the strain in the specimen, but without dispersion correction.

The three wave signals are now often saved using high-speed analog-to-digital recorders or high-speed analog-to-digital data acquisition computer modules that are directly interfaced with a personal computer. Using either platform, it is now possible to digitize the raw data directly and perform the integration numerically. With the improved availability of the digitized data, more robust Hopkinson bar data reduction software is now regularly used by research labs conducting split-Hopkinson-bar testing. This software can be used to adjust the timing between the transmitted and reflected waves to account for the transit time through the specimen as well as calculate the elastic wave dispersion in the measured data.

Strain-gauge signal conditioners range in signal responses and gain factors; all manufacturers offer units of various resolutions and gains. The same is true for the digitizing oscilloscopes used to store strain-gauge data. At this writing, experimenters must write their own software for data handling. The authors use an Ectron Model 778 strain-gauge signals conditioner, which has a variable gain from 1 to 1000 times, and Tektronix TDS 754A four-channel digitizing oscilloscope operating at 500 MHz and 2 GS/s (the four channels allow all three waves to be recorded on a single oscilloscope).

DATA ANALYSIS AND INITIAL INTERPRETATION

If we apply the basic theoretical equations given above (see Principles of the Method), the stress-strain behavior of the specimen can be computed from the measured incident, reflected, and transmitted strain pulses made on the elastic bars. By knowing the timing required for the elastic wave signals to traverse from the specimen to the gauges on the input and output bars, it is possible to synchronize the reflected and transmitted pulses to coincide at the specimen interface. If we integrate Equation 14—remembering that $\varepsilon = \ln(l_0/l_s)$ in compression—where l_0 is the initial length of the specimen, it is possible to calculate $\varepsilon(t)$ and $l_s(t)$ in the specimen. Assuming that the volume of the specimen does not change during the course of deforming the specimen, it is then possible to calculate $A_s(t)$ and therefore $\sigma(t)$ from Equation 14. The average true stress-true strain curve is then produced by removing time as a variable between the stress and strain data.

The one- and three-wave analyses used to calculate average stress (see Principles of the Method) implicitly assume that the strain-time pulses measured at the strain gauges are identical with those at the ends of the bars in contact with the specimen. This assumption is, however, not correct, because the specimen is normally smaller in diameter than the bar such that the bar ends are not uniformly loaded across their diameter and therefore will indent elastically. In addition, as the bars are compressed longitudinally, albeit elastically, they expand radially in response to the applied force, i.e., the Poisson effect. While the exact mathematical descriptions of these effects on the propagation of an elastic pulse through a large elastic rod are very complex, the result is that different frequencies of

pulses induced in the bars disperse with distance traveled in the bar. The result of this dispersion is that the pulse induced in the input bar through the impact of the striker bar does not immediately rise into a steady square-wave impulse of fixed amplitude, but rather rings up. These end effects quickly dampen after the wave has propagated ~10-bar diameters (Follansbee and Frantz, 1983).

The wave propagation behavior thereafter becomes fully described by the equation of motion in an infinite cylindrical solid. The solution of the equation of motion for these boundary conditions was derived independently during the late 19th century by Pochhammer (1876) and Chree (1889). These relations were later specifically applied to address dispersion in the SHPB (Davies, 1948a; Kolsky, 1964; Gorham, 1980; Follansbee and Frantz, 1983). The pressure bars have been determined to vibrate predominantly in a fundamental mode: i.e., while there are an infinite number of potential solutions to the equation of motion according to the vibrational mode, only one frequency appears to dominate in long elastic bars.

This vibration of the bar leads to wave dispersion that can mask resolution of fine details in the stress-strain data of interest, particularly at higher strain rates where the period of the oscillations can be a large fraction of the total strain measured. While this elastic dispersion cannot be eliminated, techniques have been adopted based on the Pochhammer-Chree equation of motion to correct for any additional dispersion that occurs during the wave propagation from the specimen–bar interfaces to the gauge locations (Gorham, 1980; Follansbee and Frantz, 1983).

At any position, z, along the pressure bar, the wave, $f(t)$, may be represented by an infinite cosine Fourier series:

$$f(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} D_n \cos(n\omega_0 t - \delta_n) \qquad (21)$$

where ω_0 is the frequency of the longest wavelength component ($n = 1$) and δ_n is the phase angle of component $n\omega_0$. The wave dispersion in the pressure bars occurs because higher-frequency components travel more slowly than lower-frequency components and thus lag behind the leading edge of the wave. This dispersion alters the phase angle, δ, such that at a position $z + \Delta z$, the new phase angle can be calculated:

$$\delta(z + \Delta z) = \delta(z) + \frac{n\omega_0 \Delta z}{C_0}\left(\frac{C_0}{C_n} - 1\right) \qquad (22)$$

where C_0 is the longitudinal wave speed in the bar and C_n is the velocity of component $n\omega_0$. The value of C_n depends on the wavelength and on the mode of vibration; in the SHPB case, this is dominated by the fundamental mode (Davies, 1948a). The phase angle at any position along the bar can be calculated using Equation 22 and the wave reconstructed at that new position using Equation 21. Accordingly, the raw stress-versus-strain data calculated using either the one- or three-wave analysis can be dispersion corrected by mathematically moving the wave to a common point on the bar (e.g., the bar–sample interface). This correction removes a large amount of the inherent wave dispersion, yielding smoother final stress-strain curves (Gorham, 1980; Follansbee and Frantz, 1983; Frantz et al., 1984; Gary et al., 1991).

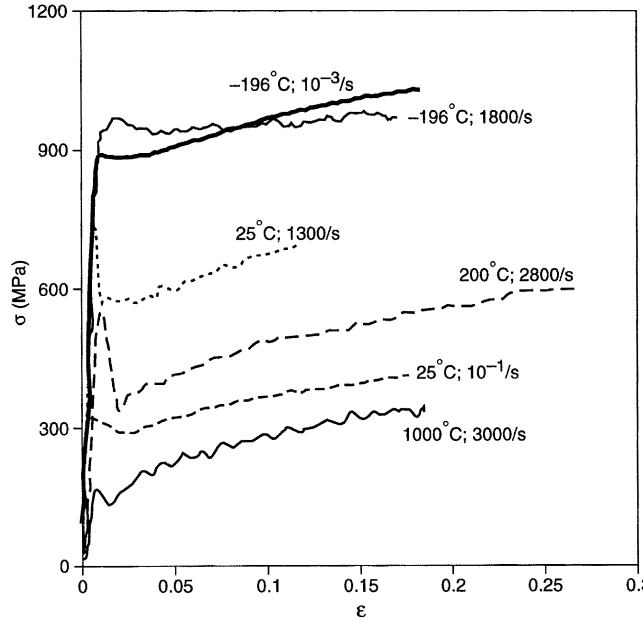


Figure 7. Stress-strain response of high-purity tantalum as a function of strain rate and temperature.

High-rate testing is currently performed on a large number of materials of scientific and engineering interest (Harding, 1983; Follansbee, 1986; Nicholas and Rajendran, 1990; El-Magd et al., 1995; Chen and Gray, 1996). Figures 7 and 8 show the stress-strain responses of high-purity tantalum and copper measured in compression at various strain rates and temperatures. The

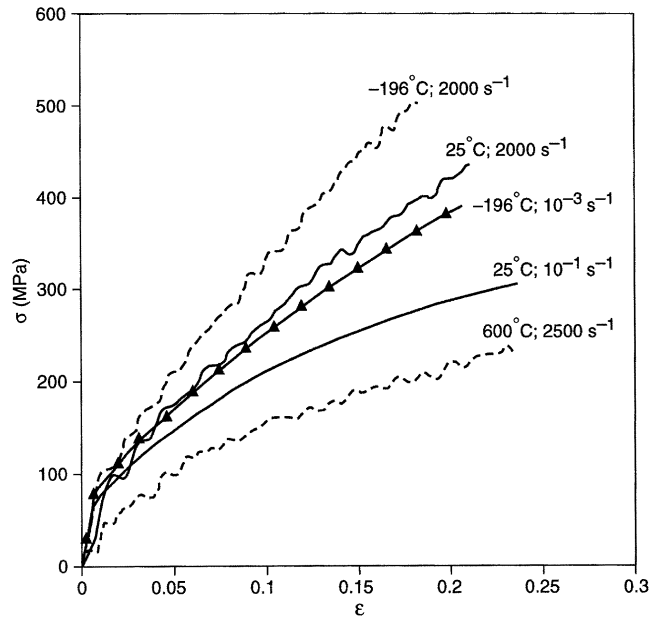


Figure 8. Stress-strain response of oxygen-free high-conductivity (OFHC) copper as a function of strain rate and temperature.

lower-strain-rate tests were performed on a standard screw-driven Instron testing machine using the same specimen geometry as the dynamic tests. All the stress-strain comparisons, as a function of temperature and strain rate, are true values. While the SHPB data at strains below ~1% to 2% are not in stress-state equilibrium, the curves are plotted in their entirety from load inception. Increasing strain rate at ambient temperature, 298 K, is seen to increase the flow stress of tantalum equivalent to a decrease in temperature at a constant strain rate.

The response of Ta is similar to that of other body-centered cubic (bcc) [Fe, W, Mo, Nb] metals and alloys and hexagonal close-packed metals and alloys [Zr, Be, Zn, Ti] (Kocks et al., 1975; Gray, 1991). Contrarily, in this class of metals strain-hardening responses after yielding are nearly invariant as a function of strain rate; i.e., the stress-strain curves are all nearly parallel in slope although offset in their initial yields.

Other materials, including high-purity face-centered cubic (fcc) metals in an annealed condition, such as Cu, Ni, Al, and Ag, exhibit nearly strain-rate-independent yielding behavior while their post-yield strain hardening is strongly rate dependent. The stress-strain behavior of oxygen-free-high-conductivity (OFHC) copper given in Figure 8 is typical of this class of materials. In-depth knowledge of the simultaneous influence of temperature and strain rate is utilized as the basis for advanced materials model development to describe high-strain-rate material response.

In addition to the direct influence of strain rate on stress-strain behavior caused by modification of defect generation and storage processes, high-rate deformation additionally alters the measured mechanical response due to the adiabatic heating accompanying high-rate plastic deformation. While adiabatic heating can be neglected during quasistatic deformation, the effect of adiabatic heating on the measured stress-strain behavior of materials must be considered. To extract an isothermal curve of the material response at high rate, a relationship between temperature and stress must be assumed and the data corrected accordingly. The temperature increase, ΔT, for mechanical tests at strain rates above $500\,s^{-1}$ can be calculated assuming a certain percentage (ψ) of the work of plastic deformation is converted into heat (Taylor and Farren, 1925; Quinney and Tayor, 1937):

$$\Delta T = \frac{\psi}{\rho C_p} \int \sigma(\varepsilon)\delta\varepsilon \qquad (23)$$

where σ and ε are the true stress and strain, ρ is the density, and C_p is the temperature-dependent heat capacity. Adiabatic heating makes a large difference to stress-strain curves measured at high strain rates, particularly at higher strains. Careful consideration of adiabatic effects on overall constitutive response is crucial to accurate material model development.

SAMPLE PREPARATION

Errors due to radial and longitudinal inertia as well as friction effects can be reduced by choosing a sample size that (1) minimizes the areal mismatch between the sample and the bar diameter, and (2) maintains a l/d ratio of 0.5 to 1.0.

For a given bar diameter, the sample diameter is typically chosen to be ~80% of the bar diameter; such a ratio allows ~30% true strain to be imparted into the sample before the sample exceeds the bar diameter. Samples for split-Hopkinson-bar testing, similar to conventional low-rate compression testing, need to be machined such that the two loading faces are flat, parallel to 0.01 mm (0.001 in.) or better, and the sides of the sample are orthogonal to the loading faces. For brittle materials this tolerance must be an order of magnitude greater. Orthogonality, as well as precision machining of parallel flat loading faces, is crucial to attaining uniform elastic loading and thereafter achieving a uniform stress state in the sample.

While most investigators utilizing the split-Hopkinson bar routinely use right-regular cylindrical samples, cubes, as well as other square-sided shapes, can be utilized. Ease of machining on a lathe to achieve accurate and reproducible sample specifications also favors cylindrical samples. The selection of the optimal sample diameter and l/d ratio is dependent on the maximum strain rate desired as well as on the sample size requirement necessary to have specimens sufficiently large to assure measurement of "bulk" properties of the material of interest. Coarse-scaled microstructures or composite materials require larger sample sizes than fine-scaled ones. An approximate rule of thumb is that the specimen diameter needs to be at least 10 times the "representative" microstructural unit size for polycrystalline metals and alloys. Coarse-scaled materials, in particular engineering composites such as concrete (Albertini et al., 1997) or polymeric lay-ups, require careful sample size selection as well as specifically designed split-Hopkinson bars to achieve valid high-strain-rate stress-strain data.

Sample design for split-Hopkinson pressure-bar testing of materials exhibiting low resistance to brittle axial cracking, such as ceramics and cermets and/or composite matrix/reinforcement debonding, require specialized sample designs. The use of "dog-bone" samples, originally designed by Tracey (see Blumenthal, 1992), makes it possible to achieve stable uniaxial stress in ceramics during Hopkinson-bar testing. By suppressing axial cracking, and/or brittle fracture processes in ceramic or cermet samples, due to the larger diameter ends of the samples, valid Hopkinson-bar tests can be achieved.

SPECIMEN MODIFICATION

Following testing in the split-Hopkinson pressure bar, samples can be examined and subsequently analyzed to evaluate the influence of high-strain-rate prestraining on microstructure evolution, stored energy, and damage evolution. This is particularly informative when repeated reloadings are done as a means of increasing the total strain, as discussed below (see Problems, section on friction). For substructure analysis, samples can be cross-sectioned, polished, and etched for examination using

optical metallography or scanning electron microscopy, and/or thinned to electron transparency for high-magnification substructure studies using transmission electron microscopy. Detailed structural studies of this type, as a function of SHPB testing over a range of strain rats and temperatures, are used to provide physical insight into the deformation mechanisms (such as dislocation slip, deformation twinning, and stress-induced phase product) to allow the construction of physically based constitutive materials models. Samples that undergo damage evolution processes, such as intergranular or transgranular cracking and/or ductile void formation (if loaded in tension), during SHPB testing can also be analyzed to quantify damage initiation and evolution processes in materials subjected to high-rate deformation loading paths.

PROBLEMS

In addition to the care needed to assure the achievement of stress equilibrium during SHPB testing, several additional problems can be encountered during experimental investigations studying the mechanical behavior of materials at high rates of loading. These are associated with inertia effects in the test machine, sample constraint due to friction at the points of contact, and issues related to assuring accurate strain gauge measurements on the pressure bars.

Inertia

Even when the specimen has been evaluated to be deforming uniformly, longitudinal and radial inertia due to the rapid particle accelerations imposed during high strain rate testing can influence the measured stress-strain behavior (Follansbee, 1985; Gorham, 1991). The errors due to both longitudinal and radial inertia have been analyzed and corrections have been derived (Davies and Hunter, 1963). This analysis predicts that errors are minimized if the strain rate is held constant or if the term inside the brackets is set to zero by choosing specimen dimensions are selected such that

$$\frac{l_s}{d} = \sqrt{\frac{3\nu_s}{4}} \tag{24}$$

where l_s is the specimen length, d is the diameter, and ν_s is the Poisson ratio.

For a Poisson ratio of 0.33, Equation 24 suggests that the optimum sample l_s/d ratio to minimize errors due to inertia is 0.5. This l_s/d ratio is less than that determined to be the most favorable for the minimization of errors due to friction in ASTM Standard E9, which specifies that $1.5 < l_s/d < 2.0$ (Follansbee, 1985). However, because the total strain in a split-Hopkinson pressure bar test is typically limited to ~25% to reduce the area mismatch between the specimen and the pressure bar, samples with an l_s/d of 0.5 are not expected to introduce any serious errors.

Radial inertia considerations limit the strain rate for which the SHPB technique is valid to $\sim 10^5\,s^{-1}$ for a high-sound-speed material. This is because as the strain rate is increased the specimen size must be decreased accordingly. Eventually the specimen required becomes so small that it no longer reflects bulk material response. Only very fine-grained materials, such as nanocrystalline structured materials with high sound speeds, can be tested with valid results at the very highest strain rates using this technique.

Friction

Friction is an important consideration in determining the validity of all compression testing (Klepaczko and Malinowski, 1978). The optimum l_s/d ratio to minimize friction for a split-Hopkinson pressure-bar compression test specimen is approximately one-half that determined to be most favorable for minimizing errors due to friction for low-rate tests. Lubrication is therefore required at the specimen–pressure bar interfaces.

Early workers studying the SHPB were concerned about longitudinal inertial effects (i.e., how long it takes the forces on the ends of the specimen to become equal), or the ringing-up time. Accordingly, they tested thin wafers of material with a thickness/diameter ratio of 0.1. Such extreme sample-aspect ratios are known to maximize the effect of friction, as the measured yield pressure is related to the material flow stress (Gorham et al., 1984).

Care must therefore be exercised in lubricating samples prior to testing as well as in limiting the strain in a single test to strains less than ~20% to 25%. Repeated reloading of a sample, with intermediate remachining to assure flat and parallel sample loading surfaces, can be used to achieve higher total strains. Repetitive reloading offers the added benefit of minimizing adiabatic heating effects on the measured stress-strain behavior.

The use of an oil-based molybdenum disulfide lubricant has been shown to be effective for room temperature SHPB testing (Follansbee, 1985). For elevated-temperature tests, a thin layer of fine boron nitride powder can be used to lubricate the specimen/pressure bar interfaces. So far, no lubricant has been found that completely eliminates friction for metallic specimens in this geometry, though the frictional stresses can be reduced to ~4% of the metal's shear strength. Friction can, in principle, be measured using annular specimens, the idea being that the higher the friction, the smaller the ratio of the internal to external diameter for a given strain (Gorham et al., 1984; Walley et al., 1997). However, the presence of a layer of lubricant at these interfaces can influence the timing between the waves recorded on the incident and output pressure bars. Consequently, it is important to keep the layer of lubricant on the bar ends as thin as possible. Routine examination of the surface finish and flatness of the pressure-bar loading faces, in addition to checking for any possible cracking or erosion of the bar ends, are also critical to valid SHPB data acquisition.

Strain-Gauge Measurements

Since the stress-strain behavior of the material of interest is deduced from the elastic strain signals in the pressure bars, the details of the conditions controlling the accuracy and reproducibility of the strain gauges are crucial. A number of considerations related to strain-gauge installation and use can affect the elastic strain measurements obtained. Two gauges generally are mounted at diametrically opposite positions on each bar and connected so as to average out most bending strains in the bars, and increase the magnitude of the strain signal measured by a factor of 2. The use of four gauges arranged equidistantly around the bar diameter, comprising a complete Wheatstone strain-gauge bridge, will totally eliminate any bending effects on the strain data. The four-gauge arrangement additionally corrects for any magnetostriction effects (induced voltage in the strain gauges due to stress wave propagation in a ferromagnetic bar material). Utilization of nonmagnetic pressure-bar materials, such as Ti-6Al-4V or Mg, eliminates any potential magnetostriction errors in the strain gauge outputs. In addition to the elimination of bending forces through the use of multiple gauge locations, shielding of the wire leads from the strain gauges to the signal conditioners is important to minimize external noise given the small magnitude of the absolute strain signals and the level of amplification required to boost the signals to the levels required by the data acquisition system.

Finally, bonded strain gauges have a finite time response capability that is linked to the stiffness of the epoxy utilized to bond the gauges to the bars. Using an epoxy that is too compliant can dampen the signals transferred from the bar surface to the strain gauges. Conversely an epoxy that is too stiff and brittle will require frequent replacement. Careful selection, application, and maintenance of the gauges bonded to the SHPB is required to assure accurate bar operation.

ACKNOWLEDGMENTS

The author would like to thank Bill Blumenthal, Gilles Naulin, and George Kaschner for offering suggestions and proofreading this work. This work was supported in part by the U.S. Department of Energy.

LITERATURE CITED

Albertini, C., Cadoni, E., and Labibes, K. 1997. Impact fracture process and mechanical properties of plain concrete by means of an Hopkinson bar bundle. *J. Phys. C3 France (DYMAT)* 7:C3-915–C3-920.

Blumenthal, W. R. 1992. High strain rate compression and fracture of B_4C-aluminum cermets. *In* Shock-Wave and High-Strain-Rate Phenomena in Materials (M. A. Meyers, L. E. Murr, and K. P. Staudhammer, eds.) pp. 1093–1100. Marcel Dekker, New York.

Blumenthal, W. R. and Gray, G. T., III. 1989. Structure-property characterization of shock-loaded B_4C-Al. *Inst. Phys. Conf. Ser.* 102:363–370.

Carrington, W. E. and Gayler, M. L. V. 1948. The use of flat ended projectiles for determining yield stress. III: Changes in microstructure caused by deformation at high striking velocities. *Proc. R. Soc. London Ser. A* 194:323–331.

Chen, S. R. and Gray, G. T., III. 1996. Constitutive behavior of tantalum and tantalum-tungsten alloys. *Metall. Mater. Trans. A* 27A:2994–3006.

Chree, C. 1889. The equations of an isotropic elastic solid in polar and cylindrical coordinates: Their solution and application. *Trans. Cambridge Philos. Soc.* 14:250–369.

Davies, E. D. H. and Hunter, S. C. 1963. The dynamic compression testing of solids by the method of split Hopkinson pressure bar (SHPB). *J. Mech. Phys. Solids* 11:155–179.

Davies, R. M. 1948a. A critical study of the Hopkinson pressure bar. *Philos. Trans. R. Soc. London Ser. A* 240:375–457.

Davies, R. M. 1948b. A simple modification of the Hopkinson pressure bar. *In* Proceedings of the 7th International Congress for Applied Mechanics, London, Vol. 1, p. 404. Organizing Committee, London.

Duffy, J., Campbell, J. D., and Hawley, R. H. 1971. On the use of a torsional split Hopkinson bar to study rate effects in 1100-0 aluminum. *Trans. ASME J. Appl. Mech.* 38:83–91.

El-Magd, E., Scholles, H., and Weisshaupt, H. P. 1995. The stress-strain behaviour of pure metals and composite materials in the region of the Lüders deformation under dynamic load conditions. *DYMAT J.* 2:151–166.

Field, J. E., Walley, S. M., and Bourne, N. K. 1994. Experimental methods at high rates of strain. *J. Phys. IV France C8 (DYMAT)* 4:3–22.

Follansbee, P. S. 1985. The Hopkinson bar. *In* Metals Handbook, Vol. 8, 9th ed., pp. 198–203. American Society of Metals, Metals Park, Ohio.

Follansbee, P. S. 1986. High strain rate deformation of FCC metals and alloys. *In* Metallurgical Applications of Shock Wave and High Strain Rate Phenomena (L. E. Murr, K. P. Staudhammer, and M. A. Meyers, eds.) pp. 451–480. Marcel Dekker, New York.

Follansbee, P. S. and Frantz, C. 1983. Wave propagation in the SHPB. *Trans. ASME J. Eng. Mater. Technol.* 105:61–66.

Follansbee, P. S. and Kocks, U. F. 1988. A constitutive description of the deformation of copper based on the use of the mechanical threshold stress as an internal state variable. *Acta Metall.* 36:81–93.

Frantz, C. E., Follansbee, P. S., and Wright, W. T. 1984. Experimental techniques with the SHPB. *In* High Energy Rate Fabrication—1984 (I. Berman and J. W. Schroeder, eds.) pp. 229–236. Marcel Dekker, New York.

Gary, G., Klepaczko, J. R., and Zhao, H. 1991. Correction de dispersion pour l'analyse des petites déformations aux barres de Hopkinson. *J. Phys. IV France Colloq. C3 (DYMAT 91)* 1: 403–410.

Gary, G., Klepaczko, J. R., and Zhao, H. 1995. Generalization of split Hopkinson bar technique to use viscoelastic materials. *Int. J. Impact Eng.* 16:529–530.

Gary, G., Rota, L., and Zhao, H. 1996. Testing viscous soft materials at medium and high strain rates. *In* Constitutive Relation in High/Very High Strain Rates (K. Kawata and J. Shiori, eds.) pp. 25–32. Springer-Verlag, New York.

Gorham, D. A. 1980. Measurement of stess-strain properties of strong metals at very high strain rates. *Inst. Phys. Conf. Ser.* 47:16–24.

Gorham, D. A. 1991. The effect of specimen dimensions on high strain rate compression measurements of copper. *J. Appl. Phys. D Appl. Phys.* 24:1489–1492.

Gorham, D. A., Pope, P. H., and Cox, O. 1984. Sources of error in very high strain rate compression tests. *Inst. Phys. Conf. Ser.* 70:151–158.

Gourdin, W. H., Weinland, S. L., and Boling, R. M. 1989. Development of the electromagnetically launched expanding ring as a high-strain-rate test technique. *Rev. Sci. Instrum.* 60:427–432.

Gray, G. T. 1991. Deformation substructures induced by high rate deformation. *In* Modeling the Deformation of Crystalline Solids (T. C. Lowe, A. D. Rollett, P. S. Follansbee, and G. S. Daehn, eds.) pp. 145–158. The Metallurgic Society of the American Institute of Mining Engineers, Warrendale, Pa.

Gray, G. T., III, Blumenthal, W. R., Trujillo, C. P., et al. 1997. Influence of temperature and strain rate on the mechanical behavior of Adiprene-L100. *J. Phys. C3 France (DYMAT 97)* 7:C3-523–C3-528.

Harding, J. 1983. High-rate straining and mechanical properties of materials. *In* Explosive Welding, Forming and Compaction (T. Z. Blazynski, ed.) pp. 123–158. Applied Science Publishers, London.

Harding, J., Wood, E. O., and Campbell, J. D. 1960. Tensile testing of materials at impact rates of strain. *J. Mech. Eng. Sci.* 2: 88–96.

Hoggatt, C. R. and Recht, R. F. 1969. Stress-strain data obtained at high rates using an expanding ring. *Exp. Mech.* 9:441–448.

Hopkinson, B. 1914. A method of measuring the pressure produced in the detonation of high explosives or by the impact of bullets. *Philos. Trans. R. Soc. London Ser. A* 213: 437–456.

Johnson, G. R. and Holmquist, T. J. 1988. Evaluation of cylinder-impact test data for constitutive model constants. *J. Appl. Phys.* 64:3901–3910.

Klepaczko, J. R. 1988. Constitutive modeling in dynamic plasticity based on physical state variables—A review. *J. Phys. France C3 (DYMAT 88)* 49:553–560.

Klepaczko, J. and Mainowski, Z. 1978. Dynamic frictional effects as measured from the split Hopkinson pressure bar. *In* High Velocity Deformation of Solids (K. Kawata and J. Shioiri, eds.) pp. 403–416. Springer-Verlag, New York.

Kocks, U. F., Argon, A. S., and Ashby, M. F. 1975. Thermodynamics and kinetics of slip. *Prog. Mater. Sci.* 19:1.

Kolsky, K. 1949. An investigation of the mechanical properties of materials at very high rates of loading. *Proc. Phys. Soc. London* 62B:676–700.

Kolsky, K. 1964. Stress waves in solids. *J. Sound Vibration* 1:88–110.

Lewis, J. L. and Goldsmith, W. 1973. A biaxial split Hopkinson bar for simultaneous torsion and compression. *Rev. Sci. Instrum.* 44:811–813.

Lindholm, U. S. 1971. High strain rate testing. *In* Techniques of Metals Research, Vol. 5, Part 1: Measurement of Mechanical Properties (R. F. Bunshah, ed.) pp. 199–271. John Wiley & Sons, New York.

Lindholm, U. S. 1974. Review of dynamic testing techniques and material behaviour. *Inst. Phys. Conf. Ser.* 21:3–21.

Lindholm, U. S. and Yeakley, L. M. 1968. High strain rate testing: Tension and compression. *Exp. Mech.* 8:1–9.

Maudlin, P. J., Foster, J. C., Jr., and Jones, S. E. 1997. A continuum mechanics code analysis of steady plastic wave propagation in the Taylor test. *Int. J. Impact Eng.* 19:231–256.

Maudlin, P. J., Wright, S. I., Gray, G. T., III, et al. 1995. Application of faceted yield surfaces for simulating compression tests of textured materials. *In* Metallurgical and Materials Applications of Shock-Wave and High-Strain-Rate Phenomena (L. E. Murr, K. P. Staudhammer, and M. A. Meyers, eds.) pp. 877–886. Elsevier/North-Holland, Amsterdam, The Netherlands.

Nemat-Nasser, S., Isaacs, J. B., and Starrett, J. E. 1991. Hopkinson techniques for dynamic recovery experiments. *Proc. R. Soc. London Ser. A* 435:371–391.

Nicholas, T. and Rajendran, A. M. 1990. Material characterization at high strain rates. *In* High Velocity Impact Dynamics (J. A. Zukas, ed.) pp. 127–296. John Wiley & Sons, New York.

Pochhammer, L. 1876. Über Fortpflanzungsgeschwindigkeiten kleiner Schwingungen in einem unbegrenzten isotropen Kreiszylinder. *J. Reine Angew. Math.* 81:324–336.

Quinney, H. and Taylor, G. I. 1937. The emission of the latent energy due to previous cold working when a metal is heated. *Proc. R. Soc. London Ser. A* 163:157–181.

Ramesh, K. T. and Narasimhan, S. 1996. Finite deformations and the dynamic measurement of radial strain on compression Kolsky bar experiments. *Int. J. Solids Structures* 33:3723–3738.

Sizek, H. W. and Gray, G. T., III. 1993. Deformation of polycrystalline Ni_3Al at high strain rates and elevated temperatures. *Acta Metall. Mater.* 41:1855–1860.

Taylor, G. I. 1948. The use of flat ended projectiles for determining yield stress. I: Theoretical considerations. *Proc. R. Soc. London Ser. A* 194:289–299.

Taylor, G. I. and Farren, W. S. 1925. The heat developed during plastic extension of metals. *Proc. R. Soc. London Ser. A* 107: 422–451.

Valle, V., Cottron, M., and Lagarde, A. 1994. Dynamic optical method for local strain measurements: Principle and characteristics. *J. Phys. IV France Colloq. C8 (DYMAT 94)* 4:59–64.

Walley, S. M., Church, P. D., and Furth, M. 1997. A high-speed photographic study of the rapid deformation of metal annuli: Comparison of theory and experiment. *J. Phys. C3 France (DYMAT 97)* 7:C3-317–C3-322.

Walley, S. M. and Field, J. E. 1994. Strain rate sensitivity of polymers in compression from low to high rates. *DYMAT J.* 1:211–228.

Walley, S. M., Field, J. E., and Safford, N. A. 1991. A comparison of the high strain rate behaviour in compression of polymers at 300K and 100K. *J. Phys. IV France Colloq. C3 (DYMAT 91)* 1:185–190.

Wu, X. J. and Gorham, D. A. 1997. Stress equilibrium in the split Hopkinson pressure bar test. *J. Phys. C3 France (DYMAT 97)* 7:C3-91–C3-96.

KEY REFERENCES

Lindholm, 1971. See above.
The seminal publication that established the technique.

Follansbee, 1985. See above.

Follansbee and Frantz, 1983. See above.

Frantz et al., 1984. See above.
Four robust reviews of the state of the art in this area.

Kolsky, 1949. See above.

GEORGE T. (RUSTY) GRAY III
Los Alamos National Laboratory
Los Alamos, New Mexico

FRACTURE TOUGHNESS TESTING METHODS

INTRODUCTION

With increasing use of steel structures such as railroads, bridges and tanks in the nineteenth century, the number of casualties due to the catastrophic failure of such structures also increased, and hence, attention was drawn to the brittle fracture of steels (Parker, 1957; Anderson, 1969). The problem of brittle failure was fully recognized when over 20% of all-welded ships built during World War II developed serious cracks (Rolfe and Barsom, 1977). The weld defects produced a tri-axial stress state that promoted the initiation and propagation of fracture by the cleavage mechanism. The implementation of notch-toughness testing methods (e.g., Charpy V-notch and Izod impact testing; ASTM, 1997a) for identifying the ductile-to-brittle transition temperature was helpful in selection of the low-strength steels used at that time. However, after World War II, new ultra-high-strength steels and aluminum alloys were developed for automobile and aircraft applications, for which the determination of the ductile-to-brittle transition was neither adequate nor relevant. Simultaneously, the fail-safe design, which is a damage-tolerant philosophy, was applied more to load-bearing structures. While the basic ideas leading to fracture mechanics can be traced to A. A. Griffith (Griffith, 1920), the concept of fracture toughness testing based on linear elastic fracture mechanics (LEFM) was developed by Irwin in the 1950s (Irwin, 1948, 1957). In the early 1960s, the American Society for Testing and Materials (ASTM) developed the first draft for plane strain fracture toughness testing. The oil crisis in the 1970s encouraged extensive off-shore drilling and resulted in the development of fracture testing techniques based on elastic-plastic fracture mechanics (Off-Shore Welded Structures, 1982). Presently, fracture toughness testing methods are applied to various types of materials including composites, intermetallics, plastics, and ceramics, as well as metals.

Fracture toughness is a measure of resistance to cracking in notched bodies. In general, during fracture toughness testing, the load-versus-displacement behavior of a pre-cracked specimen is recorded. The load-displacement curve depends on the size and geometry of the specimen as well as the loading mode. In practice, the goal of fracture toughness testing is to measure a single parameter (e.g., similar to yield strength) that is a material property and can be directly used in design. Fracture mechanics is a tool to translate the results of laboratory fracture toughness testing into engineering design. Depending on the material and the size and geometry of the specimen, different parameters can be evaluated. Stress intensity factor, K, is a parameter that describes the stress level as well as the stress distribution ahead of a crack (Westergaard, 1939; Irwin, 1957). Crack driving force, G, is a measure of the elastic energy released per unit thickness of the specimen and per unit length of the crack (Irwin, 1948). When extensive plastic deformation precedes crack initiation, the flow of energy to the crack tip can be evaluated using the J-integral approach (Rice, 1968). The amount of strain at the crack tip, known as crack tip opening displacement (CTOD), δ, is another parameter that can be calculated from the fracture toughness testing results (British Standard Institution, 1979).

To correlate the results of fracture toughness testing with a parameter that represents a material property, a fracture criterion needs to be established. Ideally, fracture toughness should be correlated with the onset of unstable fracture (fast fracture). However, depending on the type of material and the size of the specimen, catastrophic failure may not occur during testing. In these cases, the point of crack initiation can be chosen as a point value or the increase in resistance to cracking as a function of crack length can be evaluated (R-curve analysis).

Fracture toughness testing is a relatively expensive method in comparison to notched impact testing (ASTM, 1997a). However, the latter method provides only a qualitative measure of the toughness of materials, which cannot be used directly in engineering design or modeling of the fracture processes. Presently, many industries, including aerospace, automobile, and shipbuilding, follow codes and specifications that have material fracture toughness requirements. In research, the information obtained by fracture toughness testing in conjunction with fractographic and microstructural characterization techniques has been used for modeling the fracture behavior and developing new fracture-resistant materials.

There are many standard procedures established by ASTM (ASTM Standards, 1997b) where the reader can find the details of fracture toughness testing procedures. The purpose of this unit is to briefly summarize various fracture toughness testing techniques and to provide the basic principles of fracture behavior and fracture mechanics necessary for understanding and selecting the proper testing technique. Emphasis will be placed on macroscopic standard testing techniques for metallic materials. Special techniques for mixed-mode fracture, indentation fracture, crack arrest, creep fracture, and stress corrosion cracking as well as the atomistic concepts will not be discussed here. Furthermore, the discussion is limited to elastically isotropic materials.

There are mechanical testing methods that yield data related to the toughness of a material, but they do not provide an actual toughness value. For example Charpy V-Notch impact testing (ASTM, 1997a) and notched tensile testing (ASTM, 1997l) measure the energy absorbed during fracture and the stress required for fracture, respectively. Such tests may be used for screening materials based on toughness.

To conduct fracture toughness testing, a mechanical testing system, apparatus for measuring and recording load and displacement, as well as fixtures for holding the specimens, are required. The following are the two major companies in the United States that provide all the necessary equipment: Instron Corporation and MTS System Corporation.

PRINCIPLES OF THE METHOD

Fracture toughness is a measure of the resistance of a material to crack extension and is evaluated by loading a

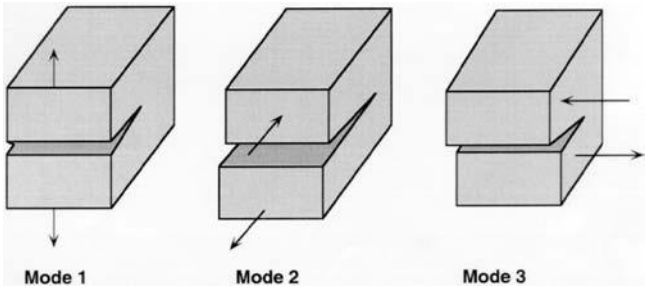


Figure 1. A schematic showing the three modes of loading.

pre-cracked specimen and recording the load-versus-displacement curve. The loading is usually performed under a displacement-controlled condition and, in most cases, is pursued until complete fracture of the specimen. A cracked solid may be loaded in three modes, as shown in Figure 1. While all three loading modes may be experienced during the fracture of a structural component, conventionally only mode 1 is used for fracture toughness testing. The rationale behind this is that almost pure mode 1 can be achieved easily while testing laboratory-size specimens, and furthermore, it represents the most severe condition for promoting brittle fracture without extensive permanent deformation near the crack tip.

When cracked bodies are loaded, very high elastic stresses are developed near the crack tip, as illustrated two-dimensionally in Figure 2A. The high density of line forces near the crack tip represents the development of high stresses. No load lines go through the crack because the atomic bonds are severed and cannot carry the load. Hence, the shaded triangles are not loaded. When the crack propagates, the density of the force lines in the vicinity of the crack increases and the "not-loaded" areas increase as shown in Figure 2B. These two phenomena—the increase in stresses at the crack tip and the release of the elastic energy due to unloading—provide the impetus or drive for crack extension. The level of the stress is evaluated by the stress intensity factor, K, and the rate of elastic energy released per thickness and unit crack extension

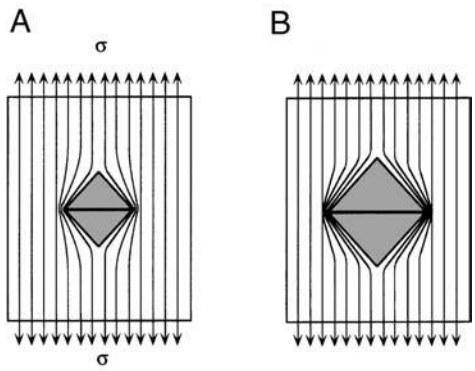


Figure 2. A representation of stress intensification near cracks, demonstrating how the volume of the material that does not carry load increases with an increase in the crack length (**A**) to (**B**). Note that σ is the applied remote stress.

is known as crack driving force, G. These two parameters are used for measuring the fracture toughness of materials.

In most materials, the very high stresses near the crack tip are relaxed by permanent (plastic) deformation. As long as the size of this plastic zone is small relative to the dimensions of the stressed body, the elastic stresses outside the plastic zone control the level and distribution of the stresses inside the plastic zone, and similarly the relationship between elastic energy that would be released upon crack propagation and the applied remote load is modified insignificantly. This condition is known as small-scale yielding and the theoretical treatment of the stresses is referred to as linear elastic fracture mechanics (LEFM). In this regime, K and G are related by (Irwin, 1948, 1957).

$$G = \frac{K^2}{E'} \tag{1}$$

In this equation, $E' = E$ (for plane-stress condition) or $E' = E/(1-\nu^2)$ (for plane-strain condition), where E is Young's modulus and ν is the Poisson's ratio of the material. If the crack extension occurs when the load is beyond the LEFM regime, the single parameter K does not represent the stress level and distribution near the crack tip, and the relationship between G and K fails. This regime is known as elastic-plastic fracture mechanics and the two parameters that have been used to evaluate fracture toughness in this regime are the J-integral, J, and the crack tip opening displacement, δ. Under the small-scale yielding condition, J and δ are related to G by (Rice, 1968; Wells, 1961):

$$G = \frac{K^2}{E'} = J = m\delta\sigma_{\text{flow}} \tag{2}$$

where m varies between 1 to 2 and depends on the plastic constraint at the crack tip (for plane stress, $m = 1$; for plane strain, $m = 2$). σ_{flow} is the flow stress of the material and is usually defined as $\sigma_{\text{flow}} = 1/2(\sigma_{\text{ys}} + \sigma_{\text{UTS}})$, where σ_{ys} and σ_{UTS} are the yield strength and the ultimate tensile strength of the material, respectively.

In the following sections, the load-displacement behaviors of notched specimens are discussed and a brief description of G, K, J, and δ are presented.

Load-Displacement Behaviors

In the absence of plasticity and slow (or stable) crack growth, the load-displacement curve of linear elastic materials is a straight line as depicted in Figure 3. The inverse of the slope of the curve, i.e., Δ/P, is known as the compliance of the specimen. As shown in Figure 3, for a given material and specimen geometry, compliance increases with an increase in the crack length. The nonlinearity in load-displacement curves can arise from two sources: one is the extension of the plasticity across the specimen ligament, and the other is the growth of a stable (or slow) crack. In the latter case, as illustrated in Figure 4, ideally, the displacement returns to zero upon unloading of the sample and the amount of crack extension is related to the decrease in the load-displacement slope (or the

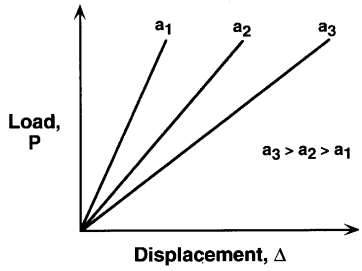


Figure 3. Linear load-displacement curves showing that crack growth results in an increase in the compliance (inverse of the slope of load-displacement curve).

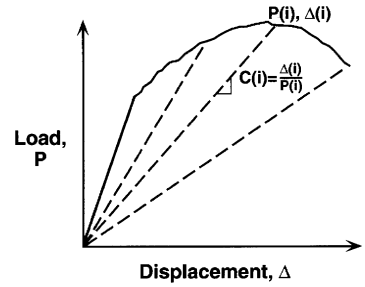


Figure 4. Nonlinearity in load-displacement curve due to slow crack growth. Note that in the absence of plasticity, displacement returns to zero upon unloading.

increase in the compliance). Figure 5 depicts the development of plasticity that leads to the nonlinear load-displacement behavior. As the extent of plastic deformation at the crack tip increases, the deviation from linear behavior becomes more prominent. At net-section yielding, the whole ligament of the specimen is plastically deformed. This load is known as the limit load, P_L, which depends on the yield strength of the material as well as its strain hardening behavior. Further increases in load would cause plastic instability, which is associated with a decrease in load and necking of the ligament.

Crack instability is associated with a drastic increase in the crack velocity, and hence is detected as a sudden decrease in the load. Fast fracture may occur anywhere on the load-displacement curve and can be preceded by plasticity and/or stable crack growth as illustrated in Figure 6. If the unstable crack propagation results in the complete fracture of the specimen, then the load drops to zero. Another phenomenon is crack arrest, which occurs after fast fracture and is characterized as a partial load drop (Figs. 6B and 6C). If the fast fracture develops only a short distance and then arrests, the phenomenon is known as "pop-in" (Fig. 6B).

The competition between fast fracture and plasticity for a given material and specimen geometry depends on many factors including loading rate, sharpness of the notch, and thickness of the specimen. The last parameter controls the stress state at the crack tip. A plane-strain condition is achieved in thick specimens and is associated with the development of high normal stresses (high tri-axial stress state) near the crack tip in the plastic zone (large plastic constraint). Consequently, the probability of brittle fracture, which is a stress-controlled fracture, is enhanced. In thin specimens, the stress state at the crack tip is a plane stress condition which results in a lower crack driving force at a given applied stress level (see discussion of Crack Driving Force, below) and lower levels of stress within the plastic zone. This stress state encourages plasticity and slow crack growth at the crack tip. In addition to the modification of the stress state at the crack tip, an increase in the specimen thickness increases the probability of finding defects that can potentially initiate brittle fracture, and hence the measured toughness will be lowered.

Crack Driving Force, *G*

Fracture mechanics, and hence fracture toughness testing, is based on the idea presented by Griffith on fracture of glass (Griffith, 1920). Griffith states that "if a body

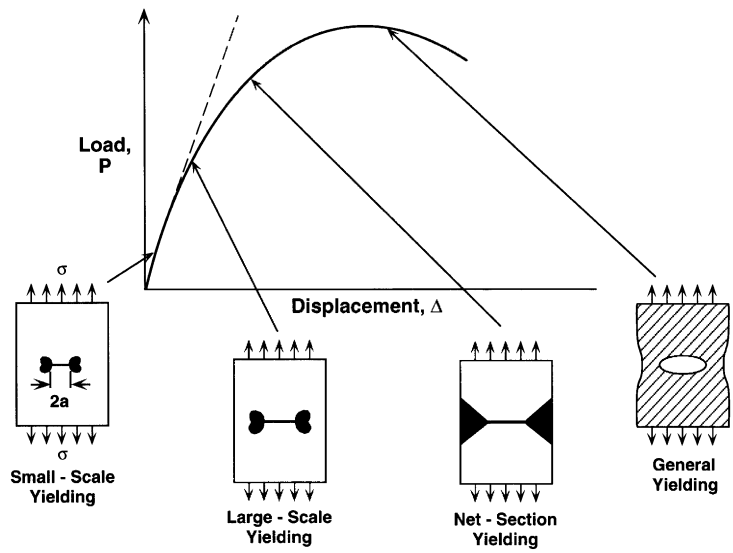


Figure 5. A schematic presenting the development of plasticity in a center-cracked panel. The black and cross-hatched areas represent plastic deformation.

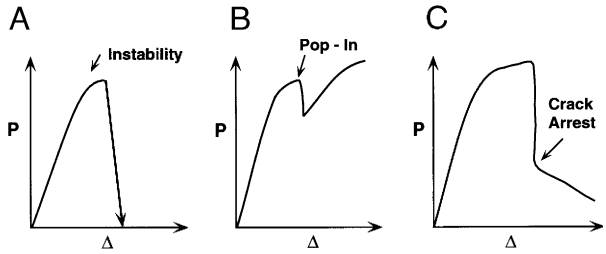


Figure 6. Manifestations of unstable crack growth as a sudden load drop. (**A**) Complete fracture of the specimen. (**B**) Development of a small unstable crack (pop-in phenomenon). (**C**) Arrest of an unstable crack.

having linear stress-strain relationships be deformed from the unstrained state to equilibrium by given (constant) surface forces, the potential energy of the latter is diminished by an amount equal to twice the surface energy. It follows that the net reduction in the potential energy is equal to the strain energy, and hence the total decrease in potential energy due to the formation of crack is equal to the increase in strain energy less the decrease in surface energy." Consider an infinite plate of unit thickness made of an homogeneous isotropic material loaded uniformly in tension by stress σ (refer to Fig. 2). The development of a through-thickness sharp central crack of length $2a$ is accompanied by an increase in energy equivalent to twice of $2a\gamma$ (note that two surfaces are produced), where γ is the surface energy, and by a decrease in energy due to a reduction in strain energy (unloading of the triangles in Fig. 2), which is given by $\sigma^2\pi a^2/E'$. The total change in the potential energy (ΔU) of the system due to the presence of the crack is given by

$$\Delta U = \frac{\sigma^2\pi a^2}{E'} - 4a\gamma \tag{3}$$

Note that the work done on the system is considered negative and the energy released by the system is taken as positive. The crack will extend in an unstable manner when:

$$\frac{\partial(\Delta U)}{\partial a} = 0 \quad \text{or} \quad \frac{\sigma^2\pi a}{E'} = 2\gamma \tag{4}$$

Equation 4 shows that whenever the elastic energy released per unit length of the crack, $G = \sigma^2\pi a/E'$ becomes equal or larger than the resistance to fracture, $R = 2\gamma$, the crack will propagate catastrophically. Since γ is a material property, the critical elastic energy release rate, $G_C = 2\gamma$ will also be a material property, and hence it can be used to represent the toughness of the material. G has units of energy/(length)^2 and can be interpreted as force per unit length of the crack. Except for a few intrinsically brittle materials (e.g., glasses, silicon, and alumina) plastic deformation always precedes propagation of fracture from existing cracks. Irwin and Orowan (Irwin, 1948; Orowan, 1955) modified the Griffith criterion for the case of elastic-plastic materials by adding a plastic term, γ_P, to the surface energy γ. The γ_P term is representative of the amount of plastic deformation that occurs at the crack tip before the crack can be extended. In this case the critical elastic energy release rate is given as:

$$G_c = 2(\gamma_P + \gamma) \tag{5}$$

It should be noted that although γ_P is much larger than γ, the former value is a function of the latter value (McMahon and Vietek, 1979). However, the plastic energy term is a material property only under the small-scale-yielding condition. Therefore, as will be discussed later, there are restrictions on the size of the fracture toughness testing specimens to ensure the validity of a test.

Other mechanisms besides plastic deformation can dissipate energy within the volume near the crack tip (usually referred to as the process zone; Ritchie, 1988) including microcracking, phase transformation, and bridging by second-phase particles, whiskers, or fibers. The dissipative processes shield the crack tip from the remote applied loads. The shielding results in an increasing crack resistance force with crack extension, known as R-curve behavior. In this case the instability criterion is defined as the point where $G = R$ and $\partial G/\partial a = \partial R/\partial a$, as depicted in Figure 7.

For finite-size specimens the value of G depends on the specimen geometry, and it can be shown to be related to the change in the compliance of the sample as (Broek, 1982).

$$G = \left(\frac{P^2}{2B}\right)\left(\frac{dC}{da}\right)_a, \qquad \sigma = \frac{P}{WB} \tag{6}$$

In this relationship, P is the applied load, B and W are the thickness and the width of the specimen, respectively, and C is the compliance of the specimen. Compliance increases with the crack length nonlinearly, and hence dC/da is also dependent on the crack length.

Stress Intensity Factor, *K*

The stress field in the vicinity of the crack tip for isotropic homogeneous materials is given as (Paris and Sih, 1965).

$$\sigma_{ij} = \left\{\frac{K}{(2\pi r)^{1/2}}\right\} f_{ij}(\theta) \tag{7}$$

where r is the radial distance from the crack tip and the function $f_{ij}(\theta)$ represents the variation of the stress

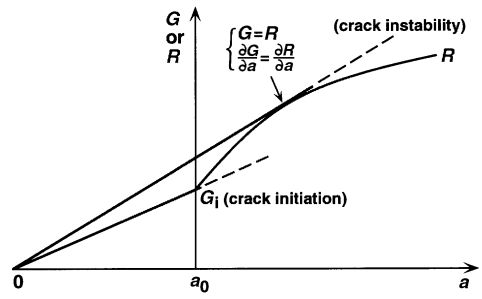


Figure 7. The R-curve behavior.

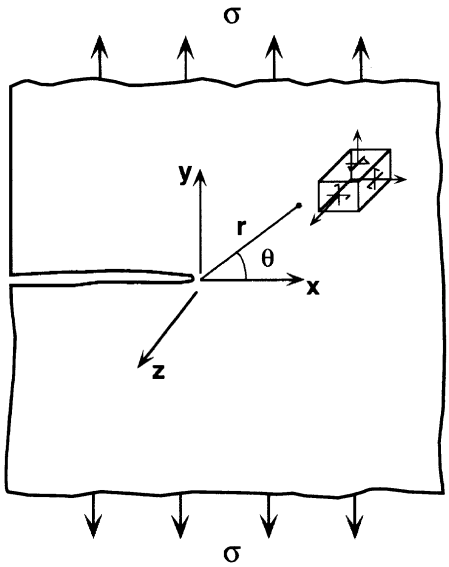


Figure 8. General mode 1 problem.

components with angle θ as illustrated in Figure 8. Each stress component is proportional to a single parameter, K, known as stress intensity factor. Values of σ_{ij} for the three loading modes, as well as the displacements, are given in Appendix A. Equation 7 is valid only for sharp cracks, and therefore ASTM standards for fracture toughness testing require that the fracture specimens have a sharp crack tip.

The K-value depends on the remote uniform applied stress, σ, and the crack length, a, as well as the specimen or component geometry. Several methods, including stress functions and numerical analysis, have been used to calculate K-values for different geometries. Many solutions can be found in Sih (1973). For an infinite plate containing a central through thickness crack of length $2a$, the stress intensity factor in mode 1 is given as:

$$K_1 = \sigma(\pi a)^{1/2} \tag{8}$$

In general, K_1 can be written as:

$$K_1 = Y\sigma a^{1/2} \tag{9}$$

where the function Y depends on the a/W ratio and is known for specific specimen geometries. K has units of force/(length)$^{3/2}$ or stress. $\sqrt{\text{length}}$, and is usually expressed in Ksi$\sqrt{\text{in}}$ or MPa$\sqrt{\text{m}}$.

The size of the plastic zone developed at a crack tip has been related to the applied crack using different approaches (Knott, 1973). In general, the plastic zone size, r_P, is given by

$$r_P = \alpha\left(\frac{K}{\sigma_{ys}}\right)^2 f(\theta) \tag{10}$$

where $f(\theta)$ is the function in Equation 7 and α depends on the stress state at the crack tip, the sharpness of the crack tip, and the strain-hardening rate of the material. The plane stress condition and a blunted crack tip promote larger plastic zone sizes. Within the small-scale yielding regime, $\alpha = 1/6\pi = 0.053$ for the plane strain condition and $\alpha = 1/\pi = 0.318$ for plane stress condition at $\theta = 0[f(\theta) = 1]$ in mode 1.

Since K and G are related in the LEFM regime, when G reaches the critical value at the point of crack instability, so does the $K(K_C = [G_C E']^{1/2})$. The K-value is also used for evaluation of slow crack initiation and propagation in thin sheets under the small-scale-yielding condition. The ASTM standards based on the stress intensity factor concept are E-399 (ASTM, 1997c), E-561 (ASTM, 1997d), E-812 (ASTM, 1997e,f), E-992 (ASTM, 1997g), E-1304 (ASTM, 1997h), and D-5045 (ASTM, 1997i).

J-Integral

Based on Eshelby's ideas (Eshelby, 1974), Rice (1968) introduced an integral, known as the J-integral, which bypasses the boundary-value problems associated with evaluating the stress (or strain) concentrations. This method is particularly useful in evaluating materials that show nonlinear load-displacement behavior. The integral J is defined as:

$$J = \int_\Gamma \{W\,dy - \mathbf{T}(\partial\mathbf{u}/\partial x)\,ds\} \tag{11}$$

where Γ is a curve surrounding the notch, W is the loading work per unit volume or strain energy density ($W = \int \sigma_{ij}\,d\varepsilon_{ij}$, for elastic bodies), $\mathbf{T}$ is the surface traction vector, $\mathbf{u}$ is the displacement vector, ds is an arc element along the curve, and x and y are the coordinates. This integral is independent of the path for elastic materials (linear or nonlinear) and therefore, its value away from the crack, which can be evaluated easily, also represents the value of the integral close to the crack tip. For linear elastic materials it can be shown that $J = G$. As long as unloading (slow crack growth) does not occur, plasticity can be dealt with as a nonlinear elastic behavior. When the crack extends upon loading, parts of the plastic zone also unload. This creates a closure effect which adds to the J-value needed to continue crack growth. Consequently, dJ/da has a positive value, but it decreases with crack extension. Begley and Landes (1972) demonstrated that J can be computed from the area under the load displacement curve. For a deeply notched specimen in which plasticity is limited to the ligament ahead of the crack, the J-integral can be calculated from (Turner, 1979).

$$\begin{aligned} J &= J_e + J_P = (\eta_e U_e/Bb) + (\eta_P U_P/Bb) \\ &= K^2/E' + (\eta_P U_P/Bb) \end{aligned} \tag{12}$$

where J_e and J_P are the elastic and plastic components of J, respectively, U_e and U_P are the areas under the elastic and plastic components of the load-displacement curve, respectively (see Fig. 9), $b = W - a$ is the ligament size, and η_e and η_P depend on the specimen geometry and size. J-integral is used for evaluating crack instability as well as crack initiation and crack propagation in materials.

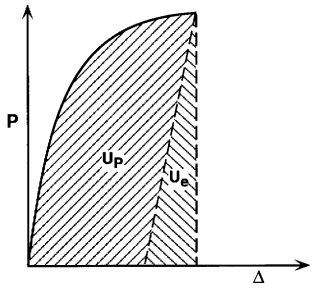


Figure 9. The division of the area under load-displacement curve to elastic, U_e, and plastic, U_p, components.

Following the ASTM nomenclature, J_{IC} refers to the toughness of materials near the onset of stable crack growth, J_C is the fracture toughness of materials at the point of crack instability with no prior slow crack extension, and J_u is defined as fracture toughness of materials at the point of crack instability with extensive prior slow crack growth. The E813 (ASTM, 1997e,f) and E1737 (ASTM, 1997j) are two ASTM standards for measuring the fracture behavior of materials under plane strain condition using J-integral.

Crack Tip Opening Displacement (CTOD), δ

The occurrence of plasticity at the tip of a sharp crack results in the blunting of the crack tip as shown in Figure 10. Within the LEFM regime, the CTOD is related to K and G according to Equation 2. In the case of large-scale deformation at the crack tip, the CTOD can be calculated from its simple geometrical relationship with the plastic component of the mouth opening displacement V_P, as illustrated in Figure 11. During CTOD testing usually the mouth opening displacement is recorded instead of the load line displacement. The CTOD is given by (British Standard Institution, 1979):

$$\delta = \delta_e + \delta_P = K^2/m\sigma_{flow}E + [r_p(W-a)V_P]/[r_p(W-a)+a+Z] \quad (13)$$

where δ_e and δ_P are the elastic and plastic components of the CTOD, respectively, r_p is the rotation factor, and Z is

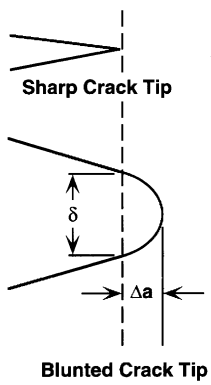


Figure 10. A schematic showing the blunting process and the development of the crack tip opening displacement, δ. (Top) Sharp crack tip. (Bottom) Blunted crack tip.

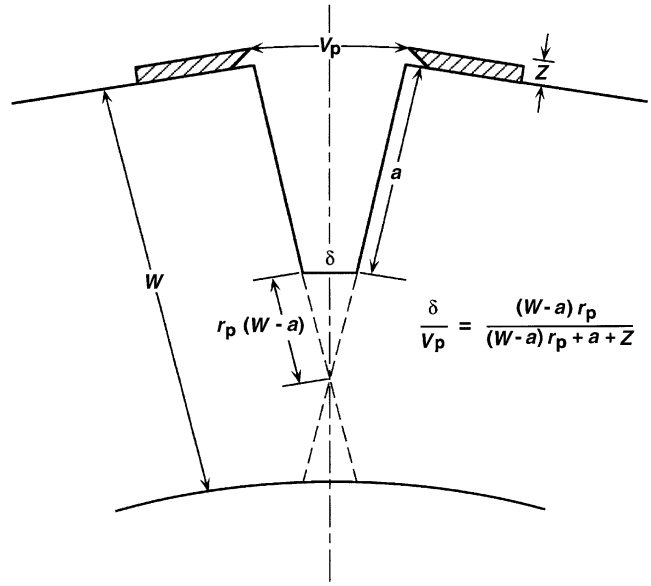


Figure 11. A geometrical representation of the relationship between the crack tip opening displacement, δ, and the plastic component of the mouth opening displacement, V_p.

the knife-edge thickness for cases where knife edges are used for mounting the clip gauge. The rotational factor changes slightly with CTOD and specimen geometry and has a value between 0.4 to 0.5. The CTOD can be evaluated at (1) the onset of stable crack growth, δ_i, (2) the onset of unstable fracture with no prior slow crack growth, δ_C, (3) the onset of unstable fracture with extensive prior slow crack extension, δ_u, and (4) the maximum load, δ_m. The E-1290 ASTM standard (ASTM, 1997k) is the standard used in the U.S. for evaluating the toughness of metallic materials by the CTOD method.

PRACTICAL ASPECTS OF THE METHOD

The most important practical aspects of fracture toughness testing are the preparation of the specimen and the instrumentation for measuring and recording the load displacement curve. Techniques for monitoring crack length extension during loading become important in cases where it is desirable to establish the J-R curve behavior. For specimen preparation, see Sample Preparation.

Apparatus for Measuring and Recording Load-Displacement Curves

To conduct fracture toughness testing, a stiff mechanical testing system capable of applying constant displacement rates, an apparatus for measuring and recording load and displacement, and fixtures for holding the specimens are required.

The mechanical testing system can be screw- or hydraulic-driven. The unloading of the test machine upon specimen fracture releases elastic energy, which adds to the crack-driving force. Therefore, it is very important for the compliance of the system to be very small compared

to the specimen's compliance. Similarly, the fixture used for holding the specimens should be bulky enough and made of a stiff material such as steel, so that the elastic deformation of the fixture does not contribute to the onset of the unstable fracture. For fatigue precracking of the metallic specimens, the testing system should be capable of operating under load control, and it should be capable of applying cyclic load-time functions (e.g., sinusoidal or square).

The load-measuring apparatus consists of a load cell, an amplifier, and a signal conditioner to amplify, balance, and calibrate the load signal. The load cell capacity and the full-scale range should be selected based on a guessed fracture load. The displacement is usually measured by attaching a clip gauge to the crack mouth opening of the specimen using a pair of knife edges (see Fig. 11), which can be machined with the specimen or can be attached to it with screw. In the case of compact tension (CT) specimens, the crack mouth opening is usually measured along the load line, and therefore represents the load-line displacement. For bend specimens, the load-line displacement can either be calculated from the mouth-opening displacement (Lin et al., 1982) or can be measured directly by placing an extensometer on the mid-span point. The load-displacement curve can be recorded using an analog *x-y* plotter or digitally using a computer via an A/D (analog/digital) convertor. In more recent models of mechanical testing systems the signal outputs are digital, and data acquisition as well as system control is performed using a computer. In the case of *J-R* curve analysis a very high resolution of data acquisition (16-bit or higher) is required.

Crack Extension Measurement

In the cases where one wants to establish the *J-R* curve (or δ-R curve), e.g., for ductile materials and composites, it is necessary to separate the contributions of plasticity and slow crack growth to the nonlinearity of the load-displacement curve. Therefore, the extent of the crack growth at several points on the load-displacement curve should be determined. This task can be achieved either by using multiple specimens or by detecting the crack length periodically during the loading of a single specimen.

The multiple-specimen technique involves loading similar specimens to various displacement levels and unloading them. In order to measure the crack length, the tip of the crack needs to be marked such that after complete fracture of the specimen, the tip can be recognized. Various methods can be used for marking the crack tip, including: (1) tinting the crack by heating metallic samples to elevated temperatures and fracturing the sample at room temperature, (2) fracturing samples at a low temperature where fracture proceeds by the cleavage mechanism and hence will be distinguishable from the slow crack (applicable to materials that can be cleaved), or (3) extending the crack further by cyclic loading of the specimen. After the complete fracture of a specimen, the extent of the slow crack growth can be measured using an optical or a scanning electron microscope (SEM). Because of the plane stress condition on the sides of the specimen, the front of slow cracks is usually not straight and has a thumb-nail shape. Therefore, crack length should be measured at many points on the crack front and an average should be used in calculating the toughness. The ASTM standards impose limitations on the crack length variation along the thickness of the specimen. These limitations guarantee the repeatability of the measurement, and hence the validity of the equations used for calculating the toughness of the material. In order to achieve a straight crack front, side grooving of the specimens is usually recommended.

The crack growth can be monitored in a single specimen using the unloading compliance technique or the electric potential drop technique (ASTM, 1997j). The unloading compliance technique involves periodic unloading of the specimen at various points on the load displacement curve. As was mentioned previously, the compliance of a specimen is dependent on its crack length. The a/W ratio can be expressed as a function of the product $E'BC$ for a given geometry (see Appendix B for functions for SENB and CT specimens). Since the change in crack length is usually very small, the load and displacement during unloading has to be measured with a very high resolution in order to accurately evaluate the change in the compliance. Presently, there are software packages available in conjunction with computer-controlled testing systems that allow unloading and measuring of the crack length via the evaluation of the slope of the load-displacement curve upon unloading. The electrical potential drop is applicable only to electrically conductive materials; however, it can provide many more data points than the elastic unloading compliance method. This technique involves applying a DC current to the specimen and measuring the potential drop across the crack plane. This difference in voltage depends on the geometry as well as the crack length of the specimen. The basic components needed are a constant current power supply, an amplifier, a precise voltmeter, and a recording device. Closed form solutions that relate the potential drop to the crack length are available for SENB and CT specimens. Details of this technique can be found in ASTM (1997j).

DATA ANALYSIS AND INITIAL INTERPRETATION

Load displacement curves obtained in mode 1 depend on the material as well as the specimen geometry and size. In order to analyze them, the elastic modulus (Young's modulus) and tensile (or compression) properties (yield strength and strain hardening behavior) of the material, for the temperature and loading rate of the test, should be known. The fracture behavior and the associated load-displacement response of the material may fall into three general categories.

Type A

Unstable fracture occurs with no, or an insignificant amount of, slow crack growth preceding it. Within this category the unstable fracture may occur on any point on the load displacement curve shown in Figure 5. If nonlinearity is observed in the load-displacement curve, its onset should correlate with the limit load of the specimen.

For many materials, the morphology and fracture markings of the slow crack growth are distinctly different from the unstable fracture. Therefore, microscopic observations using optical microscopy (see OPTICAL MICROSCOPY and REFLECTED-LIGHT OPTICAL MICROSCOPY) or scanning electron microscopy (see SCANNING ELECTRON MICROSCOPY) can assist in detecting the extent of the slow crack growth. Fracture load, P_f, is selected as the point of crack instability as shown in Figure 6. The K-value at this point can be calculated from a knowledge of load, crack length, and specimen geometry (see Appendix B for calculation of K for SENB and CT specimens). If fracture occurs within the linear region of the load-displacement curve and the specimen dimensions and the precracking procedure follow the ASTM E-399 standard (ASTM, 1997c), then the K-value can be qualified as the plane-strain fracture toughness, K_{IC}. As mentioned previously, these qualifications ensure that fracture occurs under contained small-scale yielding and plane strain conditions, and hence the toughness value is independent of the specimen geometry and size. Figure 12 presents an actual load-displacement curve for a SENB specimen of a structural steel (A36 steel) tested at -105°C. Note that the curve is almost linear; however, the specimen thickness and width do not satisfy the ASTM requirements for small scale yielding and plane-strain conditions (Equation 15).

If the load displacement curve shows nonlinearity, the J_C- and δ_C-values can be calculated from Equations 12 and 13, respectively. An equivalent K_C-value can be calculated from the J_C-value using the relationship given in Equation 2. These toughness values are usually referred to as the apparent fracture toughness, and it is understood that they depend on the specimen dimensions.

Type B

In this category the stable crack is initiated within the small-scale yielding regime, and hence the onset of nonlinearity in the load-displacement curve occurs at a load much smaller than the limit load of the specimen. In some materials the slow crack growth may lead to crack instability; otherwise the sample is usually unloaded after the attainment of a maximum in the load. The latter behavior is usually observed in relatively thin sheet materials. If fast fracture happens immediately after a short crack extension, plane strain fracture toughness may be calculated from the results. Figure 13 demonstrates the three possible scenarios. The ASTM E-399 standard (ASTM, 1997c) suggests that a line with a slope equal to 95% of the linear portion of the curve be drawn. The intersection of this point with the load displacement curve is designated as P_5, and represents the point at which the crack has grown such that the compliance is increased by 5%. If this load is less than the maximum load (load at the point of unstable fracture), then P_5 should be selected for the calculation of K (type I). If the instability appears as a pop-in (type II) then the load at this point is used for calculating K. If P_5 is past the point of instability (type III), then the maximum load is obviously the fracture load. The ASTM E-399 standard requires the P_{max}/P_5 to be equal or less than 1.1 in order for the K-value to qualify as the

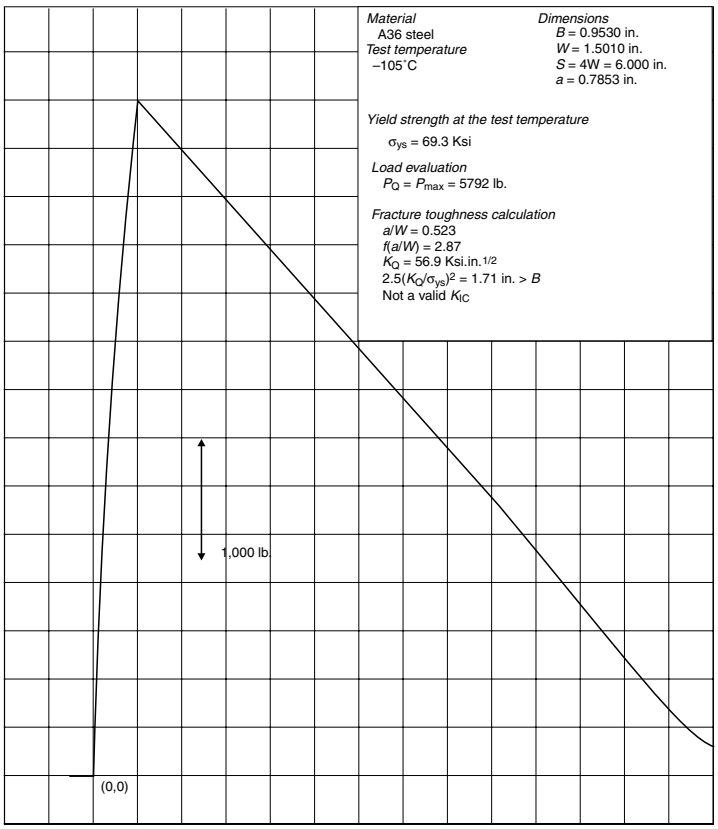


Figure 12. Actual load versus mouth-opening displacement curve obtained from testing a SENB specimen of an A36 steel at -105°C.

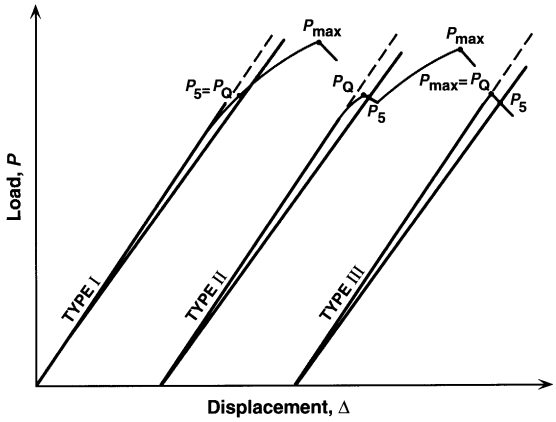


Figure 13. The three types of load-displacement behavior considered in the ASTM E-399 standard for calculating the plane-strain fracture toughness, K_{IC}.

plane strain fracture toughness. This requirement ensures that the extent of the slow crack growth prior to crack instability is insignificant. If this requirement is not fulfilled, the fracture toughness, K_C, is considered to be an apparent fracture toughness. Similar to type A behavior, J_C and δ_C can be calculated from the K_C-value. Note that while for type A behavior the K_C-value is calculated from the measured J_C-value, in type B behavior the J_C-value should be calculated from the measured K_C-value. Figure 14 presents an actual load-displacement curve for a SENB specimen of an ultrahigh strength steel (quenched and tempered 4150 steel) tested at room temperature. In contrast to the A36 steel (see Fig. 12), this material has a very high yield strength and slow crack growth occurs easily in it. Therefore, the load-displacement curve becomes nonlinear within the small-scale yielding regime. As demonstrated by the calculations shown on Figure 14, a valid K_{IC}-value could be calculated based on the ASTM requirements.

If fast fracture is not observed upon loading of the specimen, the fracture behavior can be analyzed using the *K-R* curve method. With reference to Figure 4, at each point, $P_{(i)}$ and $\Delta_{(i)}$, the compliance $C_{(i)}$ and the stress intensity factor $K_{(i)}$ can be calculated. The average crack length at each point, $a_{(i)}$, is estimated from the relationship between crack length and compliance for a given material and geometry (see Appendix B). The accuracy of the calculation can be evaluated by comparing the initial (a_0) and the unloading point (a_u) (or the point of the onset of the fast fracture) crack lengths, with the lengths measured directly on the fracture surface. The crack length used for the calculation of $K_{(i)}$ can be corrected for crack advancement due to plasticity by adding the plastic zone size (see Equation 10) to the crack length (ASTM, 1997d). This crack length is known as the effective crack length. The

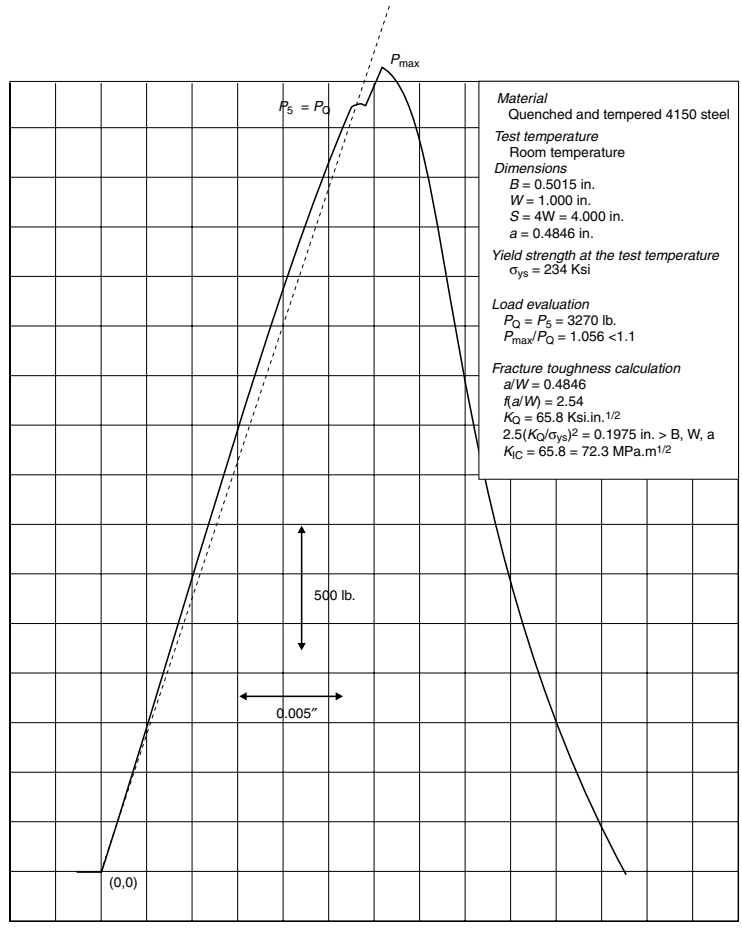


Figure 14. Actual load versus mouth-opening displacement curve obtained from testing a SENB specimen of a quenched and tempered 4150 steel at room temperature.

calculated $K_{(i)}$-values are plotted as a function of crack growth, $\Delta a = a_{(i)} - a_0$. As long as the crack propagates under small-scale yielding, J- or G-values can also be calculated from the K values using Equation 2, and hence, the J-R curve (or G-R curve), can be constructed. In the absence of fast fracture, the toughness at the instability point, K_C, may be found as demonstrated schematically in Figure 7. It should be noted that the applied G usually does not vary linearly with the crack length. Another parameter that can be calculated from the K-R curve (or J-R curve) is the toughness at the point of slow crack initiation. Since the exact point of crack initiation may be difficult to identify, an offset method, similar to the determination of the yield strength during tensile testing, is usually used. For example, an offset line at 0.2-mm crack growth can be chosen as the point of crack initiation. If the specimen geometry is such that the crack grows under the plane strain condition and the requirements set forth by the ASTM E-813 standard (ASTM, 1997e,f,) are met, then the fracture toughness value is reported as J_{IC}. This value, similar to K_{IC}, is believed to be a material property and independent of the specimen geometry and dimensions.

Type C

In many cases both plasticity and slow crack growth contribute to the non-linearity of the load-displacement curve. Consequently, it cannot be assumed that the unloading path would extend to the origin. Therefore, it is necessary to physically measure the crack length on various points on the load-displacement curve as was described in the section on crack extension measurement techniques. The best method for evaluation of the toughness for this type of behavior is the J-integral approach. At a point $[P_{(i)},\Delta_{(i)}]$, the crack length, $a_{(i)}$, is measured. From $a_{(i)}$ and $P_{(i)}$ the $K_{(i)}$ value is calculated. The $J_{p(i)}$ is calculated from the relationships given in Appendix B for single-edge notched bend (SENB) and compact tension (CT) specimens, and the total J is obtained using Equation 13. The J-R curve is established by plotting J as a function of crack extension. Similar to type B, the crack initiation point is established by an offset method. The fulfillment of ASTM E-1737 standard requirements (ASTM, 1997j) leads to the calculation of J_{IC}, which is a measure of fracture toughness at the point of crack initiation. The slope of the J-R curve, dJ/da, is representative of the resistance of the material to crack extension, and has been used to define a parameter called the tearing modulus, which is related to the tearing instability in materials exhibiting type C behavior (Paris et al., 1979).

Fracture, like all instability phenomena, is probabilistic in nature. Another characteristic of fracture is that it propagates discontinuously, i.e., fracture occurs as isolated microcracks that eventually join each other. Material processing defects, microstructure, and measurement inaccuracies contribute to the scatter in toughness data. Processing flaws—e.g., voids in sintered ceramics or solidification defects in cast or welded metallic components, can promote unstable fracture. Within the microstructure of a material, there are features that act as microcrack initiation sites. These features are usually the weakest component of the microstructure, for example, grain boundaries in ceramics, second phases in metallic alloys, or fibers in composites. In semibrittle materials plasticity precedes fracture. The inhomogenity in plastic deformation may result in formation of deformation bands, which locally raise the stresses high enough for initiation of brittle fracture (Ebrahimi and Shrivastava, 1998). In materials where unstable fracture occurs after some slow crack growth, the distribution of the stable microcracks at the crack tip influences the probability of unstable fracture (Ebrahimi and Seo, 1996). As mentioned previously, specimen dimensions affect the stress distribution at the crack tip, and consequently the toughness. Following the stringent requirements of the ASTM standards reduces the toughness variation but does not completely eliminate the scatter (Wallin, 1984; Ebrahimi and Ali, 1988). The scatter is particularly large near the ductile-to-brittle transition temperature of materials (Ebrahimi, 1988).

The number of specimens needed for toughness evaluation depends on the type of the material. Brittle materials, such as ceramics, intermetallics, and glasses show a wide scatter in data and require a large number of specimens ($\sim$10) per condition. Usually the Weibull distribution function (Weiss, 1971) is used to evaluate the variation in toughness of brittle materials. Three specimens per condition suffice for evaluation of fracture toughness of less brittle materials such as steels.

SAMPLE PREPARATION

The choice of geometry and size of the specimen depends on the availability of the material form (e.g., plate, tube, disc, or rod), toughness level (the tougher the material the larger the specimen size required to guarantee plane-strain and/or small-scale-yielding conditions), and the desired information to be extracted from the test (e.g., plane strain fracture toughness or R-curve behavior). Theoretically, one can use any notched specimen geometry as long as the compliance of the sample (or the Y function in Equation 9) as a function of crack length is known. Establishing this relationship requires loading of many specimens of the same size with different crack lengths from a material of known elastic modulus. However, these functions have been accurately established for many geometries (Sih, 1973). Figure 15 presents the two most widely used specimen geometries, single edge-notched bend (SENB) and CT specimens. The SENB specimen can be loaded in three points or in four points. The latter loading mode creates less constraint at the crack tip and is suggested for relatively brittle materials. The SENB specimens are easier to machine than the CT specimens. However, the latter specimens are more suitable for determination of the R-curve behavior. Other common specimen geometries include wedge-loaded compact tension, disc-shaped compact tension, arc-shaped tension, double-notched tensile, double-cantilever beam (tapered and straight) specimens, and center-cracked panels.

The specimen thickness and crack length size limitations are based on achieving contained small-scale yielding and plane-strain conditions at the crack tip. In many

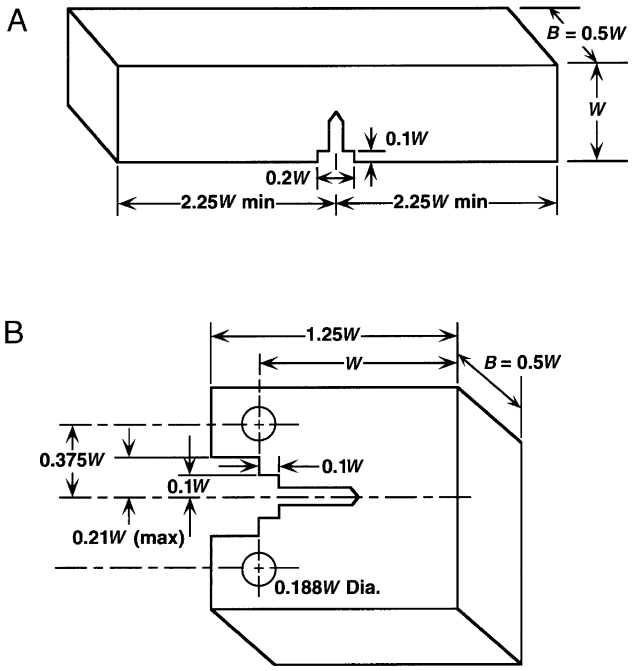


Figure 15. Standard (**A**) single-edge notch bend, SENB, and (**B**) compact tension, CT, specimens.

cases the size of the available material dictates the type of fracture toughness evaluation that can be performed. In order for the results obtained by plane strain fracture toughness testing to be valid according to the ASTM E-399 standard (ASTM, 1997c), the following requirement should be satisfied:

$$a \text{ and } B \geq 2.5(K_{IC}/\sigma_{ys})^2 \qquad (14)$$

For low-strength and tough materials, very large samples are needed to satisfy the above requirement. Usually, if the thickness is smaller than the above requirement, the load-displacement curve shows nonlinearity. In this case the toughness can be evaluated using J-integral or CTOD concepts.

The notch length, mouth opening, and shape should be carefully chosen, particularly for materials that may show extensive permanent deformation at the crack tip prior to crack extension. For example, short and wide notches may lead to the extension of plastic deformation to the surface of the specimen. According to the ASTM standards the a/W ratio is usually required to be within 0.45 to 0.55 (up to 0.7 for J-integral testing). The notch front can be straight through or have a chevron form. The latter is suitable for metallic materials in which it is difficult to initiate fatigue precracks. The loading direction as well as the position of the notch becomes important in anisotropic materials. Many processing methods, such as casting and forming, create directionality in the microstructure; therefore, the toughness of the material depends on the position of the crack front. In fiber-reinforced or laminated composites, the toughness varies considerably with the orientation of the notch. Similarly, the properties of single crystals depends on the crystallographic orientation.

Specimens are usually machined using conventional techniques. EDM (electric discharge machine) is an easy and accurate method to machine small metallic specimens. Since EDM cuts by locally melting the material, the damaged surface layer should be removed by mechanical polishing. Samples should have a fine finish, particularly on the sides. The specimens can be machined with built-in knife-edges for clipping the displacement gauge in between them, or knife-edges can be screwed on the face of the specimen (ASTM, 1997c).

As mentioned previously, the stress levels and the degree of plastic relaxation depend on the sharpness of the crack tip. For metallic materials, a sharp crack is produced by cyclic loading (fatigue) of notched specimens in tension-tension, with the maximum load much lower than the fracture load. In the case of plastics, a sharp crack tip can be produced by tapping or sliding a fresh razor blade in the notch. Sharp cracks are difficult to produce in brittle materials such as ceramics. Conventionally, unnotched bend specimens are precracked by using Vickers or Knoop indenters (Chantikul et al., 1981). Compression-compression fatigue loading has also been used to create sharp cracks in ceramics (Ewart and Suresh, 1987). Sharp straight crack fronts are also difficult to obtain in anisotropic and/or inhomogeneous materials such as single crystals and composites. Techniques using chevron-notched double torsion, and tapered double-cantilever beam specimens, are suggested as appropriate methods for measuring the toughness of brittle and composite materials (Freeman, 1984).

PROBLEMS

Most of the problems that may lead to misleading results have been mentioned throughout this article and explained in detail in the relevant ASTM standards. In general, the load and displacement measuring devices should be calibrated frequently. Load measurement and control of the applied displacement becomes crucial when dealing with small brittle specimens. If testing is conducted at a temperature other than room temperature, the thermocouple should be placed as close as possible to the notch tip without interfering with the stresses at the crack tip.

An important source of error is the alignment of the specimens, particularly bend specimens. Articulated bend fixtures with aligning guides are essential for obtaining accurate toughness values for brittle materials. Usually, if the alignment is not good, the crack may start preferentially on one corner of the specimen and/or deviate from the notch plane. A misaligned specimen experiences loading modes other than mode 1, and consequently the fracture toughness measured will be higher than the true mode 1 toughness.

A soft (compliant or not very stiff) fixture can cause premature fracture during fracture toughness testing by releasing elastic energy upon cracking (Paris et al., 1979). This problem usually arises in testing large specimens of relatively tough materials, whose fracture occurs at large

loads. To avoid this complication, short bulky fixtures should be used. For example, the rod that connects the load cell to the grip should be made of a high-elastic-modulus material, with maximum diameter and minimum length.

A frequent problem with fracture toughness testing is to obtain a straight, sharp crack tip. Fatigue precracks tend to be much shorter at the sides than at the mid-thickness (thumbnail shape). This is due to a change in the stress state from plane stress on the surfaces to plane strain in the midsection. Side grooving of the specimen usually helps with this problem. Uneven fatigue cracks may develop due to misalignment of the fixture or owing to the presence of a surface layer due to special surface treatments. If a sharp crack cannot be made, then it is understood that the toughness measured depends on the notch root radius and how it is made. At very small notch root radii (approximately less than the distance between microstructural features that cause crack initiation), the fracture toughness is not very sensitive to notch root radius; however at larger values the toughness measured can be much higher than the true value measured with a sharp crack.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation under contract No. DMR 9527624.

LITERATURE CITED

Anderson, W. E. 1969. An Engineer Views Brittle Fracture History. Boeing Report.

ASTM. 1997a. Standard Methods for Notched Bar Impact Testing of Metallic Materials: ASTM E-23. *In* Annual Book of ASTM Standards, Vol. 03.01. American Society for Testing and Materials, Philadelphia.

ASTM. 1997b. Annual Book of ASTM Standards, Vol. 03.01. American Society for Testing and Materials, Philadelphia.

ASTM. 1997c. Standard Methods for Plane-Strain Fracture Toughness of Metallic Materials: ASTM E-399. *In* Annual Book of ASTM Standards, Vol. 03.01. pp. 408–438. American Society for Testing and Materials, Philadelphia.

ASTM. 1997d. Standard Practice for R-Curve Determination: ASTM E-561-94. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 489–501. American Society for Testing and Materials, Philadelphia.

ASTM. 1997e. Standard Test Method for Crack Strength of Slow-Bend Precracked Charpy Specimens of High-Strength Metallic Materials: ASTM E-812-91. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 623–626. American Society for Testing and Materials, Philadelphia.

ASTM. 1997f. Standard Test Method for J_{IC}, A Measure of Fracture Toughness: ASTM E-813-89. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 627–641. American Society for Testing and Materials, Philadelphia.

ASTM. 1997g. Standard Practice for Determination of Fracture Toughness of Steels Using Equivalent Energy Methodology: ASTM E-992-84. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 693–698. American Society for Testing and Materials, Philadelphia.

ASTM 1997h. Standard Test Method for Plane-Strain (Chevron-Notch) Fracture Toughness of Metallic Materials: E-1304-97. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 838–848. American Society for Testing and Materials, Philadelphia.

ASTM. 1997i. Standard Test Methods for Plane-Strain Fracture Toughness and Strain Energy Release Rate of Plastic Materials: ASTM D-5045-96. *In* Annual Book of ASTM Standards, Vol. 08.01, pp. 313–321. American Society for Testing and Materials, Philadelphia.

ASTM. 1997j. Standard Test Method for J-Integral Characterization of Fracture Toughness: ASTM E-1737-96. *In* Annual Book of ASTM Standards, Vol. 03.01, pp. 968–991. American Society for Testing and Materials, Philadelphia.

ASTM. 1997k. Standard Test Method for Crack-Tip Opening Displacement (CTOD) Fracture Toughness Measurement: ASTM E-1290-93. Annual Book of ASTM Standards, Vol. 03.01, pp. 752–761. American Society for Testing and Materials, Philadelphia.

ASTM. 1997l. Standard Test Methods for Sharp-Notch Tension Testing with Cylindrical Specimens, ASTME 602-91. In Annual Book of ASTM Standards, Vol. 03.01. pp. 508–514, American Society for Testing and Materials, Philadelphia.

Begley, J. A. and Landes, J. D. 1972. The J-integral as a fracture criterion. *In* Fracture Mechanics, ASTM STP 514, pp. 1–20. American Society for Testing and Materials, Philadelphia.

Broek, D. 1982. Elementary Engineering Fracture Mechanics. Martinus Nijhoff Publishers, Boston.

British Standard Institution. 1979. Methods for Crack Opening Displacement (COD) Testing: BS5762:1979. British Standards Institution, Hemel Hempstead, U.K.

Chantikul, P., Anstis, G. R., Lawn, B. R., and Marshall, D. B. 1981. A critical evaluation of indentation techniques for measuring fracture toughness. *J. Am. Ceramic Soc*. 64:533–538.

Ebrahimi, F. 1988. A study of crack initiation in the ductile-to-brittle transition regime. *In* Fracture Mechanics, ASTM STP 945, pp. 555–580. American Society for Testing and Materials, Philadelphia.

Ebrahimi, F. and Ali, J. A. 1988. Evaluation of published data on ductile initiation fracture toughness of low-alloy structural steels. *J. Testing Evaluation* 16:113–123.

Ebrahimi, F. and Seo, H. K. 1996. Ductile crack initiation in steels. *Acta Mater*. 44:831–843.

Ebrahimi, F. and Shrivastava, S. 1998. Brittle-to-ductile transition in NiAl single crystals. *Acta Mater.* 46:1493–1502.

Eshelby, J. D. 1974. Calculation of energy release rate. *In* Prospects of Fracture Mechanics (G.C. Sih, H. C. van Elst, and D. Broek, eds.) pp. 69–84. Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands.

Ewart, L. and Suresh, S. 1987. Crack propagation in ceramics under cyclic loads. *J. Mater. Sci.* 22:1173–1192.

Freeman, S. W. 1984. Brittle fracture behavior of ceramics. *Ceramic Bull*. 67:392–401.

Griffith, A. A. 1920. The phenomena of rupture and flow in solida. *Philos. Trans. R. Soc. London* 221:163–198.

Irwin, G. R. 1948. Fracture dynamics. *In* Fracturing of Metals, pp. 147–166. American Society for Metals, Materials Park, Ohio.

Irwin, G. R. 1957. Analysis of stresses and strains near the end of a crack traversing a plate. *J. Appl. Mech*. 24:361–364.

Knott, J. F. 1973. Fundamentals of Fracture Mechanics. John Wiley & Sons, New York.

Lin, I.-H., Anderson, T. L., deWit, R., and Dawes, M. G. 1982. Displacement and rotational factors in single-edge notched-bend specimens. *Int. J. Fracture* 20:R3.

McMahon, C. J., Jr. and Vietek, V. 1979. Effects of segregated impurities on intergranular fracture energy. *Acta Metall.* 27:507–515.

Off-Shore Welded Structures. 1982. 2nd International Conference Proceedings 1982. Welding Institute, London.

Orowan, E. 1955. Energy criteria of fracture. *Welding J.* 34:157–160.

Paris, P. C. and Sih, G. C. 1965. Stress analysis of cracks. *ASTM STP* 391:30–81.

Paris, P. C., Tada, H., Zahoor, A., and Ernst, H. 1979. The theory of the instability of the tearing mode of elastic-plastic crack growth. *ASTM STP* 668:5–36.

Paris, P. C., Tada, H., Ernst, H., and Zahoor, A. 1979. Initial experimental investigation of tearing instability theory. *STM STP* 668:251–265.

Parker, E. R. 1957. Brittle Behavior of Engineering Structures, National Research Council, Ship Structure Committee Report. John Wiley & Sons, New York.

Rice, J. R. 1968. A path independent integral and the approximate analysis of strain concentrations by notches and cracks. *J. Appl. Mech.* 35:379–386.

Ritchie, R. O. 1988. Mechanisms of fatigue crack propagation: Role of crack tip shielding. *Mater Sci. Eng.* 103:15–28.

Rolfe, S. T. and Barsom, J. M. 1977. Fracture and Fatigue Control in Structures: Application of Fracture Mechanics. Prentice-Hall, Englewood Cliffs, N.J.

Sih, G. C. 1973. Hand Book of Stress-Intensity Factors. Lehigh University, Institute of Fracture Mechanics, Bethlehem, Penn.

Turner, C. E. 1979. Methods for post-yield fracture safety assessment. *In* Methods for Post-Yield Fracture Mechanics, pp. 23–210. Applied Science Publishers, London.

Wallin, K. 1984. The size effect in K_{IC} results. *Eng. Fracture Mech.* 196:1085–1093.

Weiss, V. 1971. Notch analysis of fracture. *In* Fracture: An advanced Treatise, Vol. III. (H. Liebowitz, ed.), pp. 227–264. Academic Press, New York.

Wells, A. A. 1961. Unstable crack propagation in metals-cleavage and fast fracture. *In* Proceedings of Crack Propagation, Vol. 1, Paper B4, pp. 210–230. Canfield, United Kingdom.

Westergaard, H. M. 1939. Bearing pressures and cracks. *J. Appl. Mech.* 61:A49–A53.

KEY REFERENCES

ASTM, 1997b. See above.

This volume of ASTM standards includes all the testing techniques that have been discussed in this article.

Broeck, 1982. See above.

This book provides an introduction to the principles of fracture mechanics with an emphasis on metallic materials.

Lawn, B. 1993. Fracture of Brittle Solids, 2nd ed. Cambridge University Press, Cambridge.

This is an advanced book that presents a continuum as well as an atomistic approach to fracture mechanics, with an emphasis on highly brittle materials, principally ceramics.

APPENDIX A: STRESS FIELD OF A SHARP CRACK

This appendix presents the stress field of a sharp crack at the distance r and the angle θ (see Figure 8) for the three modes of loading (see Fig. 1).

Mode 1

$$\sigma_x = \frac{K_1}{(2\pi r)^{1/2}}\cos\frac{\theta}{2}\left[1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right] \tag{15}$$

$$\sigma_y = \frac{K_1}{(2\pi r)^{1/2}}\cos\frac{\theta}{2}\left[1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right] \tag{16}$$

$$\tau_{xy} = \frac{K_1}{(2\pi r)^{1/2}}\sin\frac{\theta}{2}\cos\frac{\theta}{2}\cos\frac{3\theta}{2} \tag{17}$$

$$\sigma_z = \nu(\sigma_x + \sigma_y) \tag{18}$$

$$\tau_{xz} = \tau_{yz} = 0 \tag{19}$$

$$u = \frac{K_1}{G}\left[\frac{r}{2\pi}\right]^{1/2}\cos\frac{\theta}{2}\left[1 - 2\nu + \sin^2\frac{\theta}{2}\right] \tag{20}$$

$$v = \frac{K_1}{G}\left[\frac{r}{2\pi}\right]^{1/2}\sin\frac{\theta}{2}\left[2 - 2\nu + \cos^2\frac{\theta}{2}\right], \quad w = 0 \tag{21}$$

Mode 2

$$\sigma_x = -\frac{K_2}{(2\pi r)^{1/2}}\sin\frac{\theta}{2}\left[2 + \cos\frac{\theta}{2}\cos\frac{3\theta}{2}\right] \tag{22}$$

$$\sigma_y = -\frac{K_2}{(2\pi r)^{1/2}}\sin\frac{\theta}{2}\cos\frac{\theta}{2}\cos\frac{3\theta}{2} \tag{23}$$

$$\tau_{xy} = -\frac{K_2}{(2\pi r)^{1/2}}\cos\frac{\theta}{2}\left[1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right] \tag{24}$$

$$\sigma_z = \nu(\sigma_x + \sigma_y) \tag{25}$$

$$\tau_{xz} = \tau_{yz} = 0 \tag{26}$$

$$u = \frac{K_2}{G}\left[\frac{r}{2\pi}\right]^{1/2}\sin\frac{\theta}{2}\left[2 - 2\nu + \cos^2\frac{\theta}{2}\right] \tag{27}$$

$$v = \frac{K_2}{G}\left[\frac{r}{2\pi}\right]^{1/2}\cos\frac{\theta}{2}\left[-1 + 2\nu + \sin^2\frac{\theta}{2}\right], \quad w = 0 \tag{28}$$

Mode 3

$$\tau_{xz} = -\frac{K_3}{(2\pi r)^{1/2}}\sin\frac{\theta}{2} \tag{29}$$

$$\tau_{yz} = -\frac{K_3}{(2\pi r)^{1/2}}\cos\frac{\theta}{2} \tag{30}$$

$$\sigma_x = \sigma_y = \sigma_z = \tau_{xy} = 0 \tag{31}$$

$$w = \frac{K_3}{G}\left[\frac{2r}{\pi}\right]^{1/2}\sin\frac{\theta}{2}, \quad u = v = 0 \tag{32}$$

APPENDIX B: RELATIONSHIPS FOR CALCULATING *K* AND *J* FOR SENB AND CT SPECIMENS

This appendix provides the relationships needed for calculation of K and J at a given point ($P_{(i)}$, $\Delta_{(i)}$) and the associated crack length, a, for SENB (loaded in three points) and CT specimens (ASTM, 1997c,j).

SENB Specimen (Three-Point Bending)

Calculation of K(i)

$$K_{(i)} = (P_{(i)}S/BW^{3/2}) \cdot f(a/W) \tag{33}$$

where:

$$f(a/W) = \frac{3(a/W)^{1/2}[1.99 - (a/W)(1 - a/W)(2.15 - 3.93a/W) + 2.7a^2/W^2]}{2(1 + 2a/W)(1 - a/W)^{3/2}} \quad (34)$$

Calculation of Compliance

$$C_{(i)} = \frac{\Delta_{(i)}}{P_{(i)}} = (S/E'BW) \cdot q(a/W) \quad (35)$$

where:

$$q(a/W) = 6(a/W)[0.76 - 2.28(a/W) + 3.87(a/W)^2 - 2.04(a/W)^3 + 0.66/(1 - a/W)^2] \quad (36)$$

Calculation of Crack Length

$$a_{(i)}/W = 0.997 - 3.95D + 2.982D^2 - 3.214D^3 + 51.52D^4 - 113.0D^5 \quad (37)$$

where:

$$D = \{1 + [(E'B\Delta_{(i)}/P_{(i)})(4W/S)]^{1/2}\}^{-1} \quad (38)$$

Calculation of $J_{(i)}$

$$J_{(i)} = \frac{(K_{(i)})^2}{E'} + J_{p_{(i)}} \quad (39)$$

$$J_{p_{(i)}} = \left[J_{p_{(i-1)}} + \left(\frac{2}{b_{(i-1)}}\right)\left(\frac{U_{p_{(i)}} - U_{p_{(i-1)}}}{B}\right)\right]\left[1 - \frac{a_{(i)} - a_{(i-1)}}{b_{(i-1)}}\right] \quad (40)$$

where:

$$b_{(i-1)} = W - a_{(i-1)} \quad (41)$$

Compact Tension (CT) Specimen

Calculation of $K_{(i)}$

$$K_{(i)} = (P_{(i)}/BW^{1/2}) \cdot f(a/W) \quad (42)$$

where:

$$f(a/W) = \frac{(2+a/W)[0.866 + 4.64a/W - 13.32(a/W)^2 + 14.72(a/W)^3 - 5.6(a/W)^4]}{(1 - a/W)^{3/2}} \quad (43)$$

Calculation of Compliance

$$C_{(i)} = \frac{\Delta_{(i)}}{P_{(i)}} = (1/E'B) \cdot q(a/W) \quad (44)$$

where:

$$q(a/W) = [19.75/(1 - a/W)^2][0.5 + 0.192(a/W) + 1.385(a/W)^2 - 2.919(a/W)^3 + 1.842(a/W)^4] \quad (45)$$

Calculation of Crack Length

$$a_{(i)}/W = 1.000 - 4.500D + 13.157D^2 - 172.551D^3 + 879.944D^4 - 1514.671D^5 \quad (46)$$

where:

$$D = 1/[1 + (E'B\Delta_{(i)}/P_{(i)})^{1/2}] \quad (47)$$

Calculation of $J_{(i)}$

$$J_{(i)} = \frac{(K_{(i)})^2}{E'} + J_{p_{(i)}} \quad (48)$$

$$J_{p_{(i)}} = \left[J_{p_{(i-1)}} + \left(\frac{\eta_{(i-1)}}{b_{(i-1)}}\right)\left(\frac{U_{p_{(i)}} - U_{p_{(i-1)}}}{B}\right)\right] \times \left[1 - \gamma_{(i-1)}\left(\frac{a_{(i)} - a_{(i-1)}}{b_{(i-1)}}\right)\right] \quad (49)$$

where:

$$\eta_{(i-1)} = 2.0 + 0.522b_{(i-1)}/W \quad (50)$$

and

$$\gamma_{(i-1)} = 1.0 + 0.76b_{(i-1)}/W \quad (51)$$

and

$$b_{(i-1)} = W - a_{(i-1)} \quad (52)$$

APPENDIX C: GLOSSARY OF TERMS AND SYMBOLS

P	load
B	thickness
W	width
a	crack length
$K_{(i)}$	stress intensity factors (see Appendix A)
S	load span (for SENB specimens)
Δ	mouth opening displacement (measured on the surface. If a attachable knife edges are used, as shown in Fig. 11, the mouth opening should be corrected for the height, Z.)
E'	effective Young's modulus (= E plane-stress condition, = $E/(1 - \nu^2)$ plane-strain condition
ν	Poisson's ratio
$Up_{(i)}$	the plastic component of the area under the load displacement curve (see Fig. 9)

Note that the $Up_{(i)}$ should be calculated from the load-line displacement rather than the mouth opening displacement. Therefore, the SENB specimen may require two displacement gauges. A load-line displacement measurement is required for J computation. The crack mouth opening displacement can be used to estimate the crack length using the elastic compliance technique.

FERESHTEH EBRAHIMI
University of Florida
Gainesville, Florida

HARDNESS TESTING

INTRODUCTION

Static indentation hardness tests such as Brinell, Rockwell, Vickers, and Knoop are frequently used methods for determining hardness. The basic concept utilized in all of these tests is that a set force is applied to an indenter in order to determine the resistance of the material to penetration. If the material is hard, a relatively small or shallow indentation will result, whereas if the material is soft, a fairly large or deep indentation will result.

These tests are often classified in one of two ways: either by the extent of the test force applied or the measurement method used. A "macro" test refers to a test where a load >1 kg is applied; similarly "micro" refers to a test where a load of ≤ 1 kg of force is applied. Additionally, some instruments are capable of conducting tests with loads as light as 0.01 g and are commonly referred to as ultralight or nanoindentation testers. Rockwell and Brinell testers fall into the macro category, whereas Knoop testers are used for microindentation tests. Vickers testers are employed for both macro and microindentation tests. The measurement methods available include a visual observation of the indentation or a depth measurement of the indentation. Rockwell and some nanoindentation testers are capable of determining the depth of the indentation, whereas Brinell, Knoop, and Vickers testers require an indentation diameter measurement. These visual measurements can be automated, as will be discussed later in this unit.

Hardness is not a fundamental property of a material, yet hardness testing is considered a useful quality-control tool. Many properties are predicted from hardness values when combined with additional information such as alloy composition. The following is a list of such properties: resistance to abrasives or wear, resistance to plastic deformation, modulus of elasticity, yield strength, ductility, and fracture toughness. Some of these properties, such as yield strength, have numerical relationships with hardness values, whereas others such as fracture toughness are based on observations of cracks surrounding the indentations. Data analysis and conversions will be discussed in greater detail later in this unit.

Other relationships have developed over time by empirical observations such as the link between machinability and hardness values. In general, 300 to 350 HB (Brinell scale) is considered to be the maximum tolerable hardness for production machining of steels. For the majority of machining operations, the optimum hardness is 180 to 200 HB. If the material is too soft—$<$160 HB—poor chip formation and a poor surface finish will result. The relationship between hardness and machinability, however, is not linear. Equations were developed by Janitzky (1938) and Henkin and Datsko (1963) for determining machinability if the Brinell hardness and tensile reduction of area are known.

Competitive and Related Techniques

Many techniques have been used historically to determine hardness. The tests focused on here—static indentation hardness test methods—are widely used because of the ease of use and repeatability of the technique. Rebound and ultrasonic tests are the next most common, due to portability. Several hardness techniques are listed below with an emphasis placed on either the specific applications for which these were developed or the limitations of these techniques in comparison to static indentation tests.

Rebound tests, routinely done with Scleroscope testers, consist of dropping a diamond-tipped hammer, which falls inside a glass tube under the force of its own weight from a fixed height, onto the test specimen and reading the rebound travel on a graduated scale. The advantage of such a method is that many tests can be conducted in a very short time. However, there are several limitations to consider. The column must be in an upright position, so that even if the tester is portable it must be positioned correctly. While newer testers have a digital readout, on the older models the height of the rebound had to be closely observed by the operator (Boyer et al., 1985).

In ultrasonic microhardness testing, a Vickers diamond is attached to one end of a magnetostrictive metal rod. The diamond-tipped rod is excited to its natural frequency. The resonant frequency of the rod changes as the free end of the rod is brought into contact with the surface of the test specimen. The area of contact between the indenter and the test material can be determined by the measured frequency. However, the Young's modulus of the material must be known in order to accomplish this calculation. Only a small indent is left on the surface, so the test is classified as nondestructive. The disadvantage of this is that it is difficult to confirm the exact location of the test (Meyer et al., 1985).

One of the earliest forms of hardness testing, scratch testing, goes back to Reaumur in 1722. His scale of testing consisted of a scratching bar, which increased in hardness from one end to the other. The degree of hardness was determined by the position on the bar that the metal being tested would scratch (Boyer, 1987). The next development was the Mohs scale, which has a series of ten materials used for comparison ranging from diamond with a hardness of 10, to talc with a hardness of 1 (Petty, 1971). Further developments include a file test where a series of hardened files of various Rockwell C values (HRC values; see Table 1) are used to determine the relative

Table 1. Standard Static Indentation Hardness Tests

	ASTM	ISO
Brinell	E10	6506, 156, 726, 410
Rockwell	E18	6508, 1024, 716, 1079, 674, 1355
Vickers	E92, E384	6507, 640, 146, 409
Knoop	E384	4545, 4546, 4547

surface hardness. With this particular test, it is up to the operator to determine how much pressure to apply, at what speed to drag the file, and the angle at which to hold the file. A more controlled method was developed which uses a diamond tip and a set force on a mechanical arm to drag across the material. The width of the resulting groove is examined to determine the hardness (Bierbaum, 1930). The advantage is that a single trace can be made through a microstructure and the relative hardness of the different phases and constituents can be assessed. For example, variation at grain boundaries or case-hardened surfaces would be observed. However, it is more difficult to relate this information to other properties or hardness scales.

Abrasion and wear tests are used to evaluate the life of a component under service conditions. Typically, abrasive is applied to the surface by various means such as a rotating disc, an abrasive and lubricant mixture, or steel shot impinged at a known velocity (see TRIBOLOGICAL AND WEAR TESTING). The material-removal rate is monitored to determine the hardness (Khruschov, 1957; Richardson, 1967). This method is explained in detail in TRIBOLOGICAL AND WEAR TESTING.

Instrumented indentation is one of the newer developments in hardness testing. This method takes dynamic hardness testing one step further. Not only is a loading and unloading curve developed, but also a continuous stiffness measurement is conducted throughout the time of contact between the indenter and the material. The record of the stiffness data along with the load displacement data allows the hardness and Young's modulus to be calculated as a function of depth (Oliver et al., 1992). This method is under development as are standards for the methodology.

PRINCIPLES OF THE METHOD

The basis of static indentation tests is that an indenter is forced into the surface of the material being tested for a set duration. When the force is applied to the test piece through contact with the indenter, the test piece will yield. After the force is removed, some plastic recovery in the direction opposite to the initial flow is expected, but over a smaller volume. Because the plastic recovery is not complete, biaxial residual stresses remain in planes parallel to the free surface after the force is removed. The hardness value is calculated by the amount of permanent deformation or plastic flow of the material observed relative to the test force applied. The deformation is quantified by the area or the depth of the indentation. The numerical relationship is inversely proportional, such that as the indent size or depth increases, the hardness value decreases.

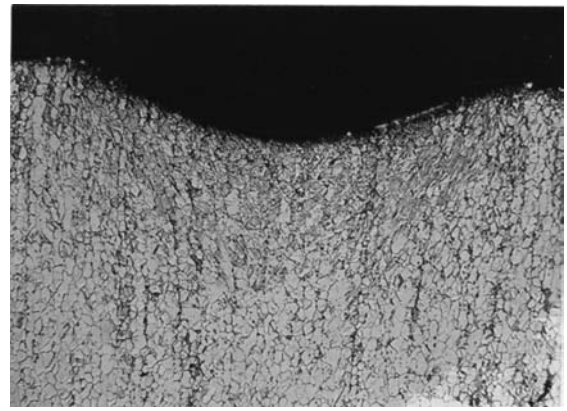


Figure 1. The cross-section of an indentation in a brass specimen demonstrates the deformation and material flow that occurs as the result of the applied force. Aqueous ferric chloride etch, 100× magnification.

The hardness is derived from two primary components: (1) a constraint factor for the test and (2) the uniaxial flow stress of the material being tested. The value of the constraint factor depends mainly on the shape of the indenter used in the hardness test. For relatively blunt indenters such as Brinell, Vickers, and Knoop, the constraint factor is approximately three. Prandtl first explained the origin of the constraint factor (Prandtl, 1920). He compared the blunt hardness indenters to the mean stress on a two-dimensional punch required for the onset of plastic flow beneath the punch. The material beneath the punch was assumed to flow plastically in plane strain, and the material surrounding the flow pattern was considered to be rigid. Hill generalized Prandtl's approach into what is now known as the slip line field theory. Hill calculated a constraint factor very similar to Prandtl. According to these theories, the material displaced by the punch is accounted for by upward flow (Hill et al., 1947). Both calculated values closely match the empirical data. The photomicrograph in Figure 1 demonstrates the stress and flow observed in the region around an indentation.

Hardness values can be directly compared only if the same test is used, since the geometry of the indenter and force applied influence the outcome of the test. For each type of hardness test conducted, a different equation is used to convert the measured dimension, depth or diameter, to a hardness value. The Brinell hardness value is calculated by dividing the test force by the surface area of the indentation. The test parameters taken into account are the test force and ball diameter while the indentation diameter is measured. For Rockwell tests, the hardness value is determined by the depth of indentation made by a constant force impressed upon the indenter. The test parameters taken into account are the test force (major and minor load) and the indenter geometry (ball or diamond cone), while the depth of penetration between the application of the minor load and major load is measured. Vickers hardness values are calculated in the same manner as Brinell tests. The projected area, instead of the surface area, is used when computing Knoop values. The

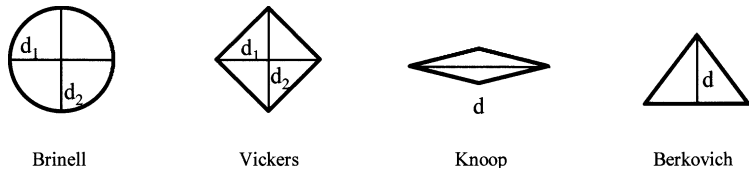


Figure 2. For the tests described in this unit, the hardness values are calculated based on the diameter of the indentation, *d*, which is measured differently for the different tests. Note that for the Brinell and Vickers tests, that diameter is an average of two measurements. See the Appendix for the equations in which these numbers are used.

test parameters taken into account for Vickers and Knoop tests are identical and include the test force and diamond indenter geometry while the indentation diameter is measured. In addition some ultralight tests are conducted with a Berkovich indenter, which has its own unique geometry. The illustrations in Figure 2 demonstrate the indentations that are measured by visual means and the dimensions of interest. The equations used to calculate the hardness values can be found in the Appendix.

The majority of hardness tests are conducted to verify that a particular processing step was done correctly. Typical processes that are evaluated include annealing, hardenability, cold working, recrystallization, surface treatments, and tempering. While an etchant could be used for visual examination of the microstructure variation, there might be other factors such as chemical composition or porosity level that would also influence the hardness value. For a known composition, the hardness associated with a particular structure will vary. For example, an alloy with a carbon content of 0.69 (wt. %) and a martensitic structure would have a hardness value of 65 HRC while an alloy with a carbon content of 0.25 and a martensitic structure would have an Rockwell C value of only 47 HRC (ASTM Standard A255, 1986).

A very common application for microindentation hardness tests is in verifying or derterminging case depth as a result of a heat-treatment process—i.e., case hardening. Case hardening may be defined as a process by which a ferrous material is hardened so that a surface layer, known as the case, becomes substantially harder than the remaining material, known as the core. The graph in Figure 3 is representative of an evaluation to determine an "effective" case depth. "Effective" case depth is the depth at which 50 Rockwell C is obtained. The "total" case depth is where hardness becomes constant. Often the visual transition is observed at a depth near the total case depth.

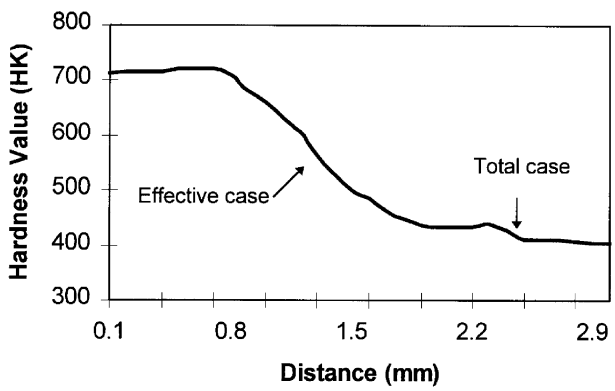


Figure 3. The effective and total case depth are noted in this typical case-depth determination graph. Even though the effective case is evaluated at 50 HRC, a Knoop (HK) test is required for this application.

PRACTICAL ASPECTS OF THE METHOD

General test methods for Brinell, Rockwell, Vickers, and Knoop tests can be found in ASTM and ISO standards. ASTM, ISO, DIN, SAE, APMI, and various other organizations also have written standards specific to certain materials, products, or processes. The standards listed in Table 1 contain information relative to the type of tester, general method, and calibration requirements.

The method described in the following paragraph has been generalized for any static indentation test. Prior to conducting the test, a specimen will typically undergo a certain level of preparation. The extent of preparation required is a function of the test force applied and is described in greater detail later in this unit (see Sample Preparation). Next, the specimen is secured through the use of a vise or an anvil. In the case of a portable tester, the tester must be secured. The main objective is to insure that the only movement observed during the course of the test is the impression of the indenter. The test force is then applied for a set duration. The measured dimension, depth, or diameter can then be used to calculate the hardness using the appropriate equation in the Appendix at the end of this unit. However, in most cases the instrument will provide a direct readout of the hardness value; otherwise a reference table containing solutions for set measurements/input values can be used.

In order to compare hardness values from one material to another, it is important for the same test conditions to be in place. Therefore, certain information needs to be provided with the hardness number. For example, 600 HBW 1/30/20 represents a Brinell hardness of 600 determined with a tungsten carbide ball 1 mm in diameter and a test force of 30 kgf (kilogram-force; 1 kgf ≡ 9.80665 N) applied for 20 s. In general, the key pieces of information to provide, in addition to the hardness value, are the test method used and test force applied (if not dictated by the method). Values such as 60 HRC or 850 HV 10, where 10 represents the test force applied in kg, are typical of the notations that would be observed on blueprints.

Of the static indentation test methods discussed, each has its advantages, intended applications, and limitations. The selection criteria to consider are as follows: hardness

Table 2. Common Applications and Nomenclature for Hardness Tests

Test	Abbreviation	Indenter	Test Load (kg)[a]	Application
Brinell	HBW	10-mm ball: tungsten carbide	3000	cast iron and steel
Brinell	HBS	10-mm ball: steel	500	copper, aluminum
Rockwell A	HRA	brale	60	very hard materials, cemented carbides
Rockwell B	HRB	$\frac{1}{16}$-in. ball	100	low-strength steel, copper alloys, aluminum alloys, malleable iron
Rockwell C	HRC	brale	150	high-strength steel, titanium, pearlitic malleable iron
Rockwell D	HRD	brale	100	high-strength steel, thin steel
Rockwell E	HRE	$\frac{1}{8}$-in. ball	100	cast iron, aluminum, and magnesium alloys
Rockwell F	HRF	$\frac{1}{16}$-in. ball	60	annealed copper alloys, thin soft metals
Superficial Rockwell T	30 T	$\frac{1}{16}$-in. ball	30	materials similar to Rockwell B, F, and G, but of thinner gauge
Superficial Rockwell N	30 N	brale	30	materials similar to Rockwell A, C, and D, but of thinner gauge
Vickers	HV	diamond	10	hard materials, ceramics, cemented carbides
Vickers	HV	diamond	0.5	all materials
Knoop	HK	diamond	0.5	all materials, case-depth determination

[a]The test load listed is specific to the application and not the only force available for the tester.

range of material to be tested, work environment, shape and size of workpiece, surface condition and whether or not the workpiece can be modified prior to testing, heterogeneity/homogeneity of the material, number of tests to be performed, and level of automation available. A majority of the factors listed above can be correlated with the test force applied and the corresponding size of the indentation.

Brinell testers can be used on a wide range of materials. When test forces in the range of 500 to 3000 kgf are applied with a 10-mm-diameter ball, a diameter will be created with an indentation between 2 and 7 mm. The large impression has its advantages with heterogeneous microstructures or segregation in that it averages out the variation. The disadvantage is that it is not sensitive enough to define a gradient in hardness and not suitable for testing small parts or thin sections. The thickness of the test piece must be ten times the depth of the indent.

Rockwell testers accommodate different materials through the use of various test forces and indenters. Each combination of indenter and force is given a specific scale designation, A to Z. For example, HRC tests are conducted with a brale indenter and 150 kgf while HRB tests are conducted with a 1/16-in. ball and 100 kgf. Superficial tests are conducted at three different forces and are designated accordingly. A 15T test is accomplished with a 1/16-in. ball and 15 kgf; likewise 30T and 45T tests use the same ball indenter with 30 and 45 kgf, respectively. Rockwell tests are used to determine bulk hardness, with the exception of superficial tests. These tests are used to evaluate coatings or surface treatments such as nitriding.

The advantage of microindentation hardness tests is the ability to monitor hardness as a function of position—e.g., placing an indent in a specific microconstituent in a multiphase material, determining case depth, or determining particle hardness. The Vickers test is considered to be relatively independent of test force; this is due to the geometrically similar impressions, as made by pyramidal indenters (Vander Voort, 1984). Testers are available that accommodate a range of test forces from 50 to 0.01 kgf. This enables direct comparison of bulk and phase-specific hardness values. Knoop indenters, on the other hand, are used most often when determining case depth, due to the elongated shape of the indenter. Table 2 lists common applications for the tests discussed (Lysaght et al., 1969; Boyer, 1987).

METHOD AUTOMATION

Automation is available on different levels for the hardness test equipment. Two types of automation are typical: (1) placement of the indentations and (2) image-analysis measurement methods. The placement of indentations can be automated with the use of a motorized stage. The most common application is with microindentation hardness tests, where as many as forty indentations may be required in a series to monitor a surface treatment. Also, the smaller specimen size lends itself more readily to being placed on a stage. Automated reading of the indentations is applied in situations where the operator would typically read the indent, for example, Knoop, Vickers, and Brinell tests. The systems are based on a specialized sequence of image analysis processes. Most systems utilize a grayscale camera system with a computer system. The indentations are detected based on either the level of contrast with the matrix material or the assumption that the indent will be made up of the darker gray pixels on the screen. The accuracy of the dimensional measurements will be based on the resolution of the system as a whole, which will be determined by a combination of the optics in the measuring device/microscope and the camera.

DATA ANALYSIS AND INITIAL INTERPRETATION

The majority of modern equipment displays the hardness value directly after the measurement has been made. For

Rockwell testers this means the value will be immediately displayed either on a digital readout or a dial. For a test where a visual measurement is conducted, the tester will either display the hardness value or simply the diameter value. In either case, most new testers are equipped with an RS232 interface either to automate the tester or to output the data for further analysis by the operator. For microindentation tests, one of the considerations is which data points to include and which are questionable. Due to the small size of the indentations, the values can be significantly altered by the presence of cracks, pores, and inclusions. One of the criteria to examine is if the shape of the indent is similar to that of the indenter. If a corner of the indent is missing or distorted or forms an asymmetrical indent, the value is highly questionable.

In some cases, creating cracks is actually the intent of the test (Palmquist, 1957). One case involves simply observing the force at which cracking begins. Typically, this method is employed when the material lacks enough ductility for other mechanical tests such as compression or tensile testing, or when there is a lack of material, since the hardness test requires only a small surface area. Crack-length observations are also used to calculate fracture toughness. A plot is constructed of the applied force versus the crack length, and a linear relationship is produced. The inverse of the slope in kg/mm is a measure of the fracture toughness.

Conversions from one scale to another are commonplace; however, a single conversion relationship does not hold true for all materials. Charts and equations are available for the following materials; nonaustenitic steels, nickel and high-nickel alloys, cartridge brass, austenitic stainless steel, copper, alloyed white iron, and wrought aluminum products in ASTM E140. Converted values should be used only when it is impossible to test the material under the condition specified.

Other properties of interest are tensile strength, yield strength, and hardenability (Siebert et al., 1977). Tensile-strength conversions have typically been developed around a particular material or class of materials. For example, with equations developed by Tabor (1951), a different constant is inserted for steel in comparison to copper. These findings have been duplicated in some cases and refuted in others (Taylor, 1942; Cahoon et al., 1971). Several other equations have been developed for specific hardness methods such as Brinell (Greaves et al., 1926; MacKenzie, 1946) or Vickers (Robinson et al., 1973). Likewise, yield strength conversions have been shown to vary with the material of interest (Cahoon et al., 1971; George et al., 1976). This observation was linked with the strain hardening coefficient. With aluminum alloys, the strain hardening coefficient is dependent on the strengthening mechanism, whereas for carbon steels the strain hardening coefficient varies directly with hardness.

SAMPLE PREPARATION

The degree of sample preparation required is inversely related to the depth of the indentation. For a relatively deep indentation, as is the case with Brinell tests, surface condition is less of a factor. The primary concern is that the indentation not be obscured when measuring the diameter. However, for a shallow indentation, a rough surface finish will cause a high level of variation in the readings. When conducting a superficial Rockwell test, for example, if the indenter were to slip into the valley of a groove, the depth measurement would be a combination of the impression and the slip, and the hardness value would be underestimated. Typically, for Rockwell tests, surface grinding to at least a 400-grit abrasive paper is recommended. For microindentation tests, such as Vickers and Knoop, rough polishing to a finish of 3 μm or better is recommended. With any test where the indentation diameter must be measured, the amount of deformation or scratches on the surface must not interfere with the operator's ability to determine the diameter. The surface finish requirements tend to be more stringent when automation is employed for the visual measurements.

SPECIMEN MODIFICATION

As a result of the material being forced aside by the indenter to make an impression in the specimen, the material surrounding the impression is disturbed and possibly work-hardened. For this reason, a minimum spacing requirement between indentations can be found for each type of hardness test in a corresponding standard. The spacing is specified in terms of indentation diameters, rather than units such as micrometers, to account for the greater amount of cold working that often occurs in soft materials that produce larger indentations. If indentations are too closely spaced, the hardness values can become erratic. For example, when a porous specimen is examined, the area around the indent is compressed as shown in Figure 4. Another test conducted in the compressed region would result in a higher hardness value. However, it is also possible for values to decrease, since, on contact with an existing indentation, the indenter may actually have less material to force aside and the result may be larger or deeper indentation.

Typically, loads are recommended that will result in an indent of sufficient size to be accurately measured, while

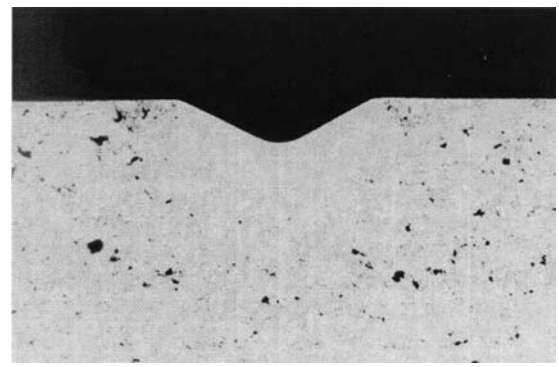


Figure 4. The cross-section of an indentation in a porous specimen demonstrating the compression of the porosity by the applied force (35× magnification).

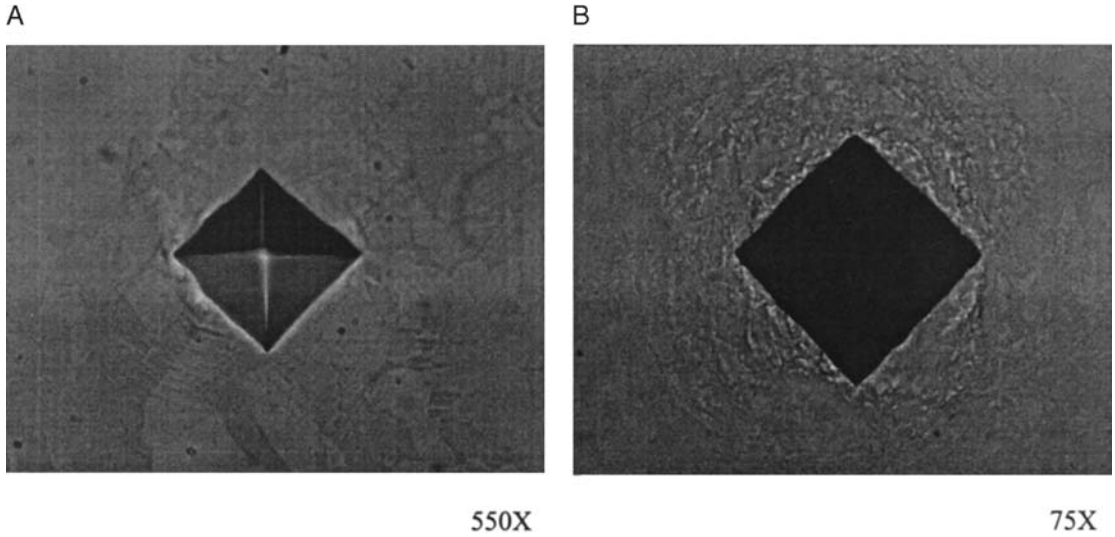


Figure 5. A comparison of the deformation around an indentation as a function of the force applied. For (**A**), a 100-g load was applied, resulting in a 41-μm-diameter indent, while for (**B**), a 10-kg load was applied, resulting in a 410-μm-diameter indent.

limiting the extent of the deformation. For example, Figure 5 displays a brass specimen with a hardness value of ~100 HV where loads of both 100 g and 10 kg have been applied. The 100-g load would be recommended. In the case of more brittle materials, such as ceramics, using too heavy a load can result in cracking of the specimen, evident at the corners of the indents, as well as chipping of the material around the indentation perimeter. Examples of materials, recommended loads, test methods, and typical hardness values are shown in Table 3.

Other concerns relate to the velocity of the indenter as it approaches the specimen, and the dwell time of the applied load. If the indenter impacts the specimen instead of applying the force, the repeatability of the test is compromised. This is also the case if the machine is set for short dwell times, since material creep rate is fastest at the beginning of the cycle. In general, creep is most readily observed during testing of low-melting-point metals at room temperature and in many metals at elevated temperatures. Hardness standards recommend a temperature range and dwell times to provide repeatable results. However, when working with low-melting point alloys and other materials more prone to creep, such as plastics, longer dwell times are suggested. In general, when creep occurs during indentation, the operator should permit the indenter to reach equilibrium before removing the load.

Table 3. Examples of Materials with Recommended Loads, Test Methods, and Typical Hardness Values[a]

	HRC (150 kgt)	HRB (100 kgf)	Brinell (10-mm steel ball)	HV (500 gf)
Nonaustenitic Steel			3000 kgf	500 gf
	60	NA	NA	697
	48	NA	451	484
				300 gf
	25	NA	253	266
	NA	93	200	200
	NA	60	107	107
Nickel and high-nickel alloys				
	36	NA	329	344
	NA	54	100	100
Cartridge brass (70% Cu/30% Zn)			500 kgf	100 gf
	NA	92.5	164	190
	NA	89	150	177
Wrought aluminum				
	NA	89	150	177
	NA	28	70	80

[a]NA, not available.

PROBLEMS

Problems are best detected by routine verification procedures. Calibrated test blocks are available to determine if the tester is in working condition; these can also serve as a training aid for new operators. The acceptable error observed from machine to machine using a known standard is dictated by the test standards. Some of the common problems observed are outlined in the following sections.

Instrument Errors

Concerns with the instrument are as follows: indenter-shape deviations, test-force deviations, velocity of force application, vibrations, angle of indentation, and indentation time. If the tester has passed calibration, the indenter shape, test force, and force velocity should be known. Vibrations arise from a combination of the work environment and robustness of the tester; often a vibration table will eliminate this concern. The indenter should be perpendicular to the specimen at the point of contact. The angle of indentation is determined by a combination of the machine and how well the specimen is secured and prepared. For some testers, this will be evident by an asymmetric indentation. Time should be held constant from test to test, but in most cases is a variable controlled by the operator.

Measurement Errors

The most common error is simply operator bias. It is common for each individual to measure an indent slightly undersized or oversized in comparison to another operator. Operators who do this work routinely, however, are self-consistent. Other measurement errors tend to be due to limitations of the equipment. In order to accurately measure the diameter or depth, the measuring device should be calibrated. For the visual measurement of small indentations, additional concerns are the resolving power of the objective (Buckle, 1954 and Buckle, 1959) or camera, and adequate lighting or image quality.

Material Errors

The quality of polish, poor surface finish, and low reflectivity can limit the feasibility of conducting a test, particularly if the indent diameter needs to be measured. Hardness values of highly porous specimens are referred to as apparent hardness values, since the measurement includes the compression of the pores along with the materials. When thickness is a concern, one should examine the backside of the test piece after conducting the test. If there is any sign of the indent's position, such as a bulge, the test piece was too thin. In some cases, after the removal of the indenter, elastic recovery can change the size and shape of the indentation. Changes tend to be more substantial in hard materials than in soft ones, as far as elastic recovery is concerned (O'Niell, 1967). However, distortion of the indent can also occur in the form of ridging or sinking around the indentation, making accurate visual measurements more difficult (Brown et al., 1951).

LITERATURE CITED

ASTM Standard A255. 1986. In Annual Book of ASTM Standards. ASTM. West Conshohocken, Pa.

Bierbaum, C. H. 1930. The microcharacter: Its application in the study of the hardness of case-hardened, nitrided and chrome-plated surfaces. *Trans. Am. Soc. Steel Treat*. 13:1009–1025.

Boyer, H. E. 1987. Hardness Testing. ASM, Metals Park, Ohio.

Boyer, H. E. and Gall, T. L. (eds). 1985. Mechanical Testing. Vol. 34, pp. 4–11 *In* Metals Handbook, Mechanical Testing. ASM, Metals Park, Ohio.

Brown, A. R. and Ineson, E. 1951. Experimental survey of low-load hardness testing instruments. *J. Iron Steel Inst*. 169: 376–388.

Buckle, H. 1954. Investigations of the effect of load on the Vickers microhardness. *Z. Metallkund*. 45:623–632.

Buckle, H. 1959. Progress in micro-indentation hardness testing. *Met. Rev*. 4:49–100.

Cahoon, J. R., Broughton, W. H., and Kutzak, A. R. 1971. The determination of yield strength from hardness measurements. *Metall. Trans*. 2:1979–1983.

George, R. A., Dinda, S., and Kasper, A. S. 1976. Estimating yield strength from hardness data. *Met. Prog*. 109:30–35.

Greaves, R. H. and Jones, J. A. 1926. The ratio of the tensile strength of steel to the brinell hardness number. *J. Iron Steel Inst.* 113:335–353.

Henkin, A. and Datsko, J. 1963. The influence of physical properties on machinability. *Trans. ASME J. Eng. Ind*. 85:321–328.

Hill, R., Lee, E. H., and Tupper, S. J. 1947. Theory of wedge-indentation of ductile metals. *Proc. R. Soc. London, Ser.* A188:273–289.

Janitzky, E. J. 1938. Taylor speed and its relation to reduction of area and Brinell hardness. *Trans. Am. Soc. Met*. 26:1122–1131.

Khruschov, M. M. 1957. Resistance of metals to wear by abrasion, as related to hardness. Proceedings of the Conference On Lubrication and Wear, pp. 655–659. Institute of Mechanical Engineers, London.

Lysaght, V. E. and DeBellis, A. 1969. Hardness Testing Handbook: American Chain and Cable. Page-Wilson, Bridgeport, Conn.

MacKenzie, J. T. 1946. The Brinell hardness of gray cast iron and its relation to some other properties. *Proc. Am. Soc. Test. Mater*. 46:1025–1038.

Meyer, P. A. and Lutz, D. P. 1985. Ultrasonic Microhardness Testing. In Metals Handbook, ASM. Metals Park, OH: 98–103.

Oliver, W. C., Pharr, G. M. 1992. An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. *J. Mater. Res.* 7:1564–1583.

O'Niell, H. 1967. Hardness Measurement of Metals and Alloys. Chapman & Hall, London.

Palmquist, S. 1957. Method of determining the toughness of brittle materials, particularly sintered carbides. *Jernkontorets Ann*. 141:300–307.

Petty, E. R. 1971. Hardness Testing, Techniques of Metals Research. Vol. V, pt 2, Interscience Publishers, New York: 157–221.

Prandtl, L. 1920. Über die Härte Plastischer Körper. Nachr. Akad. Wiss. Göttingen. Math-Physik. Kl.

Richardson, R. C. 1967. The wear of metals by hard abrasives. *Wear*. 10:291–309.

Robinson, J. N. and Shabaik, A. H. 1973. The determination of the relationship between strain and microhardness by means of visioplasticity. *Metall. Trans*. 4:2091–2095.

Siebert, C. A., Doane, D. V., and Breen, D. H. 1977. The hardenability of steels. ASM, Metals Park, Ohio.

Tabor, D. 1951. The hardness and strength of metals. *J. Inst. Met.* 79:1–18, 465–474.

Taylor, W. J. 1942. The hardness test as a means of estimating the tensile strength of metals. *J. R. Aeronaut. Soc.* 46:198–209.

Vander Voort, G. F. 1984. Metallography Principles and Practice, pp. 350–355. McGraw-Hill, New York.

KEY REFERENCES

Blau, P. J. and Lawn, B. R. (eds). 1985. Microindentation Techniques in Materials Science and Engineering, STP889. American Society for Testing and Materials, West Conshohocken, Pa.

A mixture of theoretical and application notes on microindentation techniques.

Boyer, 1985. See above.

A general overview of all the test methods available including detailed schematics of system components.

Boyer, 1987. See above.

A practical overview of hardness methods and applications is contained in this book as well as an appendix of equipment and manufacturers.

Vander Voort, 1984. See above.

A historical overview of the development of hardness methods as well as application notes.

APPENDIX: CALCULATIONS OF THE HARDNESS VALUES

This appendix provides the equations with which the measured dimension, depth or diameter, is used to calculate the hardness value for each test.

Brinell

$$\text{HBS or HBW} = 0.102 \times \frac{2F}{\pi D(D - \sqrt{D^2 - d^2})} \qquad (1)$$

where D = diameter of the ball in mm, F = test force in N, and d = mean diameter of the indentation in mm. The Brinell hardness is denoted by the following symbols: HBS in cases where a steel ball is used or HBW in cases where a tungsten carbide ball is used.

Rockwell

Rockwell tests scales, A to Z, correlate with the choice of indenter and test force applied. The equations however are based on three cases as shown below.

1. Rockwell Test with Brale Indenter

$$\text{hardness} = 100 - e \qquad (2)$$

where e = permanent increase in depth of penetration under preliminary test force after removal of the additional force, the increase being expressed in units of 0.002 mm.

2. Rockwell Test with Ball Indenter

$$\text{hardness} = 130 - e \qquad (3)$$

where e = permanent increase in depth of penetration under preliminary test force after removal of the additional force, the increase being expressed in units of 0.002 mm.

3. Superficial Rockwell Test

$$\text{hardness} = 100 - e \qquad (4)$$

where e = permanent increase in depth of penetration under preliminary test force after removal of the additional force, the increase being expressed in units of 0.001 mm.

Vickers, Knoop, and Berkovitch

The same tester can be used for all three tests below. The equation is determined specifically by the indenter employed.

Vickers

$$\text{HV} = \frac{1854.4P}{d^2} \qquad (5)$$

where P = test force in gf, d = mean diagonal of the indentation in μm, and a square-based, pyramidal indenter with a 136° angle is used.

Knoop

$$\text{HV} = \frac{14{,}229.4P}{d^2} \qquad (6)$$

where P = test force in gf, d = long diagonal of the indentation in μm, and a rhombic-based, pyramidal indenter with included longitudinal edge angles of 172°, 30 min, and 130°, 0 min, is used.

Berkovich

$$\text{HV} = \frac{1569.7P}{d^2} \qquad (7)$$

where P = test force in gf, d = diagonal of indentation in μm, and a triangular pyramid indenter with an angle of 115° is used.

JANICE KLANSKY
Buehler, Ltd.
Lake Bluff, Illinois

TRIBOLOGICAL AND WEAR TESTING

INTRODUCTION

Tribology is the science of friction, wear, and lubrication. More fundamentally, tribology is concerned with the interaction of contacting surfaces in relative motion. Friction is the resisting tangential force that occurs between two surfaces in contact when they move or tend to move relative to one another. For solid surfaces the magnitude of this force is characterized by the coefficient of friction, defined as the dimensionless ratio of the magnitude of the friction force to the magnitude of the normal force. In the absence of an externally applied normal force, this tangential force is sometimes referred to as sticktion. Wear is damage to a solid surface due to the relative motion between that surface and a contacting substance or substances. It involves the progressive loss of material. Lubrication is the ability of an interposing material to reduce either the friction between or the wear of surfaces.

The tribological behavior that a material exhibits is dependent on both the properties of the material and the contact conditions. As a result, parameters used to characterize the friction, wear, and lubrication of a material are system properties of the material, not intrinsic properties such as an elastic modulus or coefficient of thermal expansion. The collection of all those elements that can influence tribological behavior is referred to as the tribosystem. The primary elements of a tribosystem are (1) materials, (2) shapes and contours, (3) surface roughness, (4) motions, (5) loading, (6) lubrication, and (7) environment. More information on tribological behavior and phenomena can be found in Czichos (1978), Peterson and Winer (1980), Blau (1992), and Bayer (1994).

PRINCIPLES OF THE METHOD

The system nature of tribological properties result from the wide range of mechanisms that contribute to friction, wear, and lubrication. Not only the properties of the materials but also the nature and properties of the various elements of the tribosystem can influence these mechanisms, the mixture of these mechanisms that will operate in a given situation, and their interactions. In the case of wear, many mechanisms tend to have different relationships to and dependencies on material properties and contact conditions. These mechanisms can be divided into four classifications: adhesive mechanisms, single-cycle deformation mechanisms, repeated-cycle deformation mechanisms, and oxidative, or more general, chemical, mechanisms. While one mechanism often tends to predominate, these mechanisms can coexist and interact to form complex patterns of wear behavior. Adhesive mechanisms are those associated with the bonding between surfaces and the pulling out of material from the surface. Examples of single-cycle deformation mechanisms are cutting or chip formation and ploughing. These are often referred to as abrasive wear mechanisms. Fatigue or ratcheting mechanisms are examples of wear mechanisms requiring repeated cycling to produce damage. Oxide wear mechanisms are ones in which wear results from the formation and removal of oxide layers as a result of relative motion.

The situation is similar for friction. Friction mechanisms can be classified into three types, which are related to the first three types of wear mechanisms: adhesive mechanisms; abrasive mechanisms, such as ploughing and cutting; and hysteresis, associated with stress cycling during motion. In addition to these fundamental mechanisms, the friction and wear behavior of a material or system can also be influenced by the formation of tribofilms on or between the surfaces in contact. These films are composed of wear debris from either or both of the surfaces involved. Once formed, these films can act as lubricants, significantly modifying friction and wear behavior associated with a material. In general, the formation of these films depends not only on the properties of the materials involved but also on contact conditions, such as the presence or absence of lubricants, geometry, loading, and speed.

Materials can provide lubrication in a variety of ways. As with friction and wear, these mechanisms involve different types of dependencies on material parameters and contact conditions. Materials can function as lubricants by separating the surfaces, supporting some of the load, physically or chemically altering the surfaces, and providing cooling. Frequently, a combination of these mechanisms is involved. The degree of lubrication that a material can provide and the manner by which it does this depend not only on the properties of the material but also on the other elements of the tribosystem. There are four broad categories of lubrication: dry, boundary, fluid lubrication, and mixed lubrication. Dry lubrication is provided by solids, such as poly(tetrafluoroethylene) (PTFE or Teflon), MoS_2, and graphite particles and composite coatings. Solids frequently provide good lubrication when they are easily sheared and form tribofilms on the counterface. Boundary and fluid lubrications are provided by fluids, such as oils and greases. Fluid lubrication refers to situations where the surfaces are completely separated by the formation of hydrodynamic films. Boundary lubrication refers to situations where separation by hydrodynamic films is negligible and the lubrication is related to the existence of boundary layers formed by the lubricant. Mixed lubrication refers to situations where the two surfaces are not completely separated by hydrodynamic films and boundary-layer phenomena are still significant. With boundary and mixed lubrications, the ability of the fluid to form and maintain boundary layers and react with the surface is a major factor in its ability to lubricate. In the fluid range, fluid flow parameters such as viscosity and Reynolds number become the major factors, as do the shape of the counterfaces and the relative speed.

Tribological situations are classified in terms of the nature of the contact, the type of motion involved, the presence or absence of abrasives, and the type of lubrication. The nature of the contact may be categorized as one- versus two-body contact. In the former the surface is being worn by a fluid or a stream of particles. The term “erosion” is used to describe such situations. There are many subcategories of erosion, such as particle erosion, drop erosion,

cavitation erosion, and slurry erosion, that apply to specific types of erosive situations. When the friction and wear in a two-body situation are primarily associated with hard particles entrained between the two surfaces or hard protuberances on a surface, the situation is referred to as "abrasion." Abrasion is further subdivided into two-body abrasion, in which the abrasive is attached to a surface, and three-body abrasion, in which it is not. A further subdivision is based on the nature of the abrasion. Scratching abrasion, gouging abrasion, and slurry abrasion are examples of this classification.

For two-body contact the primary classification is done on the basis of the motion involved. Sliding, rolling, and impact are the major categories. However, these can be further divided in terms of mixtures of these primary motions, for example, sliding with impact or rolling with slip. For vibratory sliding, when the sliding amplitude is small, less than a few thousand micrometers, the situation is called fretting. Two-body wear situations can be further classified in terms of the nature of the contact, such as point, line, or conforming; the nominal stress level, e.g., above or below the elastic limit; and the relative amount of motion or contact each body experiences. Examples of this last aspect are the classification in terms of the ratio of the contact area to area of the wear track used by Czichos (1978) and the loaded/unloaded designation for each body proposed by Bayer et al. (1962). Wear situations can also be categorized in terms of lubrication, such as whether they are lubricated or not and the type of lubrication, e.g., dry or fluid.

Table 1 provides a classification for wear systems based on operational elements.

Friction, wear, and lubrication are interrelated but distinct phenomena. Thus, different parameters are needed to characterize their behavior. Since tribological parameters are system parameters and not fundamental material properties, a large number of different tribological and wear tests must be performed in order to quantify them, and new tests and modifications of existing tests are continually being developed. Tribological tests for friction, wear, and lubrication vary with the type of material involved, the application, and the interests of the tester. However, these tests share many features, differing only in the parameters used or how the data are analyzed. The discussion that follows will focus on these similarities and present the general features of these tests, rather than provide specific information on any one test (Bayer, 1994, pp. 147–156).

Table 1. Operational Classification of Wear Situations

One-body contact	
Impingement	Low or high angle
	With or without abrasives
Flow	Streamlined or turbulent
	With or without abrasives
Two-body contact	
Rolling	With or without slip
	With or without abrasives
	Type of lubrication[a]
Impact	With stationary or moving body
	With or without abrasives
	Type of lubrication[a]
Sliding	Unidirectional or cyclic
	Large or small amplitude
	Equal or unequal rubbing
	With or without abrasives
	Type of lubrication[a]

[a]None, dry, boundary, mixed, or fluid.

Simulation

The primary element of the common methodology for tribological testing is the need to simulate the tribological system for which the data are to be applied. This means that sliding tests should be used to determine values of tribological parameters for sliding situations, rolling tests for rolling situations, and abrasion tests for abrasive situations. In order to have good correlation between the test and the actual application, other elements of the tribosystem need to be simulated as well, and all elements of the tribosystem need to be considered for simulation. The extent to which each element needs to be simulated depends on the nature of materials, the nature of the wear or friction situation, and the purpose for which the test is intended. For example, with polymers, whose properties are significantly different above and below their glass transition temperatures, it is necessary to simulate those conditions that affect temperature to obtain good correlation with an application (ASTM D3702). Similarly, for lubricated systems, it is necessary to ensure that the same type of lubrication occurs in the test as in the application. For example, with an oil, it is necessary to select appropriate test geometries, loads, and speeds to ensure boundary lubrication if the application involves boundary lubrication. The simulation required for abrasive situations provides another example. Abrasives used in tribological tests should be similar to the abrasives encountered in the application, and the loading should be similar to that in the application. Size, shape, and hardness should be considered, as well as whether or not the loads are sufficient to fracture the abrasives. Because of the need for high correlation between tests and applications in engineering, tribological testing for engineering purposes often requires closer simulation than that used for fundamental research purposes.

A common technique for ensuring adequate simulation in wear tests is to compare the morphology of the wear scar produced in the test to that produced in the application. Another approach that can be used for both friction and wear tests is to compare rankings obtained with the tests and those found in the application.

Measures of Tribological Properties

The basic parameter used to characterize the friction behavior of a material or a material couple in two-body tribosystems is the coefficient of friction, that is, the ratio of the friction force to the normal force pressing the two bodies

together. Two coefficients are used: the coefficient of friction for sliding and the coefficient for rolling. In addition, a distinction is made between the friction force associated with initiation of motion and the force associated with sustained, uniform motion. For the former the coefficient of friction is referred to as the static coefficient of friction; for the latter, as the kinetic coefficient of friction. Since friction also results in the dissipation of energy, the energy lost in a tribosystem is sometimes used as an indirect measure of friction, such as in test methods by the American Society for Testing and Materials used to characterize the friction between rubbers and pavement (ASTM E303 and E707). Both use a pendulum-type device.

The situation for wear is not as simple. There is no equivalent parameter for wear that is as universally used and accepted as the coefficient of friction. However, the ability of materials to resist wear or to cause wear is often characterized by normalized parameters. In the case of sliding and abrasive wear, for example, this is often done by dividing the volume of material removed or displaced by the product of the load and sliding distance. The severity of the wear is sometimes characterized by multiplying this value by the hardness of the wearing surface to obtain a dimensionless wear coefficient. In erosion, wear volume is often divided by the amount of erodent used to cause the wear to provide a dimensionless measure. Other methods and parameters are also frequently used. Examples are the amount of motion or exposure required to produce a certain state of wear or the stress or load required to produce a given amount of wear for a fixed amount of motion, which is used to characterize the rolling wear resistance of metals (Morrison, 1968). In addition, parameters of specific models for wear are also used to characterize wear behavior; examples are those for characterizing wear behavior between silicon chips and various anodized surfaces (Bayer, 1981). Another example of this approach is the method used to evaluate the abrasivity of slurries (Miller and Miller, 1993).

The lubricating characteristics of a material are determined by comparing wear or friction with and without the material present. For example, the coefficients of friction with and without a particular lubricant present can be used as a measure of the material's ability to lubricate a surface. Relative lubricating characteristics are often obtained from friction or wear tests with different lubricants by comparing the friction or wear measured in tests with different lubricants. For example, in an ASTM test for evaluating the antiwear properties of lubricating greases, the size of the wear scar is used to rank lubricants (ASTM D2266). An additional important property of a material's ability to lubricate is the range of conditions under which it is able to provide lubrication. This might be the oxidative life of the lubricant, in the case of an oil, or the wear life of a bonded-solid lubricant. It can also be the maximum pressure or speed the lubricant can withstand before breakdown occurs. Parameters such as these are also used to characterize lubricant behavior and are determined in friction and wear tests designed to determine the limits of their ability. Such approaches are used in several ASTM tests for solid film and fluid lubricants (ASTM D2625; ASTM D3233).

PRACTICAL ASPECTS OF THE METHOD

General Tribological Testing

While somewhat lower in friction tests than in wear tests, the potential for scatter in tribological tests is very high. This results from the complex nature of tribological behavior and the range of friction and wear coefficients that are associated with such behavior. Examples of these ranges are shown in Tables 2 and 3. The characteristic range of the scatter depends on the nature of the test, the parameter measured, and the controls incorporated into the test. Scatter obtained with multiple test apparatuses tends to be higher than that obtained with a single apparatus. Coefficients of variation of $<30\%$ are considered good for the coefficient of friction and representative of well-designed and controlled tests. With many standard wear tests, variations by a factor of ≈ 2 are not uncommon and are often considered good. When there is inadequate control in wear tests, variations by an order of magnitude, and sometimes larger, are common. With proper controls the characteristic coefficient of variation for a wear test can often be reduced to 10% or 20% and generally $<50\%$. Because of this scatter, it is necessary to perform duplicate tests to obtain statistically significant data. Even in situations where this is not a major issue, a minimum of three replicates should be used. This is recommended because of the probability of obtaining results that are not typical (outliers). In tribological tests the probability of encountering outliers is generally considered to be high.

Another practical aspect of tribological testing is the need to clean specimens and counterfaces used in the tests. While cleaning tends to be less of a factor in erosion and lubricated tests than in unlubricated two-body wear tests, some cleaning is required to reduce scatter in all tribological tests. This element is discussed in more detail below.

Test Equipment and Measurement Techniques

The test equipment used for tribological testing consists of those elements needed to simulate the friction or wear situation and to perform the necessary measurements. For two-body situations, the essential elements consist of a mechanism for pressing the two bodies together with a known force, a mechanism for moving the two bodies relative to each other, and means of measuring friction, wear, or both. Typical methods for applying the normal force are dead-weight methods, often involving a system of levers,

Table 2. Friction Coefficients

Condition	Coefficient
Unlubricated sliding	
Metal/metal	0.3–1.4
Metal/ceramic	0.1–0.8
Metal/plastic	0.1–0.7
Plastic/plastic	0.05–0.8
Boundary-lubricated sliding	0.1–0.3
Fluid-lubricated sliding	<0.01
Rolling	
Steel/steel	≈ 0.001

Table 3. Wear Coefficients

Condition	Coefficient	
Sliding wear	(*wear volume* × *hardness*/*load* × *distance*)	
Self-mated metals, unlubricated	$2 \times 10^{-4} - 2 \times 10^{-1}$	
Self-mated metals, lubricated	$9 \times 10^{-7} - 9 \times 10^{-4}$	
Non-self-mated metals, unlubricated	$6 \times 10^{-4} - 2 \times 10^{-3}$	
Non-self-mated metals, lubricated	$9 \times 10^{-8} - 3 \times 10^{-4}$	
Plastics or metals, unlubricated	$8 \times 10^{-5} - 3 \times 10^{-7}$	
Plastics or metals, lubricated	$1 \times 10^{-6} - 5 \times 10^{-6}$	
Abrasion	(*wear volume* × *hardness*/*load* × *distance*)	
	Dry	Wet
Two-body abrasion		
File	5×10^{-2}	10^{-1}
New abrasive paper	10^{-2}	2×10^{-2}
Used abrasive paper	10^{-3}	2×10^{-3}
Coarse polishing	10^{-4}	2×10^{-4}
<100-μm particles	10^{-2}	—
>100-μm particles	10^{-1}	—
Three-body abrasion		
Coarse particles	10^{-3}	5×10^{-3}
Fine particles	10^{-4}	5×10^{-4}
Particle erosion	(*wear volume* × *hardness*/*particle volume* × *particle density* × *particle velocity*^{2})	
Soft steel	$8 \times 10^{-3} - 4 \times 10^{-2}$	
Steel	$1 \times 10^{-2} - 8 \times 10^{-2}$	
Hard steel	$1 \times 10^{-2} - 1 \times 10^{-1}$	
Aliminum	$5 \times 10^{-3} - 1.5 \times 10^{-2}$	
Copper	$5 \times 10^{-3} - 1.3 \times 10^{-2}$	

load cells, or deflection beams with strain gauges attached. For one-body situations or erosion, the essential elements consist of some means of providing the proper erosive action, holding the wear specimen, and measuring the wear. Typical ways of providing the erosive action are to use jets to create a stream of erodent or to move the wear surface through an erodent field. In addition to these essential elements, tribological tests often involve some means of controlling the environmental conditions of the test and maintaining an adequate supply of lubricant. The former is done to improve repeatability or to simulate the environmental conditions of applications, particularly for applications in hostile environments.

Friction is generally measured by the use of strain gauges, load cells, or some other force-measuring device. Perhaps the most common method is to directly measure the friction force and to obtain the coefficient of friction by dividing by the normal load. However, other options also exist. For example, in a test used to characterize the friction between a web material and a capstan or roller, the ratio of the tensions in the web on either side of the capstan is used to determine the coefficient of friction (ASTM G143). For this configuration the coefficient is obtained by dividing the natural logarithm of the ratio of the tensions by the angle of wrap. In an inclined-plane method for determining the coefficient of friction, such as is used for papers, the angle of inclination is used to determine the coefficient of friction (ASTM D4521). In this case, the static coefficient of friction is given by the tangent of the angle at which slip occurs. In friction tests that use dissipated energy as a measure of friction, other appropriate techniques are used. Power consumption in an electric motor is one; another is the height that the pendulum reaches after frictional contact is made in a pendulum test.

Volume of material lost or displaced is generally considered the more fundamental measurement of wear. However, it is not the only measure that may be used in wear tests. Different features of wear scars have been used, as well as functional characteristics of devices, such as bearing concentricity or roller-bearing noise and vibration. There are several reasons for this, including the ease of performing the measurement and the accuracy obtainable with the materials and geometries used in the test. Another is the significance of a particular mode, feature, or degree of damage to an application. For example, with optically transparent material or decorative finishes, the formation of fine scratches or haze on the surface is more significant than the volume of material removed. The analytical expressions used to analyze the data and relate the data to an application may also be a factor.

The wear measures that tend to be used most frequently can be grouped into four types based on the following properties: appearance, wear scar dimensions, volume lost or displaced, and mass loss. For some geometries volume can be determined from dimensional measurements or by scanning or three-dimensional profilometer techniques. If the density of the material is known, mass loss can also be used to determine volume. In tests that utilize appearance, wear scars are commonly evaluated by eye, by optical microscopy, or by scanning electron microscopy. For example, the unaided eye is used to examine the wear scar for characteristic damage in a standard test to evaluate galling resistance (ASTM G98). Galling is a sliding wear mode characterized by roughening and the creation of protrusions above the original surface.

Tribological Test Categories

The term "friction test" is applied to tribological tests designed solely to measure friction. The amount of wear typically generated in these tests are negligible. The term "wear test" is applied to tests designed to cause wear. However, two-body wear tests can also be used to measure friction, provided the apparatus provides some means of measuring the friction force. With this type of instrumentation, both initial friction and changes in friction that occur with use can be measured. Since wear is associated with progressive loss or damage, wear tests involve repeated cycles or exposure. Friction tests, on the other hand, often involve only a single cycle. However, neither are restricted to these situations. Some wear tests, such as those used to characterize galling resistance and scratch tests used to characterize abrasion resistance, measure wear after one cycle. Friction tests used to measure

the kinetic or dynamic coefficient of friction often involve repeated cycles or sustained motion but with negligible wear.

Since the ability of a material to lubricate is determined by comparison of friction and wear behavior with and without the material present, friction and wear tests are both used to evaluate lubrication. Friction tests provide a measure of a material's ability to reduce friction, generally with negligible wear. However, wear tests can provide a measure of a material's ability to reduce both friction and wear. When wear tests are used in this fashion, friction is monitored during the progression of the test. Lubrication tests to determine the limits of a material's lubricating ability often involve the progressive increase of some stressing element, such as load, speed, or temperature, coupled with either a single cycle or continuous motion. In these tests the failure point is identified by either a dramatic increase in friction, onset of catastrophic wear, or both, e.g., the occurrence of seizure. Direct and indirect methods are used for detecting such a point. For example, the fracture of a shear pin is used in a test to determine the failure point of oils as a function of load (ASTM D3233).

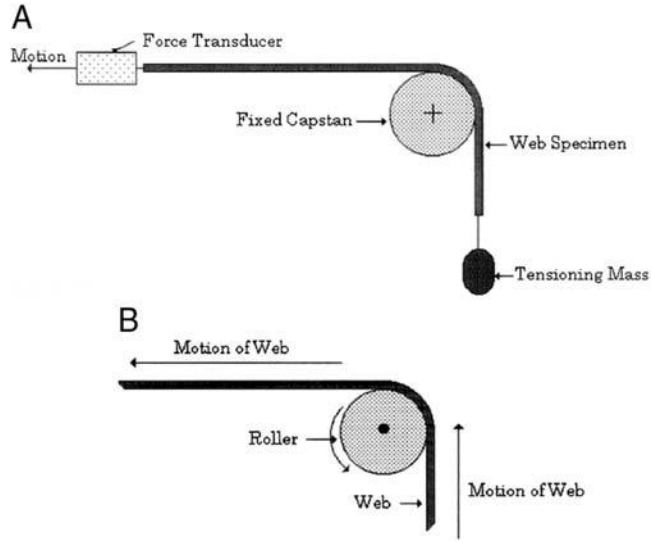


Figure 1. Similarity between test configuration used to characterize friction of web materials and contact configuration in applications involving transportation of web materials. (**A**) Test configuration used in ASTM G143. (**B**) Typical application.

Friction Tests

The basic element of a friction test is the measurement of the force required to initiate motion, maintain motion, or both. When the force is measured directly, these data are typically converted to a coefficient of friction. The static and kinetic coefficients of friction and fluctuations in friction can be measured with appropriate instrumentation and analysis techniques. This is possible only when friction is measured directly. Indirect approaches can provide information on either maximum friction or average friction, but not both. For example, the inclined plane test measures the static coefficient of friction (the force required to initiate motion), but not the kinetic coefficient of friction (the retarding force on a moving body). When dissipated energy is used as a measure of friction, as in a pendulum test, it is related to the average force needed to maintain motion. With such methods, it is not possible to observe fluctuation in friction behavior during the motion.

Simulation consists of replicating the nominal type of motion, shapes of bodies, nominal loading conditions, counterface conditions, and environmental conditions of the application. In the case of friction only minimal simulation of loading and contact geometry is needed to obtain significant correlation with applications. Provided the motion is appropriate and test specimens of the materials can be obtained, most friction tests provide information about trends, such as effects of speed and environmental conditions and coarse rankings. For example, most sliding tests for friction show that unlubricated metal pairs have coefficients of friction from 0.5 to 1, which are reduced to 0.1 to 0.3 when lubricated by oils; metal-polymer couples have coefficients between 0.1 and 0.3; and, depending on fillers and types, coefficients for elastomers can range from 0.5 to >1. This trend is consistent with what is generally found in applications.

However, coefficients of friction obtained from different tests with the same materials are not always identical, since friction is not completely independent of loading and geometry. As a result, some tests provide better correlation than others. For improved correlation, which allows more specific characterizations and finer rankings, it is necessary to replicate the nature of the shapes and the loading involved in the application. For example, consider the tests shown in Figures 1 and 2. Both are sliding friction tests that can be used to evaluate the coefficient of friction between a web material and materials used for capstans and rollers in web-handling systems. Figure 1 also shows the nominal contact situation between a roller and a capstan in these applications. It is evident that the capstan test provides a better simulation of this application.

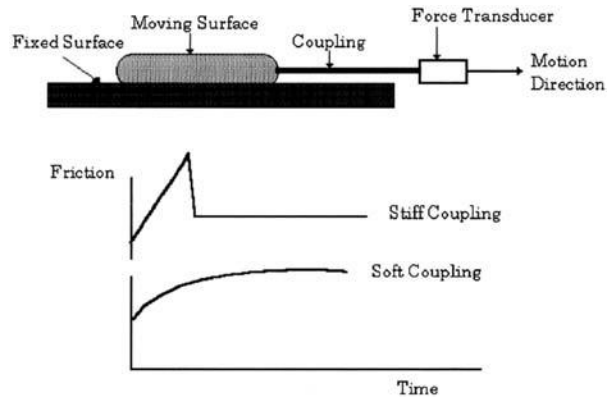


Figure 2. Effect of drive stiffness on measurement of friction. With a stiff coupling, such as a steel wire, there can be a pronounced difference between static and kinetic friction. With a flexible coupling, such as a piece of string, the difference can disappear.

Engineering experience with these two tests has shown that better correlation is obtained with the capstan test than with the sled test (Budinski, 1989).

Some applications require a high degree of simulation to obtain adequate correlation. Examples of this are some of the tests used for characterizing the friction of brake materials (Jacko et al., 1984; Tsang et al., 1985). In these not only are the geometries, loads, speeds, and environment replicated, but other characteristics of brake applications, such as pumping and prolonged braking, are incorporated into the test protocols.

Observed friction behavior is influenced by the instrumentation used to measure friction and by the design of the drive system. The detection and appearance of fluctuations in friction may be modified by sampling rate, instrument sensitivity, and frequency response of the measurement system. Short-term fluctuations can be missed and the appearance of sharp transitions modified as a result of the instrumentation used. The stiffness of the drive can also be a factor, as illustrated in Figure 2 in the case of a sled test. As the stiffness is reduced, differences between static and kinetic friction tend to be reduced and stick-slip behavior can be modified.

In many friction tests the normal load tends to fluctuate as a result of dynamic effects or variations in surface contours. In these situations it is desirable to measure the normal force in the same manner as the friction force, so that instantaneous values of the coefficient of friction can be obtained.

Wear Tests

Wear testing often requires more simulation than friction testing to obtain good correlation with applications. A useful technique for investigating the adequacy of the simulation provided by a wear test is by comparison of wear scar features obtained in the test with those observed in the application. The more similar they are, the better the simulation, and the more likely there will be good correlation. In these comparisons, the magnitude of the wear is not the important element. Appearance and morphological features, which are related to or indicative of mechanisms, are the more important aspects to be compared.

The overall design of a test apparatus tends to have characteristic influences on wear behavior that need to be considered for adequate control and simulation. For example, wear behavior can be influenced by the stiffness of the test apparatus as a result of dynamic effects and superimposed vibrations. Wear results can also be influenced by differences in the precision obtainable with different specimen mounting techniques, such as the difference between the use of a four- or three-jawed chuck to hold the cylindrical specimens in a crossed-cylinder test configuration (ASTM G83). In addition, such considerations as different specimen shapes or methods for applying loads, generating erosive streams, or providing motion tend to have unique characteristics that can affect wear behavior in different ways. Included in these characteristics is the effect that specimen wear has on test parameters. These issues also need to be considered for simulation and control. For example, when load cells or strain-gauged beams are used to generate the load in two-body wear tests, the load tends to decrease with wear. Consequently, in such situations it is necessary either to select load cells and beams so that the change in load is negligible over the wear range of interest or to use some monitoring and feedback system to maintain a constant load. Dead-weight methods do not have this problem. If not carefully designed, however, such systems, which often involve the use of levers, are prone to errors and variations in either the magnitude of the actual force applied or the line of action. For example, friction effects at bearing points can influence the magnitude, or tolerances may allow the line of action to vary.

Another example is the characteristics of test configurations commonly used in two-body wear tests. One such configuration involves a flat-ended specimen of uniform cross-section pressed against a flat or cylindrical surface. Initial alignment is a problem, as during the initial period of wear contact stresses tend to be high and inconsistent. However, once the surfaces are worn, the contact stress becomes much smaller, more constant, and more consistent. Another common configuration is a spherical surface against a flat or cylindrical surface. With this configuration there is no alignment problem and initial stress levels are more consistent. However, the contact stress continually decreases throughout the test at a decreasing rate. When cylindrical surfaces are used to obtain nominal line contact, both effects are present. There is an initial alignment problem and the contact stress decreases throughout the test. Since some wear mechanisms can be influenced by stress, wear behavior in tests with these different geometries may not be equivalent. The different characteristics of these contact geometries can also influence the duration and nature of any break-in or run-in period.

There are similar concerns with techniques used in erosion and abrasion testing. For example, in particle erosion tests, different designs of the nozzle used to produce particle streams can result in different dispersion characteristics, which influence erosion behavior. There can also be interaction between the shape of the wear scar and the erosion action of the stream as a result of particle rebound (Fig. 3). When this can occur, it is necessary to place a limit on the amount of wear that can be used for evaluating erosion behavior and comparing materials (ASTM G76).

While the basic element of a wear test is the measurement of wear after specified amounts of motion or duration, there are a number of different approaches used. One approach, which is frequently used to rank materials, is to measure wear after a specific duration, e.g., number of cycles, or exposure, e.g., amount of abrasive used. This approach is used in a four-ball test to evaluate the antiwear properties of greases, in a block-on-ring test for sliding wear resistance, and in a gouging abrasion test (ASTM D2266; ASTM G77; ASTM G81). The concept is to select a duration that is sufficiently long to be beyond any break-in or run-in behavior and to obtain stable wear behavior, as characterized by a constant or slowly changing wear rate. In many wear tests the amount of time required to achieve stable behavior depends on the materials being evaluated as well as the conditions of the tests. The duration or

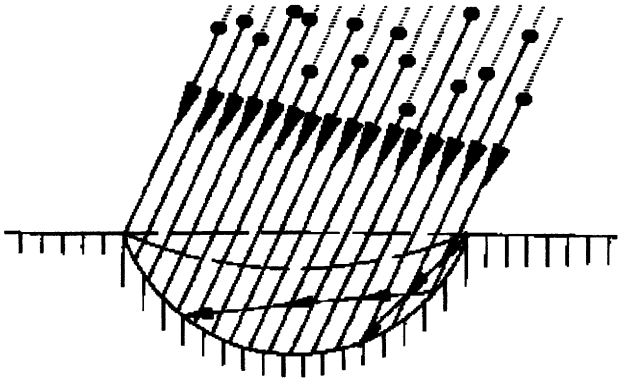


Figure 3. In particle erosion there is the possibility of deep wear scars of secondary impact from rebounding particles. As the depth of wear increases, the significance of these rebounding particles and secondary impacts tends to increase. This effect leads to a limit on the maximum amount of wear that can be used to compare materials in these types of tests.

exposure to be used in this type of testing approach must be determined empirically.

Another approach is to measure wear at different times or intervals so as to develop a wear curve, which plots wear as a function of accumulated motion or exposure. This method is used in a block-on-ring test for sliding wear resistance of plastics, in a cavitation erosion test using a jet, and in evaluating the impact wear resistance of elastomers, among others (ASTM G137; ASTM G134; Bayer, 1986). Wear curves may be developed in two ways: (1) removing and replacing the same sample and (2) using different samples for the different duration. There are concerns with both approaches. Removing and replacing the same sample can cause disruption in surface films and debris features, contamination, and a change in the position of the sample. As a consequence, wear behavior may be altered. With the second approach more samples are required and sample-to-sample variations may mask intrinsic wear behavior. The possible effects on apparent wear for both approaches are illustrated in Figure 4.

The shapes of wear curves vary with the wear test and the nature of the materials involved. However, they tend to be nonlinear, and initial wear rates are often significantly different than longer-term wear rates. The wear curves shown in Figure 4 are typical of those obtained in many sliding wear tests. Typical curves for particle and cavitation erosion are shown in Figure 5.

Wear curves are used in a variety of ways to evaluate and compare wear behavior. For semiquantitative or qualitative rankings and comparisons of materials often used for engineering, direct graphical comparisons are generally used. With this approach the wear curve is used to identify a test duration that is sufficiently long to obtain a constant or slowly changing wear rate. With this approach accumulated wear is the primary criterion for ranking. However, there is frequently some ambiguity in such comparisons, and only coarse rankings may be appropriate. This situation is illustrated in Figure 6, which shows a hypothetical but typical situation often encountered with wear data. Wear curves for six different materials are plotted. Based on accumulated wear, the materials

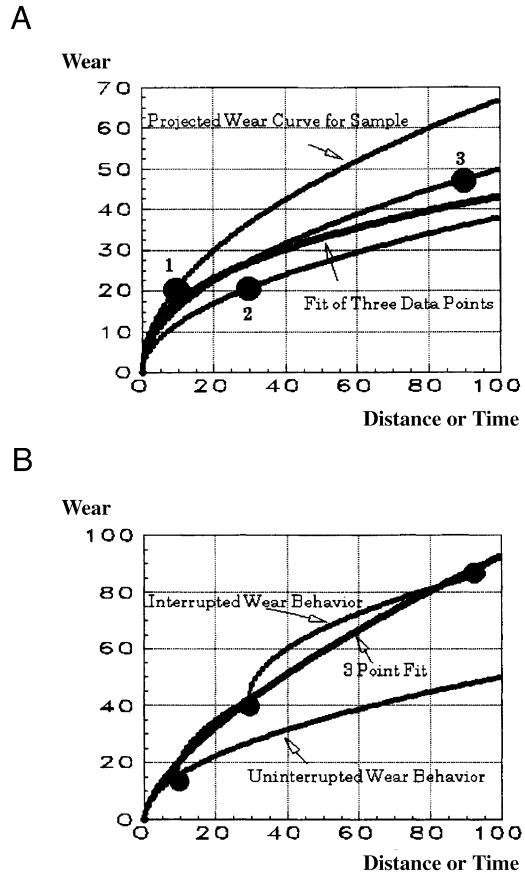


Figure 4. Errors associated with two methods of generating wear curves. Curves are typical of sliding wear data. (**A**) Possible effect of using different samples for different durations. Points 1, 2, and 3 refer to different samples. (**B**) Possible effect of using a single sample (arbitrary units).

are ranked in the order F, E, B, C, A, and D. However, the terminal slopes suggest that the order should be F, E, B, A, D, and C. It can also be seen that if a duration less than 70 units of distance is used, a different ranking would be obtained, namely, F, E, C, B, A, and D. For many practical engineering purposes the relative rankings of C, A, and D are considered immaterial and those of E and F are considered equivalent.

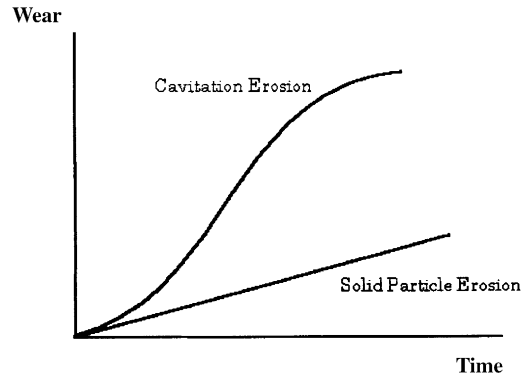


Figure 5. Typical wear curves for cavitation erosion and solid-particle erosion.

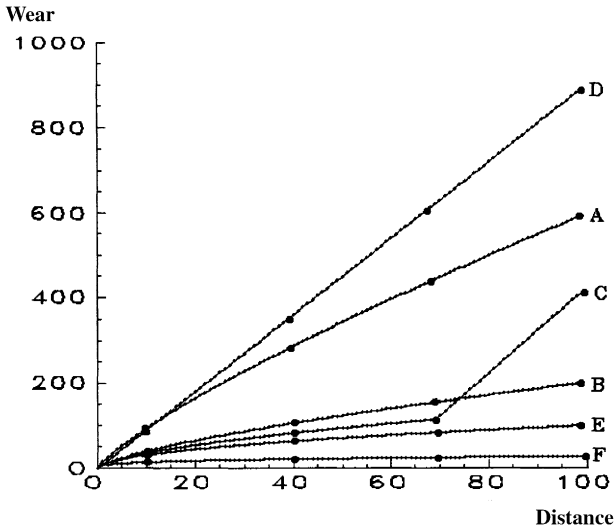


Figure 6. Possible results from sliding wear test used to compare six different materials. Units are arbitrary. Wear measurements were made at 10, 40, 70, and 100 distance units.

Another approach is to differentiate the curve to obtain a wear-rate-versus-usage curve such as shown in Figure 7. With this method wear behavior is characterized by wear rate, using either an absolute rate, such as wear volume per sliding distance or wear volume per unit volume of erodent, or a normalized rate, such as wear volume per load times sliding distance, wear volume times hardness per load times sliding distance, or wear volume times hardness per volume of erodent (see Table 3). This method is used in a block-on-ring test to evaluate the sliding wear resistance of plastics (ASTM G137). It is also used in tests for liquid impingement erosion (ASTM G73). In many wear tests the wear rates tend to stabilize with duration, becoming either constant or slowly varying. As a consequence longer term wear rates are used for comparing and evaluating materials.

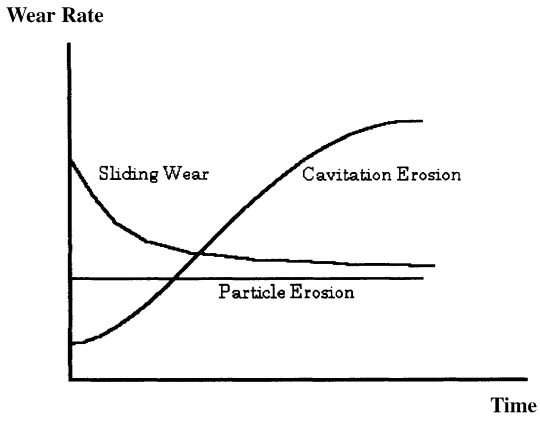


Figure 7. Wear rate curves typically obtained for sliding wear, cavitation erosion, and particle erosion. Time could be replaced by distance or volume of particles in the case of sliding wear and particle erosion, respectively.

A third approach is to analyze the wear curve by fitting the data to some empirical or theoretical model in order to obtain coefficients, which are used to compare or evaluate wear behavior. Examples of this approach are the methods associated with the implementations of engineering models for sliding and impact wear (Bayer, 1994, pp. 332–350; Engel, 1976, pp. 180–243). Another example is the method used for determining slurry abrasivity (Miller number) and the slurry abrasion response of a material (SAR number) for slurry abrasion (ASTM G75).

With either the single-point method or the wear-curve method, wear tests are run until stable behavior, as indicated by a constant or slowly changing wear rate, is achieved.

Another approach used for wear tests is to determine the value of some parameter required to produce a certain level of wear for a standard set of conditions. Duration or exposure and loading are often used for this purpose. For example, rolling wear resistance of materials is often evaluated and compared by using a wear test to determine either the number of revolutions for a fixed load or the load for a fixed number of revolutions required to produce cracking (Bayer, 1994, pp. 221–225). With this approach appearance is used as the measure of wear. Other more quantitative measures can be used, but implementation of the tests then becomes more complex.

In two-body wear tests, as in any two-body wear situation, either or both surfaces may wear. The relative potential for these surfaces to wear depends on the relative amount of motion that each surface experiences. For example, in the case of a small slider pressed against the surface of a large rotating disk, the small slider has the higher potential for wear because its surface experiences a larger amount of motion per unit area. For the thrust washer test configuration both surfaces experience the same amount of motion per unit area, and the potential for wear is the same for both.

While the potential for either body to wear in a two-body wear situation is an element that needs to be considered in simulation, it also is significant in other aspects of wear testing. One aspect is that counterfaces in two-body wear tests should be examined and measured for wear, just as the wear specimen is. A second aspect is the effect that the wear of the counterface has on the analysis method used, data interpretation, and the comparison of materials. In general, significant counterface wear makes these more complex. At the very least, the amount of wear on both members of the system needs to be considered in making any comparisons. One approach to doing this is shown in Figure 8 for a test that uses the single-point method, i.e., comparing the accumulated amount of wear. With this approach, material couples may also be compared by using the total wear of the system, i.e., the sum of the wear on both surfaces. Models that take into account simultaneous wear on both surfaces can also be used to analyze the data to obtain wear coefficients for the materials involved. Such an approach has been used to evaluate material pairs for electrical contacts (Bayer and Sirico, 1971).

Complexities associated with the wear of both surfaces can be avoided by selecting wear-test configurations that bias the wear to occur only on one member. This is done

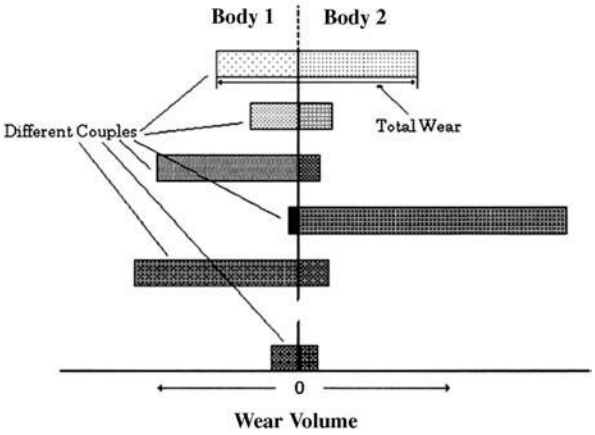


Figure 8. Example of wear data in two-body wear situations when both surfaces wear. The effect of counterface wear cannot be ignored in such cases. Total wear, the sum of wear on both surfaces, can be used to characterize the wear of the couple.

by selecting a configuration in which the wear specimen experiences much more motion than the counterface and by using hard materials for the counterface. This tends to reduce the degree of simulation between the test and application, but is often used in standard tests to evaluate different materials for generic wear situations.

The three most common ways of measuring wear volume are directly measuring the wear scar using scanning profilometer techniques; using a dimensional measurement, such as width or depth, in conjunction with some geometrical relationships; and measuring mass loss and dividing by density. While there are problems and limitations associated with all these techniques, mass loss measurement is the least preferred for several reasons. First, wear is primarily associated with volume loss, not mass loss. Therefore, for valid comparison of materials, the densities must be known, which is not always the case, particularly with coatings. Second, some wear situations result in the build-up or transfer of material, in which case mass loss does not correspond to volume loss. When this technique is used for dry sliding situations, a net mass gain, or negative wear, is frequently observed. With the other techniques it is generally possible to avoid this problem by the use of datum lines, based on unworn profiles, to determine the worn volume. These lines provide a reference for determining wear depth or area. This method is illustrated in Figure 9. A final problem with mass loss measurements is that small amounts of wear may not be detectable on large specimens.

Control in Tribological Testing

In addition to simulation, which is the most fundamental element of tribological testing, there are three other important elements associated with this type of testing: the control required, the acceleration, and the documentation of test results. Control refers to those aspects that relate to the repeatability of the test. Table 4 lists the components of a tribological test that need to be controlled to reduce scatter. The degree of control required for any of these

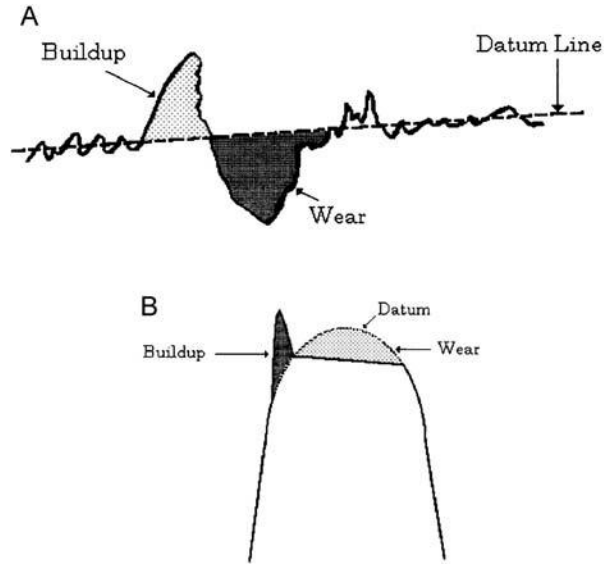


Figure 9. Examples of using datum lines to determine wear volume by profilometer techniques. (**A**) Use of a datum for a flat surface, where the average of the unworn surface is used to establish the datum. (**B**) Uses of a datum for a spherical specimen, where the datum is the unworn profile of the specimen.

components can vary, depending primarily on the nature of the materials involved and the amount of scatter that can be tolerated. However, insufficient control of any one of these can result in unacceptable scatter or erroneous data.

Three areas of control need to be stressed. (1) Since tribological behavior can be strongly influenced by surface films and the state of material near the surface, it is important to control specimen preparation and cleanliness. (2) Maintaining mechanical tolerances of specimens and equipment ensures that acceptable scatter is obtained. Loose tolerances can lead to variation in test specimen alignment and loading and to poor motion control, among other things. (3) Detailed test protocols should be developed and followed. Often, to ensure acceptable

Table 4. Test Elements Requiring Control

Apparatus related	Stability of test apparatus
	Positioning and alignment
	Motions
	Flows
	Loading
	Lubrication system
Sample related	Material supply and storage
	Specimen handling
	Specimen preparation
	Specimen and counterface tolerances
	Roughness
	Cleaning
Environmental	Temperature
	Relative humidity
	Special atmosphere
Operational	Procedures for conducting the test
	Test start-up and stopping procedures
	Measurement procedures

repeatability, it is necessary to include in the protocols not only the major elements, such as loads, speeds, and measurement techniques, but elements such specimen handling and storage.

Tests with reference materials are frequently used to ensure adequate control. Initially, repeated tests with well-controlled materials are used to develop equipment and techniques to obtain acceptable repeatability. With established tests, similar tests are performed periodically to monitor equipment and implementation. If the results fall outside an acceptable range, the various test elements are then investigated to determine the cause and appropriate corrective action identified.

Another approach used with reference materials is to establish or obtain scaling factors. This is often done with abrasion and erosion tests, where it is generally difficult to adequately control lot-to-lot variations of the abrasive and erosive material used. Sometimes the test procedure is to include the simultaneous exposure of both the reference material and the test material, such as is done in a gouging abrasion test (ASTM G81). Still another approach is to perform tests with the reference material with each batch of abrasive or erosive to develop a scaling factor. An example is the method used to investigate the effect of hardness on the wear resistance of materials sliding against paper (Bayer, 1983).

Acceleration

Except for friction tests, most tribological tests involve an element of acceleration with respect to a real-world specific application. Basically this is because long wear lives and low wear rates are typical, or at least desired, characteristics of applications, whereas short test times are desirable in testing. This element of acceleration in a test runs contrary to the need to simulate the application as closely as possible. Nonetheless, some form and amount of acceleration can usually be reconciled with maintaining adequate simulation. The maximum amount of acceleration that can be used depends on the element selected to provide the acceleration, the materials involved, and the nature of the wear situation. The basic requirement for acceptable acceleration is not to change the significance of relevant friction, wear, and lubrication mechanisms. Conditions that are likely to cause significant changes in material behavior should be avoided, as well as those that can significantly affect tribofilm formation. For example, obtaining acceleration by increasing the load so that stresses are in the plastic range is generally not a valid way of providing acceleration for an application in which the stresses are in the elastic range. Similarly, testing polymeric materials under conditions where their temperature is above the glass transition temperature should not be used in applications where the temperature will be below their glass transition temperature. Typical ways of achieving acceleration in wear tests are to use higher loads or stresses, more severe environmental conditions, e.g., higher temperatures or relative humidity and higher speeds. Other options include higher concentrations of abrasives, higher impact velocities, continuous rather than interrupted motion, and more intense streams of erodents.

I. Technical Function of Tribosystem					
II. Operating Variables					
Type of Motion			Duration of Operation		
Load Profile		Velocity Profile		Temperature Profile	
Other Operational Variables					
III. Structure of the Tribosystem					
Initial/Final Properties		Element 1	Element 2	Lubricant	Atmosphere
Volume	Geometry				
	Chemical				
	Physical				
Surface	Topographical				Other
	Surface Layer Properties				Data
Contact Area				Tribological Interactions	
Contact Area/Track Area					
Type of Lubrication					
IV. Tribological Characteristics					
Property Changes		Friction Data		Wear Data	
Element 1					
Element 2					
Lubricant		Appearance of Worn Surfaces			
Other Characteristics					

Figure 10. Outline of format for reporting tribological data, proposed by Czichos (1978).

The effect of acceleration can be evaluated by comparison of application wear scars and test wear scars, in the same manner as for the evaluation of simulation.

Reporting Test Results

An important element in tribological testing is documentation. In addition to the test results and the identification of the materials tested, test reports should include information on all the elements of the test tribosystem, as well as the test method. Figure 10 shows a form for reporting the results of tribological tests. In the figure "technical function of the tribosystem" refers to the nature of the wear situation, such as a sphere sliding on a flat or cavitation erosion by a vibrating horn. "Load profile," "velocity profile," and "temperature profile" include not only the magnitudes of these parameters but the variation over the test or test cycle. Other operating variables include other operating elements of the tribosystem, such as the method of introducing abrasives into the system or the nature of a run-in cycle. "Geometry" includes both size and shape; "chemical" refers to composition; "physical" refers to such things as hardness, modulus, and heat treatment; "topographical" refers to surface roughness characterization; and "surface layer properties" includes both chemical and physical properties. "Tribological interactions" refers to the possible ways the various elements of the tribosystem can interact and the sequence of those interactions. Other characteristics of interest can include contact resistance, features of wear debris, and the state of the lubricant.

There are ASTM guidelines for reporting friction and wear data that can be of use in determining what elements should be recorded and in what fashion (ASTM G115; ASTM G118). In addition, many standard tribological tests include a section on data reporting. For two-body wear tests it is essential to include information regarding the state of wear on the counterface, as well as on the primary surface or wear specimen.

METHOD AUTOMATION

Computer controls and automation are used in tribological testing to control loads and motions. These techniques

improve precision and consistency and also facilitate the implementation of complex test protocols such as progressively increasing the load in the test, using a raster pattern in sliding tests, or using a break-in procedure. Automated systems tend to have different characteristics than nonautomated systems, however. As a result, there is the possibility of bias between data obtained from nonautomated versus automated tests.

These techniques are also useful in measuring friction and the coefficient of friction. Automation is required to study short-term fluctuations in friction. If the normal force is simultaneously measured, instantaneous values of the coefficient of friction can be obtained. Automation can also be used to obtain average values of the coefficient of friction; this approach is used in many commercially available friction test apparatuses. When using such apparatuses, it is necessary to understand the algorithm that is used to determine the average value, as different algorithms can result in different values for complex friction profiles.

Automated *in situ* wear measurements are another application. However, at present these systems do not have the precision of off-line measuring techniques. As a result, they typically are only used to monitor coarse wear behavior during a test, and off-line techniques are used to measure wear for analysis and comparison purposes.

Automated three-dimensional scanning contact and noncontact surface profilometer methods are also used to measure and characterize wear scars. These approaches can be used to automatically determine wear scar volume and maximum depth as well as other properties of the scar, such as roughness.

DATA ANALYSIS AND INITIAL INTERPRETATION

Friction

In friction tests, where force is measured, the common method of analysis involves reducing the data to coefficients of friction, which are obtained by dividing the friction by the normal force. Typically, the data are analyzed to determine a static coefficient of friction, which is associated with the initiation of motion, and a dynamic coefficient of friction, which is associated with sustained motion. When there are fluctuations in friction, the coefficient should be determined on a point-by-point basis, rather than using average values for friction and load.

In wear tests in which friction is measured, changes in friction can frequently be correlated to changes in wear behavior. These correlations provide useful information for understanding and interpreting the friction and wear behavior of the system. However, high friction or low friction, by themselves, do not necessarily indicate high or low wear, respectively.

Wear

Analysis and interpretation methods for wear data depend on the nature of the test, the purpose of the test, and the model applied to the wear situation. Sometimes the wear data are directly compared without analysis; see Practical Aspects of the Method, Wear Tests. In other cases, the wear data are analyzed in terms of a particular model. In this case, the nature of the model dictates the method of analysis and guides the interpretation.

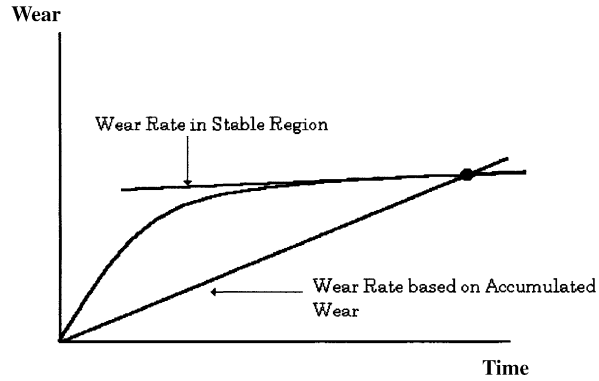


Figure 11. Possible error associated with using accumulated wear to determine wear rate.

A common way of reporting wear data, but not the only way, is to use a wear rate. For erosion this is in the form of volume lost per volume of erodent. For sliding and abrasive wear three different forms tend to be used: volume of wear per unit distance of sliding, volume of wear per unit distance of sliding per unit of load, and the former multiplied by the hardness of the wear specimen. The single-measurement method of wear testing is a poor choice if wear data are reported in this manner, as it is prone to considerable error because of the nonlinear nature of wear curves (see Fig. 11).

The last two wear rates used for sliding and abrasive wear involve normalization with respect to load and hardness, which is based on two simple models or equations for wear. These equations are the Archard equation for sliding wear and the Kruschov equation for abrasive wear (Bayer, 1994). Both have the same form:

$$V = KPS/H \tag{1}$$

where V is the volume of wear, P is the load, S is the distance of sliding, H is the hardness of the material, and K is a dimensionless wear coefficient. For sliding wear K is related to the probability of a wear fragment being formed; for abrasive wear K is related to the sharpness of the abrasives. While these equations provide a convenient way of analyzing wear data, there is considerable debate over their applicability, or that of any single equation, to all wear situations. Other models of wear often indicate different dependencies on load, distance, and hardness (Bayer, 1994, pp. 324–350; Ludema and Bayer, 1991).

As discussed earlier (see Practical Aspects of the Method, Wear Tests), the analysis and interpretation of wear data from two-body wear tests become more complex when there is significant wear on both surfaces. A common practice is to use the accumulated wear on both surfaces, rather than wear rates, for comparing and evaluating wear behavior. However, wear rates can also be obtained in these situations and used to characterize wear behavior.

SAMPLE PREPARATION

Sample preparation and sample control are key elements in reducing scatter and obtaining meaningful results. This is also true of the other materials used in tribological testing, such as counterfaces, abrasives, fluids, and lubricants. In general, it is important to control the composition, purity, dimensions, and surface topography, as appropriate, of these components. When machining is required, the method used should be the same as or equivalent to those used in the applications and be well controlled.

Surface films can have significant effects on tribological behavior. Consequently, it is necessary to clean the surfaces of wear specimens and counterfaces prior to a test and to use procedures that will maintain this state of cleanliness. Cleaning techniques vary with the type of material being cleaned as well as the type of surface film or contamination that is present. In most cases the goal of cleaning is to remove oil and grease films that result from handling, storing, and machining. This is done using detergents or solvents that will not harm the surface. Sometimes scrubbing with a mild abrasive is effective; in such cases care must be taken to avoid altering the surface topography and to clean the surface afterward to remove residual abrasives. In some cases more elaborate cleaning procedures may be required. Gloves are frequently used to avoid recontamination from handling.

PROBLEMS

The major systemic problem in tribological testing is inconsistency or variability of test results. This results from a lack of adequate control, which is a necessary and basic element of any tribological testing methodology. Inconsistency and variability can be caused by a number of factors. There are several ways of minimizing this problem. The fundamental one is to improve the precision of the various elements of the test apparatus and the consistency in the procedures used in conducting the test. One method for determining what is acceptable is by performing repeated tests with well-controlled or reference materials and systematically evaluating each of the elements that can contribute to the variability or scatter. Once the apparatus and test method are developed to a satisfactory level, it is advisable to periodically perform standard tests with these same well-controlled materials to monitor the condition of the equipment. These tests can also be used to train personnel in the techniques of tribological testing, which is a significant factor in obtaining consistent results.

Another recommendation is that the calibration of load cells, profilometers, speed controls, and other equipment be checked on a regular basis or before the start of a test. The apparatus should also be examined for signs of wear or deterioration. Protocols for handling and storage of materials should be followed, and records of the sources and batches of these materials should be kept. Such records are often a help in sorting out material problems from equipment problems. Recording the temperature and humidity of the laboratory is also good practice, since these may affect tribological behavior. It is not uncommon to find seasonal variations in data, and thus it is advisable to conduct tribological tests in environmentally controlled rooms or enclosures.

LITERATURE CITED

American Society for Testing and Materials. 1991. ASTM D2266: Standard Test Method for Wear Preventive Characteristics of Lubricating Grease (Four-Ball Method). ASTM, West Conshohocken.

American Society for Testing and Materials. 1994. ASTM D2625: Standard Test Method for Endurance (Wear) Life and Load-Carrying Capacity of Solid Film Lubricants (Falex Pin and Vee Method). ASTM, West Conshohocken.

American Society for Testing and Materials. 1993. ASTM D3233: Standard Test Methods for Measurement of Extreme Pressure Properties of Fluid Lubricants (Falex Pin and Vee Block Methods). ASTM, West Conshohocken.

American Society for Testing and Materials. 1994. ASTM D3702: Standard Test Method for Wear Rate of Materials in Self-Lubricated Rubbing Contact Using a Thrust Washer Testing Machine. ASTM, West Conshohocken.

American Society for Testing and Materials. 1988. ASTM D4521: Standard Test Method for Coefficient of Static Friction of Corrugated and Solid Fiberboard (Inclined Plane Method). ASTM, West Conshohocken.

American Society for Testing and Materials. 1993. ASTM E303: Standard Test Method for Measuring Surface Frictional Properties Using the British Pendulum Tester. ASTM, West Conshohocken.

American Society for Testing and Materials. 1990. ASTM E707: Standard Test Method for Skid Resistance Measurements Using the North Carolina State University Variable-Speed Friction Tester. ASTM, West Conshohocken.

American Society for Testing and Materials. 1998. ASTM G73: Standard Practice for Liquid Impingement Erosion Testing. ASTM, West Conshohocken.

American Society for Testing and Materials. 1995. ASTM G75: Standard Test Method for Determination of Slurry Abrasivity (Miller Number) and Slurry Abrasion Response of Materials (SAR Number). ASTM, West Conshohocken.

American Society for Testing and Materials. 1995. ASTM G76: Standard Test Method for Conducting Erosion Tests by Solid Particle Impingement Using Gas Jets. ASTM, West Conshohocken.

American Society for Testing and Materials. 1998. ASTM G77: Standard Test Method for Ranking Resistance of Materials to Sliding Wear Using Block-on-Ring Wear Test. ASTM, West Conshohocken.

American Society for Testing and Materials. 1997. ASTM G81: Standard Practice for Jaw Crusher Gouging Abrasion Test. ASTM, West Conshohocken.

American Society for Testing and Materials. 1996. ASTM G83: Standard Test Method for Wear Testing with a Crossed-Cylinder Apparatus. ASTM, West Conshohocken.

American Society for Testing and Materials. 1991. ASTM G98: Standard Test Method for Galling Resistance of Materials. ASTM, West Conshohocken.

American Society for Testing and Materials. 1998. ASTM G115: Standard Guide for Measuring and Reporting Friction Coefficients. ASTM, West Conshohocken.

American Society for Testing and Materials. 1996. ASTM G118: Standard Guide for Recommended Data Format of Sliding

Wear Test Data Suitable for Databases. ASTM, West Conshohocken.

American Society for Testing and Materials. 1995. ASTM G134: Standard Test Method for Erosion of Solid Materials by a Cavitating Liquid Jet. ASTM, West Conshohocken.

American Society for Testing and Materials. 1997. ASTM G137: Standard Test Method for Ranking Resistance of Plastic Materials to Sliding Wear Using a Block-on-Ring Configuration. ASTM, West Conshohocken.

American Society for Testing and Materials. 1996. ASTM G143: Standard Test Method for Measurement of Web/Roller Friction Characteristics. ASTM, West Conshohocken.

Bayer, R. 1981. Influence of oxygen on the wear of silicon. *Wear* 49:235–239.

Bayer, R. 1983. The influence of hardness on the resistance to wear by paper. *Wear* 84:345–351.

Bayer, R. 1986. Impact wear of elastomers. *Wear* 112:105–120.

Bayer, R. 1994. Mechanical Wear Prediction and Prevention. Marcel Dekker, New York.

Bayer, R., Clinton, W. C., Nelson, C. W., and Schumacher, R. A. 1962. Engineering model for wear. *Wear* 5:378–391.

Bayer, R. and Sirico, J. 1971. Wear of electrical contacts due to small-amplitude motion. *IBM J. R & D* 15-2:103–107.

Blau, P. (ed.). 1992. Friction, Lubrication, and Wear Technology (ASME Handbook, Vol. 18). American Society for Metals International, Metals Park, Ohio.

Budinski, K.G. 1989. Friction of plastic films. *In* Proceedings of the International Conference on Wear and Materials (K.C. Ludema, ed.) pp. 459–460. American Society of Mechanical Engineers, New York.

Czichos, H. 1978. Tribology. Elsevier/North-Holland, Amsterdam.

Engel, P. 1976. Impact Wear of Materials. Elsevier/North-Holland, Amsterdam.

Jacko, M., Tsang, P., and Rhee, S. 1984. Automotive friction materials evolution during the past decade. *Wear* 100:503–515.

Ludema, K. and Bayer, R. (eds.). 1991. Tribological Modeling for Mechanical Designers (STP 1105). American Society for Testing and Materials, West Conshohocken.

Miller, J. and Miller, J. D. 1993. The Miller number: A review. *In* ASTM 1199: Tribology: Wear Test Selection for Design and Application (A.W. Ruff and R. G. Bayer, eds.) ASTM 1199. American Society for Testing and Materials, West Conshohocken.

Morrison, R. 1968. Load/life curves for gear and cam materials. *Machine Design* 8/68:102–108.

Peterson, M. B. and Winer, W. O. (eds.). 1980. Wear Control Handbook. American Society of Mechanical Engineers, New York.

Tsang, P., Jacko, M., and Rhee, S. 1985. Comparison of chase and inertial brake dynamometer testing of automotive friction materials. *In* Proceedings of the International Conference on Wear of Materials (K. C. Ludema, ed.) pp. 129–137. American Society of Mechanical Engineers, New York.

KEY REFERENCES

American Society for Testing and Materials. 1994. ASTM D2714: Standard Test Method for Calibration and Operation of the Falex Block-on-Ring Friction and Wear Testing Machine. ASTM, West Conshohocken.

Guidelines for the conduction of block-on-ring tests and a calibration procedure for this type of tester.

ASTM G73, 1998. See above.

Guidelines for the conduction and interpretation of liquid impingement erosion tests.

ASTM G76, 1995. See above.

General guidelines for solid-particle impingement tests.

ASTM G115, 1998. See above.

General guidelines for conducting friction tests and reporting friction data; listing of ASTM friction tests.

ASTM G119, 1993. Standard Guide for Determining Synergism between Wear and Corrosion.

Methodology for distinguishing between corrosion and wear in wear tests.

ASTM G133, 1995. Standard Test Method for Linearly Reciprocating Ball-on-Flat Sliding Wear.

Guidelines for conducting ball-on-flat wear tests.

ASTM STP 474, 1974. Characterization and Determination of Erosion Resistance.

Examples of erosion test methods and problems in erosion testing.

ASTM STP 615, 1976. Selection and Use of Wear Tests for Metals.

Wear tests for metals and wear test methodology.

ASTM STP 701, 1980. Wear Test for Plastics: Selection and Use.

Wear tests for plastics and their uses.

ASTM STP 769, 1982. Selection and Use of Wear Tests for Coatings.

Methodology for the selection of coating wear tests and specific tests.

ASTM STP 1010, 1988. Selection and Use of Wear Tests for Ceramics.

Wear test methodology and specific wear tests used for ceramics.

ASTM STP 1145, 1992. Wear and Friction of Elastomers.

Friction and wear behavior of elastomers; specific wear and friction tests used for elastomers.

ASTM STP 1199, 1993. Tribology: Wear Test Selection for Design and Application.

Examples of engineering uses of wear tests and the methodology used to develop wear tests to simulate applications.

Bayer, R. 1985. Wear Testing. *In* Mechanical Testing, Metals Handbook, Vol. 8, 9th ed. (J.R. Newby, J.R. Davis, and S.K. Refnes, eds.) American Society for Metals, Metals Park, Ohio.

General methodology for wear tests.

Bayer, R. 1994. Part B: Testing. *In* Mechanical Wear Prediction and Prevention. Marcel Dekker, New York.

Overview of wear test methodology; descriptions of individual tests methods and their uses.

INTERNET RESOURCES

tribology-request@dmx.epfl.ch

Tribology List Server, a network for exchanging and obtaining information on tribology. E-mail address to request membership.

www2.shef.ac.uk/uni/academic/I-M/mpe/tribology/tribo.html

Tribology. Web site listing sources for references on tribological subjects, professional organizations, publications, and meetings, maintained by the University of Sheffield, Sheffield, UK.

RAYMOND G. BAYER
Tribology Consultant
Vestal, New Work

THERMAL ANALYSIS

INTRODUCTION

Thermal analysis is the name given to the measurement of a sample as the sample is programmed through a predetermined temperature regime in a specified atmosphere. The word sample is used to cover the substance placed in the apparatus and its solid reaction products. The temperature regime may take several forms, e.g., controlled rate at a specified heating rate, temperature jump where the temperature is held constant, or simply an isothermal run where the temperature is quickly brought to a predetermined temperature and then held constant for the remainder of the experiment. With the advent of computer control and processing, a variety of other temperature programs are available. The substance is usually studied in the condensed phase, i.e., solid or liquid, but a change from solid to gas or liquid to gas may be monitored. The gas thus produced generally escapes from the system under study. Temperature control is usually exercised over the gaseous environment containing the sample. The pressure of this environment may be controlled as well, from a vacuum through pressures in excess of an atmosphere, depending on the material under study and the intent of the measurement. At one atmosphere the gaseous environment may be stable or dynamic, i.e., the gas may either not be changed during the experiment, or one may control the flow of gas over the sample. The atmosphere in such circumstances may be oxidative (i.e., air or oxygen) or inert (e.g., helium, argon, or nitrogen). The material under study is typically placed in a crucible, and the nature of the crucible over an extended temperature range (be it of ceramic or metal), such as its expansion properties and any shape or volume changes should be carefully noted.

Over the years a network of national societies has arisen and of necessity an international confederation has been formed—the International Confederation of Thermal Analysis and Calorimetry (ICTAC). This is affiliated with IUPAC, the International Union of Pure and Applied Chemistry—which by many is regarded as the authority in such matters as formal definitions and nomenclature. Equipment for thermal analysis has been developed, and most of the apparatus is commercially available.

This chapter presents a collection of thermal analysis techniques available for the characterization of the thermal properties of materials. The general nature of the available equipment along with the formal definitions, codes of practice, and nomenclature for each of the principal methods are described early in the chapter. The most obvious property that can be measured against an imposed temperature regime is the sample mass; this is termed thermogravimetric analysis. Predating this is differential thermal analysis in which the enthalpy change or temperature change is noted. If the temperature perturbation is simply noted, then the method is known as differential thermal analysis, but if this is converted by suitable calibration to energy input or output, then the method is termed differential scanning calorimetry. With advancements in computer control and data acquisition technology comes the development of elaborate simultaneous techniques of thermal analysis, in which the sample is subjected to two or more analyses during a potentially highly sophisticated temperature program. A material's nature and composition may also be elucidated by the analysis and quantification of evolved gases released during a heating regime.

In addition to providing the materials scientist with information on the thermal stability and degradation products of a material under study, thermal analysis methods can reveal the phase properties, mechanical stabilities, thermal expansion coefficients, and electrical or magnetic properties as a function of temperature. Other properties of materials that are accessible using thermal analysis techniques include the temperatures and heats of phase transition, heat capacity, glass transition points, thermal stability, and purity. While the temperature dependence of the various methods contained throughout this volume will be presented whenever applicable, this chapter will compile the methods whose primary concern is the thermal properties and behavior of the material.

DAVID DOLLIMORE

THERMAL ANALYSIS—DEFINITIONS, CODES OF PRACTICE, AND NOMENCLATURE

INTRODUCTION

Those who practice the techniques of thermal analysis are well advised to notice and use the definitions, conventions, codes of practice, and methods of reporting data advocated by the International Confederation of Thermal Analysis and Calorimetry (ICTAC). These recommendations have been widely circulated and are to be found in the booklet *For Better Thermal Analysis* (Hill, 1991).

The recommended names and abbreviations of the various techniques in thermal analysis are listed in Table 1. The classification of thermal analysis techniques is given in Table 2. The definitions follow from these tables.

NOMENCLATURE

Distinction must be made between the terms *derivative* and *differential*. Differential methods involve the measurement of a difference in the property between the sample and a reference (e.g., differential thermal analysis). Derivative techniques involve the measurement or calculation of the first derivative, usually with respect to time (e.g., derivative thermogravimetry; DTG). With this point in mind, we can then proceed to the formal definitions.

Table 1. Recommended Terminology

Name
A. General
Thermal analysis
B. Methods associated with mass change
1. Static
Isobaric mass change determination
Isothermal mass change determination
2. Dynamic
Thermogravimetry (TG)
Derivative thermogravimetry (DTG)
3. Methods associated with evolved volatiles
Evolved gas detection (EGO)
Evolved gas analysis[a] (EGA)
C. Methods associated with temperature change
Heating curve determination[b]
Heating-rate curves
Inverse heating-rate curves[b]
Differential thermal analysis (DTA)
Derivative differential thermal analysis
D. Methods associated with enthalpy change
Differential scanning calorimetry (DSC)
E. Methods associated with dimensional change
Thermodilatometry
Derivative thermodilatometry
Differential thermodilatometry
F. Multiple techniques
Simultaneous TG and DTA, etc.

[a]The method of analysis should be clearly stated, and abbreviations such as MTA (mass spectrometric thermal analysis) and MDTA (mass spectrometry and differential thermal analysis) avoided.
[b]When determinations are performed during the cooling cycle, these become cooling curves, cooling-rate curves, and inverse cooling-rate curves, respectively.

Table 2. Classification of Thermoanalytical Techniques

Physical Property	Derived Technique(s)
Mass	Thermogravimetry
	Isobaric mass change determination
	Evolved gas detection
	Evolved gas analysis
	Emanation thermal analysis
	Thermoparticulate analysis
Temperature	Heating-curve determination[a]
	Differential thermal analysis
Enthalpy	Differential scanning calorimetry[b]
Dimensions	Thermodilatometry
Mechanical characteristics	Thermomechanical analysis (TMA)
	Dynamic thermomechanometry
Acoustic characteristics	Thermosonimetry
	Thermoacoustimetry
Optical characteristics	Thermoptometry
Electrical characteristics	Thermoelectrometry
Magnetic characteristics	Thermomagnetometry

[a]When the temperature program is in the cooling mode, this becomes cooling-curve determination.
[b]The confusion that has arisen about this term seems best resolved by separating the two modes (power-compensation DSC and heat-flux DSC), as described in the text.

Thermogravimetry (TG) is a technique by which the mass of a substrate is measured as a function of temperature while the substance is subjected to a controlled temperature program. The record is the TG curve; the mass is normally plotted on the ordinate, decreasing down toward the origin, and temperature (T) or time (t) is on the abscissa, increasing from left to right according to the basic rules for plotting any kind of graph. In decomposition reactions, the mass of reactants disappears and is often replaced by the increasing mass of solid products formed. Accordingly, in this and subsequent definitions, *substance* is to be understood to include the total mass of the system being weighed or investigated.

Derivative thermogravimetry (DTG) is a technique that yields the first derivative of the TG curve with respect to either time or temperature. The DTG curve is plotted with the rate of mass loss on the ordinate, and T or t is on the abscissa.

Isobaric mass change determination refers to the equilibrium mass of a substance at a constant partial pressure of the volatile product(s) measured as a function of temperature while the substance is subjected to a controlled temperature program.

Evolved gas detection (EGD) is a technique by which the evolution of gas from a substance is detected as a function of temperature while the substance is subjected to a controlled temperature program.

Evolved gas analysis (EGA) is a technique by which the nature and/or amount of volatile product(s) released by a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program. The method of analysis should always be clearly stated.

Emanation thermal analysis is a technique by which the release of radioactive emanation (e.g., a radioactive gas such as radon) is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Thermoparticulate analysis is a technique by which the release of particulate matter from a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Heating curve determination is a technique in which the temperature of a substance is measured as a function of the program temperature while the substance is subjected to a controlled temperature program.

Heating-rate curves plot the first derivative of the heating curve with respect to time (i.e., dT/dt) against temperature or time. The function dT/dt is plotted on the ordinate and T or t on the abscissa.

Inverse heating-rate curves plot the first derivative of the heating curve with respect to measured temperature (i.e., dt/dT) against either temperature or time. The function dt/dT is plotted on the ordinate and T or t on the abscissa.

Differential thermal analysis (DTA) is a technique in which the temperature difference between a substance and a reference material is measured as a function of temperature while the substance and reference material are subjected to a controlled temperature program. The

record is the DTA curve; the temperature difference (ΔT) is plotted on the ordinate, and T or t is on the abscissa. The temperature difference is considered to be negative in endothermic and positive in exothermic reactions.

When the equipment is calibrated, the DTA unit can be made quantitative with respect to the heat changes involved. The present tendency is to call this approach *differential scanning calorimetry* (DSC). DSC measures the difference in energy inputs into a substance and a reference material as a function of temperature while the substance and reference are subjected to a controlled temperature program.

Two modes—power-compensation DSC and heat-flux DSC—can be distinguished, depending on the method of measurement used. Two types of heat-flux calorimeters for DSC are produced commercially. The system may be provided with multiple temperature sensors (e.g., a Calvet-type arrangement) or with a controlled leak (e.g., a Boersma-type arrangement). For a general discussion on this topic see DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY or Haines and Wilburn (1995).

Thermodilatometry is a technique in which a dimension of a substance under negligible load is measured as a function of temperature while the substance is subjected to a controlled temperature program. The record is the thermodilatometric curve; the dimension is plotted on the ordinate, decreasing toward the origin, and T or t is on the abscissa. Linear thermodilatometry and volume thermodilatometry are distinguished on the basis of the dimensions measured.

Thermomechanical analysis (TMA) is a technique in which the deformation of a substance under a nonoscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature program. The mode, as determined by the type of stress applied (compression, tension, flexure, or torsion), should always be stated.

Dynamic thermomechanometry is a technique in which the dynamic modulus and/or damping of a substance under an oscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature program. Torsional braid analysis is a particular case of dynamic thermomechanometry in which the material is supported on a braid.

Thermosonimetry is a technique in which the sound emitted by a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Thermoacoustimetry is a technique in which the characteristics of imposed acoustic waves are measured as a function of temperature while the substance is subjected to a controlled temperature program.

Thermoptometry is a technique in which an optical characteristic of a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program (for a general discussion on this topic, see Reading and Haines, 1995).

Measurement of total light, light of specific wavelength(s), refractive index, and luminescence leads to thermophotometry, thermospectrometry, thermorefractometry, and thermoluminescence, respectively; observation under the microscope leads to thermomicroscopy.

Thermoelectrometry is a technique in which an electrical characteristic of a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program. The most common measurements are of resistance, conductance, and capacitance (see ELECTRICAL AND ELECTRONIC MEASUREMENT).

Thermomagnetometry is a technique in which the magnetic susceptibility of a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program (see MAGNETISM AND MAGNETIC MEASUREMENT).

Sometimes in an investigation, a sample is characterized by more than one technique. The term for this is *simultaneous techniques*, and an example is simultaneous TG and DTA. The individual techniques are separated with a hyphen; thus the appropriate abbreviation is *simultaneous TG-DTA*. Unless contrary to established practice (some journals still do not accept the International Union of Pure and Applied Chemistry's (IUPAC) recommendations), all abbreviations should be written in capital letters without periods.

Gas analysis is rarely practiced as a separate thermal analysis technique. The two instruments involved are usually connected through an interface, e.g., DTA and mass spectrometry (MS). The term *interface* refers to a specific piece of equipment that enables the two instruments to be joined together. In this case, the DTA and the gas sampling occur at 1 atmosphere but the mass spectrometer must operate under a vacuum, and an interface is necessary for this to occur. Such a process is often described as a *coupled* simultaneous technique. Although the word simultaneous is used, in fact the two measurements are rarely absolutely simultaneous. Thus in gas analysis with thermogravimetry, the mass change is detected first, followed by gas sampling and gas analysis (MS) a finite time later. In such cases, the technique used first appears first in the abbreviation, e.g., TG-MS, not MS-TG. Sometimes the second technique does not involve continuous sampling. The proper term to use in such cases is *discontinuous* simultaneous techniques. An example is DTA and gas chromatography, in which discrete portions of evolved volatiles are collected from the sample situated in the instrument used for the first technique and then are analyzed by the second instrument.

SYMBOLS

The abbreviations for each technique have already been introduced. The confusion that can arise, however, between the use of, e.g., TG and T_g (glass-transition temperature), has led a number of investigators and instrument manufacturers to use TGA for TG. There are other aspects to the use of symbols, and a report of a subcommittee chaired by J. H. Sharp (1986) and presented to the Nomenclature Committee of the ICTA is summarized here.

1. The international system of units (SI) should be used wherever possible.

2. The use of symbols with superscripts should be avoided if possible.
3. The use of double subscripts should be avoided if possible.
4. The symbol T should be used for temperature, whether expressed in degrees Celsius (°C) or in kelvin (K). For temperature intervals, the symbol °C or K can be used.
5. The symbol t should be used for time, whether expressed in seconds (s), minutes (min), or hours (h).
6. The heating rate can be expressed either as dT/dt, when a true derivative is intended, or as β in K min^{-1} or °C min^{-1}. The heating rate so expressed need not be constant and can be positive or negative. If an isothermal experiment is carried out, it is best to state this fact in full.
7. The symbols m for mass and W for weight are recommended.
8. The symbol α is recommended for the fraction reacted or changed.
9. The following rules are recommended for subscripts.
   a. When the subscript relates to an object, it should be a capital letter, e.g., m_S is the mass of the sample and T_R is the temperature of the reference material.
   b. When the subscript relates to an action, it should be a lower case letter, e.g., T_g is the glass-transition temperature, T_c is the temperature of crystallization, T_m is the melting temperature, and T_σ is the solid-state-transition temperature.
   c. When the subscript relates to a specific point in time or point on a curve, it should be a lower case letter or a figure, e.g., T_i is the initial temperature, $m_{0.5}$ is the mass at which 50% of the material reacted, $t_{0.5}$ is the time at which 50% of the material reacted, $T_{0.3}$ is the temperature at which 30% of the material reacted, T_p is the peak temperature, and T_e is the extrapolated onset temperature (see description of TG technique).

STANDARDIZATION

It must be recognized that thermal methods of analysis can be applied to all states and conditions of matter. This flexibility is demonstrated by the design of equipment and choice of experimental conditions; therefore, it is imperative to select standards. Furthermore, no single thermal analysis instrument design or set of experimental conditions is optimum for all studies. The resultant data depend heavily on procedure. The major requirements in selecting standards are

1. To provide a common basis for selecting independently acquired data.
2. To enable equipment to be calibrated independently of the design of the particular instruments.
3. To provide the means for relating thermoanalytical data to physical and chemical properties determined by conventional isothermal procedures.

Such standards must be applicable to whatever experimental design is required for a particular purpose; otherwise, the value of the results is isolated, and the full potential of thermal analysis will not be realized. If curves of standard materials obtained under the conditions of a particular study are reported, readers will be able to relate the study results to the performance of their own instrumentation, to evaluate the quality of the published data, and to derive their own conclusions.

With this as its goal, the ICTA committee on standardization produced several sets of reference materials. Each instrument manufacturer will provide detailed advice on the calibration of its equipment using the proposed standards (Hill, 1991; Charsley et al., 1993). The reference materials and methods of calibration are currently under review by the German Thermal Analysis Society (GEFTA), and have not yet been approved by ICTAC or IUPAC.

The sets of ICTA reference materials include eight inorganic materials that exhibit solid-phase transitions, two high-purity metals, and four organic compounds (mainly of interest for their melting points). All standards are for temperature calibration in the heating mode and under normal operating conditions; consequently, the temperatures quoted are generally higher than the equilibrium values given in the literature. The temperature ranges of NIST-ICTAC (formerly known as NBS-ICTAC when NIST was named the National Bureau of Standards) certified reference materials are detailed in Table 3. Table 4 lists the transition temperatures that can serve as calibration points.

REPORTING THERMAL ANALYSIS DATA

ICTAC committees have defined *good practice* for both experimentation and reporting, so the information obtained

Table 3. Sets of Reference Materials Available[a]

GM-757 180–330 K	GM-758 125–435°C	GM-759 295–575°C	GM-760 570–940°C
1,2-Dichloroethane (melting)	Potassium nitrate	Potassium perchlorate	Quartz
Cyclohexane (transition, melting)	Indium	Silver sulfate	Potassium sulfate
Phenyl ether (melting)	Tin	Quartz	Potassium chromate
o-Terphenyl (melting)	Potassium perchlorate	Potassium sulfate	Barium carbonate
	Silver sulfate	Potassium chromate	Strontium carbonate

[a]Set GM-761 (a standard polystyrene sample with $T_g = 100$°C) is also available.

Table 4. Transition Temperatures (°C) of Reference Materials[a]

Material	Transition Type	Equilibrium Transition Temperature, °C	Extrapolated Onset Temperature, °C
	Inorganic		
KNO_3	Solid-solid	127.7	128 ± 5
In	Solid-liquid	156.6	154 ± 6
Sn	Solid-liquid	231.9	230 ± 5
$KClO_4$	Solid-solid	299.5	199 ± 6
Ag_2SO_4	Solid-solid	424	242 ± 7
SiO_2	Solid-solid	573	571 ± 5
K_2SO_4	Solid-solid	583	582 ± 7
K_2CrO_4	Solid-solid	665	665 ± 7
$BaCO_3$	Solid-solid	810	808 ± 8
$SrCO_3$	Solid-solid	925	928 ± 7
	Organic		
p-Nitrotoluene	Solid-liquid	51.5	51.5
Hexachloroethane	Solid-solid	71.4	—
Naphthalene	Solid-liquid	80.3	80.4
Hexamethyl benzene	Solid-solid	110.4	—
Benzoic acid	Solid-liquid	122.4	122.1
Adipic acid	Solid-liquid	151.4	151.0
Anisic acid	Solid-liquid	183.0	183.1
2-Chloroanthraquinone	Solid-liquid	209.1	209.4
Carbazole	Solid-liquid	245.3	245.2
Anthraquinone	Solid-liquid	284.6	283.9

[a]There are many references to these standards (Hill, 1991; Charsley et al., 1993).

and published is of maximum value. The recommendations of the committee were developed for authors, editors, and referees and are summarized below.

To accompany each DTA, TG, EGA, EGD, and/or thermochemical record, the following information should be reported.

1. Identification of all substances (sample, reference, diluent) by a definitive name and empirical formula or equivalent compositional data.
2. A statement of the source of all substances and details of their histories, pretreatments, and chemical purities, so far as these are known.
3. Measurement of the average rate of linear temperature change over the temperature range involving the phenomena of interest. Nonlinear temperature programming should be described in detail.
4. Identification of the sample atmosphere by pressure, composition, and purity and whether the atmosphere is static, self-generated, or dynamic throughout the sample. When applicable, the ambient atmospheric pressure and humidity should be specified. If the pressure is other than atmospheric, full details of the method of control should be given.
5. A statement of dimensions, geometry, and materials of the sample holder.
6. A statement of the method of loading (quasi-static, dynamic) when applicable.
7. Identification of the abscissa scale in terms of time or of temperature at a specified location. Time or temperature should be plotted to increase from left to right.
8. A statement of the method used to identify intermediates or final products.
9. Faithful reproduction of all original data.
10. Identification of the apparatus used by type and/or commercial name, together with details of the location of the temperature-measuring thermocouple.

THERMODYNAMIC BASIS FOR THERMAL ANALYSIS

Thermal analysis poses a problem in that some samples will achieve equilibrium at all temperatures, whereas others will be affected by the kinetics of chemical change. In many (but not all) DSC experiments, the interpretation can be solely in terms of thermodynamics. In many thermogravimetric experiments, a kinetic approach is possible. The reason for this lies in the conditions of the experiment imposed on the sample—it is partly owing to the experimental technique and partly owing to the nature of the sample under investigation.

This unit will not discuss thermodynamic relationships in detail; such information can be found in any textbook on physical chemistry (e.g., Atkins, 1994). In brief, however, the three laws of thermodynamics may be cited (also see THERMOMETRY). The first law of thermodynamics concerns the principle of the conservation of energy. At constant

pressure, the heat q_p absorbed by the system under consideration is given by the change in enthalpy ΔH:

$$\Delta H = q_p \tag{1}$$

Standard conditions in the past have been quoted: for gases, liquids, and solids (pure material) at the standard pressure of 1 atmosphere. The standard enthalpy of formation ΔH_f^0 is the change in standard enthalpy when 1 mol of substance is formed from its elements in their standard states; however, the definition of standard conditions is in a state of change, because recent textbooks are following the recommendations of IUPAC to quote the standard state as 1 bar (1 bar = 0.09869 atm).

For example,

$$\text{Zn (solid)} + \frac{1}{2}\text{O}_2(\text{gas}, 1\,\text{bar}) = \text{ZnO(solid)} \tag{2}$$

$$\Delta H^0_{\text{fZn0(solid)}} = -348.28\ \text{kJ mol}^{-1} \tag{3}$$

The heat absorbed by a system undergoing no chemical change is defined for a change in temperature of 1 K at constant pressure as the heat capacity at constant pressure C_p, which can be written:

$$C_p = \left(\frac{\partial H}{\partial T}\right)_p \tag{4}$$

This is itself a term that varies with temperature and is often given in the form:

$$C_p = a + bT + cT^2 \tag{5}$$

Thus for a system undergoing a reaction under standard conditions, the difference in enthalpy between products and reactants is given as ΔH^0, the heat released by the system. In the decomposition of a solid, such as calcite, however, the solid might be in a nonstandard condition. Nevertheless, the enthalpy of the solid-phase materials under standard conditions can be calculated for the dissociation:

$$\text{CaCO}_3\,\text{(solid)} = \text{CaO (solid)} + \text{CO}_2\,\text{(gas)} \tag{6}$$

$$\Delta H^0 = 178\ \text{kJ mol}^{-1} \tag{7}$$

The positive value of ΔH^0 implies that heat must be supplied to the system, i.e., it is an endothermic process.

An example of an exothermic reaction is

$$\text{CH}_4(\text{gas}) + 2\,\text{O}_2(\text{gas}) \rightarrow \text{CO}_2(\text{gas}) + 2\text{H}_2\text{O(liquid)} \tag{8}$$

$$\Delta H^0 = -890\ \text{kJ mol}^{-1} \tag{9}$$

The negative sign implies the process is exothermic.

To adjust the standard-state values to other conditions of pressure and temperature requires a knowledge of C_p for all chemical species involved and of how C_p varies with temperature.

Reactions in an isolated system reach an equilibrium state. Thus for

$$\text{CaCO}_3\,\text{(solid)} \rightarrow \text{CaO (solid)} + \text{CO}_2\,\text{(gas)} \tag{10}$$

it is possible to write

$$K_p = \frac{a_{\text{CaO}} a_{\text{CO}_2}}{a_{\text{CaCO}_3}} \tag{11}$$

where K_p is the equilibrium constant (which assumes a characteristic value at each temperature) and a_{CaO}, a_{CO_2} and a_{CaCO_3} are the activities of the indicated chemical species. Usually, the activities of the solid phase are assigned a value of unity, and a_{CO_2} can be approximated to the pressure P_{CO_2}, so that:

$$K_p = P_{\text{CO}_2} \tag{12}$$

It should be noted, however, that if the carbon dioxide is kept below this value and removed as it is formed, the reaction will proceed to completion.

The tendency for reactions to proceed is covered by the second law of thermodynamics. This driving force depends both on the change in enthalpy ΔH and on the change in entropy ΔS. The third law of thermodynamics states that entropy can be assigned a zero value at the temperature of absolute zero for perfectly crystalline solids. A simple way in which enthalpy and entropy changes can be combined to demonstrate this tendency for a transition to proceed is via the Gibbs free energy (G). The relationship is

$$G = H - TS \tag{13}$$

and for a transition at constant temperature:

$$\Delta G = \Delta H - T\Delta S \tag{14}$$

For a spontaneous reaction, ΔG is negative. The standard change in Gibbs free energy is related to the equilibrium constant K as follows:

$$\Delta G^0 = -RT \ln K \tag{15}$$

where R is the gas constant. The variation with temperature is given by

$$\ln K = -\frac{\Delta H^0}{RT} + \frac{\Delta S^0}{R} \tag{16}$$

which is called the van't Hoff equation.

All these relationships find application in the interpretation of thermal analysis data. One further point is that at equilibrium,

$$\Delta G = 0 \tag{17}$$

when

$$T_{\text{transition}} = \frac{\Delta H}{\Delta S} \tag{18}$$

This sets out a thermodynamic decomposition temperature under standard conditions. Thus for the dissociation of calcium carbonate:

$$T_{\text{dissociation}} \approx 950^{\circ}\text{C} \tag{19}$$

This equation must be regarded as giving an approximate temperature for the decomposition, because both ΔH and ΔS vary with temperature. Reference data for ΔH and ΔS are usually reported at room temperature. Furthermore, the experimental data of thermal analysis probably do not conform to standard conditions. For vaporization, T_{vap} at 1 atmosphere (the normal boiling point) is given by:

$$T_{\text{vap}} = \frac{\Delta H_{\text{vap}}}{\Delta S_{\text{vap}}} \tag{20}$$

If ΔS_{vap} is assumed to be constant for most liquids, then:

$$\Delta S \; T_{\text{vap}} = \Delta H_{\text{vap}} \tag{21}$$

or

$$T_b \cdot 85\,\text{J}\,\text{K}^{-1}\text{mol}^{-1} = \Delta H_{\text{vap}} \tag{22}$$

where T_b is the normal boiling point. Expressed this way, the above equation is a mathematical statement of Trouton's rule. Most of the relationships outlined above can be used for interpreting thermal analysis data.

KINETIC BASIS FOR THERMAL ANALYSIS

In many systems investigated using a rising temperature method, it is obvious that an equilibrium condition is not found. The classical way in which kinetic data are evaluated is by way of a specific reaction rate, defined by

$$\frac{d\alpha}{dt} = kf(\alpha) \tag{23}$$

where α is the fractional extent of the reaction and t is the time in an isothermal mode. The term $f(\alpha)$ often takes an "order" form:

$$\left.\begin{array}{r} \alpha \to 1 \\ f(\alpha) \to 1 \\ \dfrac{d\alpha}{dt} = k \end{array}\right\} \tag{24}$$

where k is taken to be the specific reaction rate and α is used here instead of the conventional concentration or pressure term, because there is no concentration term in solid-state reactions. If the temperature is varied at a heating rate of β in degrees Celsius per minute, then:

$$\beta = \frac{dT}{dt} \tag{25}$$

where t is the time and T is the temperature. From plots of α against T in such circumstances, the following relationship is obtained:

$$\frac{d\alpha}{dt} = \left(\frac{d\alpha}{dT}\right)\left(\frac{dT}{dt}\right) = \left(\frac{d\alpha}{dT}\right)\beta \tag{26}$$

Thus at any point on the thermal analysis plot (i.e., at any given temperature):

$$\left(\frac{d\alpha}{dT}\right)\beta = kf(\alpha) \tag{27}$$

$$k = \frac{(d\alpha/dT)\beta}{f(\alpha)} \tag{28}$$

If an Arrhenius relation is assumed, then:

$$k = Ae^{-E_{\text{act}}/RT} \tag{29}$$

and

$$\ln k = \ln A - \frac{E_{\text{act}}}{RT} \tag{30}$$

when

$$\ln\left[\frac{(d\alpha/dT)\beta}{f(\alpha)}\right] = \ln A - \frac{E_{\text{act}}}{RT} \tag{31}$$

When E_{act} is the activation energy and A is the Arrhenius preexponential constant, Equation 31 provides a route to the calculation of A and E_{act} from thermal analysis data, if $f(\alpha)$ can be determined. There are numerous publications dealing with this topic in detail (e.g., Dollimore and Reading, 1993).

PUBLICATIONS ON THERMAL ANALYSIS

Because of the diverse applications of thermoanalytical methods, the literature is widely scattered. Two international journals for publication of papers on thermal analysis—*Journal of Thermal Analysis* and *Thermochimica Acta*—are available, but both tend to publish data outside the precise definition of thermal analysis, particularly with reference to calorimetry. Ephemeral publications such as the newsletters issued by various societies to members (e.g., *ICTA Newsletter, Bulletin de 1'Association Française de Calorimetrie et d'Analyse Thermique, NATAS Notes*, and *Aicat Notizie*), serve a useful purpose by carrying information on recent developments, meetings, books, etc., that may be of interest.

Chemical Abstracts scans all entries for articles that deal with thermal analysis and produces *Chemical Abstracts Thermal Analysis Selects* every 2 weeks. These issues are used as the basis for a biannual review, which appears in *Analytical Chemistry*. The abstracts are available in computer-compatible form. Many articles on thermal analysis are included in materials science and pharmacological journals.

Every 4 years, the International Confederation for Thermal Analysis and Calorimetry holds its international conference. In the years between international conferences, the European societies hold a thermal analysis conference. The North American Thermal Analysis Society holds a conference every year, except when it hosts the international conference. The individual societies hold regular meetings, and some hold an annual conference. The papers presented at such conferences are usually published as edited books. *Thermochimica Acta* and *Journal of Thermal Analysis* both publish special issues of conference papers, on specific topics, or in honor of well-known scientists. There are an increasing number of books and monographs on thermal analysis, some of which serve as good textbooks. Hill (1991) includes a good bibliography.

Each year *The International Thermal Analysis, Calorimetry and Rheology Buyer's Guide* is published by B&K Publishing Company. The guide contains consumables, consultants, contract analytical service providers, providers of certified reference materials, instruments (new, used and refurbished), and reference literature/resources along with advertisements from vendors of thermal analysis equipment. The guide may be requested by sending an e-mail to guiderequest@bkpublishing.com. A weekly newsletter is also published and may be received via email by sending the message SUBSCRIBE to subscribe@bkpublishing.com.

LITERATURE CITED

Atkins, P. W. 1994. Physical Chemistry, 5th ed. W. H. Freeman, New York.

Charsley, E. L., Earnest, C. M., Gallagher, P. K., and Richardson, M. J. 1993. Preliminary round-robin studies on the ICTAC certified reference materials for DTA. *J. Therm. Anal.* 40:1415–1422.

Dollimore, D. and Reading, M. 1993. Application of thermal analysis to kinetic evaluation of thermal decomposition. *In* Treatise on Analytical Chemistry, Part 1, Vol. 13, 2nd ed. Thermal Methods. (J. D. Wineforder, D. Dollimore, and J.G. Dunn, eds.). pp. 1–61. John Wiley & Sons, New York.

Haines, P. J. and Wilburn, F. W. 1995. Differential thermal analysis and differential scanning calorimetry. *In* Thermal Methods of Analysis. (P. J. Haines, ed.). pp. 63–122. Blackie Academic and Professional, London.

Hill, J. O. 1991. For Better Thermal Analysis and Calorimetry, 3rd ed. International Confederation for Thermal Analysis.

Reading, M. and Haines, P. J. 1995. Thermomechanical, dynamic mechanical and associated methods. *In* Thermal Methods of Analysis. (P. J. Haines, ed.). pp. 123–160. Blackie Academic and Professional, London.

Sharp, J. H. 1986. *Thermochim. Acta* 104:395.

KEY REFERENCES

Atkins, 1994. See above.

Outlines new standards used in thermodynamics.

Dollimore and Reading, 1993. See above.

Outlines salient features of kinetic interpretations of thermal analysis data.

Haines, P.J. (ed.), 1995. Thermal Methods of Analysis. Blackie Academic and Professional, London.

The best available primer on the subject.

INTERNET RESOURCES

http://www.rsc.org/lap/rsccom/dab/ana012list.htm

The Royal Society of Chemistry maintains a thermal methods listserver, which may be subscribed to by visiting this site.

http://www.bkpublishing.com/thermal_listserver.htm

The THERMAL listserver is another e-mail discussion group with participants from the thermal analysis community.

http://www.natasinfo.org

The North American Thermal Analysis Society is a valuable resource for individuals interested in thermal analysis methods.

DAVID DOLLIMORE
The University of Toledo
Toledo, Ohio

THERMOGRAVIMETRIC ANALYSIS

INTRODUCTION

The formal definition of thermogravimetry (TG is the preferred abbreviation, although TGA is also used) has been given by the Nomenclature Committee of the International Confederation for Thermal Analysis and Calorimetry (ICTAC) as "a technique in which the mass of a substance is measured as a function of temperature whilst the substance is subjected to a controlled temperature programme" (Mackenzie, 1979, 1983).

In the above definition, a controlled temperature program means heating or cooling the sample at some predetermined and defined rate. Although it is common to have just one constant heating or cooling rate, it is also advantageous in some cases to have different rates over different temperature ranges and in some cases even a varying rate over a specific temperature range. Thus the heating rate could be started at 20°C/min up to 200°C, followed by an isothermal hold at 200°C for 10 min and then by further heating at 10°C/min.

An apparatus called a thermobalance is used to obtain a thermogravimetric curve. By means of a thermobalance the temperature range over which a reaction involving mass change occurs may be determined.

The sample is usually a solid, or more rarely a liquid. The heating/cooling rates typically cover the range 1° to $100°C/min^{-1}$, and the atmosphere can be varied from an inert one such as nitrogen to a reactive one such as oxygen or sulfur dioxide. Vacuum can be applied, with specialized instruments being able to achieve a hard vacuum up to 10^{-6} Torr. Experiments can usually be conducted in a relatively short time, so that at a typical heating rate of 20°C/min a sample can be heated to 1000°C in 50 min. Once the experiment is set up, it requires no further attention. If the furnace is fitted with a cooling device, then turnaround time is rapid and a new experiment can be started within 15 min.

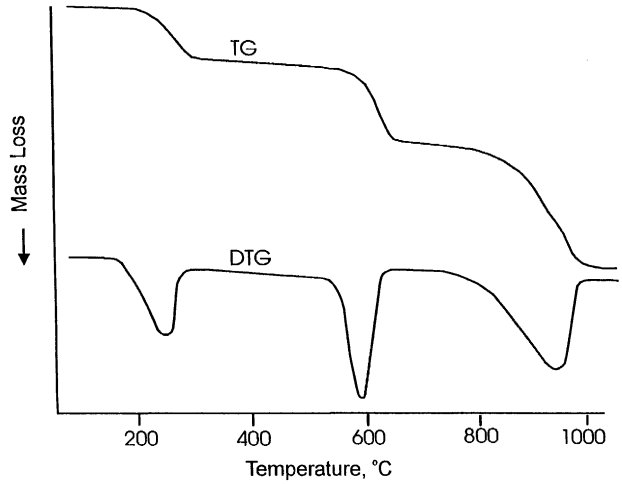


Figure 1. Schematic of TG and DTG curves.

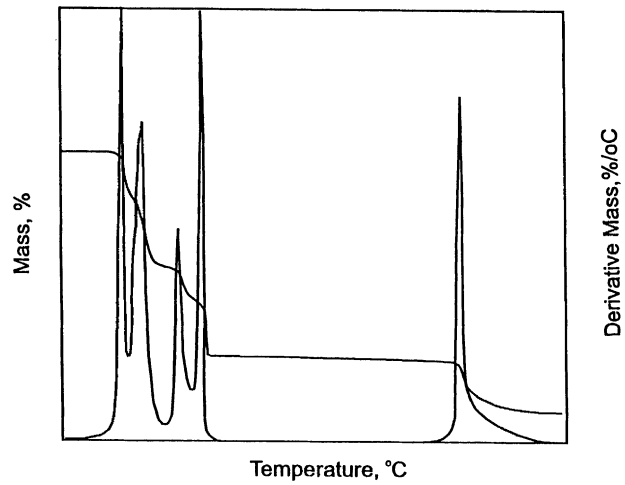


Figure 2. TG curve obtained from a controlled-rate experiment, in which the rate of mass loss is kept nearly constant,and the variable is the heating rate.

Thermogravimetric data can be presented in two ways. The TG curve is a plot of the mass against time or temperature, with the mass loss on the ordinate plotted downward and mass gains plotted upward relative to a baseline. Alternatively, data can be presented as a derivative thermogravimetric (DTG) curve, which is a plot of the rate of change of mass m with respect to time t or temperature T against time or temperature. The DTG mass losses should also be plotted downward and the gains upward. Schematic TG and DTG curves are shown in Figure 1. In practice, the mass losses may not nearly be so well resolved since the reactions are temperature-dependent rate processes that require time to reach completion. Hence the first reaction may not be complete before the second reaction commences, and the baseline between the reactions may be a slope rather than a plateau region. When a horizontal plateau is obtained in a TG curve, a minimum at which $dm/dt = 0$ is observed in the DTG curve (Fig. 1). A maximum in the DTG curve corresponds to a point of inflection in the TG curve at which mass is lost (or gained) most rapidly. Minima can also occur in a DTG curve at which $dm/dt > 0$, corresponding to a point of inflection in the TG curve rather than to a horizontal plateau.

Conventionally, thermal analysis experiments are carried out at a constant heating rate, and a property change is measured as a function of time. An alternative approach is to keep the change in property constant by varying the heating rate (Paulik and Paulik, 1971; Reading, 1992; Rouquerol, 1989). For TG, the rate of mass loss is kept constant by variation in heating rate. [This technique is given various names, although controlled-rate thermal analysis (CRTA) appears to have become the accepted terminology.] To achieve this, the mass change is monitored and the heating rate decreased as the mass loss increases, and vice versa. At the maximum rate of mass loss, the heating rate is a minimum. This gives mass losses over very narrow temperature ranges and sometimes enables two close reactions to be resolved (see Fig. 2). This method has the advantage of using fast heating rates when no thermal event is taking place and then slowing down the heating rate when a mass change is in progress.

Thermogravimetry does not give information about reactions that do not involve mass change, such as polymorphic transformations and double-decomposition reactions. Also, it is not useful for identification of a substance or mixture of substances unless the temperature range of the reaction has already been established and there are no interfering reactions. However, when a positive identification has been made, TG by its very nature is a quantitative technique and can frequently be used to estimate the amount of a particular substance present in a mixture or the purity of a single substance. The types of processes that can be investigated by TG are given in Table 1.

There is significant literature on TG, and for more extensive information the reader is referred to Keattch and Dollimore (1975), Earnest (1988), Wendlandt (1986), Dollimore (1992), Dunn and Sharp (1993), Haines (1995), and Turi (1997).

Other methods are available for the measurement of mass change as a function of temperature, although they are often limited to changes at set temperatures. Two modes of measurement can be distinguished:

Table 1. Processes That Can be Studied by TG

Process	Mass Gain	Mass Loss
Adsorption and absorption	•	
Desorption		•
Dehydration or desolvation		•
Vaporization		•
Sublimation		•
Decomposition		•
Oxidation	•	•
Reduction		•
Solid-gas reactions	•	•
Solid-solid reactions		•

1. The sample is heated to a specific temperature and the mass change followed at that temperature. After a time lapse, the sample may be heated to a higher constant temperature (or cooled to a lower constant temperature) and again the mass change observed. Quartz spring balances are often used in this application, which are cheap to construct but cumbersome to use. These systems are used to study, e.g., the adsorption/desorption of a gas on a solid at different temperatures.
2. The sample is weighed at room temperature, heated to a constant temperature for a certain time, then removed, cooled back to room temperature, and reweighed. This is the process used in gravimetric analysis and loss on ignition measurements. This method is inexpensive. However, unless the hot sample is put into an inert environment, the sample can readsorb gases from the atmosphere during cooling and hence give anomalous results.

The major disadvantage of both of the above methods is that the mass change is not followed continuously as a function of temperature.

The complementary techniques of differential thermal analysis (DTA) and differential scanning calorimetry (DSC) are dealt with in DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY. Thermogravimetry can be combined with either DTA or DSC in a single system to give simultaneous TG-DTA or TG-DSC; these techniques will be discussed in SIMULTANEOUS TECHNIQUES INCLUDING ANALYSIS OF GASEOUS PRODUCTS. To characterize the sample as well as identify reaction products at various temperatures, a range of other techniques are used, including wet chemical analysis, x-ray diffraction (XRD) (X-RAY TECHNIQUES), spectroscopic techniques such as Fourier transform infrared (FTIR) spectroscopy, optical techniques such optical microscopy (OM) (OPTICAL MICROSCOPY, REFLECTED-LIGHT OPTICAL MICROSCOPY), and scanning electron microscopy (SEM) (SCANNING ELECTRON MICROSCOPY). The information obtained from these techniques enables the reaction intermediates to be identified so that the reaction scheme can be written with some confidence.

Other techniques placed by the ICTAC Nomenclature Committee (Mackenzie, 1979, 1983) within the family of thermal analysis methods based on changes of mass are evolved gas detection, evolved gas analysis, and emanation thermal analysis. Emanation thermal analysis [developed by Balek (1991)] involves the release of radioactive emanation from a substance, which gives a change in mass if, for example, α-particles are emitted.

PRINCIPLES OF THE METHOD

The form of the TG/DTG curve obtained experimentally is dependent on the interplay of two major factors: the properties of the sample and the actual experimental conditions used, also called procedural variables. Both factors can affect the kinetics of any reaction that takes place, so that a change in either will have a subsequent effect on the form of the TG curve. It is also important to note that unless the sample is held at constant temperature, applying a heating or cooling rate produces nonequilibrium conditions.

It is possible to calculate a theoretical TG curve if the kinetic mechanism and parameters are known, on the assumption that heat transfer is instantaneous and no temperature gradient exists within the sample. Thus the kinetics of most reactions under isothermal conditions can be summarized by the general equation

$$\frac{d\alpha}{dt} = kf(\alpha) \tag{1}$$

Here α is the fraction reacted in time t and is equal to $w_i - w_t/w_i - w_f$, where w_i is the initial weight of the sample, w_t is the weight at time t, and w_f is the final weight of the sample; k is the rate of reaction; and (α) is some function of α. The various forms adopted by the function (α) and the integrated forms $g(\alpha)$ have been discussed elsewhere (Keattch and Dollimore, 1975; Satava and Skvara, 1969; Sharp, 1972; Sestak, 1972) and are given in the next section.

The temperature dependence of the rate constant follows the Arrhenius equation:

$$k = Ae^{-E/RT} \tag{2}$$

where T is the absolute temperature, A is the preexponential factor, E is the activation energy, and R is the gas constant.

For a linear heating rate,

$$T = T_0 + \beta t \tag{3}$$

where T_0 is the initial temperature and β is the heating rate.

Combination of Equations 1 and 2 give

$$\frac{d\alpha}{dt} = Af(\alpha)e^{-E/RT} \tag{4}$$

Substitution for dt using Equation 3 gives

$$\frac{d\alpha}{dT} = \frac{A}{\beta} f(\alpha)e^{-E/RT} \tag{5}$$

which, if rearranged, provides

$$\frac{d\alpha}{f(\alpha)} = \frac{A}{\beta} e^{-E/RT} dT \tag{6}$$

Equation 6 is the basic equation of the DTG curve, which when integrated is the equation of the TG curve. If the form of the function $f(\alpha)$ is known, integration of the left-hand side of the equation is straightforward and gives the associated function $g(\alpha)$. The integration limits are between the initial and final temperatures of the reaction, or between $\alpha = 0$ and $\alpha = 1$. The values of E and A have a marked influence on the temperature range over which the TG curve is observed, but they do not influence

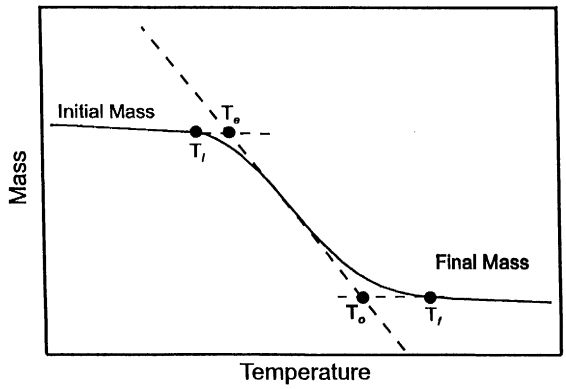


Figure 3. Various temperatures used to define the TG curve.

the shape of the curve too greatly (Satava and Skvara, 1969). The kinetic mechanism, i.e., the form of $f(\alpha)$ or $g(\alpha)$, however, determines the shape of the curve.

The TG curve can be defined in several ways. The temperature at which a change in mass is first detected—called the initial temperature T_i (see Fig. 3), or onset temperature, or procedural decomposition temperature—is not sufficiently well defined to use as a satisfactory reference point, since its detection is dependent on factors such as the sensitivity of the TG apparatus and the rate at which the initial reaction occurs. If the initial rate of mass loss is very slow, then the determination of T_i can be uncertain.

A more satisfactory approach is to use the extrapolated onset temperature, T_e, which provides consistent values (see Fig. 3). If, however, the decomposition extends over a wide temperature range and only becomes rapid in its final stages, the extrapolated onset temperature will differ considerably from the onset temperature. For this kind of reaction, it is more satisfactory to measure the temperature at which a fractional weight loss, α, has occurred (T_α) (see Fig. 4). Clearly the temperature $T_{0.05}$ is close to that at the start of the reaction and $T_{0.90}$ is close to that at the end of the reaction. To define the complete range of reaction, two further temperatures, T_0 and T_f may be identified, as shown in Figure 3. Similar comments apply to these as to the corresponding temperatures already discussed.

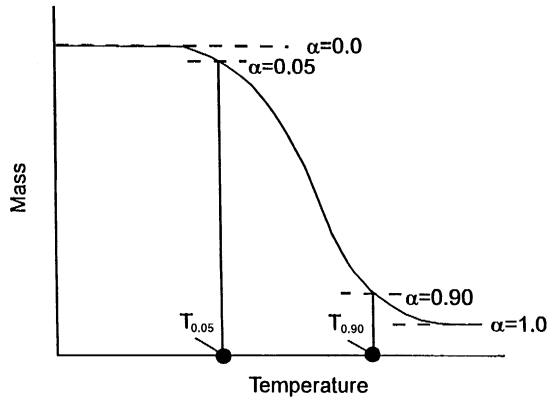


Figure 4. Measurement of the fraction reacted, α.

PRACTICAL ASPECTS OF THE METHOD

Apparatus

Thermobalance apparatus consist of the essential components of a balance and balance controller, a glass vessel to enclose the sample and balance to allow experiments to be carried out in a controlled atmosphere, a furnace and furnace controller, and a recorder system. Today it is common to purchase a thermobalance from a commercial source. Only in cases where the requirements are not met by the commercial sources should the construction of a self-designed system be contemplated, although most manufacturers will discuss possible modifications to existing systems.

Balances that monitor mass changes are of two types: deflection and null deflection. The deflection balance monitors the movement of the weight sensor, and the null-deflection balance the position of the beam. In the latter a servo system maintains the beam it in a quasi-equilibrium position, the power supply to the servo system being the measure of the mass change. An advantage of the null-deflection balance is that the sample stays in a constant position in the furnace, which assists in various ways that will be explained later. Most modern thermobalances are based on null-type electronic microbalances, which means that sample masses of typically 10 to 20 mg are sufficient to be able to detect mass changes accurately and reliably. Thermobalances capable of taking larger samples, in the gram range, are also available, although they tend to be used for specialized applications.

The automatic recording beam microbalance and ultra microbalances for use in vacuum and controlled environments have been reviewed by Gast (1974) and Czanderna and Wolsky (1980). The load-to-precision ratio (LPR) is often used as a means of comparing the performance of microbalances. A 1-g capacity beam balance with a precision of 2 μg will have an LPR of 5×10^5. This compares with LPR values for high-performance microbalances of 10^8 or better (Czanderna and Wolsky, 1980, p. 7). In the balance sensitivity category, for example, the highest quoted sensitivity for all the reviewed thermobalances is l pg, and the lowest is 50 pg. However, the lower sensitivity balances may also be capable of taking larger sample masses and so the LPR value may not vary significantly across the range. This needs to be checked for each specific balance. Details of the sensitivity or detection limit, precision, and accuracy of the balance are not always evident in the literature, and the terminology used is not always descriptive or accurate. Some degree of standardization would be helpful to the prospective buyer.

Thermobalances are enclosed in a glass envelope, or chamber, which can be sealed, partly to protect the balance against corrosion or damage but also to enable experiments to be carried out in a controlled atmosphere. Various ways are available for introducing the chosen gas. The most common way to introduce an inert gas is to

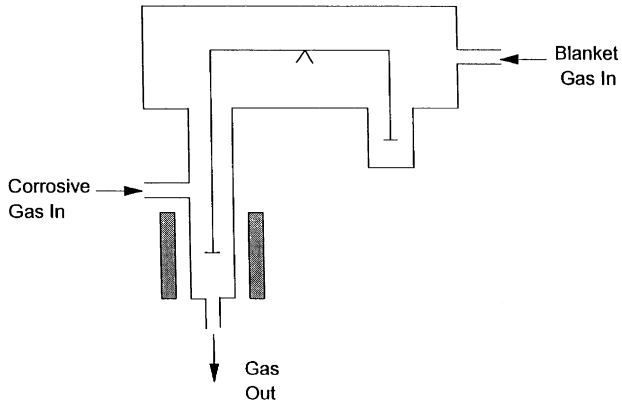


Figure 5. Protection of balance mechanism from corrosive gases by a secondary gas flow.

flow it first over the balance and then over the sample before exiting the system. This has the advantage of protecting the balance against any corrosive products generated by the decomposing sample as well as sweeping away any products that may condense on the hangdown arm of the balance and cause weighing errors.

When a corrosive reaction gas is employed, the balance needs to be protected. This can be achieved by flowing an inert gas over the balance and introducing the corrosive gas at some other point so it is swept away from the balance mechanism. One way of realizing this is shown in Figure 5. A flow of the blanket gas is established followed by the reactive (corrosive) gas. Some commercial thermobalances are equipped with flowmeters for gas control, but if they are not, then external rotameters need to be fitted. When corrosive gases are used, the flow of corrosive gas and inert gas needs to be carefully balanced in order to provide the proper protection. Examples of thermobalances that have been modified to work in different atmospheres are many, e.g., in a sulfur atmosphere (Rilling and Balesdent, 1975) or in sodium vapor (Metrot, 1975).

The problem of containing corrosive gases is considerably simplified by the magnetic balance designed by Gast (1975), where the balance and sample chambers are completely separated. This is a single-arm beam balance in which the pan is attached to a permanent magnet that is kept in suspension below an electromagnet attached to the hangdown wire. The distance between the two magnets is controlled electromagnetically. The sample chamber and the atmosphere containing the sample are thus isolated from the balance, and no attack on the mechanism can occur. This concept has been extended to microbalances (Pahlke and Gast, 1994).

The ability to have controlled flowing atmospheres is usually available, with gas flows of up to 150 mL min^{-1} being typical. Some instruments offer threaded gas valve connections, as well as a means of monitoring the gas flow. Work under vacuum down to 1×10^{-3} bar is common, and many thermobalances can operate down to 10^{-6} bar, although it is not usual to find instruments fitted with a means of monitoring the vacuum level. A commercial balance is reported to be able to operate up to 190 bars (Escoubes et al., 1984). Many examples of balances that work in high vacuum are reported at the vacuum microbalance conferences. For high-pressure work, there is a TG apparatus capable of working up to 300 bars and 350°C (Brown et al., 1972) and a versatile TG apparatus for operation between 10^{-8} and 300 bars in the temperature range −200 to 500°C (Gachet and Trambouze, 1975).

Table 2 . Common Furnace Windings and Their Approximate Maximum Working Temperature

Furnace Winding	Approximate Maximum Working Temperature (°C)
Nichrome	1000
Kanthal	1350
Platinum	1400
Platinum/10% rhodium	1500
Kanthal super ($MoSi_2$)	1600
Molybdenum	1800
Tungsten	2800

Furnaces can be mounted horizontally or vertically around the detector system. The most widely used furnaces are those wound with nichrome (nickel-chromium alloy) or platinum/rhodium heating wires. The most commonly used windings and their maximum temperatures of operation are given in Table 2. Molybdenum and tungsten windings need to be kept under a mildly reducing atmosphere to prevent oxidation.

Furnaces are frequently wound in a noninductive manner, i.e., in one direction and then the other, so that magnetic fields generated when a current is passed through the furnace windings are canceled, thus eliminating any interaction with a magnetic sample.

A requirement of a furnace is that there should be a region within which the temperature is constant over a finite distance. This distance is referred to as the hot zone. The hot zone diminishes as the temperature increases, as indicated in Figure 6. The sample should always be in the hot zone, so that all of the sample is at the same temperature. A thermocouple can be used to ensure that proper placement of the sample in relation to the hot zone is established. This is one reason to use a null-deflection balance, as the sample is always in the same position

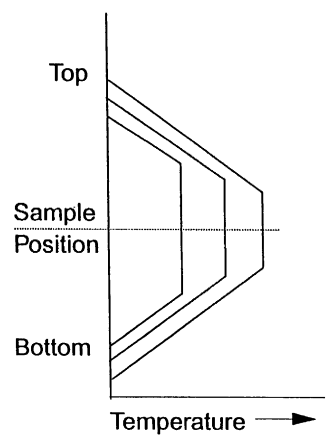


Figure 6. Schematic of the hot zone in a furnace.

in the furnace. Furnaces capable of very high heating rates have been used for TG experiments of textiles at heating rates approaching 3000°C/min (Bingham and Hill, 1975). These fast response furnaces can achieve isothermal temperatures rapidly. Alternative methods of heating include infrared and induction heating, which heat the sample directly without heating the sample chamber. Very fast heating rates on the order of 3000 to 6000°C min^{-1} can be achieved. Induction heating requires that the sample must be conducting, or placed in a conducting sample holder.

The temperature ranges for commercially available instruments given in the manufacturer's literature tend to be for the basic model, which operates typically in the temperature range ambient to 1000°C, but other models may operate from −196° to 500°C or ambient to 2400°C with steps in between.

There are three main arrangements of the furnace relative to the weighing arm of the balance. Lateral loading is a horizontal arrangement in which the balance arm extends horizontally into a furnace aligned horizontally. Two vertical arrangements are possible: one with the balance above the furnace with the sample suspended from the balance arm (bottom loading) and the other with the balance below the furnace and the sample resting on a flat platform supported by a solid ceramic rod rising from the balance arm (top loading). The following advantages are claimed for the horizontal mode:

1. Balance sensitivity is increased by virtue of the long balance arm.
2. It permits rapid gas purge rates, up to 1 L/min, since horizontal balance arms are perturbed less by a rapid gas flow than a vertical arrangement.
3. The influence of the Knudsen effect and convection errors can be ignored, i.e., chimney effects from the furnace are eliminated.
4. Evolved gas analysis is simplified.

The main advantage claimed for the vertical mode is that much higher temperatures can be achieved, since no change occurs in the length of the balance arm.

Escoubes et al. (1984) reported that the bottom-loading principle was adopted by about twice as many systems as the top-loading principle, with a small number of lateral-loading models apparent.

Temperature measurements are most commonly made with thermocouple systems, the most common being chromel-alumel, with an operating temperature range of ambient to 800°C; Pt–Pt/10% Rh at ambient to 1500°C; and Pt/6% Rh–Pt/30% Rh at ambient to 1750°C. The ideal location of the sample-measuring thermocouple is directly in contact with the sample platform. This requires that very fine wires made of the same material as the thermocouple are attached to the bottom of the platform. Such arrangements are found in simultaneous TG-DTA systems. The other alternative is to place the thermocouple at some constant distance from the sample and let the calibration technique compensate for the gap between the thermocouple and the sample. This method can only be adopted for a null-deflection balance, as this is the only system in which the distance between the sample platform and the thermocouple is constant. Two typical configurations are shown in Figure 7: one with the wires welded onto the bottom of the flat-plate thermocouple, which is kept at a constant distance from the sample pan, and the other a sideways configuration found in the horizontal balance arm arrangements.

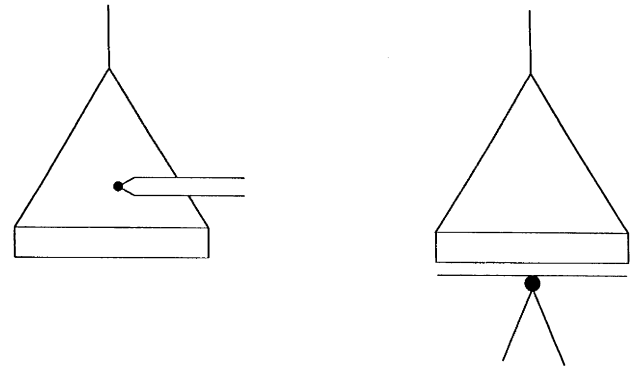


Figure 7. Typical configurations of the sample temperature-measuring thermocouple for a null-deflection balance. Left: lateral balance arm arrangement; right: bottom-loading balance.

The response of the thermocouple varies with age, and so calibration with an external calibrant is required from time to time. Thermocouples, like other devices, do not respond well to abuse. Hence, operation of the furnace above the recommended temperature or in a corrosive environment will result in relatively rapid changes to signal output and eventually render it inactive.

The chief functions of a temperature programmer are to provide a linear and reproducible set of heating rates and to hold a fixed temperature to within 1°C. The programmer is usually controlled by a thermocouple located close to the furnace windings but occasionally situated close to the sample. Modern programmers can be set to carry out several different heating/cooling cycles, although this function is being increasingly taken over by microprocessors.

The thermocouple output is usually fed to a microprocessor containing tabulated data on thermocouple readings, so that the conversion to temperature is made by direct comparison. Although heating rates often increase in steps, say, 1, 2, 5, 10, 30, 50, 100°C/min^{-1}, microprocessor-controlled systems allow increments of 0.1°C/min, which can be useful in kinetic studies based on multiple-heating-rate methods.

The recording device for most modern equipment is the computer. Thermobalances provide a millivolt output that is directly proportional to mass, and so it is possible to take this signal and feed it directly into a microprocessor via a suitable interface. This is by far the most common and best method for data acquisition and manipulation. Reading the direct millivolt output is the most accurate way of obtaining the measurement. Once stored, the data can be manipulated in various ways, and, e.g., the production of the derivative TG curve becomes a trivial task.

Data collection relies on a completely linear temperature rise uninterrupted by deviations induced by

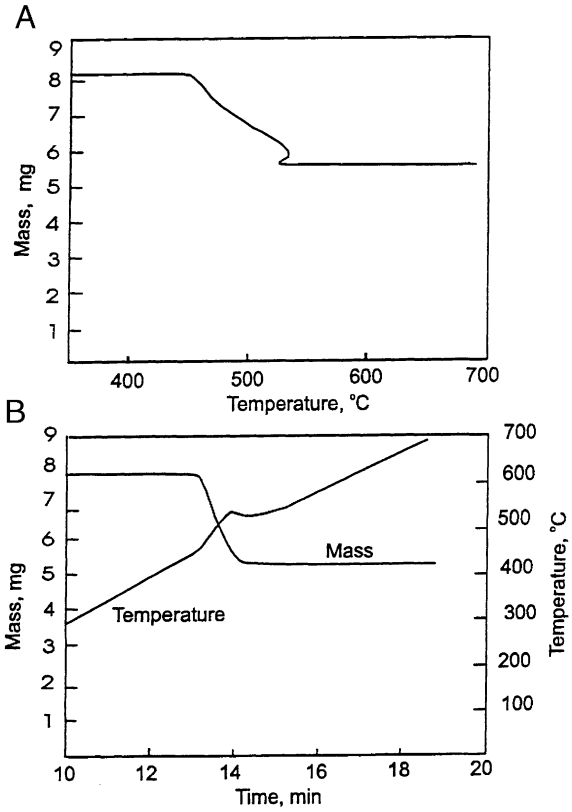


Figure 8 Distorted TG signal resulting from deviation of the linear programmed heating rate.

instrumental or experimental factors. For example, the oxidation of sulfide minerals is so exothermic that for a short time the temperature-recording thermocouple may be heated above the temperature expected for the linear heating program. If the TG plot is being recorded as a temperature-mass plot, then the TG trace will be distorted (see Fig. 8A). This can be overcome by converting the temperature axis to a time-based plot (Fig. 8B).

Experimental Variables

The shapes of TG and DTG curves are markedly dependent on the properties of the sample and the experimental variables that can be set on the thermobalance. These procedural variables include sample preparation (dealt with later), sample mass, sample containers, heating rate, and the atmosphere surrounding the sample.

The mass of the sample affects the mass and heat transfer, i.e., the diffusion of reaction products away from the sample or the interaction of an introduced reactive gas with the sample, and temperature gradients that exist within the sample. For a reaction in which a gas is produced, e.g., dehydration, dehydroxylation, or decomposition, the shape of the curve will depend on the thickness of the sample. For thick samples, gaseous products that evolved from the bottom layers of sample will take considerable time to diffuse away from the sample, and hence delay the completion of reaction. Therefore thick layers of sample will give mass losses over greater ranges of temperature than thin layers of sample.

Temperature gradients are inevitable in TG (or any thermal analysis method) because experiments are usually carried out at some constant heating or cooling rate. There is a temperature gradient between the furnace and the sample, as it takes a finite time for heat transfer to take place across the air gap between the furnace and the sample. Only under isothermal conditions will the sample and furnace temperatures be the same. At a constant heating rate this temperature gradient, commonly called the thermal lag, is approximately constant. The thermal lag increases as the heating or cooling rate increases or as the mass of the sample increases. A second and more important temperature gradient is that within the sample. Heat diffusion will be dependent on the thermal conductivity of the sample, so that heat gradients in metals will tend to be lower than those in minerals or polymers. If a large sample of a poorly conducting material is heated, then large temperature gradients can be expected, and the rate of reaction will be faster at the exterior of the sample because it is at a higher temperature than the interior. This will cause the reaction to occur over a greater temperature range relative to a smaller sample mass or one with a better thermal conductivity. These effects can be illustrated with reference to calcium carbonate (see Table 3), which decomposes above 600°C with the evolution of carbon dioxide (Wilburn et al., 1991). Both mass and heat transfer effects are operative, as the carbon dioxide has to diffuse out of the sample, which takes longer as the mass increases; also, significant thermal gradients between the furnace and sample and within the sample increase as the mass increases. The increase in $T_{0.1}$ with an increase in mass of 59°C is only related to the thermal lag of the sample with respect to the furnace, but the increase in $T_{0.9}$ with an increase in mass is the sum of the thermal lag, temperature gradients in the sample, and the mass transfer of the carbon dioxide. The increase in $T_{0.9}$ is somewhat larger at 97°C relative to the increase in $T_{0.1}$. In general, therefore, the interval between $T_{0.1}$ and $T_{0.9}$ increases as the sample mass increases.

For large samples, the heat generated during an exothermic reaction may be sufficient to ignite the sample. In this case, the mass loss will occur over a very narrow temperature range that will be different from that obtained under nonignition conditions.

In general, these problems were more significant with early thermobalances that required large sample masses to accurately record mass changes. Modern thermobalances,

Table 3. Effect of Sample Mass on Decomposition of Calcium Carbonate Heated in Nitrogen at 7°C/min

Sample Mass (mg)	$T_{0.1}$ (°C)	$T_{0.9}$ (°C)	$T_{0.9-0.1}$ (°C)
50	716	818	102
100	742	855	113
200	768	890	122
300	775	902	127
400	775	915	140

based on null-deflection microbalances, are much less troubled by these problems as the sample masses tend to be in the milligram rather than the hundreds of milligram mass ranges. Even so, it is sometimes necessary to use large samples, and so these constraints need to be considered.

Generally small samples spread as a fine layer are preferred. This ensures that (1) gaseous components can disperse quickly, (2) thermal gradients are small, and (3) reaction between solid and introduced reactive gas is rapid.

Both the geometry of the sample container and the material from which it is made can influence the shape of a TG curve. Deep, narrow-necked crucibles can inhibit gaseous diffusion processes and hence alter the shape and temperature of curves relative to a flat, open pan (Hagvoort, 1994). Sometimes these effects are quite marked, and it has been demonstrated that the rate of oxidation of pyrite decreases by about one-half as the wall height of the sample pan changes from 2 to 4 mm (Jorgensen and Moyle, 1986). Most crucibles can have a lid fitted with either a partial or complete seal, and this is useful for the prevention of sample loss as well as for holding the sample firmly in place. It is also of value to encapsulate the sample if it is being used as a calibrant, so that if the reaction is completely reversible, it can be used several times. Paulik et al. (1982) have designed what they call "labyrinth" crucibles, which have special lids that inhibit gaseous diffusion out of the cell. The partial pressure of the gaseous product reaches atmospheric pressure and remains constant until the end of the process, so leading to the technique being named quasi-isobaric.

Crucibles are usually fabricated from metal (often platinum) or ceramic (porcelain, alumina, or quartz), and their thermal conductivity varies accordingly. Care must be taken to avoid the possibility of reaction between the sample and its container. This can occur when nitrates or sulfates are heated or in the case of polymers, which contain halogens such as fluorine, phosphorus, or sulfur. Platinum pans are readily attacked by such materials. Only a very small number of experiments are required with sulfides heated to 700° to 800°C in an inert atmosphere to produce holes in a platinum crucible. Less obviously, platinum crucibles have been observed to act as a catalyst, causing change in the composition of the atmosphere; e.g., Pt crucibles promote conversion of SO_2 to SO_3. Hence, if sulfides are heated in air or oxygen, the weight gain due to sulfate formation is always greater if Pt crucibles are used, especially relative to ceramic crucibles. The reaction sequence is

$$\begin{aligned} \mathrm{MS(s)} + 1.5\mathrm{O_2(g)} &\rightarrow \mathrm{MO(s)} + \mathrm{SO_2(g)} \\ \mathrm{MO(s)} + \mathrm{SO_3(g)} &\rightarrow \mathrm{MSO_4(s)} \end{aligned} \tag{7}$$

The temperature range over which a TG curve is observed depends on the heating rate. As the heating rate increases, there is an increase in the procedural decomposition temperature, the DTG peak temperature, the final temperature, and the temperature range over which the mass loss occurs. The magnitude of the effect is indicated by the data in Table 4 for the decomposition of 50 mg of calcium carbonate heated in nitrogen at two different heating rates (Wilburn et al., 1991). The faster heating rate causes a shift to higher temperatures of $T_{0.1}$, as well as $T_{0.9}$, although the latter effect is greater than the former. Hence the mass loss occurs over a greater temperature range as the heating rate increases. Faster heating rates may also cause a loss of resolution when two reactions occur in similar temperature ranges, and the two mass losses merge.

Table 4. Effect of Heating Rate on Temperature of Decomposition of Calcium Carbonate

Heating Rate (°C/min)	$T_{0.1}$ (°C)	$T_{0.9}$ (°C)	$T_{0.9-0.1}$ (°C)
1	634	712	78
7	716	818	102

Fast heating rates combined with a large sample sometimes leads to a change in mechanism, which produces a significantly different TG curve.

The atmosphere above the sample is a major experimental variable in thermal analytical work. Two effects are important. First, the presence in the atmosphere of an appreciable partial pressure of a volatile product will suppress a reversible reaction and shift the decomposition temperature to a higher value. This effect can be achieved by having a large sample with a lid or introducing the volatile component into the inlet of the gas stream. When the atmosphere within the crucible is the result of the decomposition of the reactant, the atmosphere is described as "self-generated" (Newkirk, 1971).

An example of the decomposition of a solid material that is affected by the partial pressure of product gas is calcium carbonate, which decomposes according to the reaction

$$\mathrm{CaCO_3(s)} \rightarrow \mathrm{CaO(s)} + \mathrm{CO_2(g)} \tag{8}$$

In a pure CO_2 atmosphere, the decomposition temperature is greatly increased. This effect can be predicted from thermodynamics. If the values of the Gibbs free-energy change (ΔG°), the enthalpy change (ΔH°), and the entropy change (ΔS°), are fitted into the following equation, then the thermodynamic decomposition temperature T can be calculated for any partial pressure P of CO_2:

$$\begin{aligned} \Delta G^\circ = \Delta H^\circ - T\Delta S^\circ &= RT\ln P \\ 40{,}250 - 34.4T &= -4.6T\log P_{\mathrm{CO_2}} \\ \text{At 1 atm } \mathrm{CO_2}: \quad T &= 1170\,\mathrm{K}\,(897^\circ\mathrm{C}) \\ \text{At 0.1 atm } \mathrm{CO_2}: \quad T &= 1132\,\mathrm{K}\,(759^\circ\mathrm{C}) \end{aligned} \tag{9}$$

Hence, under equilibrium conditions, the decomposition temperature is lowered by 138°C. The temperatures found by thermal methods are always higher because these are dynamic techniques, although for small sample sizes and slow heating rates the correspondence is close. For calcium carbonate the difference between the onset temperature of decomposition and the calculated thermodynamic value was in the range 10° to 20°C (Wilburn et al., 1991).

The second major effect is the interaction of a reactive gas with the sample, which will change the course of the reaction. Under an inert atmosphere, organic compounds will degrade by pyrolytic decomposition, whereas in an oxidizing atmosphere oxidative decomposition obviously takes place. Hydrogen can be used to study reductive processes, although great care has to be taken to remove all oxygen from the system before carrying out the reaction. Other gases frequently studied are SO_2 and halogens.

It should be evident from the foregoing comments and examples that the conditions required to minimize the effects of heat and mass transfer, as well as decreasing the possibility of change in the reaction sequence, are to have (1) a small sample (10 to 20 mg), (2) finely ground material (<100 μm), (3) a slow heating rate (10° to 20°C/min); and (4) a thin film spread evenly in a flat open pan made of an inert material.

Such conditions will allow maximum resolution of close reactions and give well-defined mass losses over relatively narrow temperature ranges. However, it should be emphasized that when dealing with a sample of unknown properties, it is advisable to use a range of different conditions to ensure that the sample behaves in a similar manner. If the only effect noticed on increasing the heating rate is a modest shift to higher temperatures of the mass losses, it can be assumed that the reaction sequence is the same, and the increased temperature is just the result of an increase in the thermal lag of the sample. Any marked difference in the temperature or number of mass losses may indicate a change in reaction sequence.

In some circumstances it is not possible to use the above conditions. For example, measuring small mass changes may require large samples to give accurate mass values or in combined TG-evolved gas analysis (TG-EGA) experiments the evolved gas concentration needs to be high enough to be detected.

Applications

Thermogravimetry can be used in a wide variety of research and development and industrial applications. Rather than attempt to cover the entire range here, this section gives more general examples of the kinds of reactions that can be investigated by TG.

Kinetic Studies. The use of TG to obtain kinetic information is one of its oldest applications. Furthermore, the ability to carry out this type of work has been made even easier by the availability of software packages to take data from the TG curves and to calculate various kinetic parameters from the Arrhenius equation, such as the activation energy E and the preexponential factor A. Despite the large number of papers that have been published on this topic, there is still considerable difficulty in obtaining reliable absolute information from the technique. This is evident from the wide variety of published values for E and A for the same chemical reaction. Scientists wishing to engage in kinetic studies using TG should not only have read the relevant literature indicating the difficulties of interpreting the results (see, for example, Brown, 1997) but also preferably have spent some time in the company of people experienced in this type of application. It is, however, of relevance to point out that if absolute values of kinetic parameters are not required, then the TG method offers a very simple system for comparative studies. Thus the relative rates of oxidation of a variety of alloys with a particular gas mix or the relative oxidative stability of a range of polymers can be easily determined.

The reactions that can be studied by TG are decomposition reactions (mass loss), oxidation of metals or alloys (mass gain), reactions between a solid and a gas that produce a solid and another gas (could be mass loss or gain), and solid-solid reactions that produce a gas (mass loss). It is also obvious that there should only be one reaction taking place at one time, and overlapping reactions are difficult to treat unless the experimental variables can be manipulated to separate them. As an example of a reaction that cannot yield sensible results, the oxidation of pyrite has simultaneous oxidation and sulfation reactions taking place, and one of these gives a mass loss and the other a mass gain.

It has been demonstrated earlier that the kinetics of most reactions under isothermal conditions can be summarized by the general equation $d\alpha/dt = f(\alpha)$, where α is the fraction reacted in time t, k is the rate of reaction, and $f(\alpha)$ is some function of α (see Principles of the Method). The various mathematical equations that relate $f(\alpha)$ with the rate of some solid-state reaction are given in Table 5. The two forms of the equations, $g(\alpha)$ and $f(\alpha)$, are associated with the use of the TG data in their integral or differentiated forms, respectively. The former uses data directly from the TG curve, and the differentiated form uses data from the DTG curve. Each equation is associated with a particular model for the progress of the reaction, which may, for example, be gas diffusion into a solid particle or the appearance of product nuclei with subsequent growth of those nuclei. The shape of the TG curve will vary with the particular mechanism involved, as will the temperature at which $d\alpha/dt$ is a maximum.

The traditional method of conducting kinetic studies is the isothermal method. The method involves loading the sample into the TG apparatus, rapidly heating it to some specific temperature, and monitoring the mass change. The experiment is repeated with a new sample at another temperature, and a family of curves is generated such as those shown in Figure 9. The total mass loss for the reaction is found by heating the sample in a rising-temperature experiment until the mass loss is complete. A series of α and $d\alpha/dt$ values can then be calculated for a given temperature. These data are fitted to the expressions in Table 5, and the best fit suggests the model that is likely to apply to the reaction under test. From this expression, the average rate constant k can be calculated at each temperature. A plot of $\log k$ vs. $1/T$ should provide a straight-line graph of slope E/R from which E can be calculated, and the intercept gives the preexponential factor A.

The isothermal method is considered by some to give the most reliable data for kinetic studies but has been criticized for the following reasons.

Table 5. Broad Classification of Solid-State Rate Expressions

Classification	$g(\alpha) = kt$	$f(\alpha) = 1/k(d\alpha/dt)$
Acceleratory α-time curves		
P1 Power law	$\alpha^{1/n}$	$n(\alpha)^{(n-1)/n}$
E1 Exponential law	$\ln \alpha$	α
Sigmoidal α-time curves		
A2 Avrami-Erofeev	$[-\ln(1-\alpha)]^{1/2}$	$2(1-\alpha)[-\ln(1-\alpha)]^{1/2}$
A3 Avrami-Erofeev	$[-\ln(1-\alpha)]^{1/3}$	$3(1-\alpha)[-\ln(1-\alpha)]^{2/3}$
A4 Avrami-Erofeev	$[-\ln(1-\alpha)]^{1/4}$	$4(1-\alpha)[-\ln(1-\alpha)]^{3/4}$
B1 Prout-Tompkins	$\ln[\alpha/(1-\alpha)] + c$	$\alpha(1-\alpha)$
Deceleratory α-time curves		
1. Based on geometric models		
R2 Contracting area	$1-(1-\alpha)^{1/2}$	$2-(1-\alpha)^{1/2}$
R3 Contracting volume	$1-(1-\alpha)^{1/3}$	$3-(1-\alpha)^{2/3}$
2. Based on diffusion mechanisms		
D1 One dimensional	α^2	$1/2\alpha$
D2 Two dimensional	$(1-\alpha)\ln(1-\alpha) + \alpha$	$[-\ln(1-\alpha)]^{-1}$
D3 Three dimensional	$[1-(1-\alpha)^{1/3}]^2$	$3/2(1-\alpha)^{2/3}[1-(1-\alpha)^{1/3}]^{-1}$
D4 Ginstling-Brounshtein	$(1-2\alpha/3)-(1-\alpha)^{2/3}$	$3/2[(1-\alpha)^{-1/3}-1]^{-1}$
3. Based on order of reaction		
F1 First order	$-\ln(1-\alpha)$	$1-\alpha$
F2 Second order	$1/(1-\alpha)$	$(1-\alpha)^2$
F3 Third order	$[1/(1-\alpha)]^2$	$0.5(1-\alpha)^3$

1. It is difficult to determine exactly the time and temperature of the beginning of the reaction.
2. It requires several experiments and so is time consuming.
3. Each experiment requires a new sample, which must react in exactly the same way as all the other samples. This requires careful attention to standardization of the experimental procedures, which includes, e.g., the use of the same sample pan with the same sample packing and gas exchange. Duplicate experiments should be done to check the level of reproducibility.

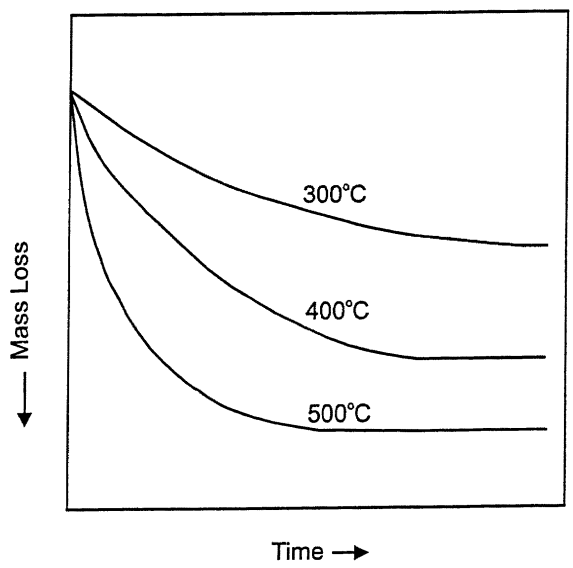


Figure 9. A family of TG curves generated at different isothermal temperatures.

The alternative to isothermal experiments is the rising-temperature method (Flynn, 1969; Sestak, 1972), which requires that the sample is heated at a constant rate over the reaction temperature range. The value of α is calculated at various temperatures, and $d\alpha/dt$ calculated for a small temperature interval. The same process of evaluation of the appropriate form of the rate-controlling equation is applied, and hence E and A are calculated. If the expression is differentiated, then $d\alpha/dt = -(dm/dt)/\Delta m$, where Δm is the total mass change and dm/dt can be obtained directly from the DTG curve.

This method has the advantage of being rapid, and duplicate experiments to check reproducibility may be all that is required. The choice of integral over differential methods is difficult. Earlier models of thermobalances had significant errors in the determination of $d\alpha/dt$, mainly because of signal noise, but sophisticated computerized data acquisition and manipulation have reduced this problem and produced renewed interest in differential methods. Integral methods have their own difficulties, because for rising-temperature experiments there is no definite integral for the right-hand side of the equation, and so approximations have to be made. Whatever the choice, some of the same criticisms apply to rising-temperature methods as apply to the isothermal method, especially the need for the development of a standardized procedure. Even small changes in sample packing, for example, will affect the results. It is also evident that multiple experiments at different heating rates are required in some cases to improve the reliability of the data.

Despite all the improvements in apparatus, technique, and data acquisition and computational methods, problems still arise. The best least-squares fit may not be given by the correct mechanism, because calculations for one model give results that are so close to another model that it is difficult to distinguish between them. It is possible to fit more than one mathematical expression to a given model. Different models can give the same mathematical expression. When any of these situations occur, analysis of the TG curve gives the mathematical relationship, but considerable judgment may be required to select the appropriate model. Efforts should be made to support the suggested reaction mechanism by other means. For example, viewing partially reacted materials by techniques such as microscopy or SEM will identify a mechanism controlled by gaseous diffusion inward or outward.

A final set of experimental methods is the controlled-rate group (CRTA), which controls one of the experimental variables, such as the temperature, evolved gas, or mass change, to keep the rate of reaction constant. The CRTA methods are claimed to have certain advantages for kinetic studies over other methods (Reading, 1988; Rouquerol, 1989).

Compositional Analysis. If the reactions that are recorded on the TG curve are for single reactions that are not interfered with by other reactions, then the mass change can be directly related to the quantity of reacting material present. Hence, TG is a technique that can be used to determine the purity of a compound or the components in a mixture. Some methods have been developed that are routinely used for quality control purposes, because TG methods can be faster and more accurate and give more information than other methods. The book edited by Earnest (1988) contains applications to areas such as drugs, polymers, inorganic compounds, minerals, fuels, and waste products.

Even when the reactions in the first experimental investigations are not separated, it may be possible to manipulate the experimental variables in such a way that separation can be achieved. Thus TG offers some flexibility to effect compositional analysis.

The simplest situation is when a single compound decomposes with an associated mass loss, and the purity of the sample is calculated. In some cases, the compound may be hydrated, and so two mass losses are observed, one associated with the loss of water and the other with the decomposition reaction. Many clays and inorganic compounds undergo this kind of reaction. Mixtures can be readily analyzed; e.g., a mixture of gypsum ($CaSO_4 \cdot 2H_2O$) and lime [$Ca(OH)_2$], which is commonly found in the plaster industry. The gypsum decomposes in an open pan in the temperature range of 100° to 150°C and the lime decomposes between 400° and 500°C, according to the equations

$$\begin{aligned} CaSO_4 \cdot 2H_2O(s) &\rightarrow CaSO_4(s) + 2H_2O(g) \\ Ca(OH)_2(s) &\rightarrow CaO(s) + H_2O(g) \end{aligned} \tag{10}$$

The first reaction is a simple dehydration process and the second a dehydroxylation reaction.

In an open pan, the dehydration of gypsum appears as a single mass loss, although the decomposition is a two-step process:

$$\begin{aligned} CaSO_4 \cdot 2H_2O(s) &\rightarrow CaSO_4 \cdot 0.5H_2O(s) + 1.5H_2O(g) \\ &\rightarrow CaSO_4(s) + 0.5H_2O(g) \end{aligned} \tag{11}$$

This reaction can be resolved into two steps by manipulation of the partial pressure of the water vapor above the sample. One way to achieve this is to seal the gypsum in a sample pan with a small leak. The water vapor trapped in the pan from the first reaction delays the commencement of the second step of the reaction by some 40° to 50°C. This method is used in the cement industry to check the composition of clinker-gypsum mixtures. When the two components are ground together, heat is generated, which may cause undesirable decomposition of gypsum to the hemihydrate. Hence phase analysis of the clinker-gypsum mixture for gypsum and hemihydrate content can be routinely carried out (Dunn et al., 1987).

Another way of separating reactions by manipulation of the atmosphere is to vary the concentration of the evolved gas in the flowing gas atmosphere. An example of this has already been given with respect to the decomposition of calcium carbonate, where a ten-fold change in the partial pressure of carbon dioxide changes the decomposition temperature by some 138°C. If some other noncarbonate decomposition reaction is interfering with the mass loss of calcium carbonate, then changing the atmospheric composition may resolve the problem.

Advantage can also be taken of changing the atmosphere from an inert one to a reactive one. In an inert atmosphere, pyrolytic decomposition of organic compounds such as polymers can occur. The analysis of rubber compounds has been carried out by TG methods for some time (Brazier and Nickel, 1980; Sirkar, 1977). Rubbers typically contain polymers, extending oils and other organic compounds, carbon black, and inorganic compounds. Heating the rubber in nitrogen volatilizes or decomposes the organic compounds and decomposes the polymer, so that the first mass loss can be related to the total content of these compounds. Changing the atmosphere to air oxidizes the carbon black to give a second mass loss, and the residue is the inorganic components converted to their oxides. This analysis can be done in less than 20 min and provides a rapid routine quality control method in the elastomer industry. The method can be automated using the computer system controlling the gas changes and heating rates. The decomposition of the pure polymer should be checked to ensure that it decomposes completely and does not form a char.

The inert/reactive atmosphere system can be used to measure the amount of an element in an inorganic material if the element has variable oxidation states. If the element is oxidized to a higher oxidation state, then a mass gain will be observed; if it is reduced to a lower oxidation state with a gas such as hydrogen or carbon monoxide, then a mass loss will take place.

Variation in the heating rate may be sufficient to separate overlapping reactions. Slowing the heating rate gives

additional time for the reaction to take place, and if necessary, an isothermal period may be required to establish a plateau in the TG curve before increasing the heating rate (Dunn and Sharp, 1993, p. 163).

In some circumstances total phase analysis can be achieved by TG methods in conjunction with other analytical methods. Thus an iron ore containing a clay, goethite (FeOOH), and hematite (Fe_2O_3) was analyzed for the first two components by TG, and the total iron was determined by dissolution and titrimetric analysis (Weissenborn et al., 1994).

Thermal Stability/Reactivity. The assessment of thermal stability has been mainly applied to polymeric materials, although the technique can be applied to a number of other materials ranging from inorganic compounds to oils (Dunn and Sharp, 1993). There are three different approaches:

1. The material is heated at some programmed rate and the temperature (either the extrapolated onset temperature or the temperature for a given fraction reacted) at which decomposition takes place is measured. This value is then used to rank the materials in relative order of thermal stability. A comparison of the stability in nitrogen and air or oxygen gives an indication of the role of oxygen in the breakdown mechanism. Some plastics are remarkably unaffected by the change in atmosphere, e.g., nylon, while others begin to decompose at much lower temperatures in an oxidizing atmosphere. The decomposition temperature provides the maximum operational temperature at which the polymer can exist but gives no indication of the long-term behavior at lower temperatures.
2. The material is heated at several different isothermal temperatures and a family of curves of mass loss against time are generated. This procedure does give information about the longer term stability at lower temperatures and sometimes reveals a critical temperature range above which decomposition is rapid.
3. The material is heated to a single isothermal temperature in an inert atmosphere, the atmosphere is changed to a reactive gas, and the time for the first major reaction to take place is measured. The oxidative stability of a material can be measured quantitatively by heating the material in an inert atmosphere to a given temperature below the decomposition temperature, establishing a stable baseline and then switching the atmosphere to oxygen. The time taken to produce the first weight loss after the admission of oxygen, called the induction period, is a measure of the oxidative stability at that temperature. By using this method, the effects of various additives on the thermal stability of materials can be compared (Hale and Bair, 1997).

Attempts have been made to use kinetic data derived from the decomposition of materials to predict their life expectancy, particularly for polymeric materials. Standard methods are time consuming and often ambiguous in interpretation. The TG methods have been developed as alternative accelerated test procedures and have achieved various degrees of success (Flynn, 1995). Flynn (1995) has argued that there are two main reasons why quantitative predictions cannot be made for polymers. First, the kinetics of decomposition below glass transition temperatures or other phase transitions will not be the same as those above the temperature of such transitions. Second, there is a temperature above which the system becomes thermodynamically unstable. If accelerated testing measurements are made above the phase transition temperature and/or the thermodynamic stability temperature when in practice the polymer use is below these temperatures, then a false picture will emerge of the lifetime capabilities.

Very reactive materials, such as organic compounds, sulfide minerals, and coal, can be ignited. Ignition can be defined as corresponding to the establishment of a self-sustaining combustion reaction after the termination of heat supply from an external source (Vilyunov and Zarko, 1989, p. 320). This definition assumes that there is a continuous supply of reactants once the ignition reaction has been initiated. A modified TG technique has been used to measure the relative ignition temperature of minerals such as sulfides and coal (Dunn et al., 1985; Dunn and Chamberlain, 1991; Dunn and Mackey, 1992). The technique involves heating the furnace to some preset temperature, establishing an air or oxygen flow, placing the sample in the pan and then raising the furnace around the sample. If the set furnace temperature is below the ignition temperature of the sample, then a slow oxidation process occurs. If the experiment is repeated with fresh samples and with increments in furnace temperature, then the sample will eventually reach a temperature at which it ignites. This gives a very rapid mass loss. The ignition temperatures of a range of minerals can be established and thus give some indication of their relative reactivity to processes such as flash smelting.

Thermogravimetry has been more recently applied to studies in cuprate family superconductors (Sale and Taylor, 1992). One of the prerequisites for a superconductor is the preparation of precursors that are homogeneous and of controlled particle size. This is achieved by making a gel followed by a drying stage and then a decomposition-oxidation stage. Thermogravimetry can be used to determine the thermal stability of the precursor as well as the reaction sequence for the decomposition-oxidation process. Furthermore, once the superconductor has been formed, TG can be used to determine the variable oxygen content as a function of temperature and partial pressure of oxygen.

Gas-Solid Reactions. The TG apparatus can be used as a minireactor to examine reactions between a solid and a gas. A number of reactions, such as oxidation and pyrolysis, have already been mentioned. The range of possibilities is obviously more extensive, and two examples are given by way of illustration.

In corrosion studies, the reaction between metals and different atmospheres can be investigated. Oxidation reactions are common (Das and Fraser, 1994; Joshi, 1992), but

gaseous mixtures containing oxygen and water (Webber, 1976), oxygen and bromine (Antill and Peakall, 1976), and various sulfur-containing atmospheres are also evident (Young et al., 1975; Dutrizac, 1977; Hunt and Strafford, 1981). Besides investigating the effect of various atmospheres on a single metal or alloy, the atmosphere can be kept constant and the composition of the alloy changed. The usual procedure is to heat the sample to a specific temperature and follow the mass change as a function of time. The data are then fitted to one of the four general equations describing the growth of corrosion products, i.e., parabolic, linear, logarithmic, or cubic.

Thermogravimetry has been extensively used to study most aspects of the life cycle of catalysts, from catalyst preparation to regeneration, and also to compare the effects of variables such as catalyst composition, operational temperatures, and gas flow rates on performance (Habersberger, 1977; Dollimore, 1981). Since most TG work is small scale and experiments can be carried out rapidly, much useful information can be gathered cheaply and quickly prior to large-scale tests. This work is further facilitated by use of TG-EGA systems to measure quantitatively the product yield. X-ray diffraction, infrared spectroscopy, and chemical analysis are frequently used to characterize the catalyst material at various stages in the cycle.

In the preparation stage, the aim is to prepare a material that has a high surface area of active sites. Many metal catalysts are prepared by adsorbing a solution of a salt onto the support material and then drying, calcining, and reducing the mixture. Each of these mass loss steps can be followed by TG, and the effect of variation in experimental conditions of heating rate, gas flow rate, and temperature range on the surface area of the metal can be determined by standard methods. Experimental studies have shown that, in general, the optimum reactivity of the catalyst is achieved at slow rates of reduction (Bartholomew and Farrauto, 1976). The lower the decomposition temperature of the parent compound, the higher the surface area of the final product (Pope et al., 1977).

Gaseous adsorption and desorption processes can be followed as mass changes, as can be catalyst efficiency and temperature stability.

Solid-Solid Reactions. The reaction between two solids to evolve a gas can be investigated by TG. An example is the reaction between a sulfide and a sulfate (Dunn and Kelly, 1977):

$$3NiS(s) + 2NiSO_4(s) \rightarrow Ni_3S_2(s) + 2NiO(s) + 3SO_2(g) \quad (12)$$

METHOD AUTOMATION

Computers that act as recorders can also be used as a control system, which enables a particular protocol to be established to carry out a repetitive test procedure. This control function can be extended to switching valves so that different atmospheres can be admitted to the TG equipment. Automatic sample loaders, which allow multiple samples to be sequentially introduced into the TG apparatus, are available.

DATA ANALYSIS AND INITIAL INTERPRETATION

In keeping with modern instrumentation, the data output from the TG apparatus is acquired through a suitable interface and the results processed and stored via appropriate software. In nearly all cases the full facility is available from the manufacturer of the thermobalance. The only drawback with this arrangement is that the algorithms are unknown and the operator has to accept what is provided. The ease of accepting the manufacturer's solutions has to be offset against the considerable time involved in writing specific in-house software. The manufacturer is in somewhat of a cleft stick, as the TG apparatus is increasingly taking on the twin roles of research and routine quality control. In the former situation the research scientist wants to be in control and make decisions on, e.g., the degree of smoothing that might be applied to a TG curve. This requires that the raw data are displayed and then various levels of smoothing applied. On the other hand, the operator in a quality control situation only wants a reliable and set system that produces comparable results from experiments, without the need for personal decision. Some companies will write user-specific software, and with some instruments it is possible for the operator to extract files and manipulate them.

There are many ways in which data can be manipulated with commercial software; some examples are as follows:

1. Rescaling either the ordinate or the abscissa, which means that it is rare for the operator to have to rerun the experiment because of an unfortunate choice of conditions.
2. Calculating the first derivative of the mass change to give the DTG curve.
3. Presenting a fully labeled TG/DTG curve on screen, with the various onset and completion temperatures identified and mass losses expressed in various units, including percentages. Labels identifying the sample, operator, and experimental conditions used can be added, and the final product can be printed out for inclusion in a report.
4. Carrying out various calculations, including a kinetic analysis of the data.

SAMPLE PREPARATION

Sample preparation in this context means both the treatments that may be applied to the sample prior to the TG experiment and the way in which the sample is introduced to the TG apparatus.

It is usual practice for a sample to be received in the laboratory and then treated in some way before any experimentation is carried out. In the simplest example, mineral samples can be received in lump form, which may be centimeters in diameter, and clearly unsuitable for experiment.

Such materials are ground and divided progressively to eventually produce a representative laboratory sample of 10 to 20 g with a particle size of typically <100 μm. Similarly, rubber samples may be presented as a sheet and then shaved down to give a number of flakes or as a circle punched out of the sheet and the experiment carried out on a flat disc.

One of the most significant characteristics of a solid sample is its particle size, as the smaller the particle size, the greater the surface area and the rate of reaction. Hence at any given temperature the value of α, the fraction reacted, will be greater for fine particles relative to large particles. Smaller particles will also heat more rapidly than larger particles. Usually the more finely divided the sample, the lower the temperature at the start and completion of the reaction. In most cases, a change in particle size will give similar TG curves that are displaced along the temperature axis. Less commonly, the actual mechanism of the reaction may change and a totally different TG curve be obtained. Overgrinding is to be avoided, as the mechanical energy may heat the sample as well as cause changes to its degree of crystallinity. In the former situation, decomposition of the sample may result, with loss of bound water and even the dehydroxylation of minerals; in the latter situation the surface of the solid may become quite active and adsorb moisture from the atmosphere to a greater extent than for an unground sample (Dunn and Sharp, 1993, p. 158).

Some materials, when heated, can be ejected from the sample pan and give a mass loss that may be incorrectly attributed to a chemical reaction rather than a physical loss. Samples in this category are those that decrepitate, i.e., dehydrate with subsequent dissolution of the solid in the water, which then boils with frothing. Advantage has been taken of this effect to assess the decrepitation of dolomites and limestones used in the glass-making industry. If the material is prone to decrepitation, then a mass loss is observed between 400° and 500°C, well before the main decomposition of the dolomite above 600°C. A nondecrepitating dolomite shows no mass loss below 600°C (Fig. 10). Organic samples, if heated under vigorous oxidizing conditions, may ignite and the solid material be carried away as a fine ash. Such problems can be solved by use of a cover on the sample pan.

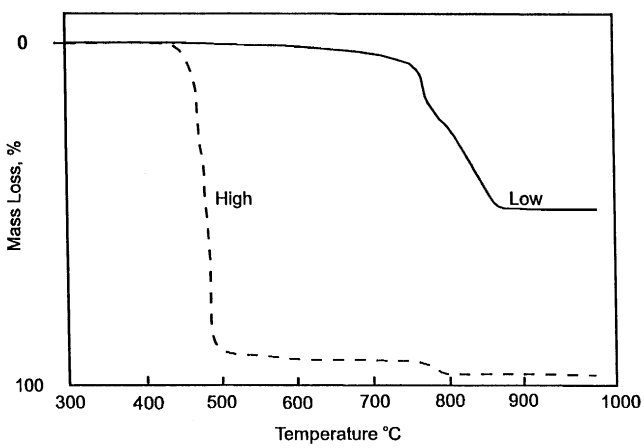


Figure 10. TG curve for a dolomite sample showing effects of decrepitation, contained in a sample crucible with and without a lid.

The use of pellets or deep beds of powder may lead to a change in mechanism from a rate-determining step based on the chemical reaction to one based on heat or mass transfer. There have been suggestions (Hills, 1967, 1968) that heat and mass transfer may be rate determining even when thin layers are used under isothermal conditions, involving reactions previously supposed to be controlled by reaction at the interface between the reactant and the product.

PROBLEMS

The major sources of errors in the TG technique are in the measurement of mass and sample temperature.

Errors in Mass Measurement

Several factors can contribute to incorrect measurement of mass. These include incorrect calibration, temperature fluctuation, electrostatic or magnetic forces, vibration, and atmospheric environment or buoyancy effects (Dunn and Sharp, 1993, p. 168).

The direct calibration of the microbalance through the use of standard weights is the simplest and the most straightforward calibration method (Schwoebel, 1980, p. 89). In this method, standardized masses are added to one arm of the balance and any necessary adjustment made. Regular checks are required. Standardized masses are often provided as part of the accessories when a commercial thermobalance is purchased, or sets of such mass calibrants can be obtained from organizations such as ICTAC, the National Physics Laboratory (United Kingdom), and the National Institutes of Standards and Technology (NIST, United States) (see MASS AND DENSITY MEASUREMENTS).

The mechanism of the balance is affected by changes in temperature. Any temperature fluctuation will result in drift in the zero point of the balance and hence will decrease the accuracy of the measurement. This problem becomes important if experiments over time periods of hours are undertaken in laboratories that are not temperature controlled. For high-accuracy work the thermobalance should be housed in a thermostatted enclosure (Sharp and Mortimer, 1968).

Spurious results can be obtained if electrostatic or magnetic forces are generated in the thermobalance and interact with the system. Electrical interference can be eliminated by grounding the system, while the magnetic effects can be overcome by shielding the furnace in a noninductive manner or winding the furnace coil in a noninductive manner.

Errors that derive from the resultant forces of the atmospheric environment include buoyancy effects, convection in the furnace atmosphere, and various molecular gas movements that act on the balance. The magnitude of the buoyancy effect, one of the major sources of error, arises from the volume of the atmosphere displaced by

the sample and its attendant sample pan and hangdown assembly. This results in an apparent increase in the mass of the sample, which increases as the temperature increases. The effect can be determined experimentally by heating sample crucibles containing an inert material such as calcined alumina as a substitute sample and using this value to adjust the results from subsequent experiments. Most software can store the buoyancy data and subtract the correction value from the experimental data. The experimental conditions must be similar to those used for the sample to be investigated. In older TG apparatus, where the volume of displaced gas was quite large because of the large sample mass and pan, the increase in mass would be of the order of several milligrams. In the systems based on the microbalance, the effect is much smaller, and for routine quality control purposes where only moderate accuracy is required, it is often ignored. It is also important to consider that during actual experiments the buoyancy of the sample itself will change, especially if there is a significant increase or decrease in sample mass by chemical reactions. This effect has been observed during the decomposition of $CaCO_3$ (Oswald and Wiedemann, 1977).

Some instruments have been manufactured to overcome problems inherent in the designs described above. The various disturbances that can occur to the mass record can be considerably reduced by use of a symmetrical thermobalance, in which sample balance pan and counter balance pan are suspended from the balance arm into two identical furnaces. Both furnaces are programmed, so that buoyancy and convection effects are almost eliminated from the weight record. Another manufacturer produces a differential motion TG apparatus, which is based on a top-loading arrangement in which the sample and counter balance pans are supported on a system of beams and subbeams so that the distance between the pans is small enough to be incorporated into the same furnace. Hence, again, the pans are subjected to the same environment, which reduces errors in the weight record.

It is evident that measuring mass changes accurately as a function of temperature is not an entirely simple matter. Regular calibration, blank runs, and buoyancy checks under conditions identical to those used in a set of experiments need to be done and the instrument needs to be calibrated at the operating temperature in order to determine the magnitude of any errors. Determining the purity of a known compound by the mass change is another way of checking the accuracy of the mass measurements.

Errors in Temperature Measurement

Accurately measuring the temperature of a sample in a TG apparatus is still a difficult task, mainly because of the inherent design of the equipment. The most significant factor is that the thermocouple measuring the sample temperature is not actually in contact with the sample, or even with the sample pan, but is usually a few millimetres from the sample, as shown in Figure 7. The actual temperature of the sample will differ from the temperature reading when the sensor is not in contact with the sample holder. The difference in temperature between the sample and thermocouple is geometry dependent as well, particularly so if the thermocouple is above or below the sample, as the difference will vary both with temperature and with heating rate (Garn et al., 1981). In addition, the sample-sensor temperature difference can be affected by the atmosphere and its flow rate. This is because gases of different thermal conductivity will carry heat away from the sample with different efficiency. For accurate work, therefore, the temperature calibration of the TG apparatus should be carried out under experimental conditions identical to those used in the actual experiments. The use of null-deflection balances is an advantage, as the sample assembly stays a constant distance from the thermocouple. The worst situation is with a deflection balance. If the sample gains mass, then the distance between the thermocouple and sample will decrease, whereas if the sample loses mass, the distance between the thermocouple and sample will increase.

One of the better calibration methods involves the use of the Curie point transition in metal samples. In this procedure, a ferromagnetic material is placed in the sample crucible as the test specimen, and a magnet is located above or below the crucible, which creates a magnetic flux aligned to the gravitational field. When the material is heated, at a particular temperature the ferromagnetic material becomes paramagnetic. The point at which the magnetic properties completely disappear is defined as the Curie temperature (see TECHNIQUES TO MEASURE MAGNETIC DOMAIN STRUCTURES). At the Curie point, a significant apparent weight change in the specimen occurs (Norem et al., 1970).

Five metals have been certified as ICTAC Certified Reference Materials for Thermogravimetry, GM761 (see Table 6; Garn et al., 1981). The materials cover the temperature range 200 to 800°C. The experimental data in a round-robin investigation have been reported by Garn (1982) and showed that intralaboratory results were reasonably precise, with a mean standard deviation of 3.6 ± 1.1°C. However, there was a wide range of results reported from interlaboratory studies and/or depending on the type of instrument used. In addition, the temperature range and standard deviation increased as the ferromagnetic temperature increased.

A more recent study by Gallagher et al. (1993) reported that more reproducible and precise results for determining transition temperature can be achieved using simultaneous thermomagnetometry/differential thermomagnetometry/DTA (TM/DTM/DTA). The results from five different laboratories, based on DTM, showed that a very high

Table 6. Measured Values for T_e Using ICTAC Certified Reference Materials for Thermogravimetry, GM761

Material	Temperature Range of Transition (°C)	Mean (°C)	Standard Deviation
Permanorm 3	242–263	253.3	5.3
Nickel	343–360	351.4	4.8
Mumetal	363–392	377.4	6.3
Permanorm 5	435–463	451.1	6.7
Trafoperm	728–767	749.5	10.9

degree of reproducibility between the laboratories was obtained with a 2σ value of 0.2°C. Gundlach and Gallagher (1997) have investigated the potential of nickel-cobalt and nickel-copper alloys, which give magnetic transitions from 1100°C to below room temperature.

The obvious advantage of the Curie point method is that the ferromagnetic material calibrates temperature directly at the sample position, and the thermobalance can be calibrated under the same conditions as used with a sample. The magnetic transitions are reversible, so that the calibrant can be encapsulated to protect it against oxidation and used repeatedly for calibration purposes. A criticism is that the thermal conductivity of the alloy is much higher than for other materials such as minerals or polymers, and so there may be some minor discrepancy in temperature measurement because of differences in thermal lag between the alloy and the sample. Another problem could be the ability to satisfactorily place the magnet in a suitable position.

McGhie et al. (1983) have suggested an alternative method that allows the melting points of metals to be used for temperature calibration. In this procedure, a mass is suspended within the sample crucible of the thermobalance by a thin wire of the temperature calibration material. When the system is heated, at the melting point of the wire the weight drops either on the sample crucible, causing a short disturbance on the weight record, or through a hole in the base of the sample crucible giving a large and rapid weight change. This method has been used to provide new estimations for the transition temperature of the ICTAC standards (Blaine and Fair, 1983). A disadvantage of this method is that the thin wire is not at the same position as the sample. This criticism has been partially met by a study that demonstrated that the temperatures determined by the Curie point method and the thin-wire method were similar (Gallagher and Gyorgy, 1986).

If the sample is visible when placed in the TG sample pan, then the melting of a pure compound can be used for calibration purposes. The method is to heat the calibrant to just below its melting point and then increase the temperature in small increments of say 1°C until the calibrant just melts. Sets of melting calibration standards are available from organizations such as ICTAC and NIST covering the temperature range ambient to 1000°C. The obvious disadvantage of this method is that the calibrant needs to be visible, and the calibration is carried out under near-isothermal conditions instead of the rising-temperature program most commonly used for experimental purposes.

There is no universally accepted way of calibrating temperature in a thermobalance at present, and there is not likely to be because of the various instrument designs and different spatial relationships between the measuring thermocouple and the sample. The choice of the above methods will depend on the accuracy required and what can be achieved with a particular TG apparatus. Even then there is likely to be significant error, so that studies undertaken by TG that require high accuracy in the temperature measurement, such as kinetic studies involving the calculation of rate constants, are likely to be subject to significant error. For these studies it is much more appropriate to use simultaneous TG-DTA, where the thermocouple is in contact with the sample crucible and the instrument can be calibrated using ICTAC standards for DTA (see SIMULTANEOUS TECHNIQUES INCLUDING ANALYSIS OF GASEOUS PRODUCTS).

LITERATURE CITED

Antill, J. E. and Peakall, K. A. 1976. The influence of bromine on the gaseous oxidation of an austenistic steel. *Corros. Sci.* 16: 435–441.

Balek, V. 1991. Emanation thermal analysis and its application potential. *Thermochim. Acta* 192: 1–17.

Bartholomew, C. H. and Farrauto, R. J. 1976. Chemistry of nickel-alumina catalysts. *J. Catal.* 45: 41–53.

Bingham, M. A. and Hill, B. J. 1975. A study of the thermal behavior of flame-resistant fibre and fabrics. *J. Therm. Anal.* 7:347–358.

Blaine, R. L. and Fair, P. G. 1983. Estimate of the "true" magnetic transition temperatures for the ICTA reference materials GM761. *Thermochim. Acta* 67:233–240.

Brazier, D. W. and Nickel, G. H. 1980. Applications of thermal analytical procedures in the study of elastomers and elastomer systems. *Rubber Chem. Technol.* 53:437–511.

Brown, H. A., Penski, E. C., and Callahan, J. P. 1972. An apparatus for high pressure thermogravimetry. *Thermochim. Acta* 3:271–276.

Brown, M. E. 1997. Steps in a minefield. Some kinetic aspects of thermal analysis. *J. Therm. Anal.* 49:17–32.

Czanderna, A. W. and Wolsky, S. P. 1980. Beam microbalance design, construction and operation. *In* Microweighing in Vacuum and Controlled Environments, Vol. 4 (A. W. Czanderna and S. P. Wolsky, eds.). pp. 1–57. Elsevier/North-Holland, Amsterdam, The Netherlands.

Das, S. and Fraser, H. L. 1994. Thermogravimetric study of the oxidation behaviour of alloys based on the Mg-Li system. *In* Thermal Analysis in Metallurgy (R. D. Shull and A. Joshi, eds.). pp. 361–379. The Minerals, Metals and Materials Society, Pittsburgh, Pa.

Dollimore, D. 1981. The use of thermal analysis in studying catalysts and the chemisorption process. *Thermochim. Acta* 50: 123–146.

Dollimore, D. 1992. Thermogravimetry. *In* Thermal Analysis: Techniques and Applications (E. L. Charsley and S. B. Warrington, eds.). pp. 31–58. Royal Society of Chemistry Special Publication No. 117, London.

Dunn, J. G. 1993. Recommendations for reporting thermal analysis data. *J. Therm. Anal.* 40:1431–1436.

Dunn, J. G. and Chamberlain, A. C. 1991. The effect of stoichiometry on the ignition behaviour of synthetic pyrrhotites. *J. Therm. Anal.* 37:1329–1346.

Dunn, J. G., Jayaweera, S. A. A., and Davies, S. G. 1985. Development of techniques for the study of flash smelting reactions using DTA and TG. *Proc. Australas. Inst. Min. Metall.* No. 290, 75–82.

Dunn, J. G. and Kelly, C. E. 1977. A TG-DTA-MS study of the oxidation of nickel sulfide. *J. Therm. Anal.* 12:43–52.

Dunn, J. G. and Mackey, L. C. 1992. The measurement of the ignition temperatures of commercially important sulfide minerals. *J. Therm. Anal.* 38:487–494.

Dunn, J. G. and Sharp J. H. 1993. Thermogravimetry. *In* Treatise on Analytical Chemistry, Part 1, Vol. 13 (J. D. Winefordner, ed.) pp. 127–266. John Wiley & Sons, New York.

Dunn, J. G., Oliver, K., Nguyen, G., and Sills, I. 1987. The quantitative determination of hydrated calcium sulfates in cement by DSC. *Thermochim. Acta* 121:181–191.

Dutrizac, J. E. 1977. The reaction of titanium with sulfur vapour. *J. Less Common Met*. 51:283–303.

Earnest, C. M. 1988. Compositional Analysis by Thermogravimetry. American Society for Testing and Materials. Philadelphia, Pa.

Escoubes, M., Eyraud, C., and Robens, E. 1984. Vacuum microbalances and thermogravimetric analysis. Part I: Commercially available instruments. *Thermochim. Acta* 82:15–22.

Flynn, J. H. 1969. The historical development of applied non-isothermal kinetics. *In* Thermal Analysis, Vol. 2 (R. F. Schwenker and P. D. Garn, eds.). pp. 1111–1123. Academic Press, New York.

Flynn, J. H. 1995. A critique of lifetime prediction of polymers by thermal analysis. *J. Therm. Anal.* 44:499–512.

Gachet, C. G. and Trambouze, J. Y. 1975. Balance for thermogravimetric measurements at low and high pressure (300 bar). *In* Progress in Vacuum Microbalance Techniques, Vol. 3 (M. Escoubes and C. Eyraud, eds.). pp. 144–152. Heyden, London.

Gallagher, P. K. and Gyorgy, E. M. 1986. Curie temperature standards for thermogravimetry: The effect of magnetic field strength and comparison with melting point standards using Ni and Pb. *Thermochim. Acta* 109:193–206.

Gallagher, P. K., Zhong, Z., Charsley, E. L., Mikhail, S. A., Todoki, M., Taniguahi, K., and Blaine, R. L. 1993. A study of the effect of purity on the use of nickel as a temperature standard for thermomagnetometry. *J. Therm. Anal.* 40:1423–1430.

Garn, P. D. 1982. Calibration of thermobalance temperature scales. *Thermochim. Acta* 55:121–122.

Garn, P. D., Menis, O., and Wiedemann, H. G. 1981. Reference materials for thermogravimetry. *J. Therm. Anal.* 20:185–204.

Gast, T. H. 1974. Vacuum microbalances, their construction and characteristics. *J. Phys. E: Sci. Instrum.* 7:865–875.

Gast, T. H. 1975. Transmission of measured values from the sample in the magnetic suspension balance. *In* Progress in Vacuum Microbalance Techniques, Vol. 3 (M. Escoubes and C. Eyraud, eds.). pp. 108–113. Heyden, London.

Gundlach, E. M. and Gallagher, P. K. 1997. Synthesis of nickel based alloys for use as magnetic standards. *J. Therm. Anal.* 49:1013–1016.

Habersberger, K. 1977. Application of thermal analysis to the investigation of catalysts. *J. Therm. Anal*. 12:55–58.

Haines, P. J. 1995. Thermal Methods of Analysis. Blackie Academic and Professional, Glasgow.

Hakvoort, G. 1994. TG measurement of gas-solid reactions. The effect of the shape of the crucible on the measured rate. *Thermochim. Acta* 233:63–73.

Hale, A. and Bair, H. E. 1997. Polymer blends and block copolymers. *In* Thermal Characterisation of Polymeric Materials P, Vol. 2 (E. A. Turi, ed.). pp. 864–870. Academic Press, New York.

Hills, A. W. D. 1967. The role of heat and mass transfer in gas-solid reactions involving two solid phases within sintered pellets. *In* Heat and Mass Transfer in Process Metallurgy, pp. 39–77. Institute of Minerals and Metals, London.

Hills, A. W. D. 1968. The mechanism of decomposition of calcium carbonate. *Chem. Eng. Sci*. 23:297–320.

Hunt, P. J. and Strafford, K. N. 1981. Metallographic aspects of the corrosion of a preoxidised Co-30 Cr alloy in an H_2-H_2O-H_2S atmosphere at 900°C. *J. Electrochem. Soc*. 128:352–357.

Jorgensen, R. A. and Moyle, F. J. 1986. Gas diffusion during the thermal analysis of pyrite. *J. Therm. Anal.* 31:145–156.

Joshi, A. 1992. Oxidation behaviour of carbon and diamond films. *In* Thermal Analysis in Metallurgy (R. D. Shull and A. Joshi, eds.). pp. 329–344. The Minerals, Metals and Materials Society, Pittsburgh, Pa.

Keattch, C. J. and Dollimore, D. 1975. An Introduction to Thermogravimetry, 2nd ed. Heyden, London.

Mackenzie, R. C. 1979. Nomenclature in thermal analysis, Part IV. *Thermochim. Acta* 28:1–6.

Mackenzie, R. C. 1983. Nomenclature in thermal analysis. *In* Treatise on Analytical Chemistry, Part 1 Vol. 12 (P. J Elving and I. M Kolthoff, eds.). pp. 1–16. John Wiley & Sons, New York.

McAdie, H. G. 1967. Recommendations for reporting thermal analysis data. *Anal. Chem.* 39(4):543.

McAdie, H. G. 1972. Recommendations for reporting thermal analysis data. Evolved gas analysis. *Anal. Chem.* 44(4):640.

McAdie, H. G. 1974. Recommendations for reporting thermal analysis data. Thermomechanical techniques. *Anal. Chem.* 46(8):1146.

McGhie, A. R., Chiu, J., Fair, P. G., and Blaine, R. L. 1983. Thermogravimetric apparatus temperature calibration using melting point standards. *Thermochim. Acta* 67:241–250.

Metrot, A. 1975. Thermobalance a reflux de sodium. *In* Progress in Vacuum Microbalance Techniques, Vol. 3 (M. Escoubes and C. Eyraud, eds.). pp. 153–158. Heyden, London.

Newkirk, A. E. 1971. Thermogravimetry in self-generated atmospheres. A decade of practice and new results. *Thermochim. Acta* 2:1–23.

Norem, S. D., O'Neill, M. J. O., and Gray, A. P. 1970. The use of magnetic transitions in temperature calibration and performance evaluation of thermogravimetric systems. *Thermochim. Acta* 1:29–38.

Oswald, H. R. and Wiedemann, H. G. 1977. Factors influencing thermoanalytical curves. *J. Therm. Anal.* 12:147–168.

Pahlke, W. and Gast, Th. 1994. A new magnetic suspension coupling for microbalances. *Thermochim. Acta* 233:127–139.

Paulik, J. and Paulik, F. 1971. "Quasi-isothermal" thermogravimetry. *Anal. Chim. Acta* 56: 328–331.

Paulik, F., Paulik, J., and Arnold, M. 1982. Kinetics and mechanism of the decomposition of pyrite under conventional and quasi-isothermal–quasi-isobaric thermoanalytical conditions. *J. Therm. Anal.* 25:313–325.

Pope, D., Walker, D. S., and Moss, R. L. 1977. Preparation of cobalt oxide catalysts and their activity for CO oxidation at low concentration. *J. Catal*. 47:33–47.

Reading, M. 1988. The kinetics of solid state decomposition reactions; a new way forward? *Thermochim. Acta* 135:37–57.

Reading, M. 1992. Controlled rate thermal analysis and beyond. *In* Thermal Analysis: Techniques and Applications (E. L. Charsley and S. B. Warrington, eds.). pp. 126–155. Royal Society of Chemistry Special Publication No. 117, London.

Rilling, J. and Balesdent, D. 1975. Thermogravimetrie continue en atmosphere de soufre pur. *In* Progress in Vacuum Microbalance Techniques, Vol. 3 (M. Escoubes and C. Eyraud, eds.). pp. 182–201. Heyden, London.

Rouquerol, J. 1989. Controlled transformation rate thermal analysis: The hidden face of thermal analysis. *Thermochim. Acta* 144:209–224.

Sale, F. R. and Taylor, A. P. 1992. Applications in metallurgy and mineral science. *In* Thermal Analysis: Techniques and Applications (E. L. Charsley and S. B. Warrington, eds.). pp. 198–216. Royal Society of Chemistry Special Publication No. 117, London.

Satava, V. and Skvara, F. 1969. Mechanism and kinetics of the decomposition of solids by a thermogravimetric method. *J. Am. Ceram. Soc.* 52:591–599.

Schwoebel, R. L. 1980. Beam microbalance design, construction and operation. *In* Microweighing in Vacuum and Controlled Environments, Vol. 4 (A. W. Czanderna and S. P. Wolsky, eds.). p. 89. Elsevier, Amsterdam, The Netherlands.

Sestak, J. 1972. Non-isothermal kinetics. *In* Thermal Analysis, Vol. 2 (H.G. Wiedemann, ed.). pp. 3–30. Birkhauser, Basel.

Sharp, J. H. 1972. Reaction kinetics. *In* Differential Thermal Analysis, Vol. 2, (R. C. Mackenzie, ed.). pp. 47–77. Academic Press, London.

Sharp, W. B. A. and Mortimer, D. 1968. Accurate thermogravimetry in flowing gases. *J. Sci. Instrum. (J. Physics E),* Series 2, 1:843–846.

Sirkar, A. K. 1977. Elastomers. *In* Thermal Characterisation of Polymeric Materials P, Vol. 2 (E. A. Turi, ed.). pp. 1219–1273. Academic Press, New York.

Turi, E. A. (ed.). 1997. Thermal Characterisation of Polymeric Materials, Vol. 2. Academic Press, New York.

Vilyunov, V. N. and Zarko, V. E. (1989). Ignition of Solids. Elsevier/North-Holland, Amsterdam, The Netherlands.

Webber, J. 1976. Oxidation of Fe-Cr alloys in O_2/3% H_2O. *Corros. Sci*. 16:99–506.

Weissenborn, P. K., Dunn, J. G., and Warren, L. J. 1994. Quantitative thermogravimetric analysis of hematite, goethite, and kaolinite in WA iron ores. *Thermochim. Acta* 239:147–156.

Wilburn, F. W., Sharp, J. H., Tinsley, D. M., and McIntosh, R. M. 1991. The effect of procedural variables on TG, DTG, and DTA curves of calcium carbonate. *J. Therm. Anal.* 37:2003–2019.

Young, D. J., Smeltzer, W. W., and Kirkaldy, J. S. 1975. The effects of molybdenum additions to nickel-chromium alloys on their sulfidation properties. *Metall. Trans. A.* 6A:205–1215.

KEY REFERENCES

Brown, M. F. 1988. Introduction to Thermal Analysis. Chapman & Hall, London.

General introductory text with a chapter on TG.

Brown, 1997. See above.

Critical assessment of the use of thermal method of analysis to kinetic studies.

Brazier and Nickel, 1980. See above.

Excellent review of applications of thermal analysis methods to the characterization of elastomers.

Charsley and Warrington (eds.)., 1992. Thermal Analysis: Techniques and Applications. Royal Society of Chemistry Special Publication No. 117, London.

Contains 15 chapters on various aspects of thermal analysis methods written by established experts.

Czanderna and Wolsky (eds.)., 1980. See above.

Extensive coverage of all the facets of weighing under different atmospheres, including design features of microbalances.

Dunn and Sharp, 1993. See above.

Comprehensive and detailed coverage of TG and its applications to a wide variety of topics.

Earnest, 1988. See above.

The use of TG for quantitative chemical analysis of a range of solid materials.

Keattch and Dollimore, 1975. See above

Although somewhat dated, still a good read, especially the section on kinetics.

Turi (ed.), 1997. See above.

A major work with significant chapters by some of the world leaders in this field.

Wendlandt, 1986. Thermal Methods of Analysis, 3rd ed. John Wiley & Sons, New York.

Comprehensive coverage of the field of thermal analysis.

Wunderlich, 1990. Thermal Analysis. Academic Press, New York.

Designed as a self-instructional text.

APPENDIX: REPORTING THERMAL ANALYSIS RESULTS

It has been stressed that TG results are very dependent on the sample properties and the experimental conditions used. To facilitate a comparison of results obtained by different workers, it follows that these various parameters, such as particle size and heating rate, should be reported.

The existing recommendations for reporting thermal methods of analysis were developed by the Standardisation Committee of the ICTAC in the late 1960s and early 1970s (McAdie, 1967, 1972, 1974) and are also in various standards, such as ASTM E472-86. The recommendations are presently in revision to take into account current practice (Dunn, 1993). The draft document is reproduced below, although at the time of writing it had not been officially ratified by the ICTAC Council. There has been some change of emphasis in the new draft, and more emphasis has been placed on accurate characterization of the sample, as well as on giving details on the way in which the computer acquires and processes data. The section relating to the experimental conditions is largely unchanged. This document does not address safety problems associated with the use of the related materials or equipment. It is the responsibility of the user to establish the appropriate safety standards and work within the validity of the given safety laws.

Information Required for All Thermal Analysis Techniques

Accompanying each thermal analysis record should be information about the:

1. Properties of substances.
   - 1.1 Identification of all substances (sample, reference, diluent) by a definitive name, an empirical formula, or equivalent compositional information.
   - 1.2 A statement of the source of all substances, details of their histories, pretreatments, physical properties, and chemical purities, as far as they are known.
     - 1.2.1 "Histories" includes method of acquisition or manufacture of the sample (how it was isolated or made, manufacturing conditions, e.g., grinding and sizing), previous thermal and mechanical treatments (e.g., repeated grinding, sintering, all

experimental conditions), surface modification, and any other physical or chemical treatment.

1.2.2 "Pretreatments" includes preparation of the sample prior to the thermal analysis experiment.

1.2.3 "Physical properties" includes particle size, surface area, porosity.

1.2.4 "Chemical purities" includes elemental analysis, phase composition, and chemical homogeneity of sample.

2. Experimental conditions.

2.1 A clear statement of the temperature environment of the sample during measurement or reaction, including initial temperature, final temperature, rate of temperature change if linear, or details if not linear.

2.2 Identification of the sample atmosphere by pressure, composition, and purity: whether the atmosphere is static, self generated or dynamic through or over the sample. Where applicable, the ambient atmospheric pressure and humidity should be specified. If the pressure is other than atmospheric, full details of the method of control should be given.

2.3 A statement of the dimensions, geometry and materials of the sample holder, and the method of loading the sample where applicable.

2.4 Identification of the abscissa scale in terms of time or temperature at a specified location. Time or temperature should be plotted to increase from right to left.

2.5 A statement of the methods used to identify intermediate or final products.

2.6 Faithful reproductions of all original records.

2.7 Wherever possible, each thermal event should be identified and supplementary supporting evidence stated.

2.8 Sample mass and dilution of sample.

2.9 Identification of the apparatus, by the provision of the manufacturer's name and model number. More detail should be provided if significant modifications to the standard commercial model have been made, or if the apparatus is new or novel, or if the apparatus is not commercially available.

3. Data acquisition and manipulation methods.

3.1 Identification of the instrument manufacturers software version, or details of any self-developed versions.

3.2 Equations used to process data, or reference to suitable literature.

3.3 The frequency of sampling, filtering and averaging of the signal.

3.4 An indication of the smoothing and signal conditioning used to convert analogue to digital signals, or reference to suitable literature. Similar information for subsequent processing of the digital data is also needed.

The following information is required specifically for Thermogravimetry (TG).

1. The ordinate scale should indicate the mass scale. Mass loss should be plotted as a downward trend and deviations from this practice should be clearly marked. Additional scales (for example, fractional decomposition, molecular composition) may be used for the ordinate where desired.
2. The method used for the calibration of the temperature and mass.
3. If derivative thermogravimetry is employed, the method of obtaining the derivative should be indicated and the units of the ordinate specified.

J. G. DUNN
Curtin University of Technology
Perth, Western Australia
Currently at: University
of Toledo, Ohio

DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY

INTRODUCTION

Differential thermal analysis (DTA) and differential scanning calorimetry (DSC) are two closely related methods in which a material under investigation is typically subjected to a programmed temp3erature change and thermal effects in the material are observed. (Isothermal methods are also possible though they are less common.) The term "differential" indicates that the difference in behavior between the material under study and a supposedly inert reference material is examined. In this manner the temperature at which any event either absorbs or releases heat can be found. This allows the determination of, e.g., phase transition temperatures and the study of order-disorder transitions and chemical reactions. Similarly, heat capacity measurements can be performed, although DTA and DSC differ significantly in the ease and precision of such measurements. These two methods are ideally suited for quality control, stability, and safety studies.

Measurement of heat capacity can be performed by other methods. Adiabatic calorimetry and drop calorimetry typically can provide heat capacity values roughly an order of magnitude more precise than those obtainable by thermal analysis methods. However, those calorimetric techniques are far more difficult to conduct than are the thermal analysis methods being discussed here. Relatively large samples are needed compared with the requirements of thermal analysis methods. In addition, commercial instruments are generally not readily available for adiabatic or drop calorimetry. Only when the highest precision is an absolute necessity would a researcher shun thermal analysis. The terminology in this area has become somewhat confused, and we will follow the recommendations of the International Confederation for Thermal Analysis Nomenclature Committee given below (Mackenzie, 1985).

In DTA, the temperature difference between a substance and a reference material is measured as a function

of temperature while the substance and reference material are subjected to a controlled temperature program.

The record is the differential thermal, or DTA, curve; the temperature difference (ΔT) should be plotted on the ordinate with endothermic reactions downward and temperature or time on the abscissa increasing from left to right.

The term "quantitative differential thermal analysis" (quantitative DTA) covers those uses of DTA where the equipment is designed to produce quantitative results in terms of energy and/or any other physical parameter.

In DSC, the difference in energy inputs into a substance and a reference material is measured as a function of temperature while the substance and reference material are subjected to a controlled temperature program.

Two modes, power compensation differential scanning calorimetry (power compensation DSC) and heat-flux differential scanning calorimetry (heat-flux DSC), can be distinguished, depending on the method of measurement used. A system with multiple sensors (e.g., a Calvet-type arrangement) or with a controlled heat leak (Boersma-type arrangement) would be heat-flux DSC, whereas systems without these or equivalent arrangements would be quantitative DTA.

These thermal analysis methods can be conducted simultaneously with other measurement methods to provide a greatly enhanced ability to understand material behavior (see Specimen Modification).

We will not dwell here on the historical development of these techniques, which has been amply summarized elsewhere (Boerio-Goates and Callanan, 1992). However, there is widespread confusion in the literature and elsewhere regarding the terms DTA and DSC that stems, in part, from how these methods developed. Early DTA procedures (classical DTA) involved thermocouples embedded in the sample and reference materials under study (Fig. 1A). Because of uncertain heat transfer considerations, quantitative measurements of, e.g., enthalpies of transformation could not be made reliably. An innovation due to Boersma (1955) led to significantly improved quantitative measurements. He recommended removing the thermocouples from direct contact with the samples and introducing a controlled heat leak between the sample and the reference containers (Fig. 1B). In subsequent years this arrangement and its various incarnations have come to be referred to as "heat-flux DSC." In 1964, the Perkin-Elmer Corporation developed and patented a differential scanning calorimeter (DSC-1) that involved separate heaters for the sample and reference containers with the differential power needed to keep the sample and the reference at the same temperature measured directly (Watson et al., 1964; Fig. 1C). This technique is currently referred to as "power-compensated DSC." As a result, the brief designation DSC is often used without clarification for both methods.

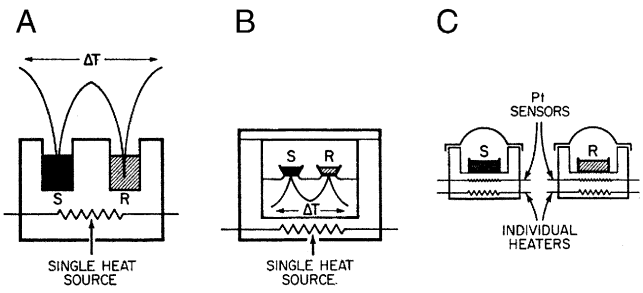


Figure 1. Schematic representation of the three principal thermal analysis systems: (**A**) classical DTA; (**B**) "Boersma" DTA; (**C**) DSC. (Courtesy of Perkin-Elmer.)

PRINCIPLES OF THE METHOD

In the extensive literature on thermal analysis methods, there are a great many explanations of the principles of the method (Boerio-Goates and Callanan, 1992). The treatment by Mraw (1988) is particularly instructive in that it enables the reader to understand the differences in principle among the three main variants of thermal analysis methods—classical DTA, heat-flux DSC, and power-compensated DSC—without dealing with endothermic or exothermic effects. Here, we develop the basic equations describing thermal analysis, following Mraw, with emphasis on the similarities and differences among the basic techniques.

Consider the schematic thermal analysis system shown in Figure 2. A sample (designated by subscript s) and an inert reference material (subscript r) are contained in pans and placed in the instrument. The notation used in the following analysis is shown in Figure 2. The total heat capacity is C_j (where j is either s or r), T_j is the actual temperature of the sample or reference, and $T_{j,\mathrm{m}}$ is the temperature indicated by the measuring thermocouple. Because of the thermal resistances R_j and R'_j, measured and actual temperatures may differ. The temperature of the heat source (T_h) may similarly be different. With these definitions we can proceed with an abbreviated derivation of the relevant relations.

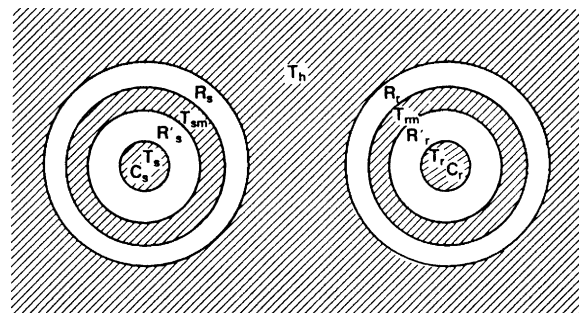


Figure 2. General considerations common to almost any type of thermal analysis instrument: C_s, total heat capacity of sample plus sample pan; T_h, temperature of the heat source; T_sm, temperature of the point where thermometer is located on the sample side; T_s, actual temperature of the sample and its pan; R_s, thermal resistance to heat flow between temperature T_h and T_sm; R'_S, thermal resistance to heat flow between temperature T_sm and T_s. The parameters C_r, T_rm, T_r, R_r, and R'_r have analogous meanings for the reference side. (From Mraw, 1988. Reproduced with permission.)

For simplicity, we consider all the heat capacities and thermal resistances to be constant. This will be approximately correct over a narrow temperature range in the determination of, for example, solidus or liquidus temperatures. Following Mraw, we take $R_s = R_r$ and $R'_s = R'_s$. Note that R and R' are not equal. Heat flow in the system is assumed to follow the simple one-dimensional law:

$$\frac{dq}{dt} = a\,\Delta T \tag{1}$$

where dq/dt is the rate of heat exchange between two bodies differing in temperature by ΔT. The heat transfer coefficient a is just $1/R$. We now examine heat exchange for the sample and reference assuming that heat flow from T_h to $T_{j,m}$ equals that from $T_{j,m}$ to T_j. This follows because we have assumed no other heat transfer processes in the system.

For the sample being subjected to a temperature scan at the rate dT_s/dt, we thus have

$$\frac{dq_s}{dt} = \frac{1}{R_s}(T_h - T_{sm}) = \frac{1}{R'_s}(T_{sm} - T_s) \tag{2}$$

and

$$\frac{dq_s}{dt} = C_s \frac{dT_s}{dt} \tag{3}$$

Similarly, for the reference,

$$\frac{dq_r}{dt} = \frac{1}{R_r}(T_h - T_{rm}) = \frac{1}{R'_r}(T_{rm} - T_r) \tag{4}$$

and

$$\frac{dq_r}{dt} = C_r \frac{dT_r}{dt} \tag{5}$$

The differences among the various thermal analysis techniques arise from differences in various parameters in the above equations. We next proceed to analyze these techniques within the framework of this simplified mathematical model.

In classical DTA (Fig. 1A), the thermocouples are imbedded within the sample. Therefore, $T_{j,m} = T_j$, i.e., there is no resistance R'. The measured signal is the difference in temperature between the reference and sample thermocouples ($T_r - T_s$). With the additional assumption that $R_s = R_r = R$, we subtract Equation 4 from Equation 2 and find an equation for the measured signal:

$$T_r - T_s = R\left(\frac{dq_s}{dt} - \frac{dq_r}{dt}\right) \tag{6}$$

By combining Equation 6 with Equations 3 and 5, we have

$$T_r - T_s = R\left(C_s \frac{dT_s}{dt} - C_r \frac{dT_r}{dt}\right) \tag{7}$$

During a programmed temperature scan at a rate dT/dt, a steady state is attained at which the sample and the reference temperatures are changing at the same rate. Equation 7 then becomes

$$T_r - T_s = R\frac{dT}{dt}(C_s - C_r) \tag{8}$$

The measured signal is thus seen to be proportional to the difference in heat capacities, but the value of R is needed to calculate the heat capacity of the sample. Because the thermocouples are embedded in the sample and reference, R will depend on, among other factors, the thermal conductivity of the sample. A calibration is, in principle, needed for every temperature range and sample configuration, making quantitative work difficult at best.

We turn next to power-compensated DSC (Fig. 1C). In this case, there are separate heaters for the sample and reference each with an associated thermometer. For this simple analysis we assume no thermal resistance, i.e., $R_j = 0$. The temperature-measuring system controls power to the two heaters to maintain the sample and reference at the same temperature; thus $T_{sm} = T_{rm}$. The measured signal in this method is the difference in power between the sample and reference sides. Thus, again assuming that steady-state heating rates have been attained, we have

$$\frac{dq_s}{dt} - \frac{dq_r}{dt} = \frac{dT}{dt}(C_s - C_r) \tag{9}$$

The measured signal for this case is again proportional to the difference in heat capacity, but the proportionality constant only includes the operator-selected heating rate and not a heat transfer coefficient. Quantitative work is therefore considerably more reliable.

Finally, using the same set of simplified equations, we turn to heat-flux DSC (Fig. 1B). Mraw's analysis is particularly valuable in dealing with this method in that he includes both R and R'. In this type of instrument the measuring thermocouples are mounted on a plate under the sample and reference pans. We can proceed as before with classical DTA but now recognize that $T_{j,m} \neq T_j$. The resulting equation, analogous to Equation 8, is

$$T_{rm} - T_{sm} = R\frac{dT}{dt}(C_s - C_r) \tag{10}$$

The difference between Equations 8 and 10 is that the R in Equation 10 is an instrument constant and can be found independent of the nature of the sample. Indeed, some manufacturers include the calibration in the electronics of the instrument so that the user obtains a differential power output.

In both power-compensated and heat-flux DSC there will be a temperature difference between T_s and T_{sm}. From Equations 2 and 3 we find

$$T_{sm} - T_s = R'_s C_s \frac{dT_s}{dt} \tag{11}$$

As an example of the temperature lag involved, the values given by Mraw for the Perkin-Elmer DSC-2 are $R'_s \approx 0.06$ K · s/mJ, $C_s \approx 50$ mJ/K, and $dT_s \approx /dt\ 0.167$ K/s and lead to a calculated temperature lag of ~0.5 K.

For the case in which an endothermic or exothermic process occurs, it is no longer acceptable to assume that $dT_s/dt = dT_r/dt$. The relevant equations are a straightforward extension of the derivation given above. These equations are given by Gray (1968), among others, and are simply reproduced here without detailed derivation. For a sample generating heat at a rate dH/dt, we have, for classical DTA or heat-flux DSC,

$$\frac{dH}{dt} = \frac{1}{R(T_s - T_r)} + (C_s - C_r)\frac{dT_r}{dt} + C_s\frac{d}{dt}(T_s - T_r) \quad (12)$$

For power-compensated DSC, assuming that the reference temperature changes at the same rate as the heat source, we have

$$\frac{dH}{dt} = -\left(\frac{dq_s}{dt} - \frac{dq_r}{dt}\right) + (C_s - C_r)\frac{dT_r}{dt} - RC_s\frac{d}{dt}\left(\frac{dq_s}{dt} - \frac{dq_r}{dt}\right) \quad (13)$$

Equations 12 and 13 are quite similar. Both contain three terms involving the signal: $T_s - T_r$ or $dq_s/dt - dq_r/dt$, a heat capacity difference $C_s - C_r$, and a derivative of the signal. Note that it is the negative of the signal that is the first term in Equation 13. Of primary importance is the presence of a thermal resistance factor in the first term of Equation 12, but the factor only appears in the third term of Equation 13.

Modulated DSC

Modulated DSC (MDSC) is a relatively new technique that improves upon the classical DSC performance in terms of increased ability to analyze complex transitions, higher sensitivity, and higher resolution. While classical DSC uses a linear heating or cooling program of the form

$$T = \alpha + \beta t \quad (14a)$$

$$\frac{dT}{dt} = \beta \quad (14b)$$

modulated DSC employs a linear heating or cooling program with a small sinusoidal component superimposed:

$$T = \alpha + \beta t + A_t \sin(\omega t) \quad (15a)$$

$$\frac{dT}{dt} = \beta + A_T\omega\cos(\omega t) \quad (15b)$$

so that the heating rate oscillates between $\beta + A_T\omega$ and $\beta - A_T\omega$ with an average underlying heating rate β. Here, A_T is the temperature modulation amplitude (in kelvin) and ω is the angular frequency of the temperature modulation (in radians per second) equal to $2\pi/P$, where P is the temperature modulation period (in seconds). To illustrate the modulated DSC capabilities, it is possible to recast Equation 13 as follows:

$$\left(\frac{dq_s}{dt} - \frac{dq_r}{dt}\right) = (C_s - C_s)\frac{dT_r}{dt} - \frac{dH}{dt} - RC_s\frac{d}{dt}\left(\frac{dq_s}{dt} - \frac{dq_r}{dt}\right) \quad (16)$$

which upon combining the last two terms on the right-hand side, denoting differences between sample and reference quantities with Δ, and replacing dT_r/dt with dT/dt transforms into

$$\frac{d\Delta q}{dt} = \Delta C\frac{dT}{dt} - F(T,t) \quad (17)$$

In modulated DSC terminology, the term on the left-hand side is the total heat flow, while the first term on the right-hand side is the reversing heat flow or heat capacity component and the second term on the right-hand side is the nonreversing heat flow or kinetic component. Classical DSC (both power compensated and heat flux) measures and puts out the total heat flow. Modulated DSC is different in that it measures and puts out the instantaneous heat flow (Fig. 3), which takes on the modulated character of the temperature. From the modulated heat flow, it is then possible to obtain both the total heat flow (roughly by averaging or, more precisely, using Fourier transform analysis; Fig. 3) and its heat capacity component (roughly by taking the ratio of the modulated heat flow amplitude A_q and A_T or, more precisely, using Fourier transform analysis; Fig. 4). An algebraic difference of these two

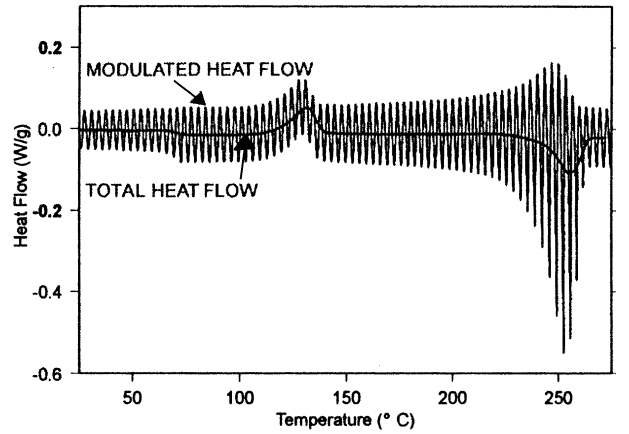


Figure 3. Modulated DSC of a quench-cooled polymer: obtaining total heat flow from modulated heat flow. (Courtesy of TA Instruments.)

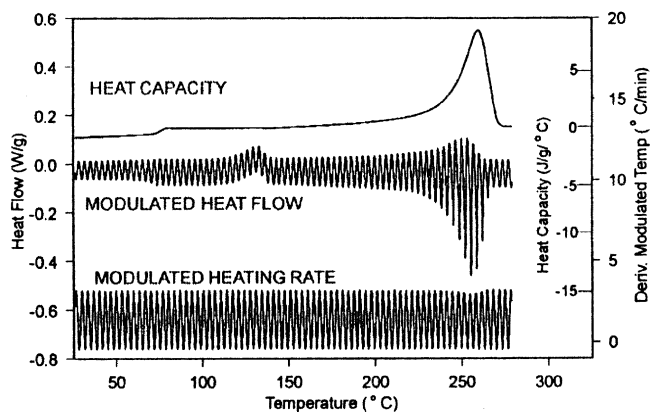


Figure 4. Modulated DSC of a quench-cooled polymer: obtaining the heat capacity (reversing) components from modulated heat flow and modulated heating rate. (Courtesy of TA Instruments.)

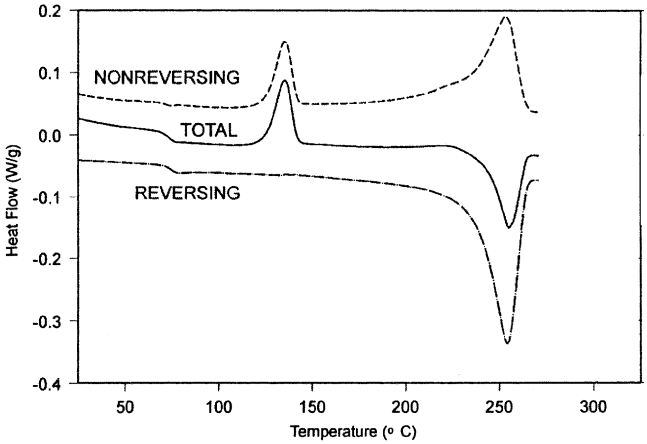


Figure 5. Modulated DSC of a quench-cooled polymer: obtaining the kinetic (nonreversing) component from the total heat flow and the heat capacity (reversing) component. (Courtesy of TA Instruments.)

quantities, yields the kinetic component of the total heat flow (Fig. 5). This significantly improves the ability to separate overlapping thermal events with different behavior, e.g., crystallization (nonreversing) and glass transition (reversing) in polymer blends. Sensitivity and resolution are also improved. In addition, the presence of noninstrumental phase lag between the modulated temperature and the modulated heat flow leads to the formulation of complex heat capacity c_p^*, which can be split into its real (c_p') and imaginary (c_p'') components and analyzed separately.

PRACTICAL ASPECTS OF THE METHOD

The DSC/DTA signal is usually recorded as a function of time, $S(t)$, or temperature, $S(T)$. Conversion between $S(t)$ and $S(T)$ can easily be made since the dependence $T(t)$ is known (see Equation 14a or 15a). In the following paragraphs, we shall use S to denote both $S(t)$ and $S(T)$, unless indicated otherwise.

Zero-Line Optimization

Following Höhne et al. (1996), we shall make a distinction here between the zero line and the baseline (see below). The zero line, S_0, is the signal recorded with both the sample and the reference holders empty. It represents the thermal behavior of the instrument itself. As the name suggests, the zero line should ideally be constant and equal to zero over the whole temperature range. This, of course, is never the case, and every effort should be made to adjust the instrument controls according to manufacturers' specifications to produce a zero line as flat and as close to zero as possible. Zero-line optimization should be followed by the zero-line repeatability check. Five or more (n) zero lines, S_0^i, are recorded over the whole temperature range at a selected scan rate. The temperature-dependent deviation (standard deviation or maximum deviation, absolute or relative) from the mean zero line, $\bar{S}_0 = \sum S_0^i/n$, is the zero-line repeatability. This uncertainty projects directly into absolute heat flow determinations, which involve calculations of the type $S - S_0$, e.g., heat capacity measurements by the scanning method (see below).

Calibration

Quantitative DSC/DTA data can only be obtained if the instrument has been properly calibrated with respect to both temperature (DSC/DTA) and heat flow or enthalpy (DSC). Calibrants should ideally be certified reference materials (values of transition temperature and/or transition enthalpy have been determined for a particular lot of material) or at least high-purity materials for which literature values exist. The most commonly used calibrants with approximate values of their transition properties are given in Table 1. Temperature calibration is a process of determining the difference, δT, between the actual sample temperature, T_s, and the indicated temperature, T_{ind}, and then either incorporating it into the final data treatment or eliminating it by instrument control adjustment. The temperature difference is, in general, a function of the indicated temperature and the heating rate:

$$\delta T = T_s - T_{ind} = f(T_{ind}, \beta) \tag{18}$$

The T_{ind} dependence of δT should be determined with at least two calibrants bracketing the temperature range of interest (two-point calibration) as closely as possible. Using one or more calibrants within the temperature range of interest is always beneficial (multipoint calibration) for checking and, if applicable, correcting the linearity assumption of the two-point calibration. The dependence of δT on β is due to increasing the thermal lag of T_s behind T_{ind} at higher scanning rates. Isothermal calibration furnishes the value of δT for a zero heating rate, δT_0, while calibration at the scanning rate of the experiment furnishes the thermal lag, δT_β. Hence, δT can be calculated as

$$\delta T = \delta T_0 + \delta T_\beta \tag{19}$$

At a given temperature, δT_0 might be of either sign but δT_β is always negative in heating and positive in cooling. A simpler method of temperature calibration is sometimes suggested by instrument manufacturers that omits the

Table 1. DSC/DTA Calibrants, Their Transition Temperatures, and Their Transition Enthalpies

Material	T_{tr}(K)	$\Delta_{tr}H°$(J/g)
Mercury (Hg)	234	11
Gallium (Ga)	303	80
Indium (In)	430	28
Tin (SN)	505	60
Lead (Pb)	601	23
Zinc (Zn)	693	108
Potassium sulfate (K_2SO_4)	858	33
Potassium chromate K_2CrO_4	944	36

isothermal calibration and yields a universal temperature correction $\delta T'$. This method is strictly valid only for very small samples.

Heat flow or enthalpy calibration is a process of determining the signal-to-heat-flow conversion factor or the area-to-enthalpy conversion factor, respectively, under scanning conditions. Heat flow calibration is used for heat capacity measurements by the scanning method (see below). Enthalpy calibration is used for measurements of heat capacity by the enthalpy method (see below) and for general enthalpy measurements. Heat flow calibration is almost universally performed with two calibrants, namely α-alumina (sapphire) for superambient operation and benzoic acid for subambient operation. Heat capacities for both materials have been accurately measured by adiabatic calorimetry, and both are available as standard reference materials. Heat flow dH/dt is assumed to be proportional to the signal $S(t)$:

$$\frac{dH}{dt} = K(t)S(t) \tag{20}$$

where K is the signal-to-heat-flow conversion factor (K is ideally temperature independent but in practice varies slightly with temperature and this dependence needs to be determined). The relation between heat capacity $c_p(t)$ and $S(t)$ is then

$$c_p(t) = \frac{dH}{dT(t)} = K(t)S(t)\left(\frac{dT}{dt}\right)^{-1} \tag{21}$$

so that $K(t)$ can be determined from the known calibrant heat capacity $c_p^{\text{cal}}(t)$, scan rate dT/dt, and calibrant signal $S^{\text{cal}}(t)$ as

$$K(t) = \frac{c_p^{\text{cal}}(t)(dT/dt)}{S^{\text{cal}}(t)} \tag{22}$$

Enthalpy calibration is usually performed with one or ideally more of the calibrants listed in Table 1. Rearranging Equation 20 leads to a relationship between the enthalpy and the integrated peak area A at some temperature of interest,

$$H = K\int S(t)\,dt = KA \tag{23}$$

so that K can be determined from the known transition enthalpy H^{cal} and the integrated calibrant peak area A^{cal} as

$$K = \frac{H^{\text{cal}}}{A^{\text{cal}}} \tag{24}$$

Note that the signal-to-heat-flow conversion factor and the area-to-enthalpy conversion factor are identical. This is only the case if the signal is considered a function of time, $S(t)$. For the signal as a function of temperature, $S(T)$, the signal-to-heat-flow conversion factor K' and the area-to-enthalpy conversion factor K'' are not identical:

$$K'' = K'\left(\frac{dT}{dt}\right)^{-1} \tag{25}$$

Instrument Parameter Selection

It is important that appropriate values of instrument parameters be selected for each experiment. To facilitate correct baseline interpolation (see below), the temperature range should extend well below and above the characteristic temperature of the thermal event under investigation. Subambient studies require a cooling accessory such as an ice-water bath, mechanical refrigeration, or liquid nitrogen cooling. Ideally, it should be possible to obtain identical results in both heating and cooling modes. However, most studies are performed in the heating mode because it is difficult to sustain higher cooling rates over large temperature ranges (especially at lower temperatures) and because temperature calibrations in the cooling mode tend to be less accurate due to nonreproducible calibrant undercooling. The scan rate β (together with sample size, see below) has a profound influence on sensitivity and resolution. High scan rates improve sensitivity because the signal is proportional to the scan rate, e.g., Equation 21, but lead to a loss of resolution and vice versa. A compromise is always necessary. Scan rates from 5 to 20 K/min are most commonly used. For modulated DSC, values for the amplitude A_T and period P of temperature modulation need to be selected. They are typically 0.5−3 K for A_T and 40−100 s for P. The sample must be able to follow the modulated temperature profile. This can be checked by analyzing the modulated heating rate profile and the modulated heat flow profile, or a Lissajous plot of these, for nonsinusoidal distortions. Purge gas compatible with the sample, sample pan, and sample holder needs to be selected. Inert gases (e.g., Ar, N_2, or He) are commonly used for superambient operation while dry He is preferred for subambient operation. Purge gas flow rates on the order of 10 cm^3/min are generally used. Higher flow rates up to 100 cm^3/min can be used with samples generating corrosive and/or condensable gases.

Typical Applications

The principal use of DSC/DTA is, and all the other uses of DSC/DTA are derived from, the measurement of transition temperature, the measurement of transition enthalpy, and the measurement of heat capacity. Transition temperature is determined as extrapolated onset temperature, and transition enthalpy is determined by integration of the signal with special methods used for the glass transition (see Data Analysis and Initial Interpretation). Heat capacity measurements are dealt with in detail below.

The simplest use of DSC/DTA is in identification and characterization of materials. Identification is the determination of sample transition temperature and/or transition enthalpy (see below), e.g., solid to liquid, for an unknown sample, followed by a comparison with a list of possible candidates with their transition temperatures

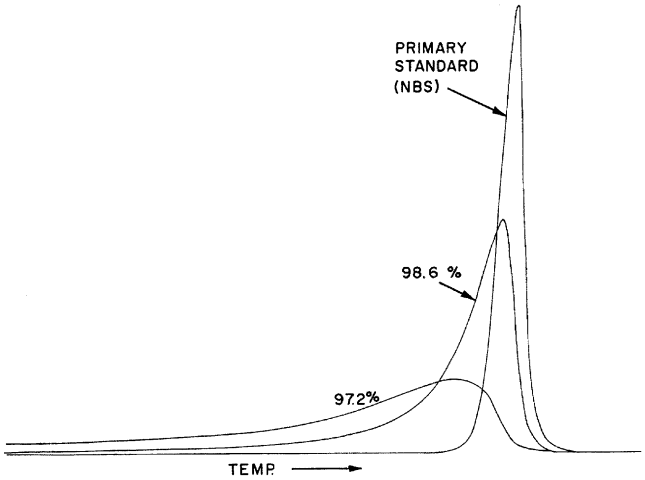


Figure 6. Effect of purity on DSC melting peak shapes. Lower purity leads to peak shift to lower temperature and peak broadening. (Courtesy of Perkin-Elmer.)

and/or enthalpies. Characterization is the determination of transition temperature and/or enthalpy for a known sample. The DSC/DTA technique has found widespread use in purity determination of materials, e.g., pharmaceuticals. This application rests upon the van't Hoff equation (Lewis and Randall, 1923) for melting point depression:

$$\Delta T = -\frac{RT_0^2 x_2}{\Delta_{\text{fus}}H} \tag{26}$$

where ΔT is the melting point depression, R is the universal gas constant, x_2 is the mole fraction of the impurity, and $\Delta_{\text{fus}}H$ is the enthalpy of fusion. Both ΔT and $\Delta_{\text{fus}}H$ are determined simultaneously and x_2 is calculated. The influence of sample purity on a DSC melting peak is shown in Figure 6.

As illustrated in Figure 7 for a simple binary eutectic system, DSC/DTA can be used in the determination of phase diagrams. In addition to scans of both pure components (A, B), several scans at different intermediate compositions are performed (C, D). These exhibit two endothermic peaks, one due to the eutectic melting ($\sim$77°C) and the other one due to the crossing of the liquidus curve. The enthalpy associated with the eutectic melting is a simple function of composition (E) and can be used to determine the eutectic composition. The final phase diagram is shown in F.

The glass transition in glasses and polymers can be successfully studied by DSC/DTA. While there is no enthalpy change associated with the glass transition, the heat capacity does change abruptly. A typical scan through a glass transition temperature T_g and its determination are shown in Figure 8.

The scanning method is the one more commonly used for heat capacity determination by DSC. Three runs, each starting and finishing with an isothermal section and each spanning the same temperature range at the same heating rate, are necessary: empty pan, calibrant, and sample, producing three raw signals $S_{\text{raw}}^{\text{emp}}(t)$, $S_{\text{raw}}^{\text{cal}}(t)$ and $S_{\text{raw}}^{\text{smp}}(t)$, respectively. Although ideally constant and equal to zero, the initial and final isothermal sections of

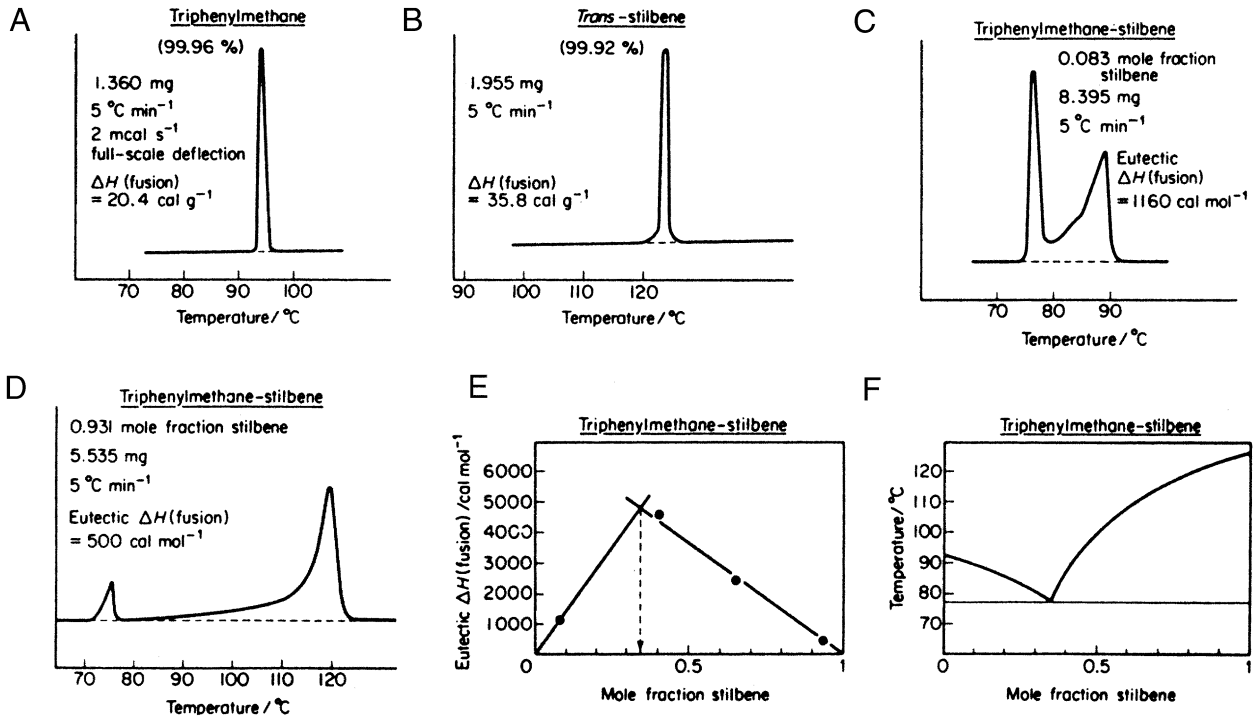


Figure 7. Binary phase information from DSC: (**A**, **B**) scans of the pure compounds; (**C**, **D**) scans of mixtures of the two compounds; (**E**) measured enthalpy change at the eutectic temperature; (**F**) binary phase diagram. (Courtesy of Perkin-Elmer.)

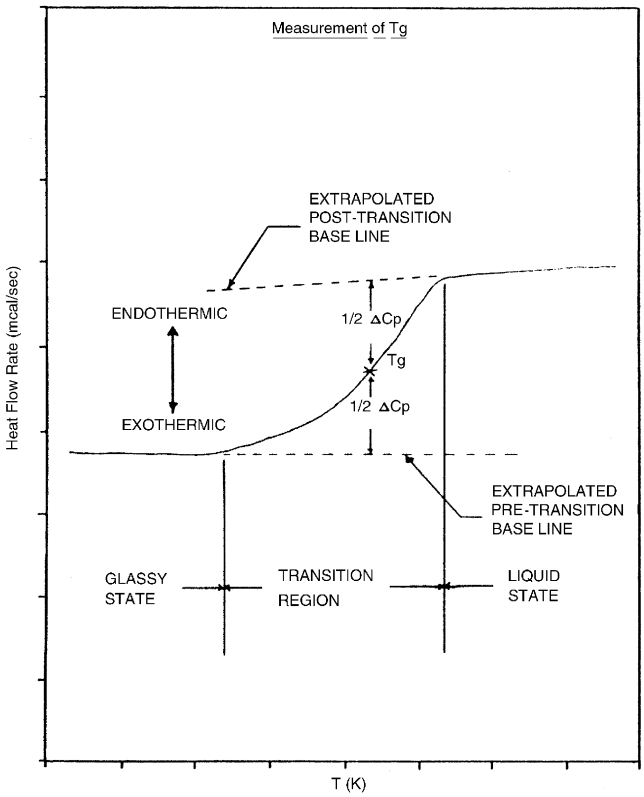


Figure 8. Measured T_g. The glass transition manifests itself as a shift, rather than a peak, in the signal. (Courtesy of Perkin-Elmer.)

the three runs differ. A baseline balancing procedure (Richardson, 1984, 1992) is needed to bring the initial and final isothermal sections into coincidence (Fig. 9), so that two net signals, $S^{\text{cal}}(t) = S^{\text{cal}}_{\text{raw}}(t) - S^{\text{emp}}_{\text{raw}}(t)$ and $S^{\text{smp}}(t) = S^{\text{smp}}_{\text{raw}}(t) - S^{\text{emp}}_{\text{raw}}(t)$, respectively, can be calculated. The signal-to-heat-flow conversion factor is determined

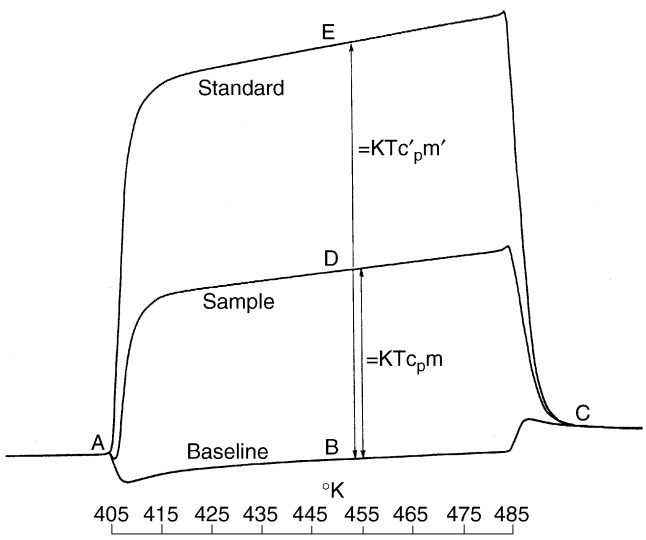


Figure 9. Typical specific heat determination. Initial and final isothermal signals have been brought into coincidence by applying a baseline-balancing procedure. (Richardson, 1984, 1992; courtesy of Perkin-Elmer.)

from Equation 22. The unknown heat capacity is then determined from Equation 21 with $S^{\text{smp}}(t)$ in place of $S(t)$. Even though it is possible to substitute Equation 22 back into Equation 21 to yield

$$c_p(t) = c_p^{\text{cal}}(t)\frac{S^{\text{smp}}(t)}{S^{\text{cal}}(t)} \tag{27}$$

valuable information about the behavior of $K(t)$, indicative of instrument performance, is lost and this method is, thus, not recommended. A continuous heat capacity curve, $c_p(t)$, can be determined over a 100 K or larger temperature range rather quickly. Heat capacity can, alternatively, be determined by the enthalpy method, which approximates the method of classical adiabatic calorimetry, although temperature scanning is still involved. The same three runs under identical conditions as in the scanning method are necessary; however, the scanning temperature interval T_{fin} to T_{ini} is narrower (e.g., 5 to 20 K). The area-to-enthalpy conversion factor is determined from Equation 24. The enthalpy involved in the scan is calculated from Equation 23. The heat capacity at the midpoint of the scanning temperature interval, T_{mid}, is then calculated as

$$C_p(T_{\text{mid}}) = \frac{H}{T_{\text{fin}} - T_{\text{ini}}} \tag{28}$$

so that, for a given temperature range, discrete values of $C_p(T)$ are obtained, their number depending on the width of the individual scanning temperature interval.

Reaction onset temperatures and enthalpies can be determined by DSC by the scanning method or by the isothermal method. In the isothermal method, the reaction mixture sample is heated rapidly from a temperature at which it is inert to some elevated temperature. The signal at that temperature is then recorded until it reaches a constant value S_{end}. The reaction enthalpy is calculated as

$$H = K\int [S(t) - S_{\text{end}}]\,dt \tag{29}$$

In the scanning method, where heat capacity changes will usually be significant due to the larger temperature range involved, it is advantageous to repeat the experiment with the already reacted sample to yield the product heat capacity signal, $S_{cp}(t)$. Assuming the validity of the Neumann-Kopp rule (Kubaschewski and Alcock, 1929) the reaction enthalpy can then be calculated as

$$H = K\int [S(t) - S_{cp}(t)]\,dt \tag{30}$$

Application of modulated DSC in determining reaction enthalpies is particularly valuable because of its ability to separate heat capacity and enthalpy changes.

The DSC/DTA technique can also be used to study the rates of thermal processes, e.g., phase transformations or chemical reactions. Measurements of the onset temperature and enthalpy of the process are analogous to the corresponding thermodynamic measurements and will not be discussed further here. However, determinations of the

rate of these processes involve somewhat different methods and are briefly discussed below. Two approaches to this problem exist: isothermal and scanning. In the isothermal method, a sample of interest is held in the instrument for a given time at a temperature at which the process will take place. The sample is then programmed through the reaction regime. This process is repeated for various times and temperatures. The observed change in the exothermal peak area can then be related to the extent of reaction undergone during the isothermal hold. From these measurements, information on the rate of the process can be obtained. In the scanning method, the reacting mixture is scanned through a temperature range in which the reaction takes place. It is assumed that the rate of reaction is proportional to the rate of heat evolution and that only one reaction is taking place. The use of this technique to determine reaction mechanisms, activation energies, and other kinetic parameters is difficult and subject to a variety of uncertainties. Some of the difficulties of using thermal analysis methods for the study of kinetics have been identified (Flynn et al., 1987). With respect to the scanning method of kinetic studies, Flynn et al. remark: "it is impossible to separate unequivocally the temperature-dependent and concentration-dependent parts of a rate expression by experiments in which temperature and concentration are changing simultaneously."

METHOD AUTOMATION

All modern DSC/DTA instruments employ computerized control, data acquisition, and data analysis. Instrumental parameters are set through and the experiment is initiated and controlled by a computer and data are acquired, saved, and analyzed (in real time or after the experiment) using sophisticated programs. Although this significantly reduces operator labor involved in producing data, it might also lead to misinterpretation of data. It is, therefore, important that the operator have access to the algorithms used in the data analysis.

The need for higher throughput in routine analyses has led to the development of autosamplers, which are now available for most commercial DSC/DTA instruments. These allow for unattended analysis of up to 60 samples.

DATA ANALYSIS AND INITIAL INTERPRETATION

An endothermic or exothermic event will give rise to a peak in the DSC/DTA curve. In power-compensated DSC, endotherms are plotted in the positive direction and exotherms are plotted in the negative direction. In heat-flux DSC and DTA, the opposite is true. The following discussion will use power-compensated DSC as an example but can readily be extended to heat-flux DSC and DTA.

A peak associated with fusion of an ultrapure metal is shown in Figure 10. The vertex of the peak is not strictly the melting temperature. Rather, the leading edge of the peak is extrapolated down to the baseline with their intersection defined as the extrapolated onset temperature (Fig. 10). It can be shown that the slope of the leading edge is equal to $(1/R_0)\,(dT/dt)$, where R_0 is the controlling thermal resistance between the sample pan and the sample holder. Since R_0 does not change from sample to sample (using the same or the same type pan), the true melting temperature of any other sample is determined as an extrapolated onset temperature using the slope determined with an ultrapure metal (Fig. 11).

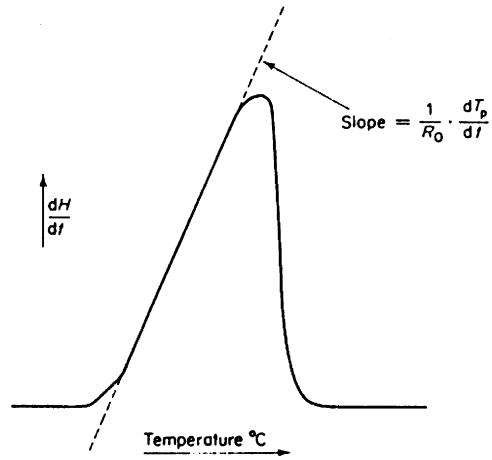


Figure 10. Thermogram for the fusion of 99.999% pure indium. (Courtesy of Perkin-Elmer.)

The enthalpy associated with a thermal event when there is little change in heat capacity before and after the transition (Fig. 11) can be calculated by drawing-in a linear baseline $S_b(t)$ and integrating:

$$H = K \int [S(t) - S_b(t)]\, dt \tag{31}$$

However, when the change in specific heat is large, a more rigorous method of establishing the baseline must be used. Several have been proposed. The principle behind the correction of Brennan et al. (1969) is outlined in Figure 12. Once the nonlinear baseline is established, Equation 29 is used to calculate the enthalpy.

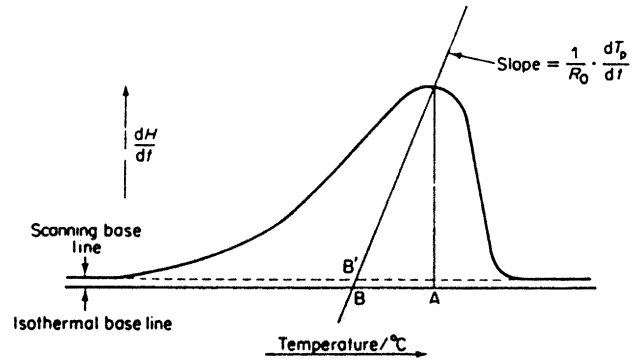


Figure 11. Thermogram for the fusion of 99.999% pure triphenylmethane. (Courtesy of Perkin-Elmer.)

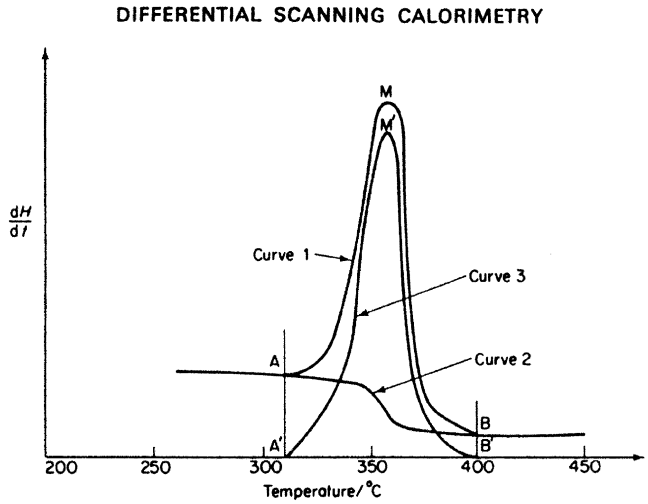


Figure 12. Correction of observed enthalpy data for the change in heat capacity: curve 1, uncorrected signal; curve 2, hypothetical signal for a reaction with zero enthalpy; curve 3, signal corrected for change in heat capacity by subtracting curve 2 from curve 1. The area under curve 3 represents the enthalpy of the reaction. (Brennan et al., 1969; courtesy of Perkin-Elmer.)

SAMPLE PREPARATION

One of the most appealing features of DSC/DTA is the small sample size and easy sample preparation. Typical sample mass is ~10 mg. While transition and reaction temperatures are independent of sample size, determination of transition and reaction enthalpies requires knowledge of accurate sample mass. For the typical sample, a $\pm$0.1-mg error in mass represents a 1% error in enthalpy. To ensure good thermal contact with the sample pan (~5 mm in diameter), disk-shaped samples lying flat on the bottom of the sample pan are preferred. Hence, a material with a density of 1 g/cm^3 would need to be formed into a disk ~0.5 mm thick. Filmlike materials, e.g., polymers, only need to be punched out to the right diameter (and stacked if the film is too thin). Solid materials, e.g., metals and alloys, need to be cut to the right size with a flat surface, if possible, toward the bottom of the pan. Powders need to be compacted into a pellet using a right diameter die. Liquids in the right amount can be measured out directly into the sample pan. Given the small DSC/DTA sample size, representative sampling of the material is crucial. Use of larger samples under otherwise identical conditions improves sensitivity but leads to deterioration of resolution (could be compensated for by lower scan rates). Using larger samples at lower scan rates is generally preferable to using smaller samples at higher scan rates.

The large variety of sample pans available from DSC/DTA instrument manufacturers (see Appendix) can be characterized by material of construction and sealability. The material of construction needs to be selected so that there are no reactions between the sample pan and the sample, and between the sample pan and the sample holder, within the temperature range of interest. At the same time, the material of construction should have the highest thermal conductivity. This makes aluminum the most frequently used sample pan material. Table 2 lists a number of commonly used sample pan materials. Most samples can be enclosed in nonhermetic sample pans. Metallic sample pans usually come with a lid that can be crimped onto the sample pan with a crimper tool; however, this makes the pans and lids single use. Graphite and ceramic sample pans also come with covers and can be reused. For special applications such as studies of volatile materials, materials generating corrosive or condensable gases, and air-sensitive materials, hermetically sealed metallic sample pans can be used. The sample pans are sealed in air, in a protective atmosphere, or under vacuum. Hermetic sample pans become pressurized at high temperatures and should be handled with caution.

SPECIMEN MODIFICATION

Specimen modification in DSC/DTA results from the exposure to high and/or low temperatures and is an integral part of the method, e.g., solid-to-liquid transformation in purity determination. In some cases, however, specimen modification can be detrimental, e.g., denaturation of biological samples, and needs to be taken into account. The nature and extent of specimen modification are frequently important for elucidation of DSC/DTA data. In such cases, auxiliary methods of analysis, e.g., x-ray diffraction analysis, are applied prior to and following a DSC/DTA experiment. If changes in the sample that occurred at higher temperature are reversible upon cooling to room temperature (for sample removal and analysis), the cooling rate has to be increased to the maximum (up to 500 K/min), effectively quenching the sample, or a simultaneous analysis technique can be applied. Simultaneous techniques are extremely valuable for elucidation of complex DSC/DTA data. The most common combinations are

Table 2. DSC/DTA Sample Pan Materials, Their Properties, and Purge Gas Compatibility

Sample Pan Material	Upper Temperature Limit[a]	Thermal Conductivity	Purge Gas
Aluminum (Al)	873 K	Excellent	Inert, reducing, oxidizing
Noble metals (Cu, Ag, Au, Pt, Ti, Ni, Ta, etc.)	>1173 K, 100 K below respective T_m	Excellent	Inert, reducing
Stainless steel	1273 K	Good	Inert, reducing
Graphite (C)	2273 K	Excellent	Inert, reducing
Oxides (SiO_2, Al_2O_3, Y_2O_3, BeO, etc.)	>1273 K, 100 K below respective T_m	Poor	Inert, oxidizing

[a]Caution! Reactions with sample and/or sample holder might occur at a lower temperature due to the existence of lower melting eutectics.

DSC/DTA + thermal gravimetric analysis (TGA; see THERMOGRAVIMETRIC ANALYSIS), DSC/DTA + mass spectrometry (MS; see SIMULTANEOUS TECHNIQUES INCLUDING ANALYSIS OF GASEOUS PRODUCTS), and DSC/DTA + Fourier transform infrared spectroscopy (FTIR; see SIMULTANEOUS TECHNIQUES INCLUDING ANALYSIS OF GASEOUS PRODUCTS). With the advent of synchrotron light sources, combinations such as DSC/DTA + x-ray absorption fine structure (XAFS; Tröger et al., 1997) and DSC/DTA + x-ray diffraction (XRD; Lexa, 1999) became possible.

PROBLEMS

Heat-flux DSC and DTA instruments use thermocouples to detect temperature. Because of interdiffusion at the junction, it is possible that thermocouple calibrations will change. This is particularly troublesome for cases of extended times of operation near the upper limit of the temperature range of the thermocouple. Periodic temperature calibration of the instrument is recommended.

Reactions with sample pans are a chronic problem that must be considered, particularly for high-temperature work. A variety of DTA/DSC sample pans are commercially available (see Table 2). It is usually possible to find suitable materials, but it is important to verify that no significant reaction has taken place. Serious errors and damage to equipment can result from ignoring this possibility.

Resolution of close peaks can present difficulties. Indeed, the experimenter may not even be aware of the existence of hidden peaks. It is important when working with unfamiliar systems to conduct scans at several heating/cooling rates. Lower rates allow resolution of closely lying peaks, at the expense, however, of signal strength. Examination of both heating and cooling traces can also be useful.

It should be obvious that caution should be observed to avoid the presence of an atmosphere in the DSC/DTA system that could react with either the sample or the crucible. Less obvious, perhaps, is the need to be aware of vapors that may be evolved from the sample that can damage components of the experimental system. Evolution of chloride vapors, e.g., can be detrimental to platinum components. Vaporization from the sample can also significantly alter the composition and the quantity of sample present. Sealable pans are commercially available that can minimize this problem.

Because DSC/DTA typically involves scanning during a programmed heating or cooling cycle, slow processes can be troublesome. In measuring melting points, e.g., severe undercooling is commonly observed during a cooling cycle. An instantaneous vertical line characterizes the trace when freezing begins. In studying phase diagrams, peritectic transformations are particularly sluggish and troublesome to define.

LITERATURE CITED

Boerio-Goates, J. and Callanan, J. E. 1992. Differential thermal methods. *In* Physical Methods of Chemistry, 2nd ed., Vol. 6, Determination of Thermodynamic Properties (B. W. Rossiter and R. C. Baetzold, eds.). pp. 621–717. John Wiley & Sons, New York.

Boersma S. L. 1955. A theory of differential thermal analysis and new methods of measurement and interpretation. *J. Am. Ceram. Soc.* 38:281–284.

Brennan, W. P., Miller, B., and Whitwell, J. 1969. An improved method of analyzing curves in differential scanning calorimetry. *I&EC Fund.* 8:314–318.

Flynn J. H., Brown, M., and Sestak, J. 1987. Report on the workshop: Current problems of kinetic data reliability evaluated by thermal analysis. *Thermochim. Acta* 110:101–112.

Gray, A.P. 1968. A simple generalized theory for the analysis of dynamic thermal measurements. *In* Analytical Calorimetry (R. S. Porter and J. F. Johnson, eds.). pp. 209–218. Plenum Press, New York.

Höhne, G., Hemminger, W., and Flammersheim, H. J. 1996. Differential Scanning Calorimetry. An Introduction for Practitioners. Springer-Verlag, Berlin.

Kubaschewski, O. and Alcock, C. B. 1929. Metallurgical Thermochemistry. Pergamon Press, New York.

Lewis, G. N. and Randall, M. 1923. Thermodynamics. McGraw-Hill, New York.

Lexa, D. 1999. Hermetic sample enclosure for simultaneous differential scanning calorimetry/synchrotron powder X-ray diffraction. *Rev. Sci. Instrum.* 70:2242–2245.

Mackenzie R. C. 1985. Nomenclature for thermal analysis—IV. *Pure Appl. Chem.* 57:1737–1740.

Mraw, S. C. 1988. Differential scanning calorimetry. *In* CINDAS Data Series on Material Properties, Vol. I-2, Specific Heat of Solids (C. Y. Ho, ed., A. Cezairliyan, senior author and volume coordinator) pp. 395–435. Hemisphere Publishing, New York.

Richardson, M. J. 1984. Application of differential scanning calorimetry to the measurement of specific heat. *In* Compendium of Thermophysical Property Measurement Methods, Vol. 1, Survey of Measurement Techniques (K. D. Maglic, A. Cezairliyan, and V. E. Peletsky, eds.). pp. 669–685. Plenum Press, New York.

Richardson, M. J. 1992. The application of differential scanning calorimetry to the measurement of specific heat. *In* Compendium of Thermophysical Property Measurement Methods, Vol. 2, Recommended Measurement Techniques and Practices (K. D. Maglic, A. Cezairliyan, and V. E. Peletsky, eds.). pp. 519–545. Plenum Press, New York.

Tröger, L., Hilbrandt, N., and Epple, M. 1997. Thorough insight into reacting systems by combined *in-situ* XAFS and differential scanning calorimetry. *Synchrotron Rad. News* 10:11–17.

Watson, E. S., O'Neill, M. J., Justin, J., and Brenner, N. 1964. A differential scanning calorimeter for quantitative differential thermal analysis. *Anal. Chem.* 36:1233–1237.

KEY REFERENCES

Boerio-Goates and Callanan, 1992. See above.

A comprehensive look at the development and current status of thermal analysis with emphasis on DSC and DTA.

Höhne et al., 1996. See above.

A detailed and up-to-date review of DSC. Sound presentation of the theoretical basis of DSC. Emphasis on instrumentation, calibration, factors influencing the measurement process, and interpretation of results.

Richardson, 1984, 1992. See above.

Two excellent reviews of heat capacity determination by DSC.

APPENDIX: ACQUIRING A DSC/DTA INSTRUMENT

Acquisition of a DSC/DTA instrument should be preceded by a definition of its intended use, e.g., routine quality control analyses vs. research and development. While in the former setting an easy-to-use model with an available autosampler might be called for, the latter setting will likely require a highly flexible model with a number of user-selectable controls. A technical specification checklist (Höhne et al., 1996, used with permission) should then be compiled for different instruments from values obtained from manufacturers:

Manufacturer	...
Type of measuring system	Heat-flux disk type Heat-flux cylinder type Power compensation
Special feature	...
Sample volume (standard crucible)	mm^3
Atmosphere (vacuum?, which gases?, pressure?)	...
Temperature range	From ... to ... K
Scanning rates	From ... to ... K/min
Zero-line repeatabilty	From ± ... μW (at ... K) to ± ... μW (at ... K)
Peak-area repeatability	± ... % (at ... K)
Total uncertainty for heat	± ... % (at ... K)
Extrapolated peak-onset temperature	
Repeatability	± ... K (at ... K)
Total uncertainty for temperature	± ... K (at ... K)
Scanning noise (pp) at ... K/min	From ± ... μW (at ... K) to ± ... μW (at ... K)
Isothermal noise (pp)	From ± ... μW (at ... K) to ± ... μW (at ... K)
Time constant with sample	... s
Additional facilities	...

The lists should then be compared with each other and with a list of minimum requirements for the intended use. (Fiscal considerations will, undoubtedly, also play a role.) Manufacturer data should be critically evaluated as to the conditions under which the same have been determined. For instance, the majority of manufacturers give the same value for isothermal noise of their instruments as 0.2 μW. This value has apparently been obtained under extremely well controlled conditions and will not be reproduced in everyday use, where isothermal noise levels of ~1 μW are more realistic.

DUSAN LEXA
LEONARD LEIBOWITZ
Argonne National Laboratory
Argonne, Illinois

COMBUSTION CALORIMETRY

INTRODUCTION

In the absence of direct experimental information, one often must rely on a knowledge of thermodynamic properties to predict the chemical behavior of a material under operating conditions. For such applications, the standard molar Gibbs energy of formation $\Delta_f G^\circ_m$ is particularly powerful; it is frequently derived by combining the standard molar entropy of formation $\Delta_f S^\circ_m$ and the standard molar enthalpy of formation ($\Delta_f G^\circ_m = \Delta_f H^\circ_m - T \cdot \Delta_f S^\circ_m$), often as functions of temperature and pressure. It follows, therefore, that the standard molar enthalpy of formation, $\Delta_f H^\circ_m$, is among the most valuable and fundamental thermodynamic properties of a material. This quantity is defined as the enthalpy change that occurs upon the formation of the compound in its standard state from the component elements in their standard states, at a designated reference temperature, usually (but not necessarily) 298.15 K, and at a standard pressure, currently taken to be 100 kPa or 101.325 kPa.

Many methods have been devised to obtain $\Delta_f H^\circ_m$ experimentally. Those include the so-called second- and third-law treatments of Knudsen effusion and mass-spectrometric results from high-temperature vaporization observations, as well as EMF results from high-temperature galvanic cell studies (Kubaschewski et al., 1993). Each of those techniques yields $\Delta_f G^\circ_m$ for the studied process at a given temperature. To derive values of $\Delta_f H^\circ_m$, the high-temperature results for $\Delta_f G^\circ_m$ must be combined with auxiliary thermodynamic information (heat capacities and enthalpy increments at elevated temperatures). Frequently, it turns out that the latter properties have not been measured, so they must be estimated (Kubaschewski et al., 1993), with a consequent degradation in the accuracy of the derived $\Delta_f H^\circ_m$. Furthermore, high-temperature thermodynamic studies of the kind outlined here often suffer from uncertainties concerning the identities of species in equilibrium. Another potential source of error arises from chemical interactions of the substances or their vapors with materials of construction of Knudsen or galvanic cells. All in all, these approaches do not, as a rule, yield precise values of $\Delta_f H^\circ_m$. Sometimes, however, they are the only practical methods available to the investigator.

Alternative procedures involve measurements of the enthalpies (or enthalpy changes) of chemical reactions, $\Delta_r H^\circ_m$, of a substance and the elements of which it is composed. Experimentally, the strategy adopted in such determinations is dictated largely by the chemical properties of the substance. The most obvious approach involves direct combination of the elements, an example of which is the combustion of gaseous H_2 in O_2 to form H_2O. In the laboratory, however, it is often impossible under experimental conditions to combine the elements to form a particular compound in significant yield. An alternative, albeit less direct route has been used extensively for the past century or more, and involves measurements of the enthalpy change associated with a suitable chemical reaction of the material of interest. Here, the compound is "destroyed" rather than formed. Such chemical reactions have

involved, *inter alia*, dissolution in a mineral acid or molten salt (Marsh and O'Hare, 1994), thermal decomposition to discrete products (Gunn, 1972), or combustion in a gas such as oxygen or a halogen. This last method is the the subject matter of this unit. As an example of the latter technique, we mention here the determination of $\Delta_f H^\circ_m$ of ZrO_2 (Kornilov et al., 1967) on the basis of the combustion of zirconium metal in high-pressure $O_2(g)$

$$Zr(cr) + O_2(g) = ZrO_2(cr) \quad (1)$$

where $\Delta_f H^\circ_m(ZrO_2) = \Delta_c H^\circ_m(Zr)$, the standard molar enthalpy of combustion of Zr in $O_2(g)$, i.e., the standard molar enthalpy change associated with the reaction described in Equation 1. It is essential when writing thermochemical reactions that the physical states of the participants be stated explicitly; thus, (cr) for crystal, (s) for solid, (g) for gas, and (l) for liquid. Enthalpies of formation of numerous oxides of metals (Holley and Huber, 1979) are based on similar measurements. These are special cases in that they involve oxidation of single rather than multiple elements, as in the combustion of aluminum carbide (King and Armstrong, 1964)

$$Al_4C_3(cr) + 6O_2(g) = 2Al_2O_3(cr) + 3CO_2(g) \quad (2)$$

Of all the materials studied thermodynamically by combustion, organic substances have been predominant, for the simple reason (apart from their numerousness) that most of them react cleanly in oxygen to give elementary and well defined products. Thus, the combustion in high-pressure oxygen of benzoic acid, a standard reference material in calorimetry, proceeds as follows

$$C_6H_5COOH(cr) + (15/2)O_2(g) = 7CO_2(g) + 3H_2O(l) \quad (3)$$

The enthalpy of formation of benzoic acid is the enthalpy change of the reaction

$$7C(cr) + 3H_2(g) + O_2(g) = C_7H_6O_2(cr) \quad (4)$$

where all substances on both sides of Equation 4 are in their standard states. It should be noted that the reference state of C is not merely the crystalline form but, more exactly, "Acheson" graphite (CODATA, 1989). Obviously, a direct determination of the enthalpy change of the reaction in Equation 4 is not possible, and it is in addressing such situations that combustion calorimetry has been particularly powerful. Similarly, with reference to Equation 3, separate equations may be written for the formation of $CO_2(g)$ and $H_2O(l)$

$$C(cr) + O_2(g) = CO_2(g) \quad (5)$$

$$H_2(g) + (1/2)O_2(g) = H_2O(l) \quad (6)$$

where $\Delta_r H^\circ_m$ for Equation 5 $= \Delta_f H^\circ_m(CO_2, g)$, and $\Delta_r H^\circ_m$ for Equation 6 $= \Delta_f H^\circ_m(H_2O, l)$. By appropriate combination of Equations 3, 5, and 6, one obtains Equation 4. In other words

$$\begin{aligned}\Delta_f H^\circ_m(C_7H_6O_2) &= -\Delta_r H^\circ_m\,(\text{Equation 3}) + 7 \cdot \Delta_r H^\circ_m\,(\text{Equation 5}) \\ &\quad + 3 \cdot \Delta_r H^\circ_m\,(\text{Equation 6}) \\ &= -\Delta_r H^\circ_m\,(\text{Equation 3}) + 7 \cdot \Delta_f H^\circ_m(CO_2, g) \\ &\quad + 3 \cdot \Delta_f H^\circ_m(H_2O, l) \end{aligned} \quad (7)$$

Thus, the standard molar enthalpy of formation of benzoic acid is obtained as the difference between its standard molar enthalpy of combustion in oxygen and the standard molar enthalpies of formation of the products, {$CO_2(g)$ and $H_2O(l)$}, the latter being multiplied by the appropriate stoichiometric numbers, 7 and 3, respectively. This is a general rule for calculations of $\Delta_f H^\circ_m$ from reaction calorimetric data.

Not surprisingly, oxygen bomb calorimetry has also been productive in thermochemical studies of organometallic materials. In that connection, the work of Carson and Wilmshurst (1971) on diphenyl mercury is taken as an example. These authors reported the massic energy of the combustion reaction

$$\begin{aligned}&Hg(C_6H_5)_2(cr) + (29/2)O_2(g) \\ &\quad = Hg(l) + 12CO_2(g) + 5H_2O(l)\end{aligned} \quad (8)$$

and derived $\Delta_f H^\circ_m$ {$Hg(C_6H_5)_2$}. Corrections were applied to the experimental results to allow for the formation of small amounts of byproduct HgO and $HgNO_3$, both of which were determined analytically.

Notice that values of $\Delta_f H^\circ_m$ are required for the products shown in Equation 3 and Equation 8. Because the value of $\Delta_f H^\circ_m$ varies as a function of temperature and pressure, it must always, for compounds interconnected by a set of reactions, refer to the same standard pressure and reference temperature.

PRINCIPLES OF THE METHOD

Thermodynamic Basis of Combustion Calorimetric Measurements

Investigations of combustion calorimetric processes are based on the first law of thermodynamics

$$U_f - U_i = \Delta U = Q - W \quad (9)$$

where U_i and U_f are the internal energies of a system in the initial and final states, respectively; Q is the quantity of heat absorbed between the initial and final states; and W is the work performed by the system. Frequently, the only work performed by a system is against the external pressure; thus

$$U_f - U_i = Q - \int_{V_i}^{V_f} p \cdot dV \quad (10)$$

where V_i and V_f are, respectively, the initial and final volumes of the system. In the case of a process at constant volume, $W = 0$, and $Q = Q_V$, from which

$$Q_V = U_f - U_i = \Delta U \quad (11)$$

Therefore, the quantity of heat absorbed by the system or released to the surroundings at constant volume is equal to the change in the internal energy of the system. Measurements of energies of combustion in a sealed calorimetric bomb are of the constant-volume type.

If a process occurs at constant pressure, Equation 10 can be rewritten as

$$U_f - U_i = Q_p - p \cdot (V_f - V_i) \tag{12}$$

whence

$$Q_p = (U_f + p \cdot V_f) - (U_i + p \cdot V_i) \tag{13}$$

where p denotes the external pressure. By substituting $H = U + p \cdot V$, one obtains

$$Q_p = H_f - H_i = \Delta H \tag{14}$$

where H denotes enthalpy. Reactions at constant pressure are most frequently studied in apparatus that is open to the atmosphere.

The relation between Q_p and Q_V is as follows

$$Q_p = Q_V + p \cdot (V_f - V_i) = Q_V + \Delta\nu^g \cdot R \cdot T \tag{15}$$

where $\Delta\nu^g$ is the change in stoichiometric numbers of gaseous substances involved in a reaction, R is the gas constant, and T is the thermodynamic temperature.

Equation 15 permits the experimental energy change determined in combustion calorimetric experiments (constant volume) for a given reaction to be adjusted to the enthalpy change (constant pressure). For example, $\Delta v^g = 1 - 2 - 1 = -2$ for the combustion of gaseous methane

$$CH_4(g) + 2O_2(g) = CO_2(g) + 2H_2O(l) \tag{16}$$

PRACTICAL ASPECTS OF THE METHOD

The Combustion Calorimeter

The combustion calorimeter is the instrument employed to measure the energy of combustion of substances in a gas (e.g., Sunner, 1979). A common theme running through the literature of combustion calorimetry is the use of apparatus designed and constructed by individual investigators for problem-specific applications. To the best of our knowledge, only two calorimetric systems are commercially available at this time, and we shall give details of those later.

In general, combustion calorimetry is carried out in a closed container (reaction vessel, usually called the bomb) charged with oxidizing gas to a pressure great enough ($\sim$3 MPa of O_2, for example) to propel a reaction to completion. The bomb rests in a vessel filled with stirred water which, in turn, is surrounded by a 1-cm air gap (to minimize convection) and a constant-temperature environment, usually a stirred-water thermostat. Occasionally, a massive copper block in thermostatically controlled air (aneroid calorimeter) is used for this purpose (Carson, 1979). Most thermostats used in combustion calorimetry are maintained at, or close to, a temperature of 298.15 K. The combination of bomb and water-filled vessel is usually referred to as the calorimetric system, but this really includes auxiliary equipment such as a thermometer, stirrer, and heater.

Apparatuses that have constant-temperature surroundings are most common, and are called isoperibol calorimeters. Rarer is the adiabatic combustion calorimeter (Kirklin and Domalski, 1983); its jacket temperature is designed to track closely that of the calorimeter so that the heat exchanged with the surroundings is negligible. For the purposes of the present discussion, we shall deal only with the isoperibol calorimeter, although what we say here, with the exception of the corrected temperature change, applies to both types.

A typical combustion bomb is illustrated in Figure 1. The bomb body (A) is usually constructed of stainless steel with a thickness sufficient to withstand not only the initial pressure of the combustion gas, but also the instantaneous surge in pressure that follows the ignition of a sample.

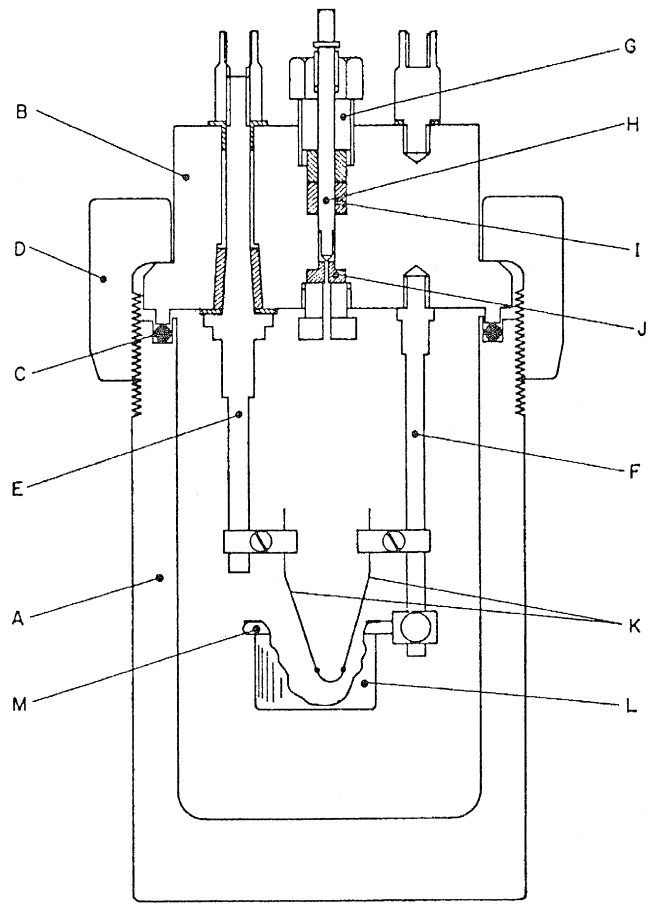


Figure 1. Cross-section of typical calorimetric bomb for combustions in oxygen. A, bomb body; B, bomb head; C, sealing O-ring; D, sealing cap; E, insulated electrode for ignition; F, grounded electrode; G, outlet valve; H, needle valve; I, packing gland; J, valve seat; K, ignition wire; L, crucible; M, support ring for crucible.

Copper has also been used as a bomb material; it has a thermal conductivity superior to that of stainless steel, but is mechanically weaker and chemically more vulnerable. In Figure 1, B is called the bomb head or lid. It is furnished with a valve (G) for admitting oxygen gas before an experiment and for discharging excess O_2 and gaseous products of combustion through the needle valve (H) after an experiment. This valve is surrounded with a packing gland (I) and is also fitted with a valve seat (J) against which it seals. The bomb head (B) also acts as a platform to support components of an electrical ignition circuit, namely, an electrically insulated electrode (E) and a grounded electrode (F). These electrodes are connected by means of a thin wire (K) that is usually made of platinum, although other wires may be used, depending on the specific application. A ring (M) attached to the grounded electrode supports a platinum crucible (L) in which the combustion sample rests (in one example mentioned above, this would be a compressed pellet of benzoic acid). The entire lid assembly is inserted in the bomb body and tightened in position by means of the sealing cap (D), which brings force to bear against the rubber O-ring (C). Thus, a properly functioning bomb can be charged with O_2 to the desired pressure while the contents are hermetically sealed.

Figure 2 gives a simplified view of the most common combustion calorimeter, that with a constant temperature environment (isoperibol). A combustion bomb (A) is supported in the water-filled calorimeter vessel (E) which, in turn, is surrounded by the air gap and thermostat (C), whose inner surface (D) can also be discerned. Note that the lid of the thermostat (F) is connected to the main thermostat; thus, water circulating at a uniform temperature surrounds the calorimetric system at all times. The thermometer used to measure the temperature of the calorimeter is inserted at G, and a stirrer is connected at J (to ensure that the energy released during the combustion is

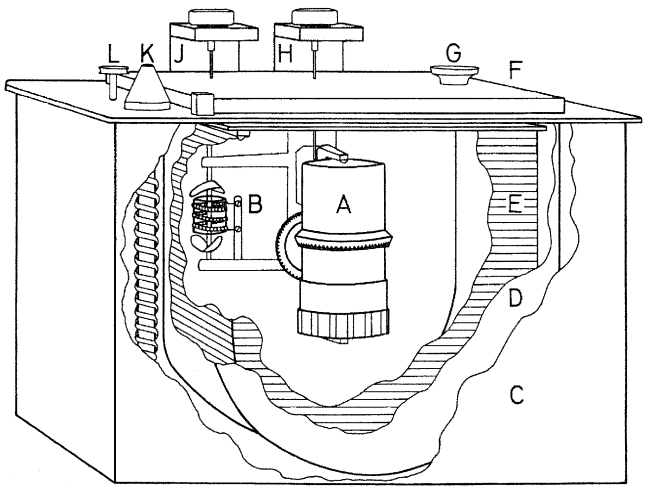


Figure 2. Cross-section of assembled isoperibol calorimeter. A, combustion bomb; B, calorimeter heater for adjusting starting temperature; C, outer surface of thermostat; D, inner surface of thermostat; E, calorimeter can; F, lid of thermostat; G, receptacle for thermometer; H, motor for rotation of bomb; J, calorimeter stirrer motor; K, thermostat heater control; L, thermostat stirrer.

expeditiously transferred to the calorimeter so that the combustion bomb, the water, and the vessel are quickly brought to the same temperature). A heater (B) is used to adjust the temperature of the calorimeter to the desired starting point of the experiment. A synchronous motor (H) can be used to rotate the bomb for special applications that are outside the scope of the present discussion. A pump (L) circulates water (thermostatted by unit K) through the jacket and lid. While the combustion experiment is in progress, the temperature of the calorimeter water is recorded as a function of time, and that of the thermostat is monitored.

The thermometer is an essential part of all calorimeters (see THERMOMETRY). In the early days of combustion calorimetry, mercury-in-glass thermometers of state-of-the-art accuracy were used. These were superceded by platinum resistance thermometers which, essentially, formed one arm of a Wheatstone (Mueller) bridge, and the change in resistance of the thermometer was determined with the help of the bridge and a galvanometer assembly. Electrical resistances of such thermometers were certified as an accurate function of temperature by national standards laboratories. Thus, a particular value of the resistance of the thermometer corresponded to a certified temperature. Thermistors have also been used as temperature sensors and, more recently, quartz-crystal thermometers. The latter instruments read individual temperatures with an accuracy of at least $\pm 1 \times 10^{-4}$ K, and give temperature differences (the quantities sought in combustion calorimetric studies) accurate to $2-3 \times 10^{-5}$ K. They display an instantaneous temperature reading that is based on the vibrational frequency of the quartz crystal. Nowadays, quartz-crystal thermometers are interfaced with computers that make it possible to record and process temperatures automatically and are, therefore, much less labor-intensive to use than their Wheatstone bridge and galvanometer predecessors.

Most ignition systems for combustion calorimetry are designed to release a pulse of electrical energy through a wire fuse from a capacitor of known capacitance charged to a preset voltage. Thus, the ignition energy introduced into the calorimeter, which usually amounts to $\sim$1 J, can be accurately calculated or preprogrammed. Overall energies of reaction must be corrected for the introduction of this "extra" energy.

As we have mentioned previously, a stirrer helps to disperse the energy released during the combustion throughout the calorimetric system. Its design is usually a compromise between efficiency and size: if the stirrer is too small, it does not dissipate the energy of reaction efficiently; if it is too bulky, an undesirable quantity of mechanical heat is introduced into the experiment. Usually, stirrer blades have a propeller design with a major axis of $\sim$1 to 2 cm in length. They are attached via a shaft to a synchronous motor to ensure that the stirring energy is constant, and also by means of a nylon (or other thermal insulator) connector to minimize the transfer of heat by this route from the calorimeter to the surroundings.

Most conventional combustion calorimeters are equipped with a heater, the purpose of which is to save time by quickly raising the temperature of the calorimeter to a value that is close to the planned starting point of the

experiment. Such heaters, typically, have resistances of $\sim 150\ \Omega$.

Outline of a Typical Combustion Calorimetric Experiment. In brief, a combustion is begun by discharging an accurately known quantity of electrical energy through a platinum wire bent in such a way (see Fig. 1) that it is close to the combustion crucible when the bomb is assembled. Usually, a short cotton thread is knotted to the wire and the free end is run beneath the material to be burned. This sparks ignition of the sample in the high-pressure O_2, and initiates a combustion reaction such as that depicted in Equation 3. Complete combustion of the sample often requires some preliminary experimentation and ingenuity. Although many organic compounds oxidize completely at $p(O_2) = 3$ MPa, some require higher, others occasionally lower, pressures, to do so. Cleaner combustions can sometimes be realized by changing, on a trial and error basis, the mass of the crucible. Some investigators have promoted complete combustion by adding an auxiliary substance to the crucible. For example, a pellet of benzoic acid placed beneath a pellet of the substance under study has often promoted complete combustion of a material that was otherwise difficult to burn. Paraffin oil has also been used for this purpose. Needless to say, allowance must be made for the contribution to the overall energy of reaction of the chosen auxiliary material.

Role of Analytical Chemistry in Combustion Calorimetry

There are, usually, two main underlying reasons for performing combustion calorimetry. One may wish to determine the energy of combustion of a not necessarily pure substance as an end in itself. An example is the measurement of the "calorific value" of a food or coal, where one simply requires a value (J/g) for the energy released when 1 g of a particular sample is burned in O_2 at a specified pressure. Here, the sample is not a pure compound and may not be homogeneous. It is usually necessary to dry a coal sample, but few, if any, analytical procedures are prerequisite.

On the other hand, the ultimate research objective may be the establishment of the enthalpy of formation of a compound, a fundamental property of the pure substance, as we have pointed out. In that case, a number of crucial operations must be an integral part of an accurate determination of the energy of combustion, and it is recommended that most of these procedures, for reasons that will become apparent, be performed even before calorimetric experiments are begun. The underlying philosophy here is governed by the first law of thermodynamics. The first law, in brief, demands exact characterization of the initial states of the reactants and final states of the products, in that they must be defined as precisely as possible in terms of identity and composition. An unavoidable consequence of these criteria is the need for extensive analytical characterization of reagents and products, which, in the experience of this writer, may require more time and effort than the calorimetric measurements themselves. Consequently, combustion processes must be as "clean" as possible, and incomplete reactions should be avoided to eliminate the time and expense required to define them. For example, in combustion calorimetric studies of a compound of iron, it is clearly preferable to produce Fe(III) only, rather than a mixture of Fe(III) and Fe(II). In dealing with organic compounds, the formation of CO_2 only, to the exclusion of CO, is a constant objective of combustion calorimetrists. Such desiderata may require preliminary research on suitable arrangements of the sample, experiments on the effects of varying the pressure of O_2, or the addition of an auxiliary kindler, as mentioned previously.

Although it may seem trivial, it is essential that the experimental substance be identified. For example, the x-ray diffraction pattern of an inorganic compound should be determined. For the same reason, melting temperatures and refractive indices of organic compounds, together with the C, H, and O contents, are used as identifiers.

Impurities above minor concentrations (mass fractions of the order of 10^{-5}) can exercise a consequential effect on the energy of combustion of a material. This is a particularly serious matter when the difference between the massic energies of combustion (J/g) of the main and impurity phases is large. In the case of organic compounds, water is a common impurity; it acts as an inert contaminant during combustions in O_2. Thus, if H_2O is present at a mass fraction level of 0.05, then, in the absence of other impurities, the experimental energy of combustion will be lower (less exothermic) than the correct value by 5 percent. For most applications, this would introduce a disastrous error in the derived enthalpy of formation.

Isomers can be another form of impurity in organic materials. Fortunately, the energy difference between the substance under investigation and its isomer may be small. In that case, a relatively large concentration of the latter may introduce little difference between the experimental and correct energies of combustion. When isomeric contamination is possible, it is essential that organic materials be examined, by methods such as gas chromatography or mass spectrometry, prior to the commencement of calorimetric measurements.

In summary then, it is vital that substances to be studied by combustion calorimetry in oxygen be analyzed as thoroughly as possible. Inorganics should be assayed for principal elements (especially when the phase diagram raises the possibility of nonstoichiometry), and traces of metals and of C, H, O, and N. In the case of organics, impurities such as H_2O and structural isomers should be sought, and elemental analyses (for C, H, O, N, etc.) should be performed as a matter of course. Fractional freezing measurements may also be used to determine contaminant levels. A further check on the purity, but not the presence, of isomers is obtained by carefully weighing the gaseous products of reaction, usually CO_2 and H_2O. This is not a trivial procedure, but has been described in detail (Rossini, 1956).

Because oxygen takes part in the combustion, its purity, too, is of importance. Gas with mole fraction $x(O_2)$=0.9999 is acceptable and can be purchased commercially. Exclusion of traces of organic gases and CO is crucial here; their participation in the combustion process introduces a source of extraneous energy that usually cannot be quantified. Efforts are also made routinely to eliminate or minimize N_2(g) because it forms HNO_3 and HNO_2 in

the combustion vessel. These acids can be assayed accurately, and are routinely sought and corrected for in carefully conducted combustion calorimetric experiments.

We have dwelt at some length on the efforts that must be made analytically to characterize calorimetric samples. From this facet of a study, one or more of the following conclusions may generally be drawn: (1) the sample can be regarded with impunity as pure; (2) trace impurities are present at such a level that they play a significant thermal role in the combustion and must be corrected for; and/or (3) major element analyses show the sample to be nonstoichiometric, or stoichiometric within the uncertainties of the analyses. By fractional freezing, recrystallization, distillation, or other methods of purification, organic substances can be freed of contaminants to the extent that none are detected by techniques such as infrared spectroscopy and gas chromatography. In such cases, the measured energy of combustion is attributed to the pure compound. It is uncommon, apart from the pure elements and specially synthesized substances such as semiconductors or glasses, to encounter inorganic materials devoid of significant levels of contamination. In applying thermochemical corrections here, a certain element of guesswork is often unavoidable. In titanium carbide (TiC), for instance, trace nitrogen is likely to be combined as TiN. But how would trace silicon be combined? As SiC or $TiSi_x$? As in this case, there are frequently no guidelines to help one make such a decision. Small amounts of an extraneous phase can be difficult to identify unequivocally. The predicament here is compounded if the massic energies of combustion of the putative phases differ significantly. In such cases, one may have no alternative but to calculate the correction for each assumed form of the contaminant, take a mean value, and adopt an uncertainty that blankets all the possibilities. Recall that an impurity correction is based not only on the amount of contaminant but on the difference between its massic energy of combustion and that of the major phase.

When major element analyses suggest that a substance is nonstoichiometric, one must be sure that this finding is consistent with phase diagram information from the literature. A stoichiometric excess of one element over the other in a binary substance could mean, for example, that a separate phase of one element is present in conjunction with the stoichiometric compound. At levels of ~0.1 mass percent, x-ray analyses (see X-RAY TECHNIQUES) may not reveal that separate phase. Microscopic analyses (OPTICAL MICROSCOPY and REFLECTED-LIGHT OPTICAL MICROSCOPY) can be helpful in such cases.

Products of combustion must be examined just as thoroughly as the substance to be oxidized. When they are gaseous, analyses by Fourier transform infrared (FTIR) spectroscopy may be adequate. If a unique solid is formed, it must be identified—by x-ray diffraction (X-RAY POWDER DIFFRACTION) and, preferably, by elemental analyses. A complete elemental analysis, preceded by x-ray diffraction examination, must be performed where two or more distinct solids are produced; thus, their identities and relative concentrations are determined.

It is wise to carry out and interpret the analytical characterization, as much as possible, before accurate determinations of the energy of combustion are begun. It would clearly be unfortunate to discover, after several weeks of calorimetric effort, that the substance being studied did not correspond to the label on the container, or that sufficient contamination was present in the sample to make the measurements worthless.

Details of Calorimetric Measurements

In practice, once the analytical characterization has been completed, a satisfactory sample arrangement devised, and the bomb assembled and charged with O_2, the remaining work essentially involves measurements, as a function of time, of the temperature change of the calorimetric system caused by the combustion. As we have already mentioned, the temperature-measuring devices of modern calorimeters are interfaced with computers, by means of which the temperature is recorded and stored automatically at preset time intervals. When a quartz-crystal thermometer is used, that time interval is, typically, ~10 sec or 100 sec.

The change in energy to be measured in an experiment, and thus the temperature change, depends on the mass and massic energy of combustion of the sample. Thus, the combustion in O_2 of 1 g of benzoic acid, whose massic energy of combustion is -26.434 kJ/g, will result in an energy release of -26.434 kJ. What will be the corresponding approximate temperature rise of the calorimeter? That depends on a quantity known as the energy equivalent of the calorimetric system, ε(calor), and we shall shortly describe its determination. Simply put, this is the quantity of energy required to increase the temperature of the calorimeter by 1 K. Thus, the combustion of 1 g of benzoic acid will increase the temperature of a conventional macro calorimeter with $\varepsilon(\text{calor}) = 13200$ J/K by ~2 K. It is worth noting that, because many of the most interesting organic materials are available in only milligram quantities, some investigators (Månsson, 1979; Mackle and O'Hare, 1963; Mueller and Schuller, 1971; Parker et al., 1975; Nagano, 2000) constructed miniature combustion calorimeters with $\varepsilon(\text{calor}) \leq 1$ kJ/K.

In principle, it is desirable to aim for an experimental temperature change of at least 1 K. However, that may be impossible in a macro system because of, e.g., the paucity of sample, and one may have to settle for a temperature change of 0.5 K or even 0.1 K. Needless to say, the smaller this quantity, the greater the scatter of the results. In short, then, the energy change of the experiment will be given by the product of the energy equivalent of the calorimetric system and the temperature change of the calorimeter. At this point, we shall discuss the determination of the latter quantity.

Figure 3 shows a typical plot of temperature against time for a combustion experiment, and three distinct regions of temperature are apparent. Those are usually called the fore period, the main period, and the after period. Observations begin at time t_i; here the calorimeter temperature (say, at $T = 297$ K) drifts slowly and uniformly upward because the temperature of the surroundings is higher ($T = 298.15$ K). When a steady drift rate has been established (fore period) and recorded, the sample is

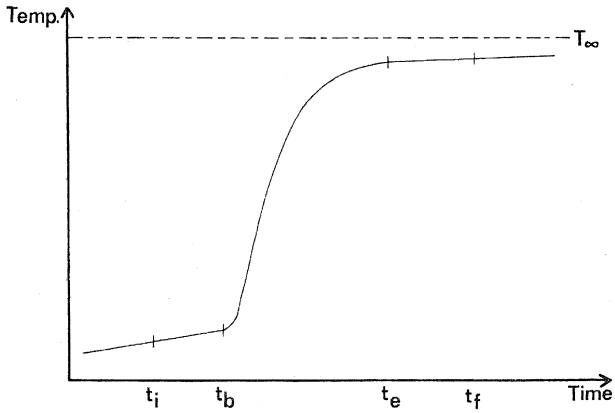


Figure 3. Typical temperature versus time curve for calorimetric experiment. t_i, initial temperature of fore-period; t_b, final temperature of fore-period; t_e, initial temperature of after period; t_f, final temperature of after period.

ignited at time t_b, whereupon the temperature rises rapidly (main period) until, at time t_e (for a representative macro combustion calorimeter, $t_e - t_b \sim 10$ min), the drift rate is once again uniform. The temperature of the calorimeter is then recorded over an after period of ~10 min, and, at t_f, the measurements are terminated. At the end of an experiment (typical duration ~0.5 hr), one has recorded the temperature at intervals of, say, 10 sec, from t_i to t_f, from which the temperature rise of the calorimeter is calculated. To be more exact, the objective is to calculate the correction that must be applied to the observed temperature rise to account for the extraneous energy supplied by stirring and the heat exchanged between the calorimeter and its environment (recall that a stirrer is used to bring the temperature of the calorimeter to a uniform state, and that there are connections between the calorimeter and the surroundings, by way of the stirrer, thermometer, ignition, and heater leads, all of which afford paths for heat leaks to the environment).

It is not within the scope of this unit to give particulars of the calculation of the correction to the temperature change. That procedure is described in detail by several authors including Hubbard et al. (1956) and Sunner (1979). Modern calorimeters, as we have mentioned, are programmed to carry out this calculation, once the time versus temperature data have been acquired.

Suffice it to say that the correction, which is based on Newton's law of cooling, can be as much as 1 percent of the observed temperature rise in isoperibol calorimeters, and therefore cannot be ignored. By contrast, adiabatic calorimeters are designed in such a way that the correction to the temperature rise is negligible (Kirklin and Domalski, 1983). Why, then, are not all combustion calorimeters of the adiabatic kind? The simple answer is that the isoperibol model is easier than the adiabatic to construct and operate.

Calibration of the Calorimetric System

Earlier in the discussion, we quickly passed over details of the determination of ε(calor), the energy equivalent of the calorimeter, defined as the amount of energy required to increase the temperature of the calorimeter by 1 K. It can also be thought of as the quantity by which the corrected temperature rise is multiplied to obtain the total energy measured during the combustion. In practice, to obtain ε(calor), one must supply a precisely known quantity of energy to the calorimeter, and, as part of the same experiment, determine the accompanying corrected temperature rise as outlined above.

The most precise determination of ε(calor) is based on the transfer to the calorimeter of an accurately measured quantity of electrical energy through a heater placed, preferably, at the same location as the combustion crucible. If the potential drop across the heater is denoted by E, the current by i, and the length of time during which the current flows by t, then the total electrical energy is given by Eit, and $\varepsilon(\text{calor}) = Eit/\Delta\theta_c$, where $\Delta\theta_c$ denotes the corrected temperature rise of the calorimeter caused by the electrical heating.

Because most laboratories that operate isoperibol calorimeters are not equipped for electrical calibration, a protocol has been adopted which, in essence, transfers this procedure from standardizing to user laboratories. Thus, standard reference material benzoic acid, whose certified energy of combustion in O_2 has been measured in an electrically calibrated calorimeter, can be purchased, e.g., from the National Institute of Standards and Technology (NIST) in the U.S.A. The certified value of the massic energy of combustion in pure O_2 of the current reference material benzoic acid (NIST Standard Reference Material SRM 39j) is -26434 J/g, when the combustion is performed within certain prescribed parameters (of pressure of O_2, for example). An accompanying certificate lists minor corrections that must be applied to the certified value when the experimental conditions deviate from those under which the energy of combustion was determined at the standardizing laboratory. It is desirable that the combustion gases be free from CO, and that soot and other byproducts not be formed. While it is possible to correct for the formation of soot and CO, for example, it is always preferable that combustions yield only those chemical entities, CO_2 and H_2O, obtained during the certification process. Thus, by simultaneous measurement of $\Delta\theta_c$, one determines ε(calor). Usually, a series of seven or eight calibration experiments is performed, and the mean value of the energy equivalent of the calorimetric system, $\langle\Delta\varepsilon(\text{calor})\rangle$, and its standard deviation, are calculated. In carefully performed calibrations with calorimeters of the highest accuracy, ε(calor) can be determined with a precision of ~0.006 percent.

It is recommended practice to examine, from time to time, the performance of combustion calorimeters. Check materials are used for that purpose; they are not standard reference materials but, rather, substances for which consistent values of the massic energies of combustion have been reported from a number of reputable laboratories. Examples include succinic acid, nicotinic acid, naphthalene, and anthracene.

Standard State Corrections in Combustion Calorimetry

So far, we have discussed the analytical characterization of the reactants and products of the combustion and the

determination of the corrected temperature rise. At this point, when, so to speak, the chemical and calorimetric parts of the experiment have been completed, we recall that we are in pursuit of the change of energy associated with the combustion of a substance in O_2, where all reactants and products are in their standard reference states at a particular temperature, usually 298.15 K. Therefore, it is essential to compute the energy difference between the state in the bomb of each species involved in the reaction and its standard reference state. If, for example, one studies the combustion of a hydrocarbon at a pressure of 3 MPa of O_2 at $T = 297.0$ K, a correction must be made that takes into account the energy change for the (hypothetical) compression of the solid or liquid from the initial $p = 3$ MPa to $p = 0.1$ MPa, the reference pressure, and to $T = 298.15$ K. Similarly, the energy change must be calculated for the expansion of product CO_2(g) from a final pressure of ~3 MPa and $T = 298.10$ K, to $p = 0$ and $T = 298.15$ K. Among other things, allowance must be made for the separation of CO_2 from the solution it forms by reaction with product H_2O and adjustment of its pressure to $p = 0.1$ MPa and temperature to $T = 298.15$ K. There are numerous such corrections. Their sum is called the standard state correction. It is beyond the scope of the present treatment to itemize them, and detail how they are calculated. This topic has been dealt with extensively by Hubbard et al. (1956) and by Månsson and Hubbard (1979).

Normally, six to eight measurements are performed. The mean value and standard deviation of the massic energy of combustion are calculated. In thermochemical work, uncertainties are conventionally expressed as twice the standard deviation of the mean (see MASS AND DENSITY MEASUREMENTS).

Combustion Calorimetry in Gases Other Than Oxygen

Several times in this unit we referred to the desirability of obtaining so-called "clean" combustions. The impression may have been given that it is always possible, by hook or by crook, to design combustion experiments in oxygen that leave no residue and form reaction products that are well defined. Unfortunately, that is not so. For example, combustions of inorganic sulfur–containing metal sulfides generally yield mixtures of oxides of the metal and of sulfur. A good example is the reaction of MoS_2 which forms a mixture of MoO_2(cr), MoO_3(cr), SO_2(g), SO_3(g) and, possibly, an ill-defined ternary (Mo, S, O). Recall that reliable calorimetric measurements require not only that the products of reaction be identified, but that they be quantified as well. Thus, a determination of the standard energy of combustion of MoS_2 in oxygen would require extensive, and expensive, analytical work. One could give additional similar examples of "refractory" behavior, such as the combustion of UN to a mixture of uranium oxides, nitrogen oxides, and N_2(g).

In such instances, O_2(g) is clearly not a powerful enough oxidant to drive the reaction to completion: MoS_2 to MoO_3 and $2SO_3$, and UN to $(1/3)U_3O_8$ and $(1/2)N_2$(g). Because of the expanding interest in the thermochemistry of inorganic materials, scientists, of necessity, began to explore more potent gaseous oxidants, among them the halogens: F_2, Cl_2, Br_2, and interhalogens such as BrF_3 and ClF_3.

Of the oxidants just listed, F_2 has been used most frequently, and it will be discussed briefly here. Its clear advantage over O_2 is illustrated by the reaction with MoS_2

$$MoS_2(cr) + 9F_2(g) \rightarrow MoF_6(g) + 2SF_6(g) \qquad (17)$$

Here, simple, well-characterized products are formed, unlike the complicated yield when O_2 is used. Comparatively little analytical characterization is required, apart from identification of the hexafluorides by infrared spectroscopy. Unfortunately, the very reactivity that makes F_2 so effective as a calorimetric reagent imposes special requirements (apart from those that arise from its toxicity), particularly with regard to materials of construction, when it is used in thermochemical studies. Combustion vessels must be constructed of nickel or Monel, and seals that are normally made of rubber or neoprene, suitable for work with oxygen, must be fashioned from Teflon, gold, lead, or other fluorine-resistant substances. Other precautions, too numerous to detail here, are listed in a recent book that deals with fluorine bomb calorimetry (Leonidov and O'Hare, 2000).

DATA ANALYSIS AND INITIAL INTERPRETATION

Computation of Enthalpies of Formation From Energies of Combustion

As we pointed out in the introduction to this article, massic energies of combustion are generally determined in order to deduce the standard molar enthalpy of formation. In this section, we will show how the calculations are performed that lead to $\Delta_f H^\circ_m$ for an organic and inorganic substance.

Ribeiro da Silva et al. (2000) reported the standard massic energy of combustion in oxygen $\Delta_c u^\circ$ of D-valine ($C_5H_{11}NO_2$) according to the following reaction

$$C_5H_{11}NO_2(cr) + (27/4)O_2(g) = 5CO_2(g) + (11/2)H_2O(l) + (1/2)N_2(g) \qquad (18)$$

$\Delta_c u^\circ = -(24956.9 \pm 10.2)\,J \cdot g^{-1}$. The molar mass of valine, based on the atomic weights from IUPAC (1996), is 117.15 $g \cdot mol^{-1}$; therefore, the standard molar energy of combustion $\Delta_c U^\circ_m$ is given by: $(0.11715) \cdot (-24956.9 \pm 10.2)\,kJ \cdot mol^{-1} = -(2923.7 \pm 1.2)\,kJ \cdot mol^{-1}$. For the combustion reaction of valine, $\Delta\nu^g \cdot R \cdot T = -3.1\ kJ \cdot mol^{-1}$ (on the basis of Equation 18), where $R = 8.31451\,J \cdot K^{-1} \cdot mol^{-1}$ and T = 298.15 K. Thus, $\Delta_c H^\circ_m = \{-(2923.7 \pm 1.2) - 3.1\}\,kJ \cdot mol^{-1} = -(2926.8 \pm 1.2)\,kJ \cdot mol^{-1}$ which, combined with the standard molar enthalpies of formation of $(11/2)H_2O$(l) and $5CO_2$(g) (CODATA, 1989), yields

$$\begin{aligned}\Delta_f H^\circ_m(C_5H_{11}NO_2, cr, 298.15\,K)/(kJ \cdot mol^{-1}) &= (11/2) \cdot (-285.83 \pm 0.04) + 5 \cdot (-393.51 \pm 0.13) \\ &\quad + (2926.8 \pm 1.2) \\ &= -(612.8 \pm 1.4)\,kJ \cdot mol^{-1}\end{aligned} \qquad (19)$$

Johnson et al. (1970) determined calorimetrically the massic energy of combustion in oxygen of a specimen of $PuC_{0.878}$, according to the following equation

$$PuC_{0.878}(cr) + 1.878O_2(g) = PuO_2(cr) + 0.878CO_2(g) \quad (20)$$

and reported $\Delta_c u^\circ = -(5412.0 \pm 10.5)\ J \cdot g^{-1}$. On the basis of the molar mass of the carbide, $M = 249.66\ g \cdot mol^{-1}$, the molar energy of combustion $\Delta_c U_m^\circ$ is calculated to be $-(1351.2 \pm 2.6)\ kJ \cdot mol^{-1}$. From Equation 20, $\Delta\nu^g \cdot R \cdot T = -2.5\ kJ \cdot mol^{-1}$ and, thence, $\Delta_c H_m^\circ = -(1353.7 \pm 2.6)\ kJ \cdot mol^{-1}$. We take $\Delta_f H_m^\circ\ (PuO_2) = -(1055.8 \pm 0.7)\ kJ \cdot mol^{-1}$ and $\Delta_f H_m^\circ(CO_2, g) = -(393.51 \pm 0.13)\ kJ \cdot mol^{-1}$ (CODATA, 1989), and calculate $\Delta_f H_m^\circ(PuC_{0.878}, Cr) = -(47.6 \pm 2.7)\ kJ \cdot mol^{-1}$.

In these examples, the original studies gave uncertainties of the $\Delta_c u^o$ results. These were, in turn, combined in quadrature by the authors with uncertainties associated with the values of $\Delta_f H_m^\circ$.

SAMPLE PREPARATION

The preparation of a sample for combustion calorimetry is governed by the desideratum that all the sample, or as much as possible of it, be reacted to completion in a well-defined combustion process.

The choice of physical form of the specimen is fairly limited: a massive piece of material (for example, a compressed pellet of an organic substance, or part of a button of an inorganic compound prepared by arc melting) should be used in initial experiments, and, if it fails to react completely, combustions of the powdered form at progressively finer consistencies should be explored. The experience of numerous combustion calorimetrists suggests, however, that organics almost always react more completely in pellet form. It is believed that most such reactions occur in the gas phase, after some material is initially vaporized from the hot zone. This behavior is also true of some inorganic substances, but others melt, however, and react in the liquid phase, in which case chemical interaction with the sample support is a distinct possibility. Even after prolonged exploration of the effect of subdivision of the material, acceptable combustions sometimes cannot be designed, at which point, as we have mentioned earlier, other variables such as sample support, auxiliary combustion aid, and gas pressure are investigated.

What the experimentalist is attempting here is to concentrate, as sharply as possible, the hot zone at the core of the combustion. That the local temperatures in a combustion zone can be substantial is demonstrated by the discovery of small spheres of tungsten, formed from the molten metal, that remained in calorimetric bombs after combustions of tungsten in fluorine to form tungsten hexafluoride (Leonidov and O'Hare, 2000).

In any case, it is important to avoid contamination during the comminution of pure materials to be used in calorimetry. Operations such as grinding should be performed with clean tools in an inert atmosphere. Needless to say, if exploratory experiments indicate the use of powder, it is that material, and not the bulk sample, that should be subjected to the exhaustive analytical characterization detailed elsewhere in this article.

PROBLEMS

The most common problems in combustion calorimetry have to do with: (1) characterization of the sample and reaction products; (2) attainment of complete combustion of the sample; (3) accurate measurement of the temperature change in an experiment; and (4) the proper functioning of the calorimeter.

It is clear from our earlier discussions that characterization of the sample and combustion products and, therefore, the problems related to it, lie entirely in the realm of analytical chemistry. Each calorimetric investigation has its own particular analytical demands, but those lie outside the scope of the present unit and will not be considered here.

The major practical problem with any combustion calorimetric study is the attainment of a complete, well-defined reaction. When burnt in oxygen, organic substances containing C, H, and O ideally form H_2O and CO_2 only. However, side-products such as carbon and CO may also be present, and their appearance indicates that the combustion did not reach an adequately high local temperature. One solution to this problem is to raise the combustion pressure, say from the typical 3 MPa to 3.5 MPa or even 4 MPa. Other solutions include lessening the mass of the combustion crucible, reducing the mass of sample, or adding an auxiliary combustion aid. The latter is usually a readily combustible material with large massic energy of combustion (numerous authors have used benzoic acid or plastic substances such as polyethylene capsules). Problems encountered with inorganic substances can also be addressed by raising or in some cases lowering the combustion gas pressure. A lower combustion pressure can moderate the reaction and prevent the melting and consequent splattering of the specimen. Complete combustions of refractory inorganic substances have been achieved by placing them on secondary supports which are themselves consumed in the reaction. In fluorine-combustion calorimetry, for example, tungsten metal serves this purpose very well, as it boosts the combustion of the sample while itself being converted to the gaseous WF_6.

It is clear that, even with well-characterized samples and products and "clean" combustions, a calorimetric experiment will not be reliable if the temperature-measuring device is not performing accurately. Thus, thermometers to be used for combustion calorimetry should be calibrated by standardizing laboratories, or, alternatively, against a thermometer that itself has been calibrated in this way, as are, for example, many quartz-crystal thermometers used in present-day studies (see THERMOMETRY). As we pointed out in the earlier part of this unit, temperature differences, not absolute temperature values, are the values obtained in calorimetric work.

Finally, we address problems related to the proper functioning of the calorimeter itself. Examples of such

problems might include the following: (1) the thermostatted bath that surrounds the calorimeter is not maintaining a constant temperature over the course of an experiment; (2) the stirrer of the calorimetric fluid is functioning erratically, so that the stirring energy is not constant (as is assumed in the calculations of the corrected temperature change of the calorimeter); or (3) the oxidizing gas is leaking through the reaction vessel gasket. All such problems can be pinpointed (as an example, by monitoring the temperature of the calorimetric jacket with a second thermometer). However, other systematic errors may not be so easily diagnosed.

Investigators traditionally check the proper functioning of a combustion calorimeter by using it to measure the energy of combustion of a "secondary reference" material. For combustions in oxygen, succinic acid is recommended. Its energy of combustion has been accurately determined with excellent agreement in numerous studies since the advent of modern calorimetry, and is thus well known. If, in check experiments, disagreement with this value (beyond the uncertainty) arises, it should be taken as an indication that a systematic error or errors are present.

LITERATURE CITED

Carson, A. S. 1979. Aneroid bomb combustion calorimetry. *In* Experimental Chemical Thermodynamics. Vol. 1. (S. Sunner and M. Månsson, eds.), Chap. 17:1. Pergamon Press, New York.

Carson, A. S. and Wilmshurst, B. R. 1971. The enthalpy of formation of mercury diphenyl and some associated bond energies. *J. Chem. Thermodyn.* 3:251–258.

CODATA Key Values for Thermodynamics. 1989. (J. D. Cox, D. D. Wagman, and V. A. Medvedev, eds.). Hemisphere Publishing Corp., New York.

Gunn, S. R. 1972. Enthalpies of formation of arsine and biarsine. *Inorg. Chem.* 11:796–799.

Holley, C. E., Jr. and Huber, E. J., Jr. 1979. Combustion calorimetry of metals and simple metallic compounds. *In* Experimental Chemical Thermodynamics. Vol. 1. (S. Sunner and M. Månsson, eds.), Chap. 10. Pergamon Press, New York.

Hubbard, W. N., Scott, D. W. and Waddington, G. 1956. Standard states and corrections for combustions in a bomb at constant volume. *In* Experimental Thermochemistry (F. D. Rossini, ed.), Chap. 10. Interscience, New York.

IUPAC. 1996. Atomic weights of the elements 1995. *Pure Appl. Chem.* 68:2339–2359.

Johnson, G. K., van Deventer, E. H., Kruger, O. L., and Hubbard, W. N. 1970. The enthalpy of formation of plutonium monocarbide. *J. Chem. Thermodyn.* 2:617–622.

King, R. C. and Armstrong, G. 1964. Heat of combustion and heat of formation of aluminum carbide. *J. Res. Natl. Bur. Stand. (U.S.)* 68A:661–668.

Kirklin, D. R. and Domalski, E. S. 1983. Enthalpy of combustion of adenine. *J. Chem. Thermodyn.* 15:941–947.

Kornilov, A. N., Ushakova, I. M., and Skuratov, S. M. 1967. Standard heat of formation of zirconium dioxide. *Zh. Fiz. Khim.* 41:200–204.

Kubaschewski, O., Alcock, C. B., and Spencer, P. J. 1993. Materials Thermochemistry. 6th ed. Pergamon Press, New York.

Leonidov, V. Ya. and O'Hare, P. A. G. 2000. Fluorine Calorimetry. Begell House, New York.

Mackle, H. and O'Hare, P. A. G. 1963. High-precision, aneroid, semimicro combustion calorimeter. *Trans. Faraday Soc.* 59:2693–2701.

Månsson, M. 1979. A 4.5 cm^3 bomb combustion calorimeter and an ampoule technique for 5 to 10 mg samples with vapour pressures below approximately 3 kPa (20 torr). *J. Chem. Thermodyn.* 5:721–732.

Månsson, M. 1979. Trends in combustion calorimetry. *In* Experimental Chemical Thermodynamics. Vol. 1. (S. Sunner and M. Månsson, eds.), Chap. 17:2. Pergamon Press, New York.

Månsson, M. and Hubbard, W. N. 1979. Strategies in the calculation of standard-state energies of combustion from the experimentally determined quantities. *In* Experimental Chemical Thermodynamics. Vol. 1. (S. Sunner, and M. Månsson, eds.), Chap. 5. Pergamon Press, New York.

Marsh, K. N. and O'Hare, P. A. G. 1994. Solution Calorimetry. Blackwell Scientific, Oxford.

Mueller, W. and Schuller, D. 1971. Differential calorimeter for the determination of combustion enthalpies of substances in the microgram region. *Ber. Bunsenges. Phys. Chem.* 75:79–81.

Nagano, Y. 2000. Micro-combustion calorimetry of coronene. *J. Chem. Thermodyn.* 32:973–977.

Parker, W., Steele, W. V., Stirling, W., and Watt, I. 1975. A high-precision aneroid static-bomb combustion calorimeter for samples of about 20 mg: The standard enthalpy of formation of bicyclo[3.3.3]undecane. *J. Chem. Thermodyn.* 7:795–802.

Ribeiro da Silva, M. A. V., Ribeiro da Silva, M. D. M. C., and Santos, L. M. N. B. F. 2000. Standard molar enthalpies of formation of crystalline L-, D- and DL-valine. *J. Chem. Thermodyn.* 32: 1037–1043.

Rossini, F. D. 1956. Calibrations of calorimeters for reactions in a flame at constant pressure. *In* Experimental Thermochemistry (F. D. Rossini, ed.), Chap. 4. Interscience, New York.

Sunner, S. 1979. Basic principles of combustion calorimetry. *In* Experimental Chemical Thermodynamics. Vol. 1. (S. Sunner, S. and M. Månsson, eds.), Chap. 2. Pergamon Press, New York.

KEY REFERENCES

Hubbard, W. N., Scott, D. W., and Waddington, G. 1959. Experimental Thermochemistry. (F. D. Rossini, ed.), Chap. 5. Interscience, New York.

This book chapter gives detailed instructions for the calculation of standard-state corrections for combustion in oxygen of organic compounds of sulfur, nitrogen, chlorine, bromine, and iodine. An excellent outline of the corrected temperature rise in a calorimetric experiment is also included.

Kubaschewski et al., 1993. See above.

Deals with the theory and practice behind the determination of thermodynamic properties of inorganic materials, and includes an appendix of numerical thermodynamic values.

Leonidov and O'Hare, 2000. See above.

Gives a comprehensive survey of the technique of fluorine combustion calorimetery, along with a completely updated (to 2000) critical evaluation of the thermodynamic properties of (mainly inorganic) substances determined by this method.

Sunner, S. and Månnson, M. (eds.). 1979. Experimental Chemical Thermodynamics, vol. 1. Pergamon Press, New York.

Contains authoritative articles on most aspects of combustion calorimetry, including the theory and practice, application to organic and inorganic substances, history, treatment of errors and uncertainties, and technological uses.

APPENDIX: COMMERCIALLY AVAILABLE COMBUSTION CALORIMETERS

We know of just two commercial concerns that market calorimeters to measure energies of combustion in oxygen. We refer to them briefly here, solely to allow interested consumers to contact the vendors.

Parr Instrument Company (Moline, Illinois, U.S.A.; *http://www.parrinst.com*) lists several isoperibol combustion calorimeters, some of which are of the semimicro kind.

IKA-Analysentechnik (Wilmington, North Carolina, U.S.A.; *http://www.ika.net*) catalogs combustion calorimeters of both isoperibol and aneroid design.

P. A. G. O'HARE
Darien, Illinois

THERMAL DIFFUSIVITY BY THE LASER FLASH TECHNIQUE

INTRODUCTION

Thermal diffusivity is a material transport property characterizing thermal phenomena significant in many engineering applications and fundamental materials studies. It is also directly related to thermal conductivity, another very important thermophysical property. The relationship between thermal diffusivity α and thermal conductivity λ is given by $\alpha = \lambda/C_p\rho$, where ρ and C_p are (respectively) density and specific heat at constant pressure of the same material. Contrary to the thermal conductivity, whose measurements involve measuring heat fluxes that are difficult to control and measure accurately, particularly at elevated temperatures, measuring thermal diffusivity basically involves the accurate recording of the temperature caused by a transient or periodic thermal disturbance at the sample boundary. It is thus frequently easier to measure thermal diffusivity than thermal conductivity. The two other properties involved in the relationship, density and specific heat, are thermodynamic properties and are either known or can be relatively easily measured.

Thermal diffusivity experiments are usually short and relatively simple. In most cases they require small samples, disks a few millimeters in diameter and less than 4 mm thick. Although the derivation of thermal diffusivity from recorded experimental data may involve complex mathematics, availability of large-capacity personal computers and the ease with which thermal diffusivity experiments can be interfaced and automated compensate for this limitation. Another important feature of thermal diffusivity methods is that the temperature variation in the sample during measurement can be quite small, so the measured property is related to an accurately known temperature. This advantage enables studies of phase transitions via thermal diffusivity throughout transition ranges, which is often not feasible with thermal conductivity measurements whose methods involve appreciable temperature gradients.

Thermal diffusivity techniques have been used in the temperature range from 4.2 to 3300 K. In general, diffusivity/specific heat techniques are not well suited for cryogenic temperatures. Diffusivity values decrease rapidly and specific heat values increase rapidly with decreasing temperature, and their product cannot be determined with great accuracy. Therefore, the most popular temperature range for thermal diffusivity measurements is from near room temperature to 2000 K. According to the shape of the temperature disturbance, diffusivity techniques may be categorized into two basic groups: the transient heat flow and the periodic heat flow techniques. Transient techniques are divided into two subgroups, depending upon whether the temperature disturbance is relatively short (pulse techniques) or long (monotonic heating). These methods are discussed in detail in several chapters of Maglic et al. (1984).

Periodic heat flow variants are based on the measurement of the attenuation or the phase shift of temperature waves propagating through the material. Periodic heat flow methods are divided into two groups. The first constitutes temperature wave techniques, which are predominantly devoted to lower and medium temperatures and are frequently called multiproperty, because they can provide data on a number of thermophysical properties within a single experiment. The second group constitutes the high-temperature variants, where the energy input is effected by modulated electron or photon beams bombarding the sample.

The laser flash technique is the most popular method. According to some statistics, 75% of thermal diffusivity data published in the 1980s were measured with this technique.

In summarizing advantages and disadvantages of the techniques, the laser flash and the wave techniques may be considered within same category. Both generally require small samples and vacuum conditions, and measurements are made within very narrow temperature intervals. The latter characteristic makes them convenient for studying structural phenomena in the materials to very high temperatures. Contrary to wave methods which need two types of apparatus for the whole temperature range, the laser flash method can cover the entire range with minor modifications in traversing from subzero to elevated temperatures. The laser flash measurements last less than 1 s and do not require very stable temperatures, while wave techniques require quasi-stationary conditions. Both methods are well studied and in both consequences of the main sources of error can be adequately compensated. There is a large volume of competent literature on this subject. Basic components of the laser flash equipment as well as complete units are commercially available from Anter Laboratories, Pittsburgh, PA, and Theta Industries, Port Washington, NY. In addition, there are well-established research and testing laboratories available for thermophysical property testing, such as TPRL, Inc., West Lafayette, IN. Temperature wave variants cover a very wide materials range and are suitable for operation under high pressures and for multiproperty measurement. Of particular advantage is the possibility of cross checking results by comparing data

derived from the amplitude decrement and the phase lag information. The wave techniques have proved particularly convenient for measuring thermal conductivity and thermal diffusivity of very thin films and deposits on the substrates. This feature is very important, as properties of very thin films typically differ from the properties of the bulk of the respective materials. The laser flash method is still establishing its place in this area.

A less attractive feature of wave techniques is that the measurements generally have to be carried out in vacuum. This limits these methods to materials in which ambient gas does not contribute to the energy transport. Mathematical procedures involved in corrections for both the laser flash and wave techniques utilize sophisticated data reduction procedures. The potential user of this equipment should be proficient in this area. Due to small samples, techniques are not well suited for coarse matrix materials where local inhomogeneities may compare with the sample thickness as well as for optically translucent or thermally transparent materials.

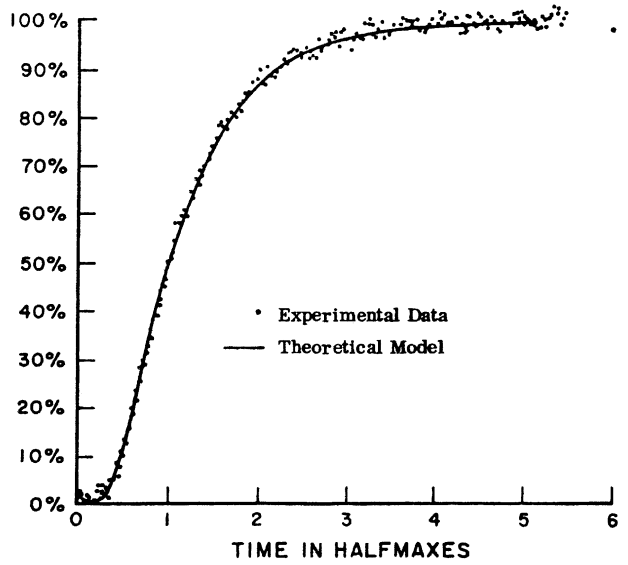


Figure 1. Comparison of normalized rear-face temperature rise with the theoretical model.

PRINCIPLES OF THE METHOD

The flash method of measuring thermal diffusivity was first described by Parker et al. (1961). Its concept is based on deriving thermal diffusivity from the thermal response of the rear side of an adiabatically insulated infinite plate whose front side was exposed to a short pulse of radiant energy. The resulting temperature rise of the rear surface of the sample is measured, and thermal diffusivity values are computed from the temperature-rise-versus-time data. The physical model assumes the following ideal boundary and initial conditions: (a) infinitely short pulse, (b) one-dimensional heat flow normal to the plate face, (c) adiabatic conditions on both plate faces, (d) uniform energy distribution over the pulse, (e) pulse absorption in a very thin layer of investigated material, (f) homogeneous material, and (g) relevant materials properties constant within the range of disturbance. As shown in the Appendix, the last two assumptions reduce the general heat diffusion equation to the form

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \tag{1}$$

where the parameter α represents the thermal diffusivity of the plate material. The solution of this equation relates the thermal diffusivity of the sample to any percent temperature rise and the square of the sample thickness (L):

$$\alpha = K_x \frac{L^2}{t_x} \tag{2}$$

where K_x is a constant corresponding to an x percent rise and t_x is the elapsed time to an x percent rise. Table 2 in the Appendix relates values of K calculated for specific x percent rises.

Parker et al. (1961) selected the point corresponding to 50% of the temperature rise to its maximum value, $t_{0.5}$, relating thermal diffusivity of the material to the square of the plate thickness, and the time needed for the rear temperature to reach 50% of its maximum:

$$\alpha = 0.1388 \frac{L^2}{t_{0.5}} \tag{3}$$

Clark and Taylor (1975) showed that the rear-face temperature rise curve could be normalized and all experimental data could be immediately compared to the theoretical (Carslaw and Jaeger, 1959) solution on-line. Figure 1 shows this comparison and Table 1 gives the corresponding diffusivity (α) values for various percent rises along with the elapsed time to reach those percentages. The calculated values are all 0.837 $\pm$ 0.010 cm^2/s, even though the calculations involved times that varied by almost a factor of 3. Thermal diffusivity experiments using the energy flash technique were one of the first to utilize rapid data acquisition to yield thermal transport data on small, simple-shaped specimens. The original techniques used a flash lamp, but the invention of the laser greatly improved the method by increasing the distance between the heat source and sample, thus permitting measurements at high temperature and under vacuum conditions.

Table 1. Computer Output For Diffusivity Experiment[a]

α (cm^2/s)	Rise (%)	Value (V)	Time (s)
0.8292	20	2.57695	0.023793
0.8315	25	2.75369	0.026113
0.8313	30	2.93043	0.028589
0.8374	33.3	3.04826	0.029275
0.8291	40	3.28391	0.033008
0.8347	50	3.63739	0.038935
0.8416	60	3.99086	0.041497
0.8466	66.7	4.22652	0.049997
0.8304	70	4.34434	0.054108
0.8451	75	4.52108	0.058326
0.8389	80	4.69782	0.065112

[a] Maximum, 5.40477 V; half maximum, 3.63739 V; baseline, 1.870 V.

The flash technique did not remain limited to conditions prescribed by the ideal model. The theoretical and experimental work of many researchers continuing from the 1960s supplemented the original concept with corrections, which accounted for heat exchange between sample and ambient, finite pulse shape and time, nonuniform heating, and in-depth absorption of the laser pulse. The result was extension of the method to, e.g., high temperatures, thin samples, layered structures, dispersed composites, and semitransparent or semiopaque materials. The review of these developments is presented elsewhere (Taylor and Maglic, 1984; Maglic and Taylor, 1992), supplemented here in further sections below with information on the literature published during the 1990s.

The capabilities of modern data acquisition and data reduction systems offer advantages over the direct approach, i.e., procedures limited to the analysis of the half-rise time or a discrete number of points. The inverse method relies on the complete transient response for thermal diffusivity measurement. This approach results in excellent agreement between theoretical and experimental curves. The parameter estimation procedure is convenient for the flash method, as it enables simultaneous determination of more than one parameter from the same temperature response. The capabilities and advantages of such determinations will be presented in Data Analysis and Initial Interpretation.

PRACTICAL ASPECTS OF THE METHOD

Implementation of the laser flash method for thermal diffusivity measurement requires the following (see Fig. 2): in a sample holder, a furnace or a cooler capable of maintaining and recording the sample reference temperature; a vacuum-tight enclosure equipped with two windows for the laser and the detector; a pulse laser with characteristics adequate for the range of materials to be studied, including available sample diameters and thicknesses; a system for measuring and recording the rear-face temperature transient; a power supply for the furnace or the cooling unit; an adequate vacuum system; and a computer for controlling the experiment, data acquisition, and subsequent data processing.

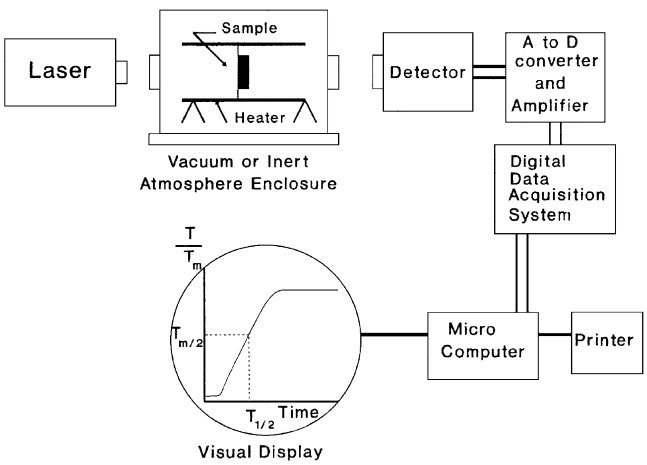


Figure 2. Schematic view of the laser flash apparatus.

Measurement times of less than 1 s are often involved. The ambient temperature is controlled with a small furnace tube. The flash method is shown schematically in Figure 2, which includes the temperature response of the rear face of the sample. This rear-face temperature rise is typically 1 to 2 K. The apparatus consists of a laser, a high-vacuum system including bell jar with windows for viewing the sample, a heater surrounding a sample holding assembly, an infrared (IR) detector, appropriate biasing circuits, amplifiers, analog-to-digital (A/D) converters, crystal clocks, and a computer-based digital data acquisition system capable of accurately taking data in the 100-μs time domain. The computer controls the experiments, collects the data, calculates the results, and compares the raw data with the theoretical model. The method is based on the Carslaw and Jaeger (1959) solution of the heat conduction equation for such a case.

The furnace should be of low thermal inertia and equipped with a programmable high-stability power supply to enable quick changes in reference temperature. Typical examples are electroresistive furnaces heated by direct passage of electrical current through a thin metallic foil or a graphite tube.

Vacuum- and/or gas-tight enclosures are mandatory for preventing heat exchange between sample and ambient by convection and gas conduction and protecting furnace and sample from chemical damage. They must have two windows along the optical axis of the sample, the front allowing entrance of the laser pulse and the rear for optical access to the sample rear side. The windows should be protected with rotating shields against evaporated deposits, as the vapor from both the heater and the sample material can condense on the cold window surfaces, reducing or completely blocking optical access. The transmittance of the rear window should be high within the infrared optical detector bandwidth.

The pulse laser [now most commonly neodymium/yttrium aluminum garnet (Nd:YAG)] should be capable of supplying pulses preferably lasting less than 1 ms and of 30 to 40 J in energy with a beam diameter of 16 mm and an energy distribution as homogeneous as possible. Usual sample diameters for homogeneous materials range from 6 to 12 mm, but some applications require diameters as large as 16 mm. Studies of thin films or samples where half times are of the order of 1 ms require much shorter pulses and *Q*-switched laser operation.

Systems for measuring reference temperature might be contact or contactless. For most metallic samples, miniature thermocouples spot welded to the specimen rear face are most convenient. For samples where this is not feasible, it is common to use a thermocouple accommodated in the sample holder or to provide a blackbody-like hole in the holder for measurement using an optical pyrometer. In both latter cases the sample temperature must be calibrated against the measured sample holder temperature.

The rear-face temperature transients should be detected with optical detectors adequate for the corresponding temperature ranges, meaning that they

should be much faster than the change they are recording and be sufficiently linear within small temperature excursions. Optical detectors should be interfaced to the computer via a suitable A/D converter.

The power supply should be matched with the furnace characteristics to easily and rapidly cover the whole temperature range of measurement. It should be capable of maintaining the prescribed temperature during the period of a few seconds of measurement for the baseline recording and up to ten half times after the laser discharge. The measurement should be effected at constant temperature or in the presence of a small and constant temperature gradient that can be compensated for in the data processing procedure. For realizing temperatures below ambient, a miniature chamber cooled by circulation of fluid from an outside cooling unit is adequate.

Thermal diffusivity values are calculated from the analysis of the temperature-time dependence of the rear face of a thin sample, whose front face has been exposed to a pulse of radiant energy. The duration of the pulse originally had to be approximately one-hundredth of the time needed for the temperature of the rear face to reach 50% of its maximum value. However, present methods for correcting deviations from the initial and boundary conditions, along with data processing possibilities, have reduced this requirement considerably. When the samples are small and homogeneous, lasers are the best means for providing energy pulses. When high accuracy is not the primary requirement and where samples have to be larger, flash lamps may sometimes be adequate, particularly at low temperatures. After establishing that the heat flow from the front to the rear sample face is sufficiently unidirectional, the accurate measurement of the rear-face temperature change is actual. Originally thermocouples were the most common device used for this purpose, but IR and other contactless detectors now provide far better service because of increased accuracy and reliability. Distortion of signals caused by contact temperature detectors may lead to significant errors in the measured thermal diffusivity. The laser flash technique has been used from 100 to 3300 K, but its most popular application has been from near room temperature to 2000 K. This technique has been successfully applied to metals in the solid and liquid states, ceramics, graphites, biological products, and many other materials with diffusivities in the range 10^{-7} to 10^{-3} m^2/s. The method is primarily applicable to homogeneous materials but has been applied successfully to certain heterogeneous materials. Even with manually driven hardware, the whole range from room temperature to the maximum temperature can be traversed within one day. With more sophisticated equipment, laser flash thermal diffusivity measurements can be quite fast.

Methods based on monotonic heating have been used from the lowest temperature limit of 4.2 K to as high as 3000 K. They proved convenient for a range of materials: ceramics, plastics, composites, and thermal insulators with thermal diffusivity values falling in the range from 10^{-8} to 10^{-5} m^2/s. They are irreplaceable for coarse-matrix and large-grain materials and porous materials in which part of thermal energy is transported via gas-filled interstices. They are also useful for property measurements over large temperature intervals carried out in relatively short times, e.g., on materials undergoing chemical or structural changes during heating. The cost is reduced accuracy, but it is usually adequate for industrial purposes. The monotonic heating techniques comprise two subgroups, for measurements in the narrow and in very wide temperature intervals, so-called regular regime and quasi-stationary regime methods.

The temperature wave variants may be applied from 60 to 1300 K, but they are generally used between 300 and 1300 K. They are convenient for metals and nonmetals and also fluids and liquid metals with thermal diffusivity in the range 10^{-7} to 10^{-4} m^2/s. This technique has a large variety of variants and their modifications, depending on the sample geometry and direction of propagation of the temperature waves. They include information on the mean temperature field and the amplitude and phase of the temperature waves. This opens the possibility of multiproperty measurements, i.e., simultaneous measurement of thermal conductivity, thermal diffusivity, and specific heat on the same sample. The high-temperature variants (modulated electron beam or modulated light beam) can be used from 330 to 3200 K, but their usual range is between 1100 and 2200 K. These high-temperature variants have been applied to refractory metals in the solid and in the molten state. They also have been used for high melting metal oxides, whose thermal diffusivity is in the range 10^{-7} to 5×10^{-5} m^2/s. Improvements in techniques, measurement equipment, and computing analysis are continuing to be made.

The measurement consists of heating or cooling the sample to the desired temperature, firing the laser, digitizing and recording the data, and possibly repeating the measurements for statistical purposes before changing to the next temperature level. The main difficulties that can be encountered include (1) deviations from assumed ideal conditions due to time of passage of the heat pulse through the sample being comparable with the pulse duration, heat exchange occurring between sample and environment, nonuniform heating of the sample surface, subsurface absorption of the laser pulse, and combined mechanisms of heat transmission through the sample; (2) interference in the transient response signal due to passage of the laser light between sample and sample holder directly to the IR detector and thermal radiation from the sample holder; (3) changes of sample chemical composition due to evaporation of alloying constituents at high temperatures and chemical interaction with the sample holder or optical barrier; and (4) errors in determining sample reference temperature.

Deviations from assumed ideal conditions manifest in the shape of the transient response, each of them deforming it in a specific manner. Procedures for correcting all deviations have been developed and are a part of standard data processing. They are included in Data Analysis and Initial Interpretation.

Transient response is also affected by the laser light passing between the sample and the sample holder or reflected laser light reaching the optical sensor. If either of these occur, the sensor may saturate, deforming the initial portion of the signal rise, thus making it difficult

to define the baseline. Precautions therefore should be taken to eliminate detector saturation.

If the thermal diffusivity of the sample holder is higher than that of the sample or the holder is partially translucent to the laser light and if stray laser pulse light hits the holder, then the optical detector could measure contributions from both the sample and its holder. This will deform the transient curve. To avoid this, the proper choice of the sample holder material and the careful design of its dimensions are necessary.

Operation at high temperatures may change the chemical composition of the sample due to the preferential evaporation of some alloying constituents. A convenient way to reduce this effect is to shorten the time that the sample is exposed to high temperatures. This is accomplished either by devoting one sample solely to high-temperature measurements and performing these as quickly as possible or by reducing the number of experiments at temperatures beyond the thermal degradation point. In addition, the sample holder material should be selected so that chemical interaction between it and the sample is minimized.

Errors in determining sample reference temperature can be significant depending upon many factors. These include the location of the temperature sensor with respect to the sample or its holder, the quality of thermal contact between the sample and the holder, and the measurement temperature and sample emissivity. It may be possible to avoid some of the errors by proper selection of the method and by establishing the relationship between the true sample temperature and that measured in the holder.

METHOD AUTOMATION

Powerful personal computers are presently available at reasonable cost. This offers ample opportunity for automation of all measurement procedures contributing to productivity and ease with which measurements are performed. Excessive automation, however, might relegate decisions to the computer that are too subtle to be automated. The following paragraph therefore reviews tasks that can be entrusted to modern computers and those that should be reserved for the operators who are knowledgeable about the limitations of the experimental technique and the physics of the properties being measured.

Desirable automation involves continuous checking of experimental conditions. This include achieving and maintaining the required level of vacuum and temperature as well as controlling the heating rates between specified temperature levels. Computers are capable of executing all these steps, as well as positioning samples when a multisample accessory is involved. Computers should also be involved in performing tasks such as firing the laser and collecting and processing the data. However, all significant phases of experiment and processing of the data should be performed by the operator. He or she should first ensure that prescribed temperature stability has been achieved. After the laser has fired and the initial data set is collected, the quality of the data should be inspected, making sure that the moment of laser discharge is recognizable, the shape of transient agrees with expectation, and possible noise on the signal will not affect the data processing. As the processing progresses, visual insight into transformation of initial experimental data is necessary before proceeding to each successive step of the procedure.

Full automation of the experiment, where the choice of accepting or rejecting the initial data set and selection of correcting procedures to be involved are left to computer, makes the experiment susceptible to various sources of systematic error. If nothing else, suspicion of their existence ought always be present. For reliable results in flash diffusivity measurements, the operator should be highly experienced.

DATA ANALYSIS AND INITIAL INTERPRETATION

After the sample has been brought to a desired temperature and its stability established, the experiment is started by initiating the program that records the baseline, fires the laser, and acquires and records data on the rear-face temperature transient. Insight into the shape of the resulting data curve indicates whether it is acceptable for extracting results or the experiment should be repeated. To increase the statistical weight of the data points, several measurements should be taken at the same temperature.

Experimental data necessary for the calculation of thermal diffusivity include baseline, which represents the equilibrium temperature of the specimen prior to laser discharge, the time mark of the laser discharge, and the transient temperature curve, extending over at least ten lengths of the characteristic half time, $t_{0.5}$. It is also useful to have a working knowledge of the shape of the laser pulse (especially its duration and intensity distribution across the beam) and the response characteristics of the temperature detector (response time, linearity). The sample thickness measurement should be done as precisely as possible prior to the diffusivity measurement.

It is important that the transient response curve be thoroughly analyzed to verify the presence or absence of laser pulse duration comparable with the time of passage of the temperature disturbance through the material (finite-pulse-time effect), heat exchange between sample and environment (heat losses or gains), nonuniform heating, subsurface heating by the laser beam, or other effects that can cause errors in thermal diffusivity calculations. This means comparing the normalized rise curve with the ideal curve for the theoretical model.

It should be noted that these four effects can be regarded merely as deviations from an ideal situation in which such effects are assumed to be negligible. It is entirely feasible to develop models that incorporate and account for these effects. This has been done for all four of these effects. Once the mathematical (analytical or numerical) expression for temperature rise is known, parameter estimation techniques can be used to calculate the thermal diffusivity of the sample. However, other unknown model parameters such as heat loss coefficients, duration and shape of the heat pulse, spatial distribution characteristics of the pulse, and effective penetration depth of the pulse must be estimated simultaneously. Even

today, when advanced parameter estimation techniques are available and high-speed, high-capacity computers can be used, estimating all parameters simultaneously is sometimes quite difficult. Careful and nontrivial mathematical analysis of so-called sensitivity coefficients [see Beck and Arnold (1977) for details] has to be conducted to find out, if it is possible to calculate all of these unknown parameters from the given model for the temperature response.

We will briefly describe how to analyze the response curve with regard to all of the above-mentioned effects.

To normalize the experimental curve and as an inherent part of determining the half time (or the time to reach any other percentage temperature rise), it is necessary to determine accurately the temperature baseline and maximum temperature rise. It is not a trivial task, especially when the signal is noisy and the temperature of the sample was not perfectly stable. Standard smoothing procedures, e.g., cubic-spline smoothing, can be used. The baseline temperature can be obtained conveniently by measuring the detector response for a known period prior to charging the laser flash tube capacitor bank and extrapolating the results to the time interval in which the rear-face temperature is rising.

The normalized experimental temperature rise (in which the baseline temperature is set to zero, the maximum temperature is equal to 1, and the time scale unit is set to the experimental half time) is then compared with the ideal dimensionless temperature rise, given by

$$V(t') = 1 + 2\sum_{n=1}^{\infty}(-1)^n \exp(-0.1388 n^2 \pi^2 t') \qquad (4)$$

where $t' = t/t_{0.5}$ is dimensionless time.

When there are no evident deviations of the experimental normalized temperature rise curve from the ideal one (analysis of residuals is a useful tool for this purpose), then any of the data reduction methods based on the simple ideal model can be used for the thermal diffusivity calculations. The result should be later checked by the other correction procedures, which should give the same answers.

The presence of the finite-pulse-time effect (without other effects being present) can be readily determined from a comparison of the normalized experimental curve to the theoretical model. The distinguishing features are that (1) the experimental curve lags the theoretical ideal curve by from an ~5% to a 50% rise, (2) the experimental curve leads the ideal theoretical curve by from an ~59% to an ~98% rise, and (3) a long, flat maximum is observed.

Of these major effects, the finite-pulse-time effect is the easiest to handle. Cape and Lehman (1963) developed general mathematical expressions for including pulse time effects. Taylor and Clark (1974) tested the Cape and Lehman expressions experimentally. Larson and Koyama (1968) presented experimental results for a particular experimental pulse characteristic of their flash tube. Heckman (1976) generated tabular values for triangular shaped pulses. Taylor and Clark (1974) showed how calculated diffusivity varied with percent rise for a triangular shaped heat pulse. They also showed how to correct diffusivities at percent rises other than the half-rise value so that the diffusivity values could be calculated over the entire experimental curve rather than at one point.

A procedure based on the Laplace transform was used to correct the finite-pulse-time effect in the data reduction method proposed by Gembarovic and Taylor (1994a). Experimental data are first transformed with the Laplace transformation and then fitted with the transform of the theoretical time-domain relation. More realistic heat pulse shapes can be calculated with this method.

Using data reduction method based on discrete Fourier transformation can eliminate errors from a noisy signal and unstable baseline conditions (Gembarovic and Taylor, 1994b).

Azumi and Takahashi (1981) proposed one of the simplest and universal methods for the correction of the finite-pulse-time effect. This correction consists of taking as the time origin the effective irradiation time (center of gravity of the pulse). Degiovanni (1987) used this technique to correct simultaneously both the finite-pulse-time and heat loss effects.

The presence of heat losses is shown by the following features: (1) the experimental curve slightly lags the theoretical curve from an ~5% to a 50% rise, (2) the experimental curve leads the theoretical curve from 50% to 100%, and (3) a relatively short maximum is observed followed by a pronounced smooth decline.

The calculated value of thermal diffusivity increases with increasing percent rise at an increasing rate.

Radiation heat losses may be corrected for by the method of Cowan (1963) or Heckman (1976). Cowan's method involves determining the values of the normalized temperature rise at $5t_{0.5}$ or $10t_{0.5}$. From these values one can estimate the radiation loss parameter and correct α.

Heat losses can also be corrected by the so-called ratio method (Clark and Taylor, 1975). In this method, the experimental data are compared at several particular points with the theoretical rear-face temperature rise with heat losses. The Clark and Taylor method has the advantage of using the data collected during the initial rise rather than the cooling data, which are subject to greater uncertainties caused by conduction to the sample holder. Clark and Taylor tested their procedure under severe conditions and showed that corrections of 50% could be made satisfactorily.

The data reduction method proposed by Degiovanni and Laurent (1986) uses the temporal moments of order zero and -1 of the defined temperature interval of the rising part of the experimental curve to correct the heat loss effect.

Koski (1982) proposed an integrated data reduction method in which both the finite-pulse-time effect and heat losses are corrected using the form

$$V(L,t) = 2\sum_{m=1}^{M} F(\tau_m) \sum_{n=1}^{\infty} \frac{\beta_n(\beta_n \cos\beta_n + L_g \sin\beta_n)}{\beta_n^2 + L_g^2 + 2L_g} \times \exp\left[\frac{-\beta_n^2 \alpha(t-\tau_m)}{L^2}\right] \Delta\tau_m \qquad (5)$$

where the pulse time τ is divided into M subintervals τ_1, τ_2, ..., τ_M, the function F represents the heat pulse shape at the time τ_m, $\Delta\tau_m = \tau_m - \tau_{m-1}$, L_g is the heat loss parameter, and β_n are roots of the transcendental equation

$$(\beta^2 - L_g)\tan\beta = 2L_g\beta \tag{6}$$

Other original methods (Balageas, 1982; Vozar et al., 1991a,b) are based on the assumption that the experimental temperature rise is less perturbed by heat losses at times closer to the origin (i.e., time of the flash). The thermal diffusivity is obtained by extrapolating the time evolution of calculated values of an apparent diffusivity to zero time.

Nonlinear least-squares fitting procedures have been developed (Takahashi et al., 1988; Gembarovic et al., 1990) in which all experimental points of the experimental temperature rise can be used for the determination of the thermal diffusivity. The methods are particularly useful in the case of noisy data, which are otherwise close to an ideal model.

A data reduction method for heat loss correction is described by Beck and Dinwiddie (1997). A parameter estimation technique was used to calculate thermal diffusivity in the case when heat loss parameters from the front and rear faces of the sample are different. Finite-pulse-time effects and radiative heat losses only occur in selective cases; i.e., finite-pulse-time effects occur with thin samples of high-diffusivity materials and radiative heat losses occur at high temperatures with thicker samples. In contrast, nonuniform heating can occur during any flash diffusivity experiment. However, very little has been published on the effects of nonuniform heating. Beedham and Dalrymple (1970), Mackay and Schriempf (1976), and Taylor (1975) have described the results for certain nonuniformities. These results show that when the heating is uniform over the central portion of the sample, reasonably good results can be obtained. However, a continuous nonuniformity over the central portion can lead to errors of at least ten times the usual error. Nonuniform heating can be corrected by properly designing the experiment. If the laser beam is nonhomogeneous in cross-section, then an optical system that homogenizes the beam can be used, the sample surface can be covered with an absorbative layer that will homogenize the beam, or a thicker sample that is less prone to this effect can be used.

If the sample is not completely opaque to the laser beam, then subsurface heating can distort experimental temperature rise. The presence of a spike and shift of baseline temperature to a new higher level after the laser flash indicates when the beam is completely penetrating the sample. It is more difficult to detect the case where the laser beam is absorbed in a surface layer of a finite thickness. It is not recommended to eliminate this effect using mathematical means. Covering the sample surface with one or more protective layers can eliminate this effect.

Data reduction procedures for layered structures are based on two- or three-layer models (see, e.g. Taylor et al., 1978). Thicknesses, densities, and specific heats of all layers have to be known, along with the thermal diffusivities of all but the measured layer. Parameter estimation techniques or other nonlinear fitting procedures are used to calculate the desired value of the thermal diffusivity of the measured layer. If the known layers are relatively thin and highly conductive and the contact thermal resistance is low, then the results of the two- or three-layer thermal diffusivity calculation are the same as for a homogeneous (one-layer) sample with the thickness given as a sum of all layers. Application of the laser flash method to very thin, highly conductive multilayer structures still remains an unsolved problem and a big challenge to experimenters.

SAMPLE PREPARATION

Preparing opaque specimens for flash diffusivity measurement is generally simple. The diameter of the sample has to conform to the size of the most homogeneous part of the laser energy pulse, and its thickness to the permissible ratio between the laser pulse duration and the characteristic time of heat passage through the specimen. Problems might arise from satisfying the requirement of plane-parallelism of the specimen flat sides. If the sample material is magnetic or if its length is sufficient to be held during machining and the material is readily machinable, no serious problem exists. However, if it has to be made thin or very thin or the material is hard and brittle or difficult to be fixed to a substrate, a lot of ingenuity on the part of the sample manufacturer will be necessary. Coping with its transparency to thermal radiation or the laser pulse may be difficult. Often it can be solved by coating the front or both sample faces with a thin metallic or graphite layer. This overlayer has to absorb the laser pulse energy within its finite thickness and convert it into a single thermal function. Layers of refractory metals are thinner and longer lasting, but their high reflectivity allows only a small portion of the pulse to be absorbed. Graphite is much better in this respect, but laser pulses, particularly those with higher energies, tend to evaporate the layer. Attaching the coating to the sample may be a real challenge, particularly if the sample material is smooth and slippery. The lifetime of a layer in terms of pulses is shorter as temperature increases.

As machining of metallic samples always involves mechanical deformation of the sample material, it is advisable to relieve strains (which may affect the measured diffusivity values) by methods well known to metallurgists.

SPECIMEN MODIFICATION

Experimenting with the flash diffusivity technique may affect the specimen in a few ways. The most obvious include modifications due to exposure to elevated temperature in vacuum, contamination from the coating layer material, and damage caused by the laser pulse action due to structural changes caused by fast heating or cooling.

Often repeated cycling of experiments will not affect the outcome of measurements, but this will depend on the material and the maximum temperature reached. If

the specimen structure is in a state that is susceptible to thermal treatment, temperature cycling will definitely involve specimen modification. If the specimen material is an alloy whose alloying components preferentially evaporate at temperatures much below its melting point, experiments above this temperature will lead to undesirable modifications of specimen composition. A vacuum environment will definitely stimulate this process. How big the damage will be and how much it will affect the outcome of the flash diffusivity results will depend on the specimen diameter-to-thickness ratio, the maximum operating temperature, and the length of time that the specimen is exposed to undesirable temperatures as well as the relative vapor pressures of the constituents.

For flash diffusivity measurements, specimens of materials that are partly or totally transparent or translucent must be coated. If the sample is porous, colloidal graphite from the coating may penetrate the specimen structure, evidenced by a change of its color, and will most likely influence its overall thermal properties. Such a change of the sample's thermal optical properties, however, does not necessarily preclude the measurement of thermal diffusivity by this technique. Within the small specimen temperature excursion caused by the laser pulse, a small amount of graphite within grain interstices will not affect the basic mechanisms governing energy transport through it.

Although the powerful laser energy pulse might cause damage to the specimen surface, in testing common materials like metals, graphites, and ceramics and with energy densities of $\sim$10 J/cm^2, adverse effects of specimen modification have not been observed. More danger lies in defining the depth of the energy pulse absorption in the case of rough surfaces typical of, e.g., composite materials, as this affects the basic geometry parameter L, which enters as a squared term in Equation 3.

PROBLEMS

Some of the most common problems are discussed below:

1. Optimum half-rise time is 40 to 100 ms. Half times are controlled by sample thickness and diffusivity value. Longer times have larger heat loss corrections. Shorter times have finite-pulse-time effects and greater uncertainty in baseline values.
2. Very thin samples have large uncertainty in sample thickness (which enters as a squared term), larger surface damage effects (sample preparation), and possibly too large a temperature rise (nonlinear IR detector response). Also, thin samples may not be sufficiently homogeneous.
3. Rear-face temperature rise may be too large (resulting in nonlinear IR response and laser damage) or too small (resulting in noisy signal). Sample emissivity and laser power should be controlled to change the energy absorbed and linearity of the IR detector response. Heat-resistive paints can be effectively used to increase sample resistivity in the case of a small signal.
4. A nonuniform laser beam can be a major problem. The uniformity should be checked using laser foot print paper with a partially absorbing solution to reduce laser power to the paper. Copper sulfate solution is very good for this purpose (for Nd:YAG primary frequency). If the beam is nonuniform, adjusting the dielectric mirrors or using optics may improve homogeneity.
5. Scattered radiation can cause spurious signals and even temporarily saturate the IR detector, making baseline determinations difficult.
6. Applying a coating to translucent samples may result in a thermal contact resistance. Thermal contact resistance may lower the value of the measured thermal diffusivity of the sample material. Coatings applied to high-diffusivity samples are especially prone to this problem.
7. Diffusivity values for porous materials can be strongly affected by the surrounding gas and its moisture content. The diffusivity values for gases are very large, even though their conductivity values are quite small.

The laser flash technique is an ASTM (1993) standard (E-1461) and the step-by-step procedures are given there. This standard is easily obtained and cannot be duplicated here. Also included in ASTM E1461-92 is a discussion of the measurement errors and the "nonmeasurement" errors. The latter arise from the nonobeyance of the initial and boundary conditions used in the data analysis. In general, the nonmeasurement errors cause greater uncertainty in the results than measurement errors, which simply involve length and time measurements. Both of these can generally be measured to a high degree of accuracy. Heat losses and finite-pulse-time effects have been studied intensively and procedures for correcting these are generally adequate. Nonuniform heating is more insidious since there is an infinite variety of nonuniformities possible and the nonuniformity can change with time as room temperature varies and the operating characteristics of the laser change. Testing with an acceptable standard such as a fine-grain, isotropic graphite AXM5Q (Hust, 1984; ASTM, 1993) is useful, but the errors in the actual tests may be significantly different due to, e.g., sample translucency, significantly different emissivity, and rise times.

LITERATURE CITED

American Society for Testing and Materials (ASTM). 1993. Standard Test Method for Thermal Diffusivity of Solids by the Flash Method, E1461–92. ASTM, Philadelphia, PA.

Azumi, T. and Takahashi, Y. 1981. *Rev. Sci. Instrum.* 52:1411–1413.

Balageas, D. L. 1982. *Rev. Phys. Appl.* 17:227–237.

Beck, J. V. and Arnold, K. J. 1977. Parameter Estimation in Engineering and Science. Wiley, New York.

Beck, J. V. and Dinwiddie, R. B. 1997. Parameter estimation method for flash thermal diffusivity with two different heat transfer coefficients *In* Thermal Conductivity 23 (K. E. Wilkes,

R. B. Dinwiddie, and R. S. Graves, eds.). pp. 107–118. Technomic, Lancaster.

Beedham, K. and Dalrymple, I. P. 1970. The measurement of thermal diffusivity by the flash method. An investigation into errors arising from the boundary conditions. *Rev. Int. Hautes Temp. Refract.* 7:278–283.

Cape, J. A. and Lehman, G. W. 1963. Temperature and finite pulse-time effects in the flash method for measuring thermal diffusivity. *J. Appl. Phys.* 34:1909.

Carslaw, H. S. and Jaeger, J. C. 1959. Conduction of Heat in Solids, 2nd ed. Oxford University Press, Oxford.

Clark III, L. M. and Taylor, R. E. 1975. Radiation loss in the flash method for thermal diffusivity. *J. Appl. Phys.* 46:714.

Cowan, R. D. 1963. Pulse method of measuring thermal diffusivity at high temperatures. *J. Appl. Phys.* 34:926.

Degiovanni, A. 1987. *Int. J. Heat Mass Transfer* 30:2199–2200.

Degiovanni, A. and Laurent, M. 1986. *Rev. Phys. Appl.* 21:229–237.

Gembarovic, J. and Taylor, R. E. 1994a. A new technique for data reduction in the laser flash method for the measurement of thermal diffusivity. *High Temp. High Pressures* 26:59–65.

Gembarovic, J. and Taylor, R. E. 1994b. A new data reduction in the laser flash method for the measurement of thermal diffusivity. *Rev. Sci. Instrum.* 65:3535–3539.

Gembarovic, J., Vozar, L., and Majernik, V. 1990. Using the least square method for data reduction in the flash method. *Int. J. Heat Mass Transfer* 33:1563–1565.

Heckman, R. C. 1976. Error analysis of the flash thermal diffusivity technique. *In* Proceedings of the Fourteenth International Thermal Conductivity Conference, Vol. 14 (P. G. Klemens and T. K. Chu, eds.). Plenum Press, New York.

Hust, J. G. 1984. Standard reference materials: A fine-grained, isotropic graphite for use as NBS thermophysical property RM's from 5 to 2500 K. *NBS Special Publication* 260–289.

Koski, J. A. 1982. Improved data reduction methods for laser pulse diffusivity determination with the use of microcomputers *In* Proceedings of the Eighth Symposium on Thermophysical Properties (A. Cezairliyan, ed., Vol. II), pp. 94–103. American Society of Mechanical Engineers, New York.

Larson, K. B. and Koyama, K. 1968. Correction for finite pulse-time effects in very thin samples using the flash method of measurement thermal diffusivity. *J. Appl. Phys.* 38:465.

Mackay, J. A. and Schriempf, J. T. 1976. Corrections for nonuniform surface heating errors in flash-method thermal diffusivity measurements. *J. Appl. Phys.* 47:1668–1671.

Maglic, K. D. and Taylor, R. E. 1984. The apparatus for thermal diffusivity measurement by the laser pulse method. *In* Compendium of Thermophysical Property Measurement Methods, Vol. 2: Recommended Measurement Techniques and Practices (K. Maglic, A. Cezairliyan, and V. E. Peletsky, eds.). pp. 281–314. Plenum Press, New York.

Maglic, K. D., Cezairliyan, A., and Peletsky, V. E. (eds.). 1984. Compendium of Thermophysical Property Measurement Methods, Vol 1: Survey of Measurement Principles. Plenum Press, New York.

Parker, W. J., Jenkins, R. J., Buttler, C. P., and Abbott, G. L. 1961. Flash method of determining thermal diffusivity, heat capacity and thermal conductivity. *J. Appl. Phys.* 32:1679.

Takahashi, Y., Yamamoto, K., Ohsato, T., and Terai, T. 1988. Usefulness of logarithmic method in laser-flash technique for thermal diffusivity measurement. *In* Proceedings of the Ninth Japanese Symposium on Thermophysical Properties (N. Araki, ed.). pp. 175–178. Japanese Thermophysical Society, Sapporo.

Taylor, R. E. 1975. Critical evaluation of flash method for measuring thermal diffusivity, *Rev. Int. Hautes Temp. Refract.* 12:141–145.

Taylor, R. E. and Clark, L. M., III. 1974. Finite pulse time effects in flash diffusivity method. *High Temp. High Pressures* 6:65.

Taylor, R. E. and Maglic, K. D. 1984. Pulse method for thermal diffusivity measurement. *In* Compendium of Thermophysical Property Measurement Methods, Vol. 1: Survey of Measurement Techniques (K. Maglic, A. Cezairliyan, and V. E. Peletsky, eds.). pp. 305–334. Plenum Press, New York.

Taylor, R. E., Lee, T. Y. R., and Donaldson, A. B. 1978. Thermal diffusivity of layered composites. *In* Thermal Conductivity 15 (V. V. Mirkovich, ed.). pp. 135–148. Plenum Press, New York.

Vozar, L., Gembarovic, J., and Majernik, V. 1991a. New method for data reduction in flash method. *Int. J. Heat Mass Transfer* 34:1316–1318.

Vozar, L., Gembarovic, J., and Majernik, V. 1991b. An application of data reduction procedures in the flash method. *High Temp. High Pressures* 23:397–402.

Watt, D. A. 1966. Theory of thermal diffusivity by pulse technique. *Br. J. Appl. Phys.* 17:231–240.

KEY REFERENCES

Taylor and Maglic, 1984. See above.

Survey of thermal diffusivity measurement techniques.

Maglic and Taylor, 1992. See above.

Specific description of the laser flash technique.

INTERNET RESOURCES

http:/www.netlib.org

Collection of mathematical software, papers, and databases.

http:/www.netlib.org/odrpack/

ODRPACK 2.01—Software package for weighted orthogonal distance regression (nonlinear fitting procedure used to calculate optimal values of the unknown parameters).

APPENDIX

The heat balance equation for transient conditions may be written as

$$\nabla \cdot \lambda \, \nabla T + (\text{internal sources and sinks}) = C_p \rho \frac{\partial T}{\partial t} \quad (7)$$

where λ is the thermal conductivity, C_p is the specific heat at constant pressure, and ρ is the density. If there are no internal sources and sinks,

$$\nabla \cdot \lambda \, \nabla T = C_p \rho \frac{\partial T}{\partial t} \quad (8)$$

For homogeneous materials whose thermal conductivity is nearly independent of temperature, we may treat λ as a constant. Then $\nabla \cdot \lambda \nabla T$ becomes $\lambda \nabla^2 T$, and Equation 8 can be written as

$$\lambda \, \nabla^2 T = C_p \rho \frac{\partial T}{\partial t} \quad (9)$$

or

$$\nabla^2 T = \frac{C_p \rho}{\lambda}\frac{\partial T}{\partial t} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{10}$$

where $\alpha = \lambda/C_{p\rho}$ is the thermal diffusivity. For one-dimensional heat flow

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \tag{11}$$

The assumed adiabatic conditions at the faces of a plate of thickness L result in boundary conditions

$$\frac{\partial T(0,t)}{\partial x} = \frac{\partial T(L,t)}{\partial x} = 0 \qquad t > 0 \tag{12}$$

The solution of Equation 11 defining temperature at a given time at position x within the plate is then given by

$$T(x,t) = \frac{1}{L}\int_0^L f(x')\,dx' + \frac{2}{L}\sum_{n=1}^{\infty} \exp\left(\frac{-n^2\pi^2\alpha t}{L^2}\right) \times \cos\frac{n\pi x}{L}\int_0^L f(x')\cos\frac{n\pi x'}{L}\,dx' \tag{13}$$

The function *f(x)* represents the temperature field in the plate resulting from a short pulse of energy Q instantaneously absorbed in a thin surface layer of thickness g.

The initial conditions defining temperature distribution in the plate at time $t = 0$ are

$$f(x) = \begin{cases} \frac{Q}{\rho C_p g} & 0 \le x \le g \\ 0 & g < x < L \end{cases} \tag{14}$$

For simplicity, plate reference temperature is assumed to be zero. For these initial conditions, Equation 13 transforms to

$$T(x,t) = \frac{Q}{\rho C_p L}\left[1 + 2\sum_{n=1}^{\infty} \exp\left(\frac{-n^2\pi^2\alpha t}{L^2}\right) \times \cos\frac{n\pi x}{L}\frac{\sin(n\pi g/L)}{n\pi g/L}\right] \tag{15}$$

For opaque solids the ratio g/L is sufficiently small that the approximation $\sin(n\pi g/L) \approx n\pi g/L$ can be valid, simplifying Equation 15 for the plate rear face to

$$T(L,t) = \frac{Q}{\rho C_p L}\left[1 + 2\sum_{n=1}^{\infty}(-1)^n \exp\left(\frac{-n^2\pi^2\alpha t}{L^2}\right)\right] \tag{16}$$

After infinite time the rear-face temperature will be reduced to the first member in the square brackets:

$$T_{L,\max} = \frac{Q}{\rho C_p L} \tag{17}$$

Table 2. Values of K_x in Equation 19

x (%)	K_x	x (%)	K_x
10	0.0661	60	0.1622
20	0.0843	66.7(2/3)	0.1811
25 (1/4)	0.0927	70	0.1919
30	0.1012	75 (3/4)	0.2105
33.3 (1/3)	0.1070	80	0.2332
40	0.1190	90	0.3035
50 (1/2)	0.1388		

For practical application it is useful to relate thermal diffusivity to the percent rise in the rear-face temperature *V(L,t)*:

$$V(L,t) = \frac{T(L,t)}{T_{L,\max}} = 1 + 2\sum_{n=1}^{\infty}(-1)^n \exp\left(\frac{-n^2\pi^2\alpha t}{L^2}\right) \tag{18}$$

From Equation 18 the thermal diffusivity of the material can be related to any percent rise and the square of the plate thickness L:

$$\alpha = K_x \frac{L^2}{t_x} \tag{19}$$

where K_x is a constant corresponding to an x percent rise and t_x is the elapsed time to an x percent rise. Table 2 relates values of K_x calculated for specific x percent rises.

The detailed theory of one-layer homogeneous samples is in Watt (1966).

RAYMOND E. TAYLOR
JOZEF GEMBAROVIC
Thermophysical Properties Research Laboratory

KOSTA D. MAGLIC
Institute of Nuclear Sciences
Belgrade, Yugoslavia

SIMULTANEOUS TECHNIQUES INCLUDING ANALYSIS OF GASEOUS PRODUCTS

INTRODUCTION

The concept of concurrently using two or more techniques of thermal analysis on a single substance has been promoted by advances in microprocessor control. This has made it possible to program and collect the data and process the information in many ways. Strictly speaking, the nomenclature "simultaneous" applied in such circumstance is not always correct. The mass change in thermogravimetry (TG) is generally the first signal received. The signal from the differential thermal analysis (DTA) probe will be received a little later as the heat has to diffuse through the sample and the crucible before reaching

the sensing thermocouple. The sample is defined as the material (usually a solid) under investigation. Likewise, in evolved gas analysis (EGA), the gases are sampled but it then takes a finite time for the gas to be analyzed. In the mass spectrometer, there is an interface to allow the gas to be analyzed. In gas chromatography (GC), the signal of gas analysis takes considerable time to process, and consequently the gas analysis signal is intermittent and not received in "real time." In the literature, there are descriptions of both TG (see THERMOGRAVIMETRIC ANALYSIS) and DTA (see DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY) being coupled with techniques such as Fourier transform infrared (FTIR) spectrometry and x-ray diffraction.

In common with all other thermal analysis techniques, the following parameters must be considered: the sample characteristics, its mass, the sample holder, the temperature range for study, the rate of heating, and the atmosphere.

It must be remembered that a single thermal analysis method may not give the investigator sufficient information. Thus DTA will indicate if a transformation is exothermic or endothermic. It needs thermogravimetric analysis to indicate if there is mass change as well as gas analysis to identify any gaseous products involved.

Although "simultaneous techniques" is the approved term used by the International Union of Pure and Applied Chemistry (IUPAC), the recommended nomenclature for such simultaneous techniques is to use a hyphen to separate the two techniques, e.g., TG-DSC. This has led to such techniques being called "hyphenated" techniques.

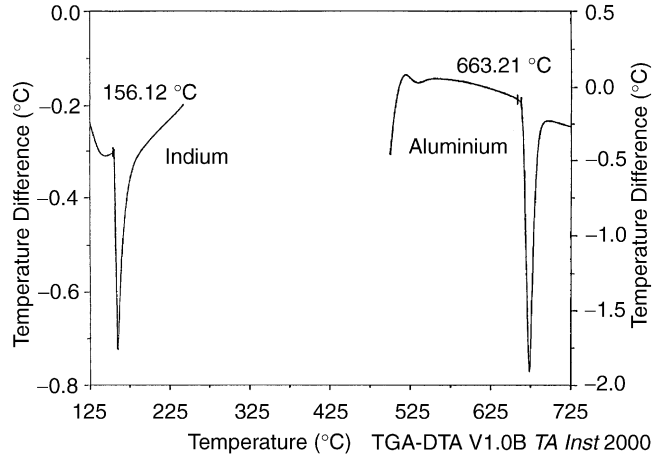


Figure 1. The melting DTA trace for indium and aluminum used to calibrate the equipment.

PRINCIPLES OF THE METHOD

New Studies Made Possible by Use of Simultaneous TG-DTA

Evaporation. The use of simultaneous TG-DTA can lead to interpretations not possible using conventional TG and DTA separately. An example is the study of evaporation. Because the DTA in the simultaneous unit cannot be operated in the usual manner with a covered crucible, evaporation may become a predominating feature (Lerdkanchanaporn and Dollimore, 1997).

Calibration. Calibration of TG is always a problem. With simultaneous TG-DTA, the calibration for temperature becomes much easier. The DTA trace can be used for temperature calibration in the simultaneous unit. The melting DTA trace for a metal can be used to calibrate the equipment. Figure 1 shows two runs in such a unit for indium and aluminum allowing the melting point to be established. Temperature for these materials and other standard calibration materials is the subject of extensive studies from the group in Germany reporting their findings to the International Confederation for Thermal Analysis and Calorimetry (ICTAC) and the IUPAC (Sarge et al., 1997).

Complex Thermal Degradations. Complex degradation patterns are observed in many salts. Magnesium nitrate hexahydrate, $Mg(NO_3)_2 \bullet 6H_2O$ melts at 90°C (see Fig. 2) and then water is lost from the system. The loss of water continues up to 350°C (42% loss), leaving an anhydrous salt that is a solid. This melts at 390°C and then decomposes to MgO at 600°C (residue 15.7%). All loss of water is endothermic and the dissociation to the oxide is also endothermic. Evolved gas analysis and hot-stage microscopy (HSM) would be additional simultaneous thermal analysis techniques that could provide details of this degradation.

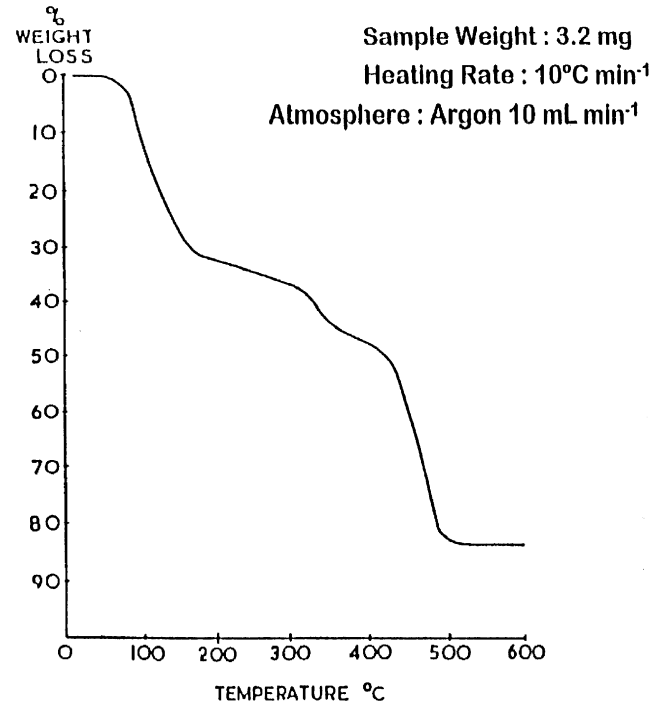


Figure 2. Complex degradation patterns of magnesium nitrate hexahydrate [$Mg(NO_3)_2 \bullet 6\ H_2O$].

PRACTICAL ASPECTS OF THE METHOD

General Comment on the Advantages and Disadvantages of Simultaneous Thermal Analysis Techniques

The obvious advantage of the simultaneous thermal analysis systems is that the same sample of the material under investigation is being studied. When two thermal analysis systems, e.g., DTA and TG, are investigated separately, the samples may vary in behavior. This is especially true for materials such as coal, geological samples, and building materials.

A further advantage of using the simultaneous techniques concerns the environment around the sample. The temperature and the gaseous environment in the simultaneous technique are the same for both the TG and the DTA signal. It is of course possible to combine more than two techniques in simultaneous thermal analysis. Such a system was devised to explore the nature of the soils on Mars (Bollin, 1969). Uden et al. (1976) have combined a variety of thermal analysis units into a complex simultaneous thermal analysis unit.

The economic advantages and disadvantages of using simultaneous thermal analysis techniques have to be considered carefully. It might be more advantageous to use a mass spectrometer in the laboratory rather than to dedicate the equipment to a simultaneous thermal analysis unit. On the other hand, if thermal analysis experiments are necessary on many samples so that a 100% usage is approached, then there is a definite economic advantage in using the simultaneous techniques.

Simultaneous TG-DTA/DSC

The most frequently used techniques of thermal analysis quoted in the literature are TG, DTA, and differential scanning calorimetry (DSC; see DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY). With the advent of the computer workstation, DTG (derivative thermogravimetry) can be produced from any TG curve, so it is not considered here as a separate technique. However, numerous commercial units involving TG-DTA or TG-DSC fulfill an obvious need in organic, pharmaceutical, and inorganic fields.

Advantages of Simultaneous TG-DTA

The Sample. The advantage of obtaining data in the same sample for both TG and DTA is obvious. There is a distinct possibility that different samples of the same material will show different signals, especially with DTA or DSC, where heat transitions may be affected by impurities present. In cases such as pharmaceutical stearic acid or magnesium stearate, the presence of other carboxylic acids can cause marked differences in both TG and DTA signals. For coal samples and other natural products (both geologically and biologically based), the use of different samples in obtaining the TG and DTA data could produce conflicting results due to the inhomogeneity of the samples.

The Mass of the Sample. Both TG and the DTA signals will be affected by the mass of material. In TG, the presence of large samples can cause kinetic features to change, usually from reaction interface kinetics for the small samples to diffusion-based kinetics for the large samples. The DTA is most sensitive when small samples are used. Thus many dehydration stages are often shown on the DTA unit when only one is seen on the TG unit. By having the same mass of material for both TG and DTA, the interpretation has to combine both TG and DTA data from the same sample.

The Sample Holder. Early work showed that in both TG and DTA the sample holder could affect the data. Thus long, narrow sample holders could cause diffusion to predominate whereas an open-tray sample would probably be best for studying kinetics based on reaction interface movement. Most obvious is the use of a crimped closed container in DTA to prevent mass loss (usually by evaporation). The TG data would be meaningless in such a situation. The use of an open crucible suitable for both TG and DTA would mean that the data from both techniques can be analyzed. It should be noted, however, that the DTA data may not correspond to the conventional DTA signal.

The Temperature Range of Study. The temperature range of study is exactly the same in the simultaneous experiment but may not be so when each technique is used separately.

The Heating Rate. It has been noted (Lerdkanchanaporn et al., 1997) that during a transition (phase change or chemical reaction) the imposed heating rate is perturbed. This phenomenon is unfortunately ignored by most investigators. It is, however, an important feature of simultaneous techniques, where one can be assured that any temperature perturbation is the same for both the TG and the DTA signal.

The Atmosphere. In the simultaneous technique the sample is subjected to the same gaseous environment while two thermal analysis signals are generated. This would not necessarily be the case if the TG and DTA signals were generated in two separate runs on two samples. It is important to note that there are two aspects to the environmental atmosphere: the nature of the gas and the rate of flow.

Description of Commercial TG-DTA Equipment

The Derivatograph manufactured in Hungary has been described by Paulik et al. (1958) and more recently by Paulik (1995) and ranks among the first commercial instruments to provide simultaneous TG, DTG, and DTA signals, as shown in Figure 3.

A TGA-DTA unit manufactured by TA Instruments that can operate up to 1650°C in controlled-temperature and gaseous environments is shown in Figure 4. It has a double-beam balance shown in the figure. Another unit described by TA Instruments as TGA-DSC operates up to 700°C and differs only in the temperature range, which allows it to be calibrated as a DSC unit. In both units

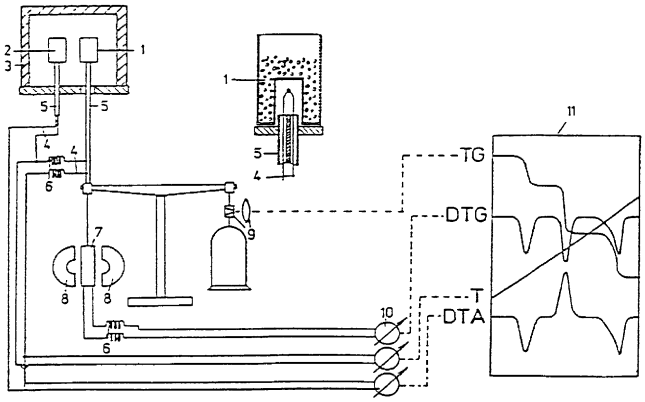


Figure 3. Derivatograph for simultaneous TG, DTG, and DTA.

special holders provide appropriate contact between the sample and the temperature sensor.

A similar TG-DSC manufactured by Netzsch contains the sample and the reference materials in a special crucible supported on a heat flux DSC plate. This system has a temperature range from −120 to 1500°C. The subambient temperature is achieved by use of a liquid nitrogen assembly. There is some difficulty in calibrating above 750°C so the equipment provided by Netzsch up to 2400°C is described as TG-DTA (Netzsch, 1997).

Setaram provides a TG-DSC system (described as SETARAM TG-DSC III) that uses symmetrical twin furnaces each with a Calvet microcalorimeter (Parlover, 1987).

The Use of Gas Analysis in Simultaneous Thermal Analysis Units

Even in the most obvious and simple process of material decomposition, there is a need to confirm the evolution of gaseous products in thermal decomposition as well as to identify the solid residue.

This applies in dehydrations,

$$CaC_2O_4 \bullet 2H_2O(s) \rightarrow CaC_2O_4(s) + 2H_2O(g) \qquad (1)$$

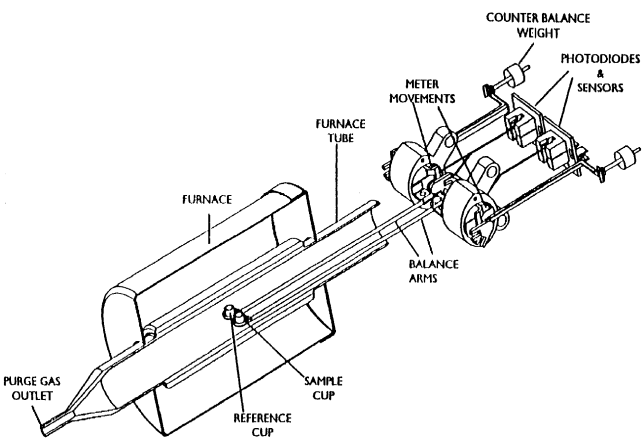


Figure 4. The SDT 2960 simultaneous TG-DTA (TA Instruments) showing a double-beam balance.

in decomposition of carbonates,

$$CaCO_3(s) \rightarrow CaO(s) + CO_2(g) \qquad (2)$$

in the dehydroxylation of hydroxides,

$$Mg(OH)_2(s) \rightarrow MgO(s) + H_2O(g) \qquad (3)$$

and in the gasification of carbon in air,

$$C(s) + O_2(g) \rightarrow CO_2(g) \qquad (4)$$

In fact, amorphous carbon gasified in air produces a mixture of CO and CO_2, the exact amounts depending on the sample and the experimental conditions. In many amorphous carbons, the surface complexes can produce water on gasification and small amounts of many volatile organic compounds.

The decomposition of calcium hydroxide often reveals small quantities of CO_2 being evolved, as a result of the $Ca(OH)_2$ partially carbonating on storage:

$$Ca(OH)_2(s) + CO_2(g) \rightarrow CaCO_3(s) + H_2O(g) \qquad (5)$$

In all combustion reactions of organic compounds, especially natural organic macromolecules such as starch or cellulose, the degradation process breaks down the sample into many simpler volatile organic compounds. Subsequent combustion then produces only CO, CO_2, and water.

The general themes outlined above serve to demonstrate the need to identify gaseous products in thermal analysis and explain why EGA is found so often in descriptions of simultaneous thermal analysis techniques.

A wide range of techniques are available to determine both the nature and the amount of gaseous products in thermal analysis, including GC, infrared (IR) spectroscopy or mass spectroscopy (MS), or using specific gas detectors. There is a real possibility that gases evolved as mixtures from the solid material degradation at high temperatures may themselves undergo gaseous reaction processes or gas condensation during transfer to the gas analysis device at room temperature. The gas analysis takes place some finite time after sampling, and in GC and certain other techniques the analysis takes place some time after the sampling process.

The following methods of analyzing gases in simultaneous thermal analysis can be noted: MS, GC, IR spectroscopy, condensation of volatile products, and chemical analysis.

The detectors used in GC may be used independently of chromatographic columns to detect gaseous product evolution. Furthermore, mass spectrometers and IR spectrometers may be used as detecting devices in gas analysis based on chromatography.

Mass Spectroscopy Used for Evolved Gas Analysis in Thermal Degradation Studies

The MS unit is generally used to identify and analyze gaseous samples from a TG, DTA, or DSC unit. The main difficulty is that the mass spectrometer works well under

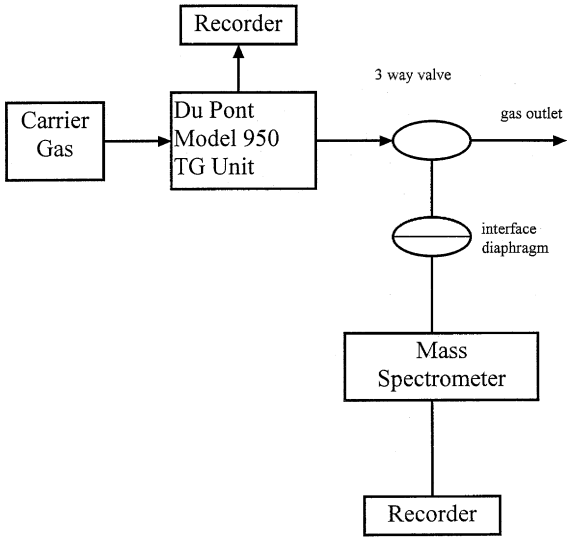


Figure 5. Schematic of a TG-MS unit.

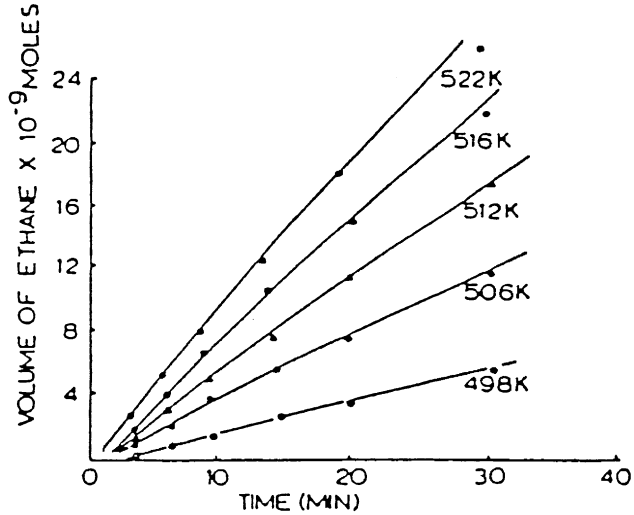


Figure 7. Volume of ethane evolved after various isothermal heating times at five temperatures of methanol chemisorbed on the nickel-silica catalyst.

vacuum but sampling is generally performed at 1 atm. There is a need for an efficient interface between the sampling of the gases at 1 atm and their analysis under high vacuum. A schematic of a typical TG-MS unit is shown in Figure 5. To eliminate side reactions, the interface should be located near the decomposing sample. Carbon monoxide presents problems because it has the same molar mass as N_2. It may be better in such cases to use IR methods (vide infra). Water also presents problems because it is difficult to degas from the equipment. A distinction must be drawn between (1) MS thermal analysis in which the samples are actually located in the mass spectrometer and (2) MS coupled to DTA, TG, or both. The latter is most often used by commercial instrument manufacturers. When the mass spectrometer is used in the laboratory for a variety of other uses, an economic solution to the problem is to collect the volatile products that have evolved and take the collected sample to the mass spectrometer. A scheme for such a set-up is shown in Figure 6. The equipment is simple, economical, and very suitable for kinetic evaluation (Galwey et al., 1975). Using this arrangement, the sample vessels (A, B, C, D, etc.) are closed, detached, and then presented to the mass spectrometer for analysis at different periods of time. Dollimore and Jones (1979) were able to study the desorption and decomposition of methanol on a supported nickel oxide oxidation catalyst. Figure 7 shows volumes of ethane evolved after various times of isothermal heating at five temperatures from methanol chemisorbed on a nickel-silica catalyst. Similar behavior was found for the production of methane and of ethene.

Equipment in which the sample is heated in the mass spectrometer has been described by Gallagher (1978) and by Price et al. (1978). This equipment had good temperature control, the sample decomposed under high vacuum, and the product gases had little time to react further, so side reactions could be eliminated. The disadvantage was that such results must not be expected to coincide with thermal decomposition studies carried out conventionally at 1 atm.

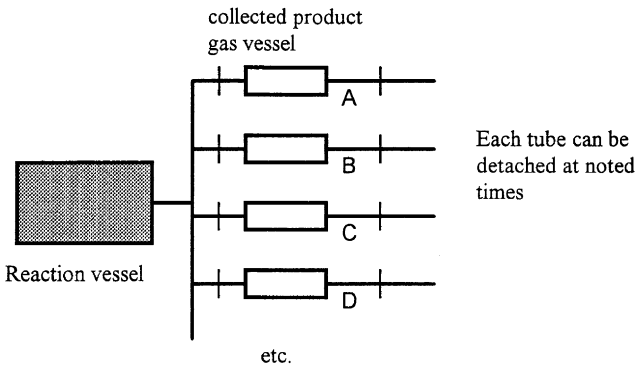


Figure 6. Schematic showing sample vessels that can be detached and presented to the mass spectrometer.

Evolved Gas Analysis and Chromatography

The choice of detector used in GC depends very much on the gases being analyzed. The four types of detectors usually used are the Katharometer (also called the thermal conductivity detector), the gas density detector, the ionization detector, and the infrared detector.

The Katharometer consists of two matched electronically heated filaments situated in the sample and reference streams and connected into a bridge circuit so that their resistances may be compared (see Fig. 8). The filaments are used as matched pairs on inlet and outlet gas systems. When the temperatures of the filaments are not equal, their resistances also differ and the unbalanced bridge gives a signal proportional to the temperature difference. The temperature of each filament depends on the operating conditions (e.g., the voltage across the filament, the temperature of the filament housing, and the

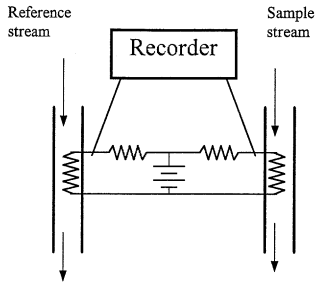


Figure 8. Katharometer.

rate of gas flow), the thermal conductivity of the gas stream, and the composition of the effluent gas.

If the operational factors are standardized, the bridge output is proportional to the difference in the thermal conductivity of the sample and reference streams; i.e., the evolved gas is detected by the change it imparts to the thermal conductivity of the carrier gas. The carrier gas is restricted because large differences are required between the thermal conductivity of the carrier gas and evolved product gases. In this context, it should be noted that hydrogen and helium have a high thermal conductivity while that for argon is low.

When using a gas density detector, the reference stream is usually split into two streams (see Fig. 9). If evolved gas is produced, the gas stream over the sample has a different density from that of the reference. The sensitivity of the measurement depends on the difference between the molecular weight of the evolved gas (M_e) and the molecular weight of the carrier gas (M_c). A peak area correction factor is given by

$$\text{Peak area} = \frac{M_c}{M_e - M_c} \tag{6}$$

Hence the molecular weight of the evolved gas must be known.

The advantage of the gas density method is that it can be used with corrosive gases. Any carrier gas can be used provided its molecular weight is different from that of the evolved gas.

Several types of ionization detectors are used in gas analysis, including argon triode and electron-capture detectors. In a flame ionization detector, the sample effluent is mixed with hydrogen and burned, and the resistance of the hydrogen flame is measured continuously with polarized (direct current) electrodes connected to an electrometer. Combustible materials introduced into the flame give enhanced ion concentrations. The resistance of the flame is reduced. These detectors are good for organic vapors but not for water vapor.

Gas detection devices utilizing IR spectroscopy have become so important that they are dealt with here as a separate subject (vide infra).

A problem with using gas-liquid chromatography (GLC) for EGA is that the gas cannot be passed continuously into the GLC column. A rapid intermittent injection procedure has to be utilized. Cold traps to trap out the volatile products or cold traps set at different temperatures to induce fractional condensation may be used. An alternative is to use an automatic feed device. However, the rate of sampling is conditioned by the retention time. Let t_r represent the retention time of the first component eluted from the column. Then the maximum number of samples that can be analyzed in GLC is given by

$$n = \frac{T_h - T_l}{\beta t_r} \tag{7}$$

where T_l and T_h are the low and high limits of the temperature range over which the gas analysis is attempted and β is the heating rate in degrees Celsius per minute. Therefore, it can be seen that the retention time determines the number of samplings.

Infrared Spectrometry

This section is labeled infrared spectrometry because equipment sold in the market allows these units to provide a wide range of gas analyses coupled with either TG or DTA. However, specialized and simple IR detectors can be used to analyze the concentration of evolved gases such as CO and CO_2. This involves nondispersive IR analyzers. Carbon monoxide and CO_2 are especially difficult gases to analyze accurately on a mass spectrometer. This type of IR spectroscopy is well suited to on-stream analysis (Morgan, 1977). It is not, however, restricted to the above gases but there is special merit in applying the method to them.

In an IR radiometer, the output from an IR source is split to pass through the sample and the reference effluent streams and the radiation intensities are then compared. Figure 10 shows this arrangement schematically.

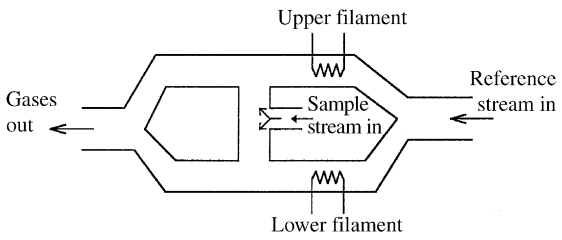


Figure 9. Gas density detector.

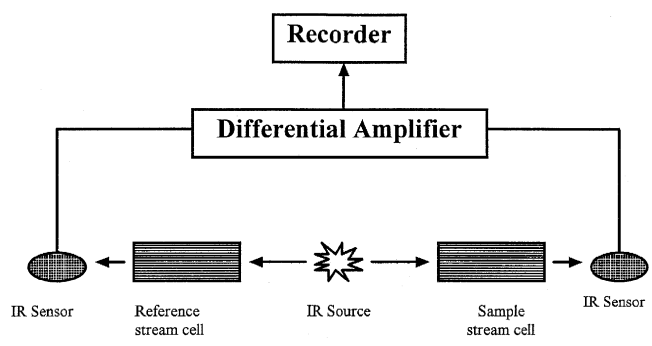


Figure 10. Schematic illustrating the IR radiometer.

Table 1. Wavelength, Path Length, and Amplification Required on a Miran 104 Portable Gas Analyzer Set to Detect CO_2, CO, and H_2O

	CO_2 Detection	CO Detection	H_2O Detection
Wavelength, μm	4.25	4.7	6.6
Path length, mm	1.5	14.5	5.5
Amplification	×2	×5	×5

An example of the use of IR detectors to analyze single gas may be quoted from Hoath (1983) where CO, CO_2, and H_2O were analyzed by such equipment using a Miran 104 portable gas analyzer. Details of the wavelength, path-length, and the signal amplification are given in the Table 1. Figure 11 shows the detection of H_2O, CO, and CO_2 in evolved gas stream passed over α-cellulose subjected to controlled heating program.

In dispersive IR spectrometers, the IR radiation from the IR source is focused by mirror optics onto the sample. To reduce IR absorption, the sample is placed in between windows of ionic materials. Dispersion of the radiation is achieved by a grating or prism and detection is usually achieved by thermocouple or pyroelectric detector units.

On the other hand, FTIR spectrometry uses similar sources, sampling units, and detectors but uses the interference of two IR beams controlled by a mirror system. The Fourier transform software then converts the interferogram to a spectrum of transmittance plotted against wave number. The software provided allows the production of a total evolution profile.

In attenuated total reflection (ATR) units the sample is in contact with a prism of high refractive index. Sample preparation is minimal, and thermal ATR instruments are available so that the sample spectrum can be monitored as a function of temperature.

Separate Collection of the Gases and Volatile Products

In many cases, the analysis of gaseous samples collected may not occur during the original thermal analysis experiment. The system can still be considered as simultaneous because the sampling was achieved at the same time as all the other measurements. The volatile products can be collected by passing the carrier gas through an absorber, a cold trap, or selective gas detectors or by fractional condensation. McNeill (1970, 1977) has described a technique called thermal volatilization analysis (TVA) in which the sample heated in vacuum (see Fig. 12) is progressively condensed by passage through traps at a series of low temperatures.

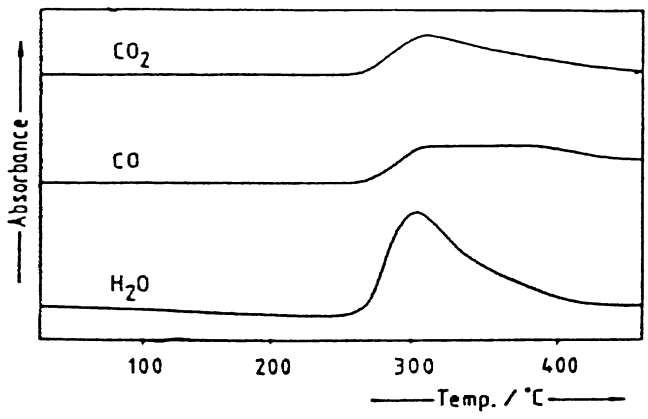
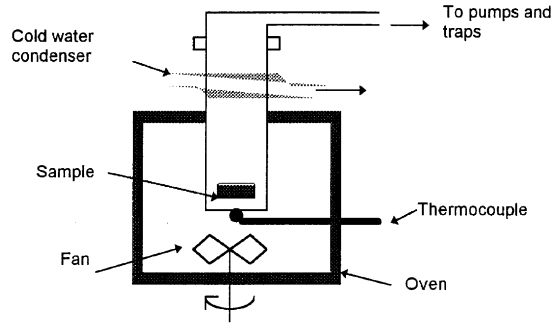
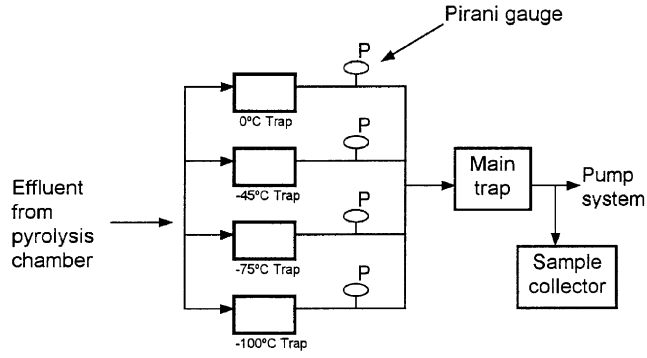


Figure 11. The EGA of α-cellulose decomposed in nitrogen.



the pyrolysis chamber



Trap system for differential condensation

Figure 12. Thermal volatilization equipment.

Chemical and Physical Methods of Evolved Gas Detection

In certain cases it has been found possible to identify and analyze product gases by chemical means. Evolved CO_2 is measured by noting the changes in conductivity of a barium hydroxide solution. Keattch (1964) used such a method to determine the CO_2 evolved on heating samples of hardened concrete through a predetermined temperature range.

Acids such as HCl from the degradation of polyvinylchloride can be detected and estimated by absorbing the gas in a solvent such as aqueous alkali and measuring the change in pH. McGhie et al. (1985) similarly dissolved evolved ammonia in water and used a continuous measurement of pH of a constantly renewed solution to monitor the evolution of the ammonia in the decomposition of $(NH_4 \bullet H_3O)_{1.67}\ MgO_{0.67}\ Al_{10.37}O_{17}$ (ammonium/hydronium beta alumina). This helped to identify multiple mass losses on the TG data as due to NH_3 below 400°C.

Gas-sensing membrane electrodes have been described for CO_2, NO_2, H_2S, SO_2, and NH_3 (Fifield and Kealey, 1990). Brinkworth et al. (1981) used a coulometric detector to measure SO_2 formed in coal combustion studies.

DuPont instruments (Hassel, 1976) made an electrochemical cell operated in a nitrogen atmosphere to analyze for water over a temperature range from 0 to 1000°C. This is a physical method. The moisture is transferred by nitrogen as a carrier gas into an electrolytic cell detector where it is absorbed by phosphorus pentoxide coated onto platinum electrodes. The water is electrolyzed and the product gases hydrogen and oxygen carried away by the gas stream. The electrolysis current is integrated and gives the amount of water directly. It forms the basis of method D4019 of the American Society for Testing and Materials (ASTM). A capacitance probe (Diefenderfer, 1979) and a dew point instrument (Gallagher et al., 1982) have been used to detect water.

METHOD AUTOMATION

The Role of the Digital Computer in Simultaneous Techniques

The computer plays many important roles in the present-day lifestyle and is a pivotal part of any research laboratory. Today laboratories are equipped with many computers, either standalone, or with network capability, or linked with one or more instruments.

The use of computers in the thermal analysis laboratory may be classified into four categories (Haines, 1995):

1. Controlling the operation of the instrument;
2. Making the collection, interpretation, storage, and retrieval of instrumental data easier for the operator;
3. Making calculation of the experimental results easy and accurate; and
4. Simulating the behavior of the instrument or the sample under special conditions.

The computer may be programmed to operate a variety of thermal analysis experiments. The tasks the computer can control may include programming the experimental conditions (received in input from the operator via a keyboard), loading the sample from a robotics autosampler into the analyzer, carrying out the run, acquiring the experimental signals, storing the data into a memory or onto disk, retrieving the stored data, analyzing the data according to a selected program, displaying the results, and then printing a report. The computer can also perform more than one task at the same time. For example, while the experiment is being carried out, the data stored in its memory can be retrieved and analyzed.

Data acquisition is far more frequent and precise than human capability and results in a huge array of signals. Again, calculations on this large amount of information are much quicker and more accurate with the aid of computer codes. The basic software may be sold as part of the instrument and the computer workstation, whereas the specific programs may be offered separately. To avoid paying for an expensive program, operators are well advised to analyze their data via a spreadsheet or other common mathematics programs or even to write their own program.

However, getting accurate results requires not only good computer programs and instruments but also proper experimental plans, sample preparation, and experimental conditions as well as critical thinking.

PROBLEMS

Disadvantages of Simultaneous TG-DTA

Equipment used for TG-DTA is generally operated at high temperatures, perhaps as high as 1650°C. This means that the normal calibration for enthalpy is inadequate as the limiting temperature for DSC is normally around 700°C. Equipment made available as TG-DSC can be used for accurate determination of heat input and output. Such measurement may not correspond with similar heat measurements on DSC or DTA alone. The reason is that DSC and DTA used alone are most often operated in a closed container to prevent loss of material during the study. Benzoic acid may be quoted as an example. This material is a standard reference material, and in a closed container the signal generated in a DSC unit can be used to indicate both the temperature and the constant required for enthalpy determinations using the known heat of fusion at the melting point. In TG-DTA or TG-DSC equipment, the DTA or DSC signal has to be generated from an open crucible, and it is easy to demonstrate that considerable sublimation occurs, which was suppressed in the conventional calibration of DTA or DSC alone when using a closed container.

A further disadvantage of this simultaneous signal is that in DTA or DSC alone, the best signal is generated with extremely small sample size. However, to accurately record the mass loss or gain in the sample, a minimum amount of sample has to be compatible with balance sensitivity and accuracy. This may mean that the mass of sample used to get a good TG signal is in excess of the smaller amount required by the DTA sensor to give the best DTA trace. This results in a compromise being needed between the optimal requirements for the two techniques.

Another disadvantage of TG-DTA is that a DTA unit requires specially shaped crucibles to fit over the thermocouple sensors. This precludes using TG in such a system to investigate shaped materials, i.e., films or granules, which will not fit into these specially shaped containers.

It should be noted that the highest sensitivity of a DTA experiment is achieved at a high heating rate whereas for TG a low heating rate will give the best result. Once again, a compromise has to be achieved in simultaneous TG-DTA units.

LITERATURE CITED

Bollin, E. M. 1969. A study in differential thermal analysis (DTA)—effluent gas analysis (EGA) for the determination of planetary environmental parameters. *In* Thermal Analysis, Proceedings of the Second International Conference for Thermal Analysis (R. F. Scwhwenker and P. D. Garn, eds.), Worcester, Mass., 1968. pp. 1387–1410. Academic Press, New York.

Brinkworth, S. J., Howes, R. J., and Mealor, S. E. 1981. Coal combustion studies using thermogravimetry with colorometric titration of evolved sulphur dioxide. *In* Proceedings of the Second European Symposium on Thermal Analysis (D. Dollimore, ed.). p. 508. Heyden, London.

Diefenderfer, A. J. 1979. Principles of Electronic Instrumentation. Saunders, Philadelphia.

Dollimore, D. and Jones, T. E. 1979. *Surf. Technol.* 8:483–490.

Fifield, F. W. and Kealey, D. 1990. Principles and Practice of Analytical Chemistry, 3rd ed. Blackie, Glasgow.

Gallagher, P. K. 1978. *Thermochim. Acta* 26:175.

Gallagher, P. K., Gyorgy, E. M., and Jones, W. R. 1982. *J. Therm. Anal*. 23:185.

Galwey, A. K., Dollimore, D., and Rickett, G. 1975. *J. Chim. Phys.* 72:1059–1064.

Haines, P. J. 1995. Thermal Methods of Analysis: Principles, Applications and Problems, p. 11. Blackie Academic & Professional, London.

Hassel, R. L. 1976. *Am. Lab. DuPont Application Brief TA 46* (8):33.

Hoath, J. M. 1983. Thermal Degradation Studies on Impregnated Cellulase. Ph.D. Thesis, University of Salford.

Keattch, C. J. 1964. Analysis of Calcareous Materials. SCI Monograph, No. 18, London.

Lerdkanchanaporn, S. Dollimore, D. and Alexander, K. S. 1997. *J. Therm. Anal*. 49:879–886.

Lerdkanchanaporn, S., Dollimore, D., and Alexander, K. S. 1997. *J. Therm. Anal*. 49:887–896.

McGhie, A. R., Denuzzio, J. D., and Farrington, G. C. 1985. A micro pH detector for thermogravimetric analysis. *In* Proceedings of the Fourteenth NATAS Conference (B. B. Chowdhury, ed.). pp. 22–26. North American Thermal Analysis Society, Sacramento, Ca.

McNeill, I. C. 1970. *Eur. Polym. J*. 6:373.

McNeill, I. C. 1977. *J. Polym. Sci. Polym. Chem. Ed*. 15:381.

Morgan, D. J. 1977. *J. Therm. Anal*. 12:245.

Netzsch. 1997. Netzsch STA 409 System Brochure, Selb, Germany.

Parlover, P. Le. 1987. *Thermochim. Acta* 121:307.

Paulik, F. 1995. pecial Trends in Thermal Analysis, p. 459. John Wiley & Sons, Chichester.

Paulik F., Paulik, J., and Erdey, L. 1958. *Z. Anal. Chem.* 160:241.

Price, D., Fatemi, M.S., Whitehead, R., Lippiatt, J. H., Dollimore, D., and Selcuk, A. 1978. Mass spectrometric thermal analysis. *In* Dynamic Mass Spectroscopy, Vol. 5 (D. Price and J. F. H. Todd, eds.). pp. 216–225. Heyden, London.

Sarge, S. M., Hemminger, W., Gmelin, E., Höhne, G. W. H., Cammenga, H. K., and Eysel, W. 1997. *J. Therm. Anal*. 49:1125.

Uden, P. C., Henderson, D. E., and Lloyd, R. J. 1976. *In* Proceedings of the First European Symposium on Thermal Analysis (ESTA) (D. Dollimore, ed.). p. 29. Heyden, London.

DAVID DOLLIMORE
SUPAPORN LERDKANCHANAPORN
The University of Toledo
Toledo, Ohio

ELECTRICAL AND ELECTRONIC MEASUREMENTS

INTRODUCTION

Electrical and electronic measurements of materials are among the most powerful techniques available for materials characterization. These measurements can reveal information that completely characterizes the electrical transport properties of a material. Furthermore, modern instrumentation makes precision measurements and computer interfacing straightforward at reasonable cost.

Conductivity measurements yield information on the conductivity (resistivity) of materials and indirectly the mobility of current carriers. Resistivity is one of the most sensitive measures of the electrical transport in materials and can vary from 10^{12} ohm-cm for the best insulators to 10^{-6} ohm-cm in pure normal metals. Measurements include both DC and AC techniques. The Hall effect is a versatile technique, yielding such information as carrier concentrations and mobility, and indirectly, estimates of scattering mechanisms, all of which affect the conductivity as a function of temperature. Careful experimental technique can also yield information about the ionization energy of dopants contributing to the electrical transport. These measurements, used in conjunction with solid-state spectroscopies, can yield information about the band structure of semiconductors.

The capacitance-voltage (C-V) measurement is the standard method of "profiling" free carrier concentrations as a function of position below the surface. A C-V measurement is the most common of the carrier profiling techniques due to its simplicity and low cost. Deep level transient spectroscopy (DLTS) is a transient capacitance measurement, which permits identification of deep traps within the forbidden gap of a semiconductor. This highly effective and elegant measurement technique yields information regarding the energy levels of these defects, their capture and emission rates, and their concentration. Equipment for measuring both C-V and DLTS are often combined in a single apparatus. The use of C-V measurements, DLTS characterization, and the Hall effect yields a potent suite of characterization techniques that can completely describe the electronic transport properties of a material.

Semiconductor materials can be prepared with either electrons or holes as the majority current carriers. Most measurement techniques are useful in characterizing a single carrier type, or, at best, the net carrier type for mixed transport. It is often useful to gain insight into transport across a junction formed between materials at which the carrier type changes abruptly from holes (p-type material) to electrons (n-type material). Transport measurements at pn–junctions yield information on the transport of both holes and electrons and can determine transport mechanisms as a function of carrier (current) density.

The carrier lifetime is one of the most sensitive measurements available for detecting impurities in semiconductors. At its most sensitive, this materials-characterization technique is capable of detecting impurity concentrations as low as one impurity atom per 10^{12} host atoms. Techniques available for carrier lifetime measurements are varied and plentiful. Optical techniques rely upon excitation of the material from its equilibrium state, typically by using a short optical (laser) pulse, then carefully measuring the return to equilibrium. Because the optical excitation can be finely focused, these techniques have the added ability to spatially map a material.

Transport measurements, particularly those techniques that yield information on carrier concentration, mobility, and scattering mechanisms, are useful educational tools for undergraduate science and engineering students. Students can perform experiments that reinforce lectures in courses such as solid state physics, physical chemistry, materials science, and device physics. Furthermore, the use of such apparatus as high input-impedance electrometers, capacitance meters, phase-sensitive detectors, lasers for optical excitation, spectrometers, and photon counting equipment greatly enriches students' knowledge of the equipment and instrumentation used in the study of materials. Thus, experiments centered upon materials characterization prepares a student for graduate study or entrance into the work force. The student is equipped with both knowledge of electrical transport in materials and the sophisticated yet inexpensive equipment used in their measurement.

PETER A. BARNES

CONDUCTIVITY MEASUREMENT

INTRODUCTION

A material's conductivity, σ, (or the inverse property, resistivity, ρ, where $\rho = 1/\sigma$,) relates to its ability to conduct electricity. In metals, conduction of electricity is tantamount to conduction of electrons, which depends on charge density and on scattering of the electrons by the crystal lattice (phonons) or by lattice imperfections. In semiconductors, conductivity is determined by the number of available charge carriers (electrons or holes) and the carrier mobilities. Because of the different mechanisms for conductivity, its dependence on temperature also differs. Conductivity increases with increasing temperature for semiconductors (more carriers are generated) and it decreases with increasing temperature for metals (more scattering by the lattice). Conductivity also depends on

physical structure. In crystals, the crystal type and orientation affect conductivity because the electronic structure is intimately tied to the crystal structure. The size of the crystallites (grains) in polycrystalline materials is also important as it affects the scattering of carriers, and, at very small sizes may also affect electronic structure.

For a rectangular slab of homogeneous material with length L, width W, and thickness t (in m), the resistance, R (in ohms, Ω) along the length of the slab is related to the resistivity ρ in Ω-m by

$$R = \frac{\rho}{t}\frac{L}{W} \tag{1}$$

Units for conductivity are most often expressed in terms of siemens per meter (S/m), or by the equivalent terms 1/(Ω-m) and (less common) mhos/m. In the semiconductor field, the quantity ρ/t is termed sheet or surface resistance, R_s, and quoted in units of Ω per square. The number of "squares" of a thin-film conductor or resistor material is given by the ratio L/W.

The accurate determination of a material's conductivity can be critical for understanding material composition or device performance. The method used to determine conductivity depends on whether the material is a bulk sample or a thin film. This unit begins with a section on bulk measurements (see the discussion of Bulk Measurements). Two-point measurement using an inexpensive ohmmeter is perhaps the simplest approach, although its accuracy is limited and fairly large samples are required. A four-point bulk measurement is only slightly more complicated, but yields much better accuracy. Both of these methods are potentially destructive to the sample, depending upon how electrical contacts are made. These measurements are described along with advice on how to make the most accurate measurements in different situations.

In the second section (see discussion of Surface Methods), two thin-film approaches are presented which are most frequently used in semiconductor characterization: four-point probing and the Van der Pauw method. These methods require more expensive test equipment but are very accurate. Four-point probing requires little sample preparation, although it can cause damage to the surface. The Van der Pauw approach generally requires the creation of a test pattern on the surface but can take up less test space on a semiconductor than four-point probing.

In the third section (see discussion of Non-Contact Methods) measurement approaches are presented that offer the possibility of measurements without damaging the material. These approaches are not as well developed as the contact approaches described in the first two sections. Specialized equipment is required, and accuracy can be rather limited. Extracting conductivity values often depends upon qualitative comparison to standard data. The most widely used non-contact approach is the eddy current method, which can be used for metals and semiconductors, and which is implemented in a variety of ways. Also described in this section is the relaxation method.

The fourth section (see discussion of Microwave Techniques) briefly describes the most common microwave techniques that can be employed to determine conductivity information at high frequencies. Expensive test equipment is required for these approaches, along with specialized test apparatus to probe or hold the samples. The conductivity is not directly measured, but is typically extracted from loss tangent determination.

Some general characteristics for the methods described in these four sections are listed in Table 1.

Electrical conductivity (resistivity) is a fundamental materials property and as such finds its value incorporated in many complex and sophisticated treatments of materials

Table 1. General Characteristics for Conductivity Measurement Methods

Method	Favoredmaterial Type	Favored Material Form	Estimated Measurement Range	Estimated Frequency Range
Two-point measurement	High-resistance metals	Solid bar	10^2–10^9Ω	DC AC (<300 Hz)
Four-point measurement	Metals	Solid bar	10^{-7}–10^4 Ω	DC AC (<300 Hz)
Four-point probe	Semiconductor surface, thin metallic films	Planar solid	10^{-3}–10^5 Ω-cm	DC
Van der Pauw	Semiconductor surface, thin metallic films	Planar solid	10^{-3}–10^5 Ω-cm	DC
Eddy current	Metals, semiconductors, Insulators	Solid wafer, cylinder, containerized liquid	10^{-10}–10^{10} Ω-cm	10–10^6 Hz
Relaxation	Metals, semiconductors	Solid wafer, containerized liquid	10^{-8}–1 Ω-cm	10^{-1}–10^7 Hz
Coaxial probe	Semiconductors, insulators	Flat surfaced solid, liquid	10^{-3}–10^5 Ω-cm	2×10^8 to 2×10^{10} Hz
Transmission line	Semiconductors, insulators	Precision-machined sample, containerized liquid	10–10^6 Ω-cm	10^8–10^{10} Hz
Free-space radiation	Metals, moderate to high loss materials	Large area, flat solid sample	10–10^5 Ω-cm	10^9–10^{11} Hz
Cavity resonator	Metals, moderate to high loss materials	Machined solid sample	10–10^8 Ω-cm	5×10^8 to 10^{11} Hz

behavior. Whereas this value is often "plugged-in" from a table of known properties, it may also be deduced from the governing relations if other pertinent factors are known. This unit does not attempt to describe the universe of indirect measurement possibilities in regard to conductivity and resistivity.

BULK MEASUREMENTS

Principles of the Method

The conductivity or resistivity of a bulk sample is based on accurate measurement of both resistance and the sample dimensions. The resistance is the ratio of the voltage measured across the sample to the current driven through the sample or of the voltage applied across the sample to the measured current.

For a homogeneous bar of length, L, and uniform cross-section, A, the resistance, R, is related to the resistivity, ρ, by

$$R = \rho L/A \tag{2}$$

In Figure 1A, the bar is connected in a "two-point" arrangement, so-called since the measurement apparatus is connected to the bar at the two end-points. The measurement apparatus is represented by an ideal current source in parallel with a high-impedance voltmeter. The apparatus can also be realized using an ideal voltage source in series with a low-impedance ammeter, as shown in Figure 1B. A more realistic view of a two-point measurement using an ohmmeter is shown in Figure 2. A voltage source and a variable range resistor (R_r) supply the current, where R_r is adjusted to provide a convenient voltage across the voltmeter. Typical values of R_r range from 100 to 10,000 Ω. R_c represents series resistance in the cable and the wire-to-sample contact resistance. The resistance in the bar is calculated as

$$R = \frac{(R_r V_m/V_s)}{[1 - (V_m/V_s)]} - 2R_c \tag{3}$$

A long bar of resistive material is desirable to minimize the effect of extra resistance in the measurement system or inaccurate length measurement.

A more accurate method, especially for low-resistance materials, is provided by the "four-point" arrangement, a simple schematic of which is shown in Figure 3. The resistance now is a function of the length between the voltage probes, L', or

$$R = \rho L'/A \tag{4}$$

The current through the high-impedance voltmeter is very small and minimizes the effect of cable and contact resistance. The four-point approach, when used with a micro-ohmmeter, is capable of measuring down to 10^{-7} Ω.

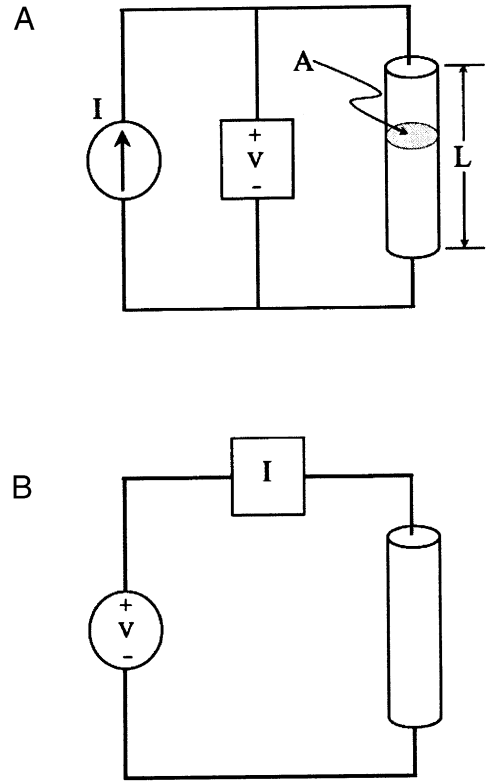


Figure 1. Two-point measurement of a resistive bar of length, L, and cross-sectional area, A, using (**A**) an ideal voltmeter in parallel with an ideal current source, and (**B**) an ideal ammeter in series with an ideal voltage source.

Practical Aspects of the Method

The two-point approach is most accurate for high-resistance measurements where the usually small R_c term (generally less than 2 Ω) can be ignored. These measurements are often made using ohmmeters contained within multimeters capable of measuring voltage, current, and resistance. The ohmmeter within the multimeter performs like the measurement circuits shown in Figure 1, and is realized using a variety of circuit implementations [see, e.g., Witte (1993) or Coombs (1995)]. Inexpensive multimeters (<$50) can measure resistance between 0.1 Ω and

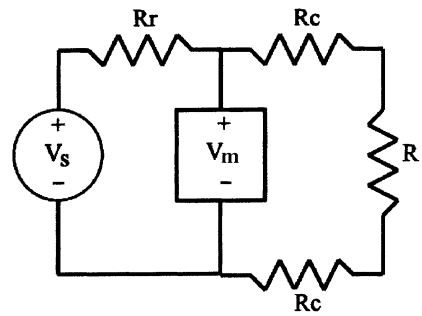


Figure 2. Two-point ohmmeter measurement circuit, which includes contact and cable resistance, R_c.

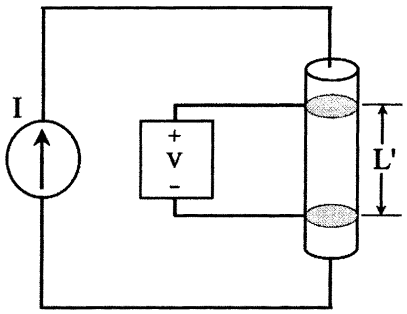


Figure 3. Four-point measurement circuit drives a known current through the sample, and a high-impedance voltmeter measures voltage drop along a designated length, L'.

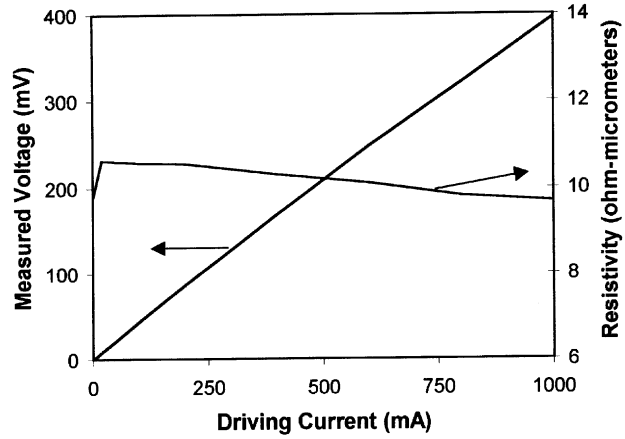


Figure 4. Measured voltage across a 2.5-cm length section of graphite rod (0.0889 cm diameter) versus the driving current, along with the extracted resistivity.

10 kΩ. More expensive meters (~$1000) can measure 100 μΩ to 100 MΩ and are provided with an extra pair of terminals for performing four-point measurements, described below. Dedicated ohmmeters can also be purchased for special measurements. For instance, high-resistance meters are capable of measurements up to 10^9 Ω. These meters require large, precision voltage sources (in the vicinity of 1000 V) along with very low−current measurement capability, perhaps as low as 10^{-15} A. Also, since ohmmeters have internal capacitance, C, the large time constant, $\tau = RC$, for a high-resistance measurement can be excessive and it may be extremely difficult to make a steady-state measurement.

At low resistance, the cable and contact resistance can seriously degrade measurement accuracy. Some ohmmeters have a "zero-adjust" feature, where the test leads are shorted together and the meter is adjusted to zero. Such an adjustment compensates for cable resistance and for measurement drift internal to the ohmmeter. In a related approach, the user can measure a short circuit, perhaps employing the same electrical contact strategy as for the actual test device, in order to determine R_c. Another approach to account for R_c is to make measurements on a pair of devices having the same cross-section but different lengths. The difference in resistance between the two measurements should result in the resistance due only to the difference in length of the bar.

In the four-point approach, it is im portant to keep the voltage contacts very thin so an accurate measurement of L' can be made. Voltage probes should be well away from the current probes to avoid "spreading resistance" interference. This spreading resistance is a consequence of the current spreading radially from the probe tip into the sample, with a considerable amount of resistance encountered near the tip. As a rule of thumb, current and voltage probes should be separated by a distance of at least 1.5× the thickness of the measured sample (ASTM B193, 1987).

As an illustration of bulk measurement, a 6-cm-long graphite rod of diameter 0.0889 cm was measured at room temperature using a two-point ohmmeter. Measurement of 1.9 Ω on a 5.8-cm length (distance between the two mechanical clips) yielded a resistivity of 20 Ω-μm. No attempt was made to remove cable and contact resistance from this measurement. Next, a four-point measurement was made using a driving current ranging from 1 mA up to 1 A that was connected to each end of the rod, and a voltmeter connected at points 2.5 cm apart near the center of the rod. The measured voltage as a function of driving current is shown in Figure 4. The resistivity from this four-point approach is ~10 Ω-μm. This value is about half of that measured using the two-point approach, indicating the presence of considerable contact and/or cable resistance that is avoided by the four-point approach. There is slight curvature on the voltage plot, resulting in a measured drop in resistivity at higher currents. This can be explained by graphite's decrease in resistance as a function of temperature, since the rod heats up as more current is driven through it.

Method Automation

A number of companies make digital multimeters, current sources, and voltage meters with GPIB (general purpose interface bus) connections suitable for test control and automatic data recording. The software package LabView (National Instruments Corp.) is routinely employed to handle such measurements.

Sample Preparation

Making an accurate resistance measurement requires good electrical contacts to the sample. A minimum spacing between voltage and current contacts of 1.5× the cross-sectional perimeter of the sample is recommended (ASTM B193, 1987), and care should be taken to prevent any alternative conductive path (for instance, a solder bridge between a current and a voltage probe). Making good electrical contacts often consists of mechanical contact (i.e., alligator clips or probing with a sharp probe tip) or soldering a wire to a clean surface on the sample. For troublesome surfaces, silver paint or indium (a soft metal) can provide a contacting layer.

Problems

Even with good contacts and a good four-point measurement setup, other factors should be considered. For instance, appreciable power dissipation in the resistive material can result in heating that will affect the conductivity value. Also, some resistive materials are non-ohmic, meaning the current is not a linear function of the voltage. This nonlinearity is most often noticed at high voltage levels. For semiconductors, carriers may be generated by light or by high-frequency radiation. This excess carrier generation can be avoided by testing semiconductors in an adequately shielded dark chamber.

SURFACE MEASUREMENTS

Principles of the Method

The two most common approaches for measuring sheet or surface conductivity are the four-point probe method and the Van der Pauw method. Both approaches are similar to four-point measurements in that current is driven between a pair of probes or connections and the voltage is measured across the other two. The four-point probe method is most often realized by contacting a flat film surface with four equally spaced in-line probes (Valdes, 1954; Hall, 1967). The Van der Pauw method can measure resistivity of small, arbitrarily shaped layers where the four contacts are typically placed around the periphery of the sample.

The four-point probe method, as indicated in Figure 5, has four equally spaced in-line probes with probe tip diameters small compared to the probe spacing, s. An ohmic contact is assumed between the probe tip and the sample. Current is most commonly passed between the outer two probes, and the voltage difference is measured between the two inner probes. Resistivity in a four-point probe measurement is given by

$$\rho = 2\pi s F V/I \tag{5}$$

where F is a correction factor. For placement of probes near the center of a medium of area large relative to the probe spacing, and of a semi-infinite thickness, the correction factor, F, is unity. However, few practical measurements are made on semi-infinitely thick media. The factor F accounts for more realistic thickness and other nonidealities (Combs and Albert, 1963; Albert and Combs, 1964; Albers and Berkowitz, 1985). This factor is itself the product of a factor to correct for sample thickness, F_1, and a factor to account for lateral dimensions of the wafer, F_2. For a sample of conductive layer thickness, t, with a conducting bottom side

$$F_1 = \frac{t/s}{2\ln(\cosh(t/s)/\cosh(t/2s))} \tag{6}$$

If the bottom side is nonconducting,

$$F_1 = \frac{t/s}{2\ln(\sinh(t/s)/\sinh(t/2s))} \tag{7}$$

For a circular wafer of diameter, d, the correction factor F_2 is

$$F_2 = \frac{\ln 2}{\ln 2 + \ln\left[\dfrac{(d/s)^2+3}{(d/s)^2-3}\right]} \tag{8}$$

Note that F_2 approaches unity for $d > 40s$.

Locating the probes closer than four probe spacings from the wafer edge can also result in measurement error. Correction factors to account for edge proximity can be found in Schroder (1990).

The Van der Pauw method can determine the resistivity of small, arbitrarily shaped layers and generally requires less surface area than the four-point probe method (Van der Pauw, 1958). It is often used in integrated circuit processing. The method considers four small contacts placed around the periphery of a homogeneous, uniform thickness (t) sample, as indicated in Figure 6. In this figure, a resistance $R_{ab,cd}$ is determined by driving a current from point a to b and measuring the voltage from point c to d, or

$$R_{ab,cd} = \frac{|V_c - V_d|}{|I_{ab}|} \tag{9}$$

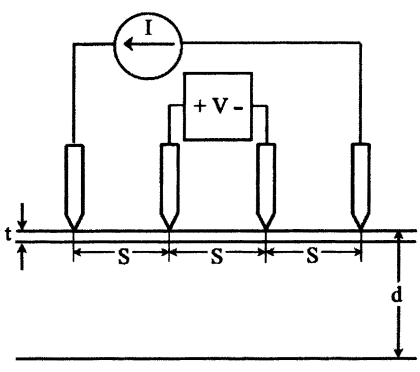


Figure 5. In-line four-point probe measurement of a conductive film of thickness, t, uses a known current source, high-impedance voltmeter, and spring-loaded sharp probes.

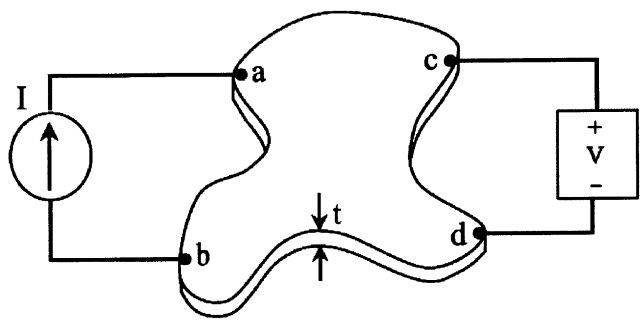


Figure 6. Van der Pauw measurement of an arbitrarily shaped sample uses a known current and a high-impedance voltmeter.

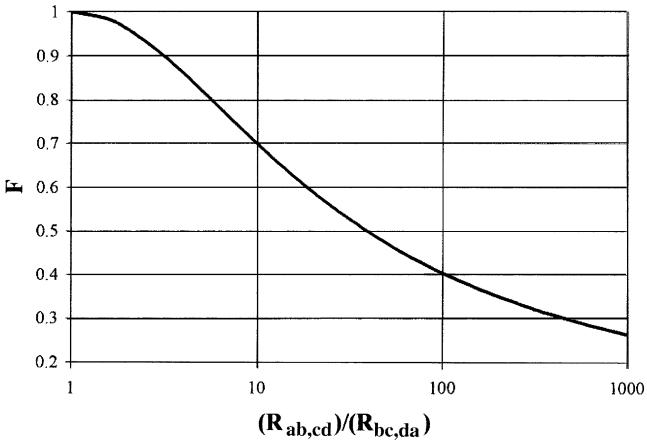


Figure 7. Van der Pauw function, F, plotted versus resistance ratio.

Using a conformal mapping approach Van der Pauw shows that

$$\exp\left[\frac{-\pi R_{\mathrm{ab,cd}} t}{\rho}\right] + \exp\left[\frac{-\pi R_{\mathrm{bc,da}} t}{\rho}\right] = 1 \qquad (10)$$

Solving for resistivity gives

$$\rho = \frac{\pi t}{\ln 2} \frac{R_{\mathrm{ab,cd}} + R_{\mathrm{bc,da}}}{2} F \qquad (11)$$

where the value of F is given by

$$\frac{R_{\mathrm{ab,cd}} - R_{\mathrm{bc,da}}}{R_{\mathrm{ab,cd}} + R_{\mathrm{bc,da}}} = \frac{F}{\ln 2} \operatorname{arccosh}\left[\frac{\exp(\ln 2/F)}{2}\right] \qquad (12)$$

This equation is used to plot F as a function of the resistance ratio $R_{\mathrm{ab,cd}}/R_{\mathrm{bc,da}}$ in Figure 7.

Equation 11 is simplified for 4-fold symmetrical shapes. For instance, the rotation of the structures in Figure 8 by 90° should result in the same measured resistance for a uniform-thickness, homogeneous film with identical contacts. In such a case, $F = 1$ and

$$\rho = \frac{\pi t}{\ln 2} R_{\mathrm{ab,cd}} = 4.532 t R_{\mathrm{ab,cd}} \qquad (13)$$

Practical Aspects of the Method

Like the four-point method used for bulk resistivity measurements, separating the current source from the high-impedance voltage meter avoids errors associated with contact resistance. Also, when considering semiconductor measurements, sufficient separation between the current and voltage probes is required so that minority carriers injected near the current probes recombine before their presence can be felt at the voltage probes. The probes

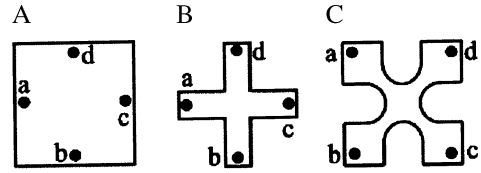


Figure 8. Common Van der Pauw structures: (**A**) square (contacts placed on edges, or placed on each corner), (**B**) "Greek Cross," and (**C**) clover leaf.

themselves have a tightly controlled spacing, typically 1 mm ± 0.01 mm, and the probe tips are individually spring-loaded with a force typically between 1 and 2 N to minimize damage to the film surface during probing. Hard tungsten probe tips are common, but osmium tips are also available, which can make good contact when heated on a semiconductor. A typical measurement range for such a system is 0.001 Ω-cm to 1000 Ω-cm.

In Van der Pauw measurements, it is common to calculate resistivity from two sets of measurements ($R_{\mathrm{ab,cd}}$ and $R_{\mathrm{bc,da}}$). For uniform samples with good contacts, the same results should be measured. The square pattern shown in Figure 8A, or a circular pattern with four points equidistant about the periphery, can be made very small for integrated-circuit film measurements and are often used in practice. However, because of probe alignment difficulties and problems making ideal probe to sample contacts, the "Greek Cross" of Figure 8B or the clover leaf structure of Figure 8C is often used. These shapes isolate the contacts and reduce the error encountered for finite-sized probe tips that are not located all the way out to the periphery of the sample.

Method Automation

As with bulk measurements, four-point probes are routinely connected to current sources and voltage meters for automated measurements. Four-point probing systems typically contain an adjustable constant current source used in conjunction with a high-impedance voltmeter. These systems automatically test and average both forward and reverse signals to reduce errors from thermal effects and rectifying contacts. More elaborate Van der Pauw measurement apparatus will make multiple measurements ($R_{\mathrm{ab,cd}}$ and $R_{\mathrm{bc,da}}$) and average the results.

Problems

The measured surface resistance depends on the doping profile, which affects both carrier concentration and carrier mobility. The conductivity of such a doped semiconductor is therefore a function of depth into the semiconductor. Measuring the surface resistance of a doped semiconductor assumes that the substrate below the junction depth is much more resistive than the layer to be measured. If the substrate is appreciably conductive, then a surface resistance measurement can still be made by forming a reverse-biased diode with the substrate.

NON-CONTACT METHODS

Principles of the Method

Contact-free test methods are useful when electrical contacts are difficult to make or when the sample must not be damaged. Also, a contactless measurement can be much quicker, since the time it takes to make contacts is eliminated. Eddy current approaches and relaxation techniques are often used for non-contact measurement. While such methods are not as accurate as the contact approaches discussed above, errors below 5% are achievable.

The most widely used approach for contactless measurement of conductivity is the eddy current method (ASTM E1004, 1991). In its most straightforward realization, an alternating current-carrying coil is placed close to a conductive sample. The magnetic fields generated by the coil induce circulatory currents, called eddy currents, in the sample which act to oppose the applied magnetic field. Small impedance changes are then measured in the coil as it is loaded by the conductive sample. The measured impedance is a function of sample conductivity, and also of the sample's mechanical structure near the surface. In fact, eddy current testing is routinely used to inspect the mechanical condition of conductive surfaces (Cartz, 1995).

Figure 9 illustrates the general measurement approach. The coil is driven by a sinusoidal current source, and measurement with a high-impedance voltmeter determines the coil impedance, $Z = V/I$. The current through the coil induces eddy currents in the conductive sample which act to oppose the change in magnetic flux. The induced eddy currents are in a plane parallel to a plane containing the coil loops and are a function of sample conductivity, magnetic permeability, and thickness, along with driving coil properties such as number of turns and distance from the sample. The coil itself may also be wrapped around or contained within a ferrite block to enhance the magnetic field.

An equivalent circuit for the simple test setup is shown in Figure 10, where the mutual inductance term M accounts for field coupling into the sample, which appears as a series RL circuit.

It can be shown (Cartz, 1995) that the impedance looking into the circuit is

$$Z = R_T + jX_T = \left[R + \frac{\omega^2 M^2 R'}{R'^2 + \omega^2 L'^2}\right] + j\left[\omega L - \frac{\omega^3 M^2 L'}{R'^2 + \omega^2 L'^2}\right] \quad (14)$$

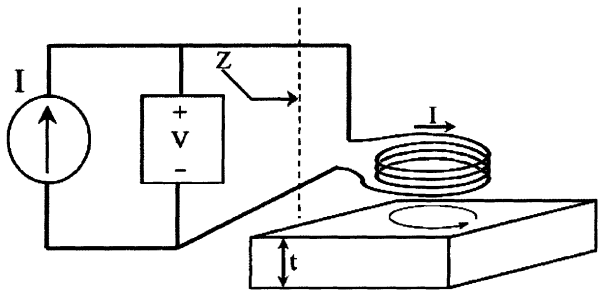


Figure 9. Single coil eddy-current measurement circuit. The impedance, Z, looking into the coil is changed by eddy currents induced in the sample.

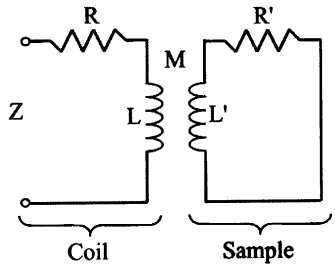


Figure 10. Equivalent circuit for the single coil eddy-current measurement approach.

where R_T and X_T represent the total real and imaginary parts of the impedance, respectively, and j is the imaginary square root of (-1). A plot of X_T versus R_T for a series of samples traces a conductivity curve as shown in Figure 11. This is a comparison method where samples of known conductivity are plotted, and the test sample behavior is then inferred from its location on the chart.

Another non-contact method is the relaxation approach, where a voltage step function applied to a resistive-reactive circuit will result in an exponential response. For instance, when the switch in Figure 12 is opened, the capacitor discharges through the resistor such that

$$V_r(\tau) = V_s \exp(-\tau/RC) \quad (15)$$

When time τ equals one "RC time constant," the voltage has dropped to e^{-1} of its initial value. Most often, this approach is used with a known value of resistance in order to measure the reactive component, but knowing the reactive component allows computation of the resistance. This is illustrated in Figure 13, where the voltage V_R from Figure 12 is initially charged to 1 V, and at time $\tau = 0$ the switch is thrown which allows this voltage to discharge through a 1-μF capacitor. The relaxation for three different resistance values is shown.

This relaxation approach has been used in conjunction with eddy currents where the sample is excited and the fields are measured as they decay (Kos and Fickett, 1994).

Practical Aspects of the Method

The eddy current approach requires two or more calibration standards covering the range of expected sample

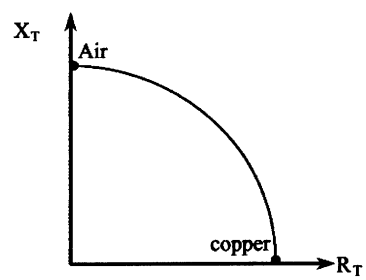


Figure 11. Impedance plane chart showing a typical conductivity curve.

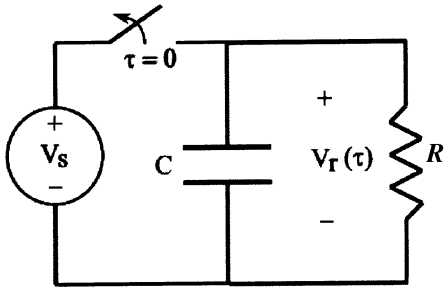


Figure 12. RC network used in relaxation method.

conductivity. Test conditions, such as driving current and frequency, must be the same for both calibration and test samples. The sample thickness and the distance between the coil and the sample surface (termed the "lift-off") must be carefully controlled. Finally, the coil should not be placed within two coil diameters of a sample discontinuity (such as a hole or edge).

In the simple method presented, the driving coil is also used for measurement. Improved eddy-current probes separate the driving and measuring coils, using them to sandwich the test sample. Since very little current is in the measurement coil, eddy currents induced in the sample are essentially all from the driving coil. The measurement coil sees fields from both the driving coil and from the eddy currents in the sample. For best sensitivity, the driving coil fields can be removed from the measurement using a lock-in amplifier circuit (Crowley and Rabson, 1976). Further improvement in the measurement circuit uses an additional set of coils for compensation. Such an approach has been used with miniature radio frequency (RF) coils to provide resistivity mapping with a 4 mm spatial resolution (Chen, 1989). An interesting variation on this method uses a driving coil and a pair of resonant coils to monitor in situ aluminum thin films deposited by chemical vapor deposition (Ermakov and Hinch, 1997).

The magnitude of the eddy currents attenuate by $e^{-y/\delta}$ as they penetrate a depth y into the sample, reaching a value of e^{-1} of the surface value at one skin depth, δ, given by

$$\delta = 1/\sqrt{\pi f \mu \sigma} \tag{16}$$

where f is frequency and μ is sample magnetic permeability. For a single-coil system, the effect of thickness is negligible for samples that are at least 2.6 skin depths thick. Note that for measuring surface defects, or the resistivity near the surface, higher frequencies may be employed. For a system where the measuring coil is on the other side of the sample from the driving coil, samples thinner than a skin depth are required.

The eddy current method is also used to measure conductivity of cylindrical samples. This is done by comparing the impedance of the coil with and without a cylindrical sample inserted (Zimmerman, 1961). Other approaches have relied on mutual inductance measured between a driving coil and measurement coil wrapped around the cylindrical sample (Rosenthal and Maxfield, 1975).

In the relaxation approach, the step function excites all possible resonances, with the one showing the longest decay time being proportional to the sample resistivity. It may be necessary to determine this time constant with a curve-fitting routine after sufficient time has passed for the other resonances to subside. Also, the time constant for the measurement coils must be much shorter than the sample time constant. This approach is most often used for low-resistivity samples (metals with $\rho < 10^{-4}$ ohm-micrometer), since otherwise the time constant is excessive.

A relaxation approach can also find semiconductor resistivity, provided the electrical permittivity is known. Here, a semiconductor wafer is inserted between conductive plates to form a parallel resistive-capacitive circuit. A step voltage is applied to the capacitors, and the current is measured versus time to arrive at an RC time constant equal to the product of the sample resistivity ρ and electrical permittivity, ε, called the "space charge relaxation time." Details of such a procedure used to map the resistivity profile across a gallium arsenide wafer are provided in Stibal et al. (1991).

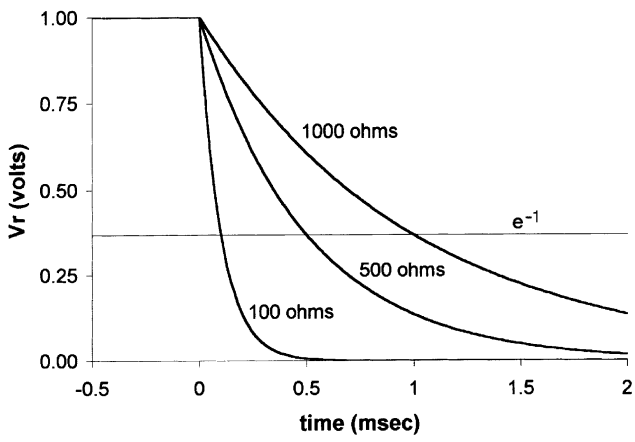


Figure 13. Response of the RC network with a 1-μF capacitor that has been initially charged to 1 V. The relaxation is shown for three values of resistance. The line at 0.368 indicates when the voltage has decayed to e^{-1} of its initial value.

MICROWAVE TECHNIQUES

Principles of the Method

A number of approaches exist to determine dielectric sample conductivity at high frequencies. The primary approaches use open-ended coaxial probes, coaxial or waveguide transmission lines, free-space radiation, and cavity resonators. Scattering parameter measurements are made using vector network analyzers, which are available at considerable expense, to perform measurements up to 10^{11} Hz. These scattering parameters can be related to the sample's electrical permittivity and loss tangent, from which the conductivity can be extracted.

From Maxwell's equations for time harmonic fields

$$\nabla \times \mathbf{H} = (\sigma + j\omega\varepsilon)\mathbf{E} \tag{17}$$

where E and H are the electric and magnetic fields, respectively, and ω is the angular frequency. Here we consider the conductivity σ and the permittivity ε within the dielectric. It is convenient to express Equation 17 as

$$\nabla \times \mathbf{H} = j\omega\varepsilon_c \mathbf{E} \tag{18}$$

where ε_c is a complex permittivity

$$\varepsilon_c = \varepsilon - j\frac{\sigma}{\omega} \tag{19}$$

Plotting the imaginary part of Equation 19 versus the real part reveals where the term loss tangent, tan δ, arises

$$\tan\delta = \frac{\sigma}{\omega\epsilon} \tag{20}$$

Here δ is the angle made with the real axis. Note that this δ has nothing to do with skin depth, which inconveniently shares the same symbol. The term loss tangent is typically applied when discussing dielectric materials, for which a small value is desirable.

Extracting the complex permittivity from scattering parameter data is far from trivial and is almost always performed using software within the network analyzer or in post-processing. In the cavity resonator, waveguide transmission line, and free space approaches, it is also possible to extract the material's complex permeability from the measured scattering parameters (Application Note 1217; see Literature Cited). This can be useful for the study of magnetic materials.

Practical Aspects of the Method

The four general approaches are briefly described below, along with suitable references. Of particular interest are the publications of Bussey (1967) and Baker-Jarvis et al. (1992).

Coaxial probes. In this approach, the open end of a coaxial probe is placed against the flat face of a solid sample as shown in Figure 14A. The probe may also be immersed in a liquid sample. The fringing fields between the center and the outer conductors are affected by the sample material. Complex permittivity is extracted from measurements of the reflection scattering parameter S_{11}. This approach is typically used to measure complex permittivity over a frequency range from 2×10^8 Hz up to 2×10^{10} Hz, and has the advantage that no special sample preparation is required. In addition, stainless steel coaxial probes have performed at temperatures up to 1000°C (Gershon et al., 1999).

Transmission lines. In the various transmission line methods, a material sample is snugly inserted into a section of coaxial transmission line, rectangular waveguide, or circular waveguide (Srivastava and Jain, 1971; Ligthart, 1983; Wolfson and Wentworth, 2000). More than one measurement is required for extraction of complex permittivity data from one port measurements. This can be accomplished by using more than one length of sample or by terminating the guide containing the sample in more than one load. The measured complex reflection coefficients can be manipulated to extract complex permittivity. Two-port measurements can be performed on a single sample where in addition to reflection, the transmission properties are measured. Precise machining of the samples to fit inside the transmission line is critical, which can be somewhat of a disadvantage for the coaxial approach (shown in Figure 14B) where an annular sample is required. An advantage of the coaxial approach is that measurements can be made over a much broader frequency range than is possible with the waveguide procedure. Common problems for the transmission line techniques include air gaps, resonances at sample lengths corresponding to half wavelengths, and excitation of higher-order propagating modes.

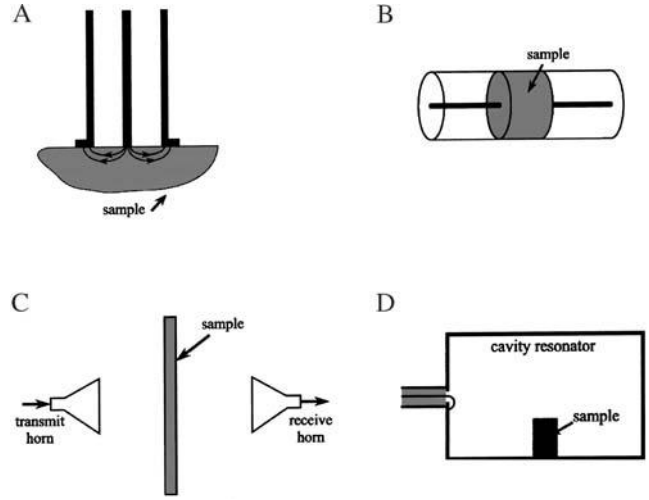


Figure 14. Microwave measurement approaches: (**A**) open-ended coaxial probe in contact with sample, (**B**) sample inserted into coaxial transmission line, (**C**) free-space approach where sample is placed between horn antennas, and (**D**) sample placed within a cavity resonator.

Free-space radiation. Microwaves are passed through material samples using transmitting and receiving antennas as shown in Figure 14C. In addition to measurements through the sample, the transmitting and receiving antennas may be placed at angles on the same side of the sample and a reflection measurement taken. The samples used in free-space techniques must be fairly large, with flat surfaces. Measurement approaches and data processing proceeds as with the transmission line approaches. The free-space approach is a noncontacting method that can be especially useful for making measurements at high temperature or under harsh environmental conditions. It has been employed in measurements ranging from 2×10^9 Hz up to 1.1×10^{11} Hz (Otsuka et al., 1999).

Cavity resonators. Here, a sample of precisely known geometry is placed within a microwave cavity and changes in the resonant frequency and the resonator Q are measured and processed to yield the complex permittivity (Kraszewski and Nelson, 1992). As shown in Figure 14D,

a coaxial probe inserted into the center of one wall of the cavity is used to supply energy for resonance. If the sample dimensions are precisely known, and if calibration standards of the same dimensions are available, then this approach can yield very accurate results for both complex permittivity and complex permeability, although it tends to have limited accuracy for low-loss materials. It has been employed in measurements ranging from 5×10^8 Hz up to 1.1×10^{11} Hz.

ACKNOWLEDGMENT

The author would like to thank Professor Peter A. Barnes for his helpful suggestions on this work.

LITERATURE CITED

Albers, J. and Berkowitz, H. L. 1985. An alternative approach to calculation of four-probe resistances on nonuniform structures. *J. Electrochem. Soc.* 132:2453–2456.

Albert, M. P. and Combs, J. F. 1964. Correction factors for radial resistivity gradient evaluation of semiconductor slices. *IEEE Trans. Electron. Dev.* 11:148–151.

Application Note 1217-1. Basics of measuring the dielectric properties of materials. Hewlett-Packard literature number 5091–3300. Hewlett-Packard, Palo Alto, Calif.

ASTM B193-1987 (revised annually). Standard test method for resistivity of electrical conductor materials. *In* Annual Book of Nondestructive Testing. American Society for Testing and Materials, Philadelphia.

ASTM E1004-1991 (revised annually). Standard test method for electromagnetic (eddy-current) measurement of electrical conductivity. *In* Annual Book of Nondestructive Testing. American Society for Testing and Materials, Philadelphia.

Baker-Jarvis, J., Janezic, M. D., Grosvenor, J. H. Jr., and Geyer, R. G. 1992. Transmission/reflection and short-circuit line methods for measuring permittivity and permeability. NIST Technical Note 1355.

Bussey, H. E. 1967. Measurement of RF properties of materials, a survey. *Proc. IEEE*. 56:1046–1053.

Cartz, L. 1995. Nondestructive Testing, pp. 173–188. American Society for Metals, Materials Park, Ohio.

Chen, M. C. 1989. Sensitive contactless eddy-current conductivity measurements on Si and HgCdTe. *Rev. Sci. Instrum*. 60:1116–1122.

Combs, J. F. and Albert, M. P. 1963. Diameter correction factors for the resistivity measurement of semiconductor slices. *Semic. Prod./Solid State Technol*. 6:26–27.

Coombs, C. F. 1995. Electronic Instrument Handbook, 2nd Ed. McGraw-Hill, New York.

Crowley, J. D. and Rabson, T. A. 1976. Contactless method of measuring resistivity. *Rev. Sci. Instrum*. 47:712–715.

Ermakov, A. V. and Hinch, B. J. 1997. Application of a novel contactless conductivity sensor in chemical vapor deposition of aluminum films. *Rev. Sci. Instrum*. 68:1571–1574.

Gershon, D. L., Calame, J. P., Carmel, Y., Antonsen, T. M., and Hutcheon, R. M. 1999. Open-ended coaxial probe for high temperature and broad-band dielectric measurements. *IEEE Transactions on Microwave Theory and Techniques*. 47:1640–1648.

Hall, R. 1967. Minimizing errors of four-point probe measurements on circular wafers. *J. Sci. Instrum.* 44:53–54.

Kos, A. B. and Fickett, F. R. 1994. Improved eddy-current decay method for resistivity characterization. *IEEE Transactions on Magnetics* 30:4560–4562.

Kraszewski, A. W. and Nelson, S. O. 1992. Observations on resonant cavity perturbation by dielectric objects. *IEEE Transactions on Microwave Theory and Techniques* (40)1:151–155.

Ligthart, L. P. 1983. A fast computational technique for accurate permittivity determination using transmission line methods. *IEEE Transactions on Microwave Theory and Techniques* 31: 249–254.

Otsuka, K., Hashimoto, O., and Ishida, T. 1999. Measurement of complex permittivity of low-loss dielectric material at 94 GHz frequency band using free-space method. *Microwave and Optical Technology Letters* (22)5:291–292.

Rosenthal, M. D. and Maxfield, B. W. 1975. Accurate determination of the electrical resistivity from mutual inductance measurements. *Rev. Sci. Instrum*. 46:398–408.

Schroder, D. K. 1990. Semiconductor Material and Device Characterization pp. 1–40. Wiley Interscience.

Srivastava, G. P. and Jain, A. K. 1971. Conductivity measurements of semiconductors by microwave transmission technique. *The Review of Scientific Instruments* 42:1793–1796.

Stibal, R., Windscheif, J., and Jantz, W. 1991. Contactless evaluation of semi-insulating GaAs wafer resistivity using the time-dependent charge measurement. *Semicond. Sci. Technol*. 6: 995–1001.

Valdes, L. B. 1954. Resistivity measurements on germanium for transistors. *Proceedings of the IRE* 42:420–427.

Van der Pauw, L. J. 1958. A method of measuring specific resistivity and Hall effect of discs of arbitrary shape. *Phil. Res. Rep*. 13: 1–9.

Witte, R. A. 1993. Electronic Test Instruments: Theory and Applications, pp. 59–63. Prentice-Hall, Englewood Cliffs N.J.

Wolfson, B. and Wentworth, S. 2000. Complex permittivity and permeability measurement using rectangular waveguide. *Microwave and Optical Technology Letters* 27:180–182.

Zimmerman, J. E. 1961. Measurement of Electrical Resistivity of Bulk Metals. *Rev. Sci. Instrum*. 32:402–405.

KEY REFERENCES

Baker-Jarvis et al., 1992. See above.

An extremely thorough reference for coaxial and rectangular waveguide transmission line techniques for measuring electrical permittivity and magnetic permeability.

Cartz, 1995. See above.

Presents eddy-current testing from a material inspection point of view.

Coombs, 1995. See above

This reference goes into considerable detail on bulk measurements and gives much practical information on measurement instrumentation.

Heaney, M. B. 1999. Electrical conductivity and resistivity. *In* The Measurement, Instrumentation, and Sensors Handbook, Vol. 43 (J.G. Webster, ed.) pp. 1–14. CRC Press.

Gives practical advice for making two point, four point, four-point probe and Van der Pauw measurements. It cites equipment for conducting measurements and discusses common experimental errors.

Schroder et al., 1990. See above.

This gives a thorough discussion of conductivity measurements as they are applied to semiconductors (four-point probing, Van der Pauw methods).

STUART M. WENTWORTH
Auburn University
Auburn, Alabama

HALL EFFECT IN SEMICONDUCTORS

INTRODUCTION

The Hall effect, which was discovered in 1879 (Hall, 1879), determines the concentration and type (negative or positive) of charge carriers in metals, semiconductors, or insulators. In general, the method is used in conjunction with a conductivity measurement to also determine the mobility (ease of movement) of the charge carriers. At low temperatures and high magnetic fields, quantum effects are sometimes evident in lower dimensional structures; however, such effects will not be considered here. Also, this unit will concentrate on semiconductors, rather than metals or insulators, although the same theory generally applies (Gantmakker and Levinson, 1987).

Three, strong advantages of Hall effect measurements are ease of instrumentation, ease of interpretation, and wide dynamic range. With respect to implementation, the only elements necessary, at the lowest level, are a current source, a voltmeter, and a modest-sized magnet. The carrier concentration can then be calculated within a typical accuracy of 20% without any other information about the material. In our laboratory, we have measured concentrations ranging from 10^4 to 10^{20} cm^{-3}. Also, the type (n or p) can be unambiguously determined from the sign of the Hall voltage.

Competing techniques include capacitance-voltage (C-V) measurements to determine carrier concentration (see CAPACITANCE–VOLTAGE (C-V) CHARACTERIZATION OF SEMICONDUCTORS); thermoelectric probe (TEP) measurements to determine carrier type; and magnetoresistance (MR) measurements to determine mobility (Look, 1989). The C-V measurements have an advantage of depth profile information but a disadvantage of requiring a Schottky barrier. The TEP measurements require a temperature gradient to be imposed and are often ambiguous in the final conclusions. The MR measurements require a high mobility μ or high magnetic field strength B to implement, because the signal varies as μ^2B^2 rather than μB, as in the Hall effect. Further comparisons of these techniques, as well as others, can be found in monographs by Runyan (1975), Look (1989), and Schroder (1990).

PRINCIPLES OF THE METHOD

A phenomenological equation of motion for electrons of charge $-e$ moving with velocity $\mathbf{v}$ in the presence of electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$ is

$$m^*\dot{\mathbf{v}} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - m^*\frac{\mathbf{v} - \mathbf{v}_{eq}}{\tau} \tag{1}$$

where m^* is the effective mass, $\mathbf{v}_{eq}$ is the velocity at equilibrium (steady state), and τ is the velocity or (momentum) relaxation time, i.e., the time in which oscillatory phase information is lost through collisions. Consider a rectangular sample, as shown in Figure 1, with an external electric field $\mathbf{E}_{ex} = E_x\mathbf{x}$ and magnetic field $\mathbf{B} = B_z\mathbf{z}$. Then, if no current is allowed to flow in the y direction (i.e., $v_y = 0$), the steady-state condition $\dot{\mathbf{v}} = 0$ requires that $E_y = -v_xB_z$, and E_y is known as the Hall field. For electron concentration n, the current density $j_x = nev_x$, and thus, $E_y = -j_xB_z/en \equiv -j_xB_zR_H$, where $R_H = -1/en$ is the Hall coefficient. Thus, simple measurements of the quantities E_y, j_x, and B_z yield a very important quantity, n, although a more detailed analysis given below slightly modifies this relationship.

The above analysis assumes that all electrons are moving with the same velocity $\mathbf{v}$ (constant τ), which is not true in a semiconductor. To relieve this constraint, we note that Equation 1 consists of three, coupled differential equations (in v_x, v_y, v_z) that can be solved by standard techniques. After averaging over energy, the steady-state currents can then be shown to be

$$j_x = \sigma_{xx}E + \sigma_{xy}E_y \tag{2}$$

$$j_y = \sigma_{yx}E_x + \sigma_{yy}E_y \tag{3}$$

where the σ_{ij} are elements of the conductivity tensor, defined by

$$\sigma_{xx} = \sigma_{yy} = \left\langle \frac{\tau}{1 + \omega_c^2\tau^2} \right\rangle \tag{4}$$

$$\sigma_{xy} = -\sigma_{yx} = -\left\langle \frac{\omega_c\tau^2}{1 + \omega_c^2\tau^2} \right\rangle \tag{5}$$

Here $\omega_c = eB/m^*$ is the cyclotronic frequency, where B is the magnitude of $\mathbf{B}$, and the brackets denote an average over energy ε taken as follows:

$$\langle F(\varepsilon)\rangle = \frac{\int_0^\infty F(\varepsilon)\varepsilon^{3/2}\frac{\partial f_0}{\partial \varepsilon}d\varepsilon}{\int_0^\infty \varepsilon^{3/2}\frac{\partial f_0}{\partial \varepsilon}d\varepsilon} \rightarrow \frac{\int_0^\infty F(\varepsilon)\varepsilon^{3/2}e^{-\varepsilon/kT}d\varepsilon}{\int_0^\infty \varepsilon^{3/2}e^{-\varepsilon/kT}d\varepsilon} \tag{6}$$

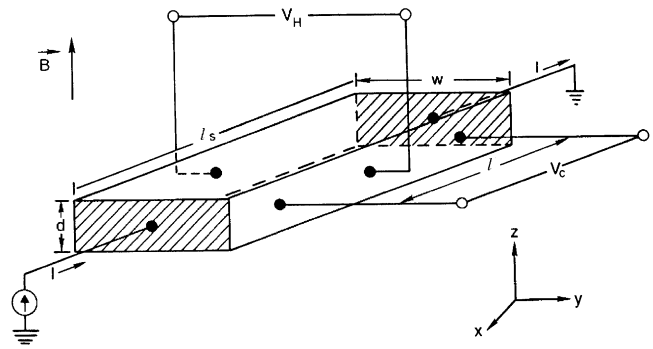


Figure 1. Hall bar configuration for resistivity and Hall effect measurements.

where $F(\varepsilon)$ is any function of ε and f_0 is the Fermi-Dirac distribution function. The second equality in Equation 6 holds for nondegenerate electrons, i.e., those describable by a Boltzmann distribution function. For small magnetic fields, i.e., $\omega_c\tau \ll 1$, and under the usual constraint $j_y=0$, it is easy to show that

$$j_x = \frac{ne^2\langle\tau\rangle}{m^*}E_x \equiv -ne\mu_c E_x \tag{7}$$

$$R_H = \frac{E_y}{j_x B} = -\frac{1}{ne}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} = -\frac{r}{en} \tag{8}$$

where r is the "Hall factor" and $\mu_c = e\langle\tau\rangle/m^*$ is known as the "conductivity" mobility, since the quantity $ne\mu_c$ is just the conductivity σ. We define the "Hall" mobility as $\mu_H = R_H\sigma = r\mu_c$ and the "Hall" concentration as $n_H = n/r = -1/eR_H$. Thus, a combined Hall effect and conductivity measurement gives n_H and μ_H, not n and μ_c; fortunately, however, r is usually within 20% of unity and almost never as large as 2. In any case, r can often be calculated or measured $[r = R_H(\mu_B \ll 1)/R_H(\mu_B \gg 1)]$ so that an accurate value of n can usually be determined.

The relaxation time $\tau(\varepsilon)$ depends on how the electrons interact with the lattice vibrations as well as with extrinsic elements, such as charged impurities and defects. For example, acoustical-mode lattice vibrations scatter electrons through the deformation potential (leading to a relaxation time τ_{ac}) and piezoelectric potential (τ_{pe}); optical-mode vibrations through the polar potential (τ_{po}); ionized impurities and defects through the screened coulomb potential (τ_{ii}); and charged dislocations, also through the coulomb potential (τ_{dis}). The strengths of these various scattering mechanisms depend upon certain lattice parameters, such as dielectric constants and deformation potentials, and extrinsic factors, such as donor, acceptor, and dislocation concentrations N_D, N_A, and N_{dis}, respectively (Rode, 1975; Wiley, 1975; Nag, 1980; Look, 1989). The total momentum scattering rate, or inverse relaxation time, is

$$\tau^{-1}(\varepsilon) = \tau_{ac}^{-1}(\varepsilon) + \tau_{pe}^{-1}(\varepsilon) + \tau_{po}^{-1}(\varepsilon) + \tau_{ii}^{-1}(\varepsilon) + \tau_{dis}^{-1}(\varepsilon) \tag{9}$$

This expression is then used to determine $\langle\tau(\varepsilon)\rangle$ and $\langle\tau^2(\varepsilon)\rangle$ via Equation 6, and thence, $\mu_H = e\langle\tau^2\rangle/m^*\langle\tau\rangle$. Since μ_H is a function of temperature T, a fit of μ_H vs. T can often be used to determine N_D, N_A, and N_{dis}. The μ_H-vs.-T curve should be solved simultaneously with the n-vs.-T curve, given by the charge-balance equation (CBE)

$$n + N_A = \frac{N_D}{1 + n/\phi_D} \tag{10a}$$

where $\phi_D = (g_0/g_1)N_C' \exp(\alpha_D/k)T^{3/2}\exp(-E_{D0}/kT)$. Here, g_0/g_1 is a degeneracy factor ($= 1/2$ for an s-state), $N_C' = 2(2\pi m_n^* k)^{3/2}/h^3$, where h is Planck's constant, E_D is the donor energy, k is Boltzmann's constant, and E_{D0} and α_D are defined by $E_D = E_{D0} - \alpha_D T$. If more than one donor exists within a few kT of the Fermi energy, then equivalent terms are added on the right-hand side of Equation 10a.

For a p-type sample, we use the nearly equivalent equation

$$p + N_D = \frac{N_A}{1 + p/\phi_A} \tag{10b}$$

where p is the hole concentration and $\phi_A = (g_1/g_0)N_V' \exp(\alpha_A/k)T^{3/2}\exp(-E_{A0}/kT)$, and where $N_V' = 2(2\pi m_p^* k)^{3/2}/h^3$ and $E_A = E_{A0} - \alpha_A T$.

The mobility analysis described above, known as the relaxation time approximation (RTA), is limited to elastic (energy-conserving) scattering processes, because for these, a relaxation time $\tau(\varepsilon)$ can be well defined. Unfortunately, this is usually not the case for polar optical-mode (po) scattering, since the energy exchanged in the po scattering event is typically $\gg kT$. However, we can often approximate τ_{po} by an analytic formula, and, in any case, at low temperatures, po scattering is not very important. Nevertheless, for the most accurate calculations, the Boltzmann transport equation (BTE) should be solved directly, as discussed in several references (Rode, 1975; Nag, 1980; Look, 1989).

Hall samples do not have to be rectangular, such as that shown in Figure 1; in fact, we will discuss arbitrarily shaped specimens below (see Practical Aspects of the Method). However, the above analysis does assume that n and μ are homogeneous throughout the sample. If n and μ vary with depth z only, then the measured quantities are

$$\sigma_{sq} = \int_0^d \sigma(z)\,dz = e\int_0^d n(z)\mu(z)\,dz \tag{11}$$

$$R_{Hsq}\sigma_{sq}^2 = \int_0^d n(z)\mu^2(z)\,dz \tag{12}$$

where d is the sample thickness and the subscript "sq" denotes a sheet (areal) quantity (in reciprocal centimeters squared) rather than a volume quantity (in reciprocal cubic centimeters). If some of the carriers are holes rather than electrons, then the sign of e for those carriers must be reversed. The general convention is that R_H is negative for electrons and positive for holes. In some cases, the hole and electron contributions to $R_{Hsq}^2\sigma_{sq}$ exactly balance at a given temperature, and this quantity vanishes.

PRACTICAL ASPECTS OF THE METHOD

Useful Equations

The Hall bar structure of Figure 1 is analyzed as follows: $E_x = V_c/l$, $E_y = V_H/w$, and $j_x = I/wd$, so that

$$\sigma = \rho^{-1} = \frac{j_x}{E_x} = \frac{Il}{V_c wd} \tag{13}$$

$$R_H = \frac{E_y}{j_x B} = \frac{V_H d}{IB} \tag{14}$$

$$\mu_H = R\sigma = \frac{V_H l}{V_c wB} \tag{15}$$

$$n_H = (eR)^{-1} \tag{16}$$

In the meter-kilogram-second (mks) system, current I is in amperes, voltage V in volts, magnetic field strength B in tesla, and length l, width w, and thickness d in meters. By realizing that $1\ \mathrm{T} = 1\ \mathrm{V \cdot s/m^2}$, $1\ \mathrm{A} = 1\ \mathrm{C/s}$, and $1\ \Omega = 1\ \mathrm{V/A}$, we find that σ is in units of $\Omega^{-1}\mathrm{m}^{-1}$, R_H in $\mathrm{m^3/C}$, μ_H in $\mathrm{m^2/V \cdot s}$, and n_H in $\mathrm{m^{-3}}$. However, it is more common to denote σ in $\Omega^{-1}\ \mathrm{cm^{-1}}$, R_H in $\mathrm{cm^3/C}$, μ_H in $\mathrm{cm^2/V \cdot s}$, and n_H in $\mathrm{cm^{-3}}$, with obvious conversion factors ($1\ \mathrm{m} = 10^2\ \mathrm{cm}$). Since B is often quoted in gauss, it is useful to note that $1\ \mathrm{T} = 10^4\ \mathrm{G}$.

Although the Hall bar configuration discussed above is the simplest and most straightforward geometry to analyze, it is not the most popular shape in use today. The reason stems from a very convenient formulation by van der Pauw (1958) in which he solved the potential problem for a thin layer of arbitrary shape. A convenient feature of the van der Pauw technique is that no dimension need be measured for the calculation of sheet resistance or sheet carrier concentration, although a thickness must of course be known for volume resistivity and concentration. Basically, the validity of the van der Pauw method requires that the sample be flat, homogeneous, isotropic, and a singly connected domain (no holes) and have line electrodes on the periphery, projecting to point contacts on the surface, or have true point contacts on the surface. The last requirement is the most difficult to satisfy, so that much work has gone into determining the effects of finite contact size.

Consider the sample shown in Figure 2A. Here, a current I flows between contacts 1 and 2, and a voltage V_c is measured between contacts 3 and 4. Let resistance $R_{ij,kl} \equiv V_\mathrm{kl}/I_\mathrm{ij}$, where the current enters contact i and leaves contact j and $V_{kl} = V_k - V_l$. [These definitions, as well as the contact numbering, correspond to ASTM (1988) Standard F76.] The resistivity ρ, with $B = 0$, is then calculated as

$$\rho = \frac{\pi d}{\ln(2)}\left(\frac{R_{21,34} + R_{32,41}}{2}\right) f \tag{17}$$

where f is determined from a transcendental equation:

$$\frac{Q-1}{Q+1} = \frac{f}{\ln(2)}\operatorname{arccosh}\left[\frac{1}{2}\exp\left(\frac{\ln(2)}{f}\right)\right] \tag{18}$$

Here, $Q = R_{21,34}/R_{32,41}$ if this ratio is greater than unity; otherwise, $Q = R_{32,41}/R_{21,34}$. A curve of f vs. Q accurate to $\approx 2\%$ is presented in Figure 3 (van der Pauw, 1958). Also useful is a somewhat simpler analytical procedure

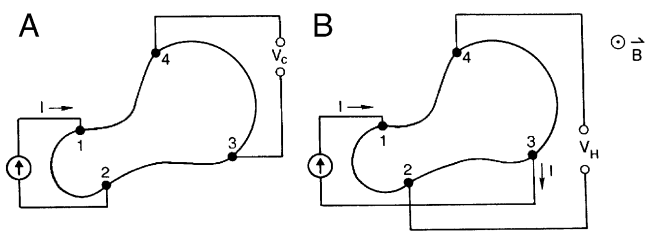


Figure 2. Arbitrary shape for van der Pauw measurements: (**A**) resistivity; (**B**) Hall effect.

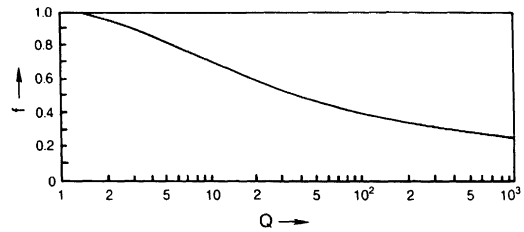


Figure 3. Resistivity ratio function used to correct the van der Pauw results for asymmetric sample shape.

for determining f due to Wasscher and reprinted in Weider (1979). First calculate α from

$$Q = \frac{\ln(1/2 - \alpha)}{\ln(1/2 + \alpha)} \tag{19}$$

and then calculate f from

$$f = \frac{\ln(1/4)}{\ln(1/2 + \alpha) + \ln(1/2 - \alpha)} \tag{20}$$

Here, it is of course required that $-1/2 < \alpha < 1/2$, but this range of α covers $Q = 0$ to $Q = \infty$. For example, a ratio $Q = 4.8$ gives a value $\alpha \approx 0.25$, and then $f \approx 0.83$. Thus, the ratio must be fairly large before ρ is appreciably reduced.

It is useful to further average ρ by including the remaining two contact permutations, and also reversing current for all four permutations. Then

$$\rho = \frac{\pi d}{\ln(2)}\left[\frac{(R_{21,34} - R_{12,34} + R_{32,41} - R_{23,41})f_\mathrm{A}}{8} + \frac{(R_{43,12} - R_{34,12} + R_{14,23} - R_{41,23})f_\mathrm{B}}{8}\right] \tag{21}$$

where f_A and f_B are determined from Q_A and Q_B, respectively, by applying either Equation 18 or 19. Here,

$$Q_\mathrm{A} = \frac{R_{21,34} - R_{12,34}}{R_{32,41} - R_{23,41}} \tag{22}$$

$$Q_\mathrm{B} = \frac{R_{43,12} - R_{34,12}}{R_{14,23} - R_{41,23}} \tag{23}$$

The Hall mobility is determined by using the configuration of Figure 2B, in which the current and voltage contacts are crossed. The Hall coefficient becomes

$$R_\mathrm{H} = \frac{d}{B}\left(\frac{R_{31,42} + R_{42,13}}{2}\right) \tag{24}$$

In general, to minimize magnetoresistive and other effects, it is useful to average over current and magnetic field polarities. Then

$$R_\mathrm{H} = (d/B)[R_{31,42}(+B) - R_{13,42}(+B) + R_{42,13}(+B) - R_{24,13}(+B) + R_{13,42}(-B) - R_{31,42}(-B) + R_{24,13}(-B) - R_{42,13}(-B)]/8 \tag{25}$$

Sensitivity

The most common magnetic field strength used by researchers for Hall effect measurements in 5 kG = 0.5 T. Also, typical semiconductor materials (e.g., GaAs) vary in resistivity ρ from 10^{-4} to 10^{9} $\Omega \cdot$ cm, depending on intrinsic factors as well as impurity and defect concentrations. Finally, typical sample thicknesses d may range from 10^{-5} to 10^{-1} cm. Consider the worst case on the high-resistivity end of the spectrum. Let $\rho = 10^{9}$ $\Omega \cdot$ cm, $d = 10^{-5}$ cm, and $l/w \approx 1$. Then, if a voltage of 100 V can be applied, the current will be 1 pA, which can be measured with an electrometer. If the mobility $\mu_H \approx 10^{3}$ cm^2/V $\cdot$ s, then the Hall coefficient R will be 10^{12} cm^3/C, and the Hall voltage V_H will be 5 V. Conversely, let $\rho = 10^{-4}$ $\Omega \cdot$ cm and $d = 10^{-1}$ cm. Then, the resistance is about 10^{-3} Ω, and we will not be able to apply more than about 10^{-3} V in order to avoid sample heating. Such a voltage will produce 1 A of current and a resulting Hall voltage of 50 μV, again easily measurable. Of course, if $\mu H \approx 1$ cm^2/V $\cdot$ s, then $V_H \approx 50$ nV, which can still probably be measured if noise levels are low. We may note that the lowest mobility that we have measured in our laboratory, by using our standard dc techniques, is 0.1 cm^2/V s.

METHOD AUTOMATION

A basic design for a high-impedance, automated van der Pauw–Hall apparatus, which can accommodate samples of up to 10^{12} Ω is given in Figure 4. All components are commercially available, and the electrometers can, of course, be replaced by high-impedance, unity-gain buffer amplifiers, although possibly at some sacrifice in input impedance. The inputs and outputs of the high-impedance scanner and the inputs to the ammeter and electrometers are of triaxial design, which has the advantage that the inner shields can be driven by the unity-gain output of the electrometers, which effectively reduces cable-charging effects. The outer shields are all grounded at a common point, although in practice the grounding may not be critical. The current source should have an effective output impedance of about 10^{12} Ω and be able to regulate currents of 10^{-10} A. At higher resistance levels, the current regulation will cross over to voltage regulation, and the currents will diminish; however, the data will still be accurate as long as the actual current and voltage are both measured. The low-impedance scanner should be mainly designed for low thermal offsets (say, a few microvolts) since current leakage is not a problem on this side of the electrometers. Note that although the ammeter needs to be of electrometer design, to be able to measure the very low currents, the voltmeter does not, since the electrometer output impedances are only a few kilohms.

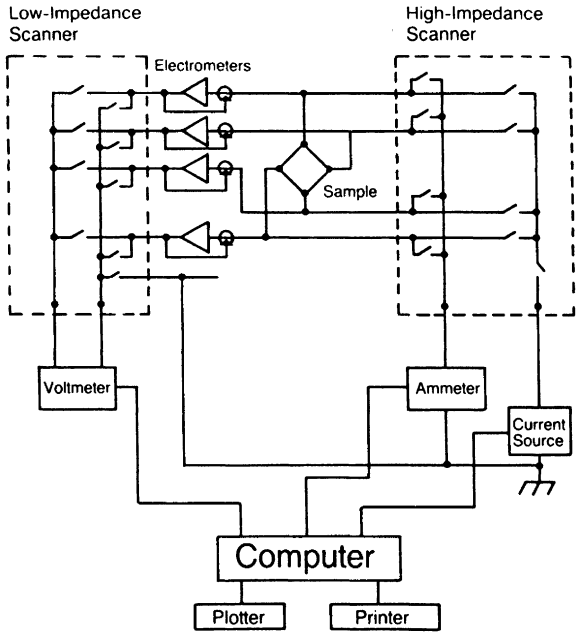


Figure 4. Schematic of an automated, high-impedance Hall effect apparatus.

The wiring to the high-impedance scanner in this design allows the current source and current sink to be individually applied to any desired sample contact. Also, the high input and low input on the voltmeter can be connected to any electrometer unity-gain output through the low-impedance scanner. Figure 4 does not represent the most efficient design in terms of the minimum number of scanner switches, but it does illustrate the main ideas quite well. Although not shown here, for very low resistance samples it is desirable, and sometimes necessary, to be able to bypass the electrometers, since their typical noise levels of 10-μV peak to peak may be unacceptably high.

We will not discuss the computer and peripherals in any detail here, because so many variations available today will work well. However, we would recommend that the system be designed around the standard IEEE-488 interface bus, which allows new instruments to be added at will and effectively eliminates serious hardware problems.

Complete Hall systems, including temperature and magnetic field control, are available from several companies, including Bio-Rad Semiconductor, Keithley Instruments, Lake Shore Cryotronics, and MMR Technologies.

DATA ANALYSIS AND INITIAL INTERPRETATION

The primary quantities determined from Hall effect and conductivity measurements are the Hall carrier concentration n_H or p_H and mobility μH. As already discussed, $n_H = -1/eR_H$, where R_H is given by Equation 24 (for a van der Pauw configuration) and $\mu_H = R_H\sigma = R_H/\rho$, where ρ is given by Equation 21. Although simple 300-K values of ρ, n_H, and μ_H are quite important and widely used, it is in temperature-dependent Hall (TDH) measurements that the real power of the Hall technique is demonstrated, because then the donor and acceptor concentrations and energies can be determined. We will illustrate the methodology with a two-layer GaN problem.

The GaN sample discussed here was a square (6 $\times$ 6-mm) layer grown on sapphire to a thickness $d = 20$ μm.

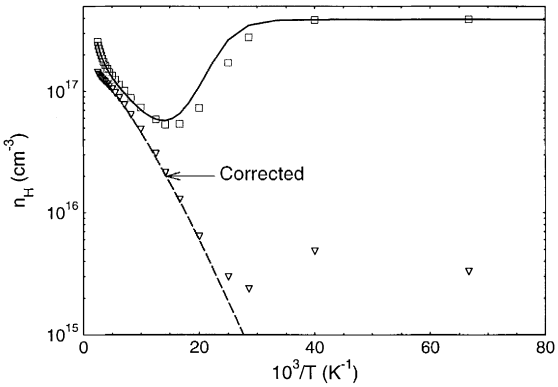


Figure 5. Uncorrected Hall concentration data (squares) and fit (solid line) and corrected data (triangles) and fit (dashed line) vs. inverse temperature.

Small indium dots were soldered on the corners to provide ohmic contacts, and the Hall measurements were carried out in an apparatus similar to that illustrated in Figure 4. Temperature control was achieved by using a He exchange-gas dewar. The temperature dependences of n_H and μ_H are shown by the squares in Figures 5 and 6, respectively. At temperatures below 30 K (i.e., $10^3/T > 33$), the carriers (electrons) in the main part of the layer "freeze out" on their parent donors and thus are no longer available for conduction. However, this sample had a very thin, strongly n-type layer between the sapphire and GaN, and the carriers in such a layer do not freeze out and, in fact, have a temperature-independent concentration and mobility, as seen at low T (high $10^3/T$) in Figures 5 and 6, respectively. Thus, we need to use the depth-profile analysis given by Equations 11 and 12.

For two layers, Equations 11 and 12 give

$$\mu_H = R_{sq}\sigma_{sq} = \frac{R_{sq}\sigma_{sq}^2/d}{\sigma_{sq}/d} = \frac{\mu_{H1}^2 n_{H1} + \mu_{H2}^2 n_{Hsq2}/d}{\mu_{H1} n_{H1} + \mu_{H2} n_{Hsq2}/d} \tag{26}$$

$$n_H = \frac{n_{Hsq}}{d} = \frac{1}{eR_{sq}d} = \frac{\sigma_{sq}^2/d^2}{eR_{sq}\sigma_{sq}^2/d} = \frac{(\mu_{H1} n_{H1} + \mu_{H2} n_{Hsq2}/d)^2}{\mu_{H1}^2 n_{H1} + \mu_{H2}^2 n_{Hsq2}/d} \tag{27}$$

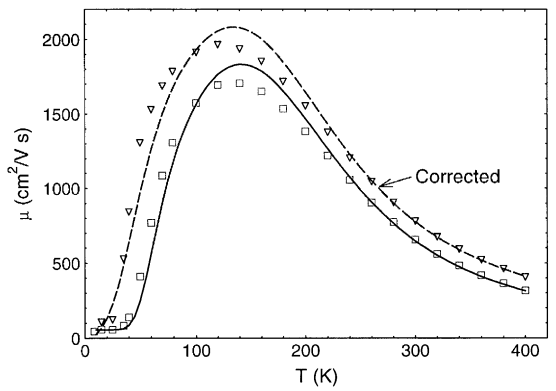


Figure 6. Uncorrected Hall mobility data (squares) and fit (solid line) and corrected data (triangles) and fit (dashed line) vs. temperature.

where layer 1 is the main, 20-μm-thick GaN growth and layer 2 is the very thin interface region. Since we do not know the thickness of layer 2, we simply normalize to thickness $d = 20$ μm for plotting purposes. From the figures, we get $\mu_{H2} = 55$ cm^2/V · s and $n_{Hsq2}/d = 3.9 \times 10^{17}$ cm^{-3}. Because these values are constant, we can invert Equations 26 and 27 and solve for μ_{H1} and n_{H1} at each temperature. (The resulting equations are given later.) To fit the uncorrected data (squares), we parametrize n vs. T by Equation 10a and μ_H vs. T by $\mu_H = e\langle\tau^2\rangle/m^*\langle\tau\rangle$, where τ is given by Equation 9. Because $n = n_H r = n_H\langle\tau^2\rangle/\langle\tau\rangle^2$ the fits of n_H vs. T and μ_H vs. T must be carried out simultaneously. In this case r varies only from about 1.2 to 1.4 as a function of T.

Formulas for $\tau_{ac}, \tau_{pe}, \tau_{po}$, and τ_{ii} are given below. For ionized impurity (or defect) scattering in a nondegenerate, n-type material,

$$\tau_{ii}(\varepsilon) = \frac{2^{9/2}\pi\varepsilon_0^2 (m^*)^{1/2}\varepsilon^{3/2}}{e^4(2N_A + n)[\ln(1+y) - y/(1+y)]} \tag{28}$$

where $y = 8\varepsilon_0 m^* kT\varepsilon/\hbar^2 e^2 n$. Here ε_0 is the low-frequency (static) dielectric constant. [If the sample is p type, let $(2N_A + n) \rightarrow (2N_D + p)$]. For acoustic-mode deformation-potential scattering,

$$\tau_{ac}(\varepsilon) = \frac{\pi\hbar^4\rho_d s^2\varepsilon^{-1/2}}{2^{1/2}E_1^2 (m^*)^{3/2} kT} \tag{29}$$

where ρ_d is the density, s is the speed of sound, and E_1 is the deformation potential. For acoustic-mode piezoelectric-potential scattering,

$$\tau_{pe}(\varepsilon) = \frac{2^{2/3}\pi\hbar^2\varepsilon_0\varepsilon^{1/2}}{e^2P^2(m^*)^{1/2}kT} \tag{30}$$

where P is the piezoelectric coupling coefficient [$P = (h_{pz}^2/\rho s^2\varepsilon_0)^{1/2}$]. Finally, for polar optical-mode scattering, a rough approximation can be given

$$\tau_{po}(\varepsilon) = (C_{po}2^{3/2}\pi\hbar^2(e^{T_{po}/T} - 1)[0.762\varepsilon^{1/2} + 0.824(kT_{po})^{1/2} - 0.235(kT_{po})^{-1/2}\varepsilon])/[e^2 kT_{po}(m^*)^{1/2}(\varepsilon_\infty^{-1} - \varepsilon_0^{-1})] \tag{31}$$

where T_{po} is the Debye temperature, ε_∞ is the high-frequency dielectric constant, and C_{po} is a fitting parameter, of order unity, to correct for the inexact nature of $\tau_{po}(\varepsilon)$. That is, if we had *only* po scattering, then the exact (i.e., BTE) calculation of μH vs. T would be almost identical with the RTA calculation (i.e., $\mu_H = e\langle\tau^2\rangle/m*\langle\tau\rangle$), with $C_{po} = 1$. However, if other scattering mechanisms are also important, then the correction factor C_{po} will be dependent on the relative strengths of these other mechanisms. As an example, to get a good fit to high-temperature (>300 K) μ_H-vs.-T data in GaN (Fig. 6), we use $C_{po} \approx 0.6$. Fortunately, below 150 K, the po mechanism

is no longer important in GaN, and the RTA approach is quite accurate.

The RTA analysis discussed above and the CBE analysis discussed earlier (Equation 10a or 10b) constitute a powerful method for the determination of N_D, N_A, and E_D in semiconductor material. It can be easily set up on a personal computer, using, e.g., MATHCAD software. In many cases, it is sufficient to simply assume $n = n_H$ (i.e., $r = 1$) in Equations 10a and 28, but a more accurate answer can be obtained by using the following steps: (1) let $n = n_H = 1/eR_H$ at each T; (2) fit μ_H vs. T using $\mu_H = \langle\tau^2\rangle/m^*\langle\tau\rangle$, and get a value for N_A; (3) calculate $r = \langle\tau^2\rangle/\langle\tau\rangle^2$ at each T; (4) calculate a new $n = rn_H$ at each T; and (5) fit n vs. T to Equation 10a and get values of N_D and E_D. Further iterations can be carried out, if desired, but usually add little accuracy.

For the benefit of the reader who wishes to set up such an analysis, we give the scattering strength parameters and fitting parameters used for the GaN data fits in Figures 5 and 6. Further discussion can be found in the work of Look and Molnar (1997). From the GaN literature we find: $E_1 = 9.2$ eV $= 1.47 \times 10^{-18}$ J; $P = 0.104$, $\varepsilon_0 = 10.4(8.8542 \times 10^{-12})$ F/m; $\varepsilon_\infty = 5.47(8.8542 \times 10^{-12})$ F/m; $T_{po} = 1044$ K; $m^* = 0.22(9.1095 \times 10^{-31})$ kg; $\rho_d = 6.10 \times 10^3$ kg/m^3; $s = 6.59 \times 10^3$ m/s; $g_0 = 1$; $g_1 = 2$; $\alpha_D = 0$; and $N'_C = 4.98 \times 10^{20}$ m^{-3}. With all the parameters in mks units, μH is in units of m^2/V$\cdot$ s; a useful conversion is μ_H (cm^2/V$\cdot$ s) $= 10^4$ μ_H (m^2/V $\cdot$ s). To fit the high-T μ_H data, we also have to set $C_{po} = 0.56$, as mentioned earlier. However, before carrying out the RTA and CBE analyses, it was necessary in this case to correct for a degenerate interface layer, having $\mu_{H2} = 55$ cm^2/V$\cdot$ s and $n_{Hsq2}/d = 3.9 \times 10^{17}$ cm^{-3}. The corrected data, also shown in Figures 5 and 6, are calculated by inverting Equations 26 and 27:

$$\mu_{H1} = \frac{\mu_H^2 n_H - \mu_{H2}^2 n_{Hsq2}/d}{\mu_H n_H - \mu_{H2} n_{Hsq2}/d} \tag{32}$$

$$n_{H1} = \frac{(\mu_H n_H - \mu_{H2} n_{Hsq2}/d)^2}{\mu_H^2 n_H - \mu_{H2}^2 n_{Hsq2}/d} \tag{33}$$

The fitted parameters are $N_D = 2.1 \times 10^{17}$ cm^{-3}, $N_A = 5 \times 10^{16}$ cm^{-3}, and $E_D = 16$ meV.

SAMPLE PREPARATION

In deciding what kind of geometrical structure to use for a particular application, several factors should be considered, including (1) available size, (2) available fabrication techniques, (3) limitations of measurement time, (4) necessary accuracy, and (5) need for magnetoresistance data.

In considering these factors, we will refer to Figure 7, which depicts six of the most popular structures.

The available size can be a severe constraint. In our laboratory, we have often measured bulk samples of dimension 2 mm or less, which virtually rules out any complicated shapes, such as (B), (D), or (F) in the figure; in fact, it is sometimes not possible to modify an existing sample shape at all. Thus, the best procedure for small samples

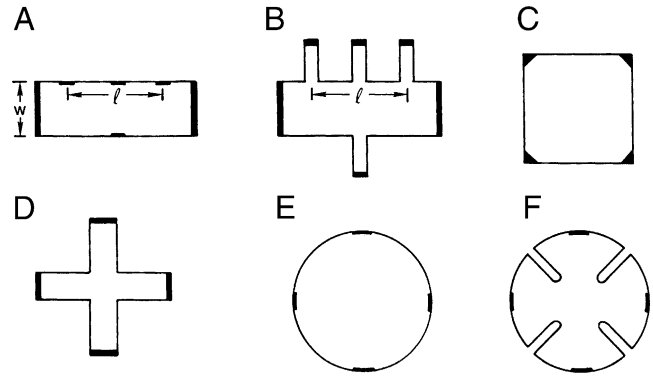


Figure 7. Various specimen/contact patterns commonly used for resistivity and Hall effect measurements: (**A**) Hall bar; (**B**) Hall bar with contact arms; (**C**) square; (**D**) Greek cross; (**E**) circle; (**F**) cloverleaf.

is simply to put four very small contacts around the periphery and apply the standard van der Pauw analysis. We have found that indium, applied with a soldering iron, works well for almost any semiconductor material. Contact size errors can be estimated if the shape is somewhat symmetrical. The simple Hall bar structure (A) is not recommended for a very small bulk sample unless the sample is already in that form, because it is then necessary to measure l and w, which can introduce large errors.

For larger bulk samples, there is a greater choice among the various structures. The availability of an ultrasonic cutting tool opens up the possibility of using structures (B), (D), (E), or (F), if desired. Here, it might be noted that (B) is rather fragile compared to the others. If the samples must be cleaved from a wafer, then the shapes are basically limited to (A) and (C). In our laboratory, for example, it is common to use a square (C) of about 6 × 6 mm and then put contacts of dimension 1 mm or less on the corners.

If photolithographic capabilities are available, then one of the more complex test structures (B), (D), or (F) should be used, because of the advantages they offer. Most manufacturers of discrete semiconductor devices or circuits include Hall bar or van der Pauw structures at several points on the wafer, sometimes one in every reticle (repeated unit).

Another possible constraint is measurement time. By comparing Equation 13 with Equation 17 and Equation 14 with Equation 24, it appears that a typical van der Pauw experiment should take about twice as long as a Hall bar experiment, and indeed, this is experimentally the case. Thus, in an experiment that involves many measurements on the same sample, such as a temperature dependence study, it may well be a distinct advantage to use a Hall bar instead of a van der Pauw pattern. Also, if a contact-switching network (Fig. 4) is not part of the available apparatus, then a van der Pauw structure cannot be conveniently used.

If high accuracy is necessary for the Hall effect measurements, then structures (B), (D), and (F) are the best, because contact size effects are much stronger for Hall effect data than for resistivity data. The same structures,

along with (A), should be used if μB must be made large, since structures (C) and (E) do not have as good a V_H-vs.-B linearity. From a heat dissipation point of view, structures (A) and (B) are the worst, since they are required to be long and narrow, and thus the Hall voltage is relatively small for a given current, since $V_H \propto w$. A high heat dissipation can lead to large temperature gradients, and thus stronger thermomagnetic effects, or may simply raise the sample temperature an unacceptable amount.

Finally, it is important to ask whether or not the same sample is going to be used for magnetoresistance measurements, because, if so, structures (A) and (B) are undoubtedly the best. The reason is that the analysis of magnetoresistance is more complicated in van der Pauw structures than in Hall bar structures, and, in general, a simple formula cannot be found. It may be noted that in Shubnikov–de Haas and quantum Hall measurements, in which magnetic field dependence is critical, the Hall bar is nearly always used.

PROBLEMS

Contact Size and Placement Effects

Much has been written about this subject over the past few decades. Indeed, it is possible to calculate errors due to contact size and placement for any of the structures shown in Figure 7. For (A), (C), and (E), great care is necessary, while for (B), (D), and (F), large or misplaced contacts are not nearly as much of a problem. In general, a good rule of thumb is to keep contact size and distance from the periphery each below 10% of the smallest sample edge dimension. For Hall bar structures (A) and (B), in which the contacts cover the ends, the ratio $l/w > 3$ should be maintained.

Thermomagnetic Errors

Temperature gradients can set up spurious electromotive forces that can modify the measured Hall voltage. Most of these effects, as well as misalignment of the Hall contacts in structure (B), can be averaged out by taking measurements at positive and negative values of both current and magnetic field and then applying Equations 21 and 25.

Conductive Substrates

If a thin film is grown on a conductive substrate, the substrate conductance may overwhelm the film conductance. If so, and if μ_{sub} and n_{sub} are known, then Equations 32 and 33 can be applied, where layer 2 is the substrate. If the substrate and film are of different types (say, a p-type film on an n-type substrate), then a current barrier (p-n junction) will be set up, and the measurement can possibly be made with no correction. However, in this case, the contacts must not overlap both layers.

Depletion Effects in Thin Films

Surface states as well as film-substrate interface states can deplete a thin film of a significant fraction of its charge carriers. Suppose these states lead to surface and interface potentials ϕ_s and ϕ_i, respectively. Then regions of width w_s and w_i will be depleted of their free carriers, where

$$w_{s(i)} = \left[\frac{2\varepsilon_0 \phi_{s(i)}}{e(N_D - N_A)}\right]^{1/2} \tag{34}$$

Here it is assumed that $\phi_{(s)i} \gg kT/e$ and $e\phi_{(s)i} \gg \varepsilon_C - \varepsilon_F$. The electrical thickness of the film will then be given by $d_{elec} = d - w_s - w_i$. Typical values of ϕ_s and ϕ_i are 1 V, so that if $N_D - N_A = 10^{17}$ cm^{-3}, then $w_s + w_i \approx 2000$ Å $= 0.2$ μm in GaN. Thus, if $d \approx 0.5$ μm, 40% of the electrons will be lost to surface and interface states, and $d_{elec} \approx 0.3$ μm.

Inhomogeneity

A sample that is inhomogeneous in depth must be analyzed according to Equations 11 and 12. In simple cases (e.g., for two layers) such an analysis is sometimes possible, as was illustrated in Figures 5 and 6. However, if a sample is laterally inhomogeneous, it is nearly always impossible to carry out an accurate analysis. One indication of such inhomogeneity is a resistivity ratio $Q \gg 1$ (Fig. 3) in a symmetric sample, which would be expected to have $Q = 1$. The reader should be warned to never attempt an f correction (Fig. 3) in such a case, because the f correction is valid only for sample shape asymmetry, not inhomogeneity.

Nonohmic Contacts

In general, high contact resistances are not a severe problem as long as enough current can be passed to get measurable values of V_c and V_H. The reason is that the voltage measurement contacts carry very little current. However, in some cases, the contacts may set up a p-n junction and significantly distort the current flow. This situation falls under the "inhomogeneity" category, discussed above. Usually, contacts this bad show variations with current magnitude and polarity; thus, for the most reliable Hall measurements, it is a good idea to make sure the values are invariant with respect to the magnitudes and polarities of both current and magnetic field.

ACKNOWLEDGMENTS

The author would like to thank his many colleagues at Wright State University and Wright-Patterson Air Force Base who have contributed to his understanding of the Hall effect and semiconductor physics over the years. He would also like to thank Nalda Blair for help with the manuscript preparation. Finally, he gratefully acknowledges the support received from the U.S. Air Force under Contract F33615-95-C-1619.

LITERATURE CITED

American Society for Testing and Materials (ASTM). 1988. Standard F76. Standard Method for Measuring Hall Mobility and Hall Coefficient in Extrinsic Semiconductor Single Crystals. ASTM, West Conshohocken, Pa.

Gantmakker, V. F. and Levinson, Y. B. 1987. Carrier scattering in metals and semiconductors. *In* Modern Problems in Condensed Matter Physics, Vol. 19 (V. M. Agranovich and A. A. Maradudin, eds.). North-Holland, Amsterdam, The Netherlands.

Hall, E. H. 1879. On a new action of the magnet on electric circuits. *Am. J. Math*. 2:287–292.

Look, D. C. 1989. Electrical Characterization of GaAs Materials and Devices. John Wiley & Sons, New York.

Look, D. C. and Molnar, R. J. 1997. Degenerate layer at GaN/sapphire interface: Influence on Hall-effect measurements. *Appl. Phys. Lett*. 70:3377–3379.

Nag, B. R. 1980. Electron Transport in Compound Semiconductors. Springer-Verlag, Berlin.

Rode, D. L. 1975. Low-field electron transport. *In* Semiconductors and Semimetals, Vol. 10 (R. K. Willardson and A. C. Beer, eds.) pp.1–89. Academic Press, New York.

Runyan, W. R. 1975. Semiconductor Measurements and Instrumentation. McGraw-Hill, New York.

Schroder, D. K. 1990. Semiconductor Material and Device Characterization. John Wiley & Sons, New York.

van der Pauw, L. J. 1958. A method of measuring specific resistivity and Hall effect of discs of arbitrary shape. *Philips Res. Repts*. 13:1–9.

Wieder, H. H. 1979. Laboratory Notes on Electrical and Galvanomagnetic Measurements. Elsevier/North-Holland, Amsterdam.

Wiley, J. D. 1975. Mobility of holes in III-V compounds. *In* Semiconductors and Semimetals, Vol. 10 (R. K. Willardson and A.C. Beer, eds.). pp. 91–174. Academic Press, New York.

KEY REFERENCES

Look, 1989. See above.

A detailed description of both theory and methodology related to Hall Effect, magnetoresistance, and capacitance-voltage measurements and analysis.

Nag, 1980. See above.

A comprehensive treatise on electron scattering theory in semiconductor materials.

Schroder, 1990. See above.

A brief, practical description of the Hall effect and many other techniques related to the measurement of semiconductor properties.

Wiley, 1975. See above.

A good reference for hole scattering in p-type semiconductors

DAVID LOOK
Wright State University
Dayton, Ohio

DEEP-LEVEL TRANSIENT SPECTROSCOPY

INTRODUCTION

Defects are responsible for many different characteristic properties of a semiconductor. They play a critical role in determining the viability of a given material for device applications. The identification and control of defects have always been among the most important and crucial tasks in materials and electronic device development. The performance and reliability of devices can be significantly affected by only minute concentrations of undesirable defects. Since the determination of the type and quality of defects in a material depends on the sensitivity of a characterization technique, the challenge in materials characterization has been to develop detection methods with improved sensitivity.

Whereas electrical characterization methods are more sensitive than physical characterization techniques, they may arguably be less sensitive than some optical techiques. However, since device operation depends largely on the electrical properties of its components, it is conceivable that a characterization from an electrical point of view is more relevant. Also, the activation of defects due to electrical processes requires scrutiny, as it has a direct impact on the performance and reliability of a device.

Deep-level transient spectroscopy (DLTS) probes the temperature dependence of the charge carriers escaping from trapping centers formed by point defects in the material. This technique is able to characterize each type of trapping center by providing the activation energy of the defect level relative to one of the energy band edges and the capture cross-section of the traps. It can also be used to compute the concentration and depth profile of the trapping centers.

Although several electrical characterization techniques exist, such as Hall effect (see HALL EFFECT IN SEMICONDUCTORS), current-voltage, capacitance-voltage (see CAPACITANCE-VOLTAGE (C-V) CHARACTERIZATION OF SEMICONDUCTROS), and carrier lifetime measurements (see CARRIER LIFETIME: FREE CARRIER ABSORPTION, PHOTOCONDUCTIVITY, AND PHOTOLUMINESCENCE), very few of them exploit spectroscopy. The spectroscopic nature of DLTS is a key feature that provides both convenience and sensitivity.

Deep-level transient spectroscopy has been widely used for many different semiconductors. This technique has distinguished itself in contributing to the resolution of many defect-related problems in several technologically important semiconductors such as silicon, the Group III to V and II to VI compounds, and alloys. Many variations of basic DLTS have also been developed for improved sensitivity and for more specialized applications in device structures different from the normal *p-n* or Schottky barrier diodes.

Deep-level transient spectroscopy is not able to determine the chemistry or the origin of a defect. Deep-level transient spectroscopy data should therefore be used in conjunction with other techniques. A successful study of defects requires the concerted efforts of many researchers using various characterization techniques in order to derive a more accurate and consistent picture of the defect structure of a given material.

Defects in Semiconductors

In a real crystal, the periodic symmetry of the lattice can be broken by defects. Lattice defects produce localized energy states that may have energy levels occurring within the

band gap. A charge carrier (electron or hole) bound to such a defect in a lattice has a localized wave function as opposed to a carrier in the allowed energy bands (conduction or valence bands) that is free to move.

Crystal imperfections known as point defects can be vacancies or impurities introduced either deliberately or unintentionally during the growth process. Processing of materials during device fabrication can also introduce point defects. Some defects are unavoidable, and they play a key role in determining the properties of a semiconductor. Chemical impurities that form point defects may exist interstitially or substitutionally in the lattice. An interstitial atom may be of the same species as the atoms in the lattice (intrinsic defects) or of a different species (extrinsic defects). Defects can also consist of vacant lattice sites. There are also defect complexes that are conglomerations of different point defects.

In addition to point defects, there are also one-dimensional defects such as dislocations, two-dimensional defects such as surfaces and grain boundaries, and three-dimensional defects such as micropipes and cavities (voids).

Defects obey the laws of thermodynamics and the law of mass action. Hence, the removal or suppression of one type of defect will enhance the effects of another type. For instance, the removal of defects such as grain boundaries and dislocations increases the significance of point defects.

The presence of defects in semiconductors can be either beneficial or detrimental, depending on the nature of the defects and the actual application of the material in devices.

Gold impurities in silicon junctions are used to provide fast electron-hole recombination, resulting in faster switching time. Impurities such as gold, zinc, and mercury in silicon and germanium produce high-quantum-efficiency photodetectors. The emission wavelength of light-emitting diodes (LEDs) is determined by the presence of deep defect levels. In undoped semi-insulating GaAs, a family of deep donor levels, commonly known as EL2, compensates the acceptor levels due to carbon impurities to impart the high resistivity or semi-insulating properties to these materials. Chromium is also used to dope GaAs to produce semi-insulating GaAs-Cr, although this is no longer in widespread use.

Device performance and reliability are greatly affected by the presence of defects. The success of fiber-optics-based telecommunication systems employing laser diodes depends critically on the lifetime of the laser diodes and LEDs. Degradation of laser diodes and LEDs has been widely attributed to formation of local regions where nonradiative recombination occurs. There is general agreement that these regions result from the motion of dislocations that interact with defect centers to promote nonradiative recombination. A dislocation array can also propagate from a substrate into active layers resulting in device failure. This is a reason why it is crucial that electronic materials have a low dislocation density.

Device fabrication processes can involve ion implantation, annealing, contact formation, mechanical scribing, and cleaving, all of which introduce dislocations and point defects. Dislocations in devices may be of little significance by themselves. The problems only arise when these dislocations are "decorated" by point defects. Impurities tend to gather around dislocations, and the diffusion of impurities is also enhanced at dislocations. It is difficult to differentiate between the effects of point defects and dislocations because the production of dislocations also generates point defects such as interstitial atoms and vacancies.

The topic of reliability and degradation of devices is wide ranging, and the above examples serve only to show the importance of the study of defects in devices.

The location of localized states within the band gap can range from a few milli-electron-volts to a few tenths of an electron volt from either the bottom of the conduction band or the uppermost valence band. The determination of whether a level can be considered deep or shallow is rather arbitrary. A depth of $\geq$0.1 eV is usually termed "deep level." It must also be noted that the accuracy in the determination of any level $<$0.1 eV is questionable in any experiment that involves the electrical or photoelectronic measurement of the thermal emission of carriers from the trapping levels.

A defect center can either be a recombination or a trapping center depending on the probabilities with which each of the two processes will occur. If a carrier bound to a defect center recombines with another carrier of opposite polarity before it can be thermally released to the nearest allowed energy band, then that center is a recombination center. However, if the carrier can escape to the nearest energy band through thermal excitation before recombination can occur, then that center is a trapping center. The probability of escape of a carrier from a defect center is therefore higher if the energy level of the center is nearer to either the valence or conduction band. This means that a donor or acceptor level near either the valence or conduction band is more likely to be a trapping center.

The defect structure in a semiconductor consists of a group of well-defined carrier trapping levels. Deep-level transient spectroscopy is an effective method to characterize the defect structure of a semiconductor. Since different semiconductors would have different combinations of defect levels with different activation energies and capture cross-sections, a DLTS curve taken at a known rate window can be used as a defect signature that is characteristic of the material. A knowledge of the various defect levels in a semiconductor is important to the understanding of the electronic properties of the material. Deep-level transient spectroscopy curves do not describe the exact nature or the chemical origin of the defects. Deep-level transient spectroscopy data require a proper and informed interpretation. They are best used in conjunction with other appropriate methods described in this book.

PRINCIPLES OF THE METHOD

Deep-level transient spectroscopy was first introduced by Lang (1974, 1979) based on the theoretical groundwork of Sah et al. (1970) and Milnes (1973). The basis of this method is the dependence of the capacitance of a space-charge region on the occupancy of the traps within the space-charge region in a semiconductor. Under

a nonequilibrium condition such as that existing in a space-charge region, a trapped carrier can escape from a trapping center by thermal excitation to the nearest energy band.

To characterize a semiconductor, it is necessary to fabricate a junction diode such as a *p-n* or Schottky barrier diode on the material of interest. A space-charge region is formed by reverse biasing the *p-n* or Schottky barrier diode. The diode is initially reverse biased to empty the traps. When the bias across the junction is reduced (or even forward biased), the width of the space-charge region is reduced. An equilibrium condition is established in the neutralized region with the majority carriers populating the traps. When the reverse bias is restored, the space-charge region is again created as before, with the only difference being that there are trapped carriers now residing in the defect centers within the space-charge region. The nonequilibrium condition thus created causes the trapped carriers to be thermally reemitted to the relevant energy band. The rate of thermal emission, or "detrapping," of a carrier is temperature dependent. The change of the occupancy of these trapping centers is reflected in the capacitance of the junction producing a capacitance transient. Minority-carrier injection can also occur when the junction is forward biased during the bias pulse. In essence, both majority and minority carriers are involved, which makes this method all the more effective because the emission of each type of carrier can be detected by monitoring the capacitance transients. The activation energy of the trap, or the depth of the trapping level from the nearest energy band edge, and the capture cross-section can be determined from the temperature dependence of the emission rate. The trap concentration can be determined from the in ensity of the capacitance peak, and the polarity of the carriers can be found from the sign of the capacitance change. The DLTS technique can be applied to the study of trapping centers along a critical current conduction path in any device structure in which a depletion region can be formed when appropriately biased. This technique is suitable not only for the study of bulk semiconductor materials but also for the characterization of actual device structures such as *p-n* diodes, Schottky barrier diodes, metal-insulator-semiconductor (MIS) device, or any devices that contain such basic structures. Examples would be bipolar junction transistors, field-effect transistors, or derivatives thereof.

Emission Rate

Consider an *n*-type semiconductor with only a single deep electron trapping level within the band gap in addition to the donor level, as shown in Figure 1. Assume also that the trapping level is located at an energy level E_T from the conduction band edge, E_C. The donor level is indicated by a shallow energy level E_D that is closer to the conduction band edge.

The probability that an electron is captured by the deep electron trap is given by

$$c_n = n v_{th} \sigma_n \tag{1}$$

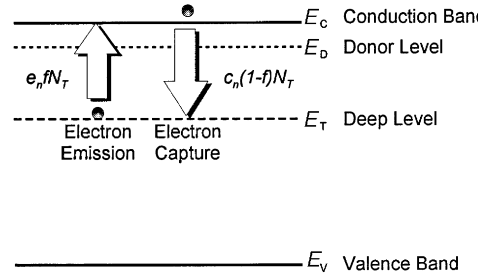


Figure 1. Electron transitions between the trapping level and the conduction band.

where n is the electron concentration, v_{th} is the thermal velocity, and σ_n is the electron capture cross-section. The rate of capture of an electron not only depends on the capture probability c_n but also on the availability of a vacant trap as measured by the concentration of unoccupied traps N_T^0. The electron capture rate can therefore be expressed as

$$\frac{dc_n}{dt} = c_n N_T^0 = n v_{th} \sigma_n N_T^0 \tag{2}$$

The concentration of unoccupied traps N_T^0 can be written in terms of the probability that a given trap out of the total concentration of traps N_T is empty, i.e.,

$$N_T^0 = (1 - f) N_T \tag{3}$$

where $1 - f$ is the probability that a trap is vacant and f is the Fermi-Dirac probability factor (defined below).

Therefore, the rate of electron capture can now be written as

$$\frac{dc_n}{dt} = n v_{th} \sigma_n (1 - f) N_T \tag{4}$$

The rate of electrons escaping from the traps is dependent on both the emission rate e_n and the concentration of occupied traps N_T^-. Therefore, the electron detrapping rate can be written as

$$\frac{de_n}{dt} = e_n N_T^- \tag{5}$$

The concentration of occupied traps N_T^- can also be expressed in terms of the probability that a fraction of the total concentration of traps N_T is filled. This probability is represented by the Fermi-Dirac probability factor f. Therefore,

$$\frac{de_n}{dt} = e_n f N_T \tag{6}$$

According to the principle of detailed balance, the rate of electron capture should balance the rate of electron emission. This means that the concentration of free electrons should remain constant, or

$$\frac{dn}{dt} = 0 = e_n f N_T - n v_{th} \sigma_n (1 - f) N_T \tag{7}$$

Therefore,

$$e_n f N_T = n \nu_{th} \sigma_n (1-f) N_T \tag{8}$$

or

$$e_n = n \nu_{th} \sigma_n \frac{1-f}{f} \tag{9}$$

The Fermi-Dirac probability factor f is given by

$$f = \frac{1}{1 + (1/g) e^{(E_T - E_F)/kT}} \tag{10}$$

where g is the degeneracy factor, E_F is the Fermi level, k is the Boltzmann constant, and T is the temperature. The degeneracy factor takes into consideration that a given energy level can be occupied by an electron of either spin or by no electron at all. The Fermi level can be defined as the energy level where the probability of being occupied at temperature T is 1/2.

The free-electron concentration n is given by

$$n = N_C e^{(E_F - E_C)/kT} \tag{11}$$

where N_C is the density of states in the conduction band and $E_C - E_F \geq 2kT$.

Substituting for f and n, the electron emission rate e_n can be written as

$$e_n = \frac{N_c \nu_{th} \sigma_n}{g} e^{-(E_C - E_T)/kT} \tag{12}$$

By using the relationships

$$N_C = \left(\frac{2\pi m_e^* kT}{h^2} \right)^{3/2} \tag{13}$$

where m_e^* is the electron effective mass and h is Planck's constant, and

$$\nu_{th} = \sqrt{\frac{8kT}{\pi m_e^*}} \tag{14}$$

the equation for the electron emission rate e_n can be expressed as

$$e_n = \frac{16\pi m_e^* k^2 \sigma_n T^2}{g h^3} e^{-(E_C - E_T)/kT} \tag{15}$$

By taking the natural logarithm, we arrive at the equation

$$\ln \frac{e_n}{T^2} = \ln \left(\frac{16\pi m_e^* k^2 \sigma_n}{g h^3} \right) - \frac{E_C - E_T}{kT} \tag{16}$$

It can been seen from this equation that the slope of the Arrhenius plot of $\ln(e_n/T^2)$ against $1/T$ gives the activation energy of the electron trap relative to the conduction band edge, $E_C - E_T$. The y intercept of this plot gives the capture cross-section σ_n of the electron trap.

The next task is to determine the emission rate e_n at a range of temperatures so that an Arrhenius plot can be made.

Junction Capacitance Transient

We consider, for simplicity, a Schottky barrier diode formed on an n-type semiconductor. The energy band diagram of a Schottky barrier diode under different bias condition is shown in Figure 2. Figure 2A shows the diode in the reverse-bias condition. In the space-charge or depletion region, denoted by the shaded area in the figure, all the electron traps above the Fermi level are ionized or empty. In Figure 2B, the reverse bias is reduced by introducing a short pulse (one can also apply a short forward-bias pulse), thus reducing the width of the depletion region. By doing so, the electrons are now allowed to be captured at the traps, thus filling up the traps. Figure 2C shows that when the bias is returned to the original reverse-bias condition, the electrons begin to be thermally excited or emitted to the conduction band by absorbing thermal energy. The capacitance of the junction will vary in response to changes in the width w_d of the space-charge region according to

$$C = \frac{\varepsilon A}{w_d} \tag{17}$$

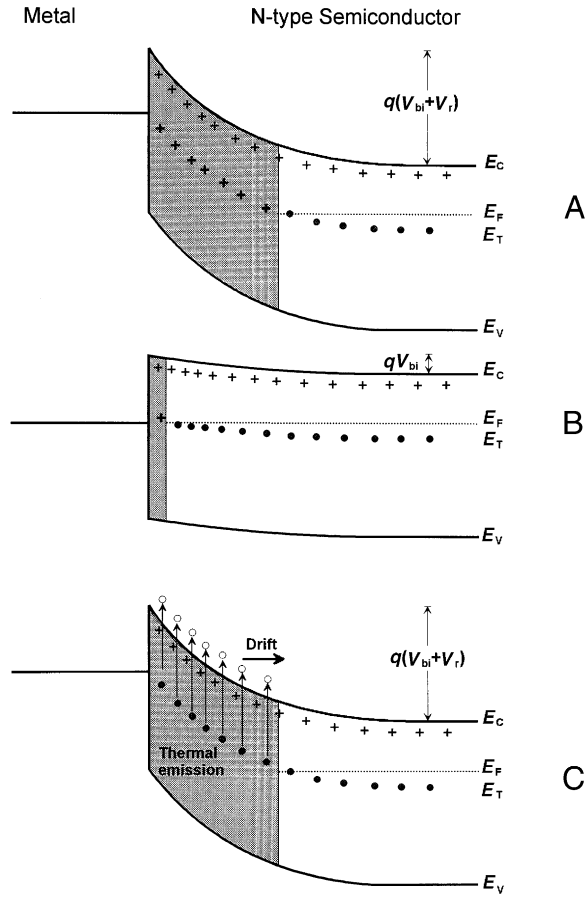


Figure 2. Energy band diagrams for a diode under (**A**) reverse bias; (**B**) reduced bias pulse; and (**C**) restored reverse bias.

where ε is the permittivity of the semiconductor and A is the surface area of the gate contact.

The width w_d is in turn dependent on the concentration of charged states or electron-trapping centers in the material as well as the applied bias across the junction, as shown by the equation

$$w_d = \sqrt{\frac{2\varepsilon[V_{bi} + V_r - (kT/q)]}{q(N_D^+ + N_T)}} \tag{18}$$

where V_{bi} is the built-in potential of the barrier, V_r is the applied reverse bias, q is the magnitude of the electronic charge, and N_D^+ is the concentration of ionized donor centers.

If we assume that the falling edge of the pulse coincides with time $t = 0$, we can denote the capacitance at time $t = 0$ as $C(0)$ and the capacitance at a sufficiently long delay ($t = \infty$) after the end of the pulse as $C(\infty)$. The capacitance $C(\infty)$ is actually the capacitance of the junction under the quiescent reverse-bias condition. It is commonly accepted, within certain restrictions (Look, 1989), that the difference in the capacitance at $t = 0$, $\Delta C(0)$, is related to the concentration of the electron traps by

$$\frac{\Delta C(0)}{C(\infty)} \cong \frac{N_T}{2N_D} \tag{19}$$

where N_D is the net dopant concentration.

The time variation of the capacitance following the applied bias pulse is shown in Figure 3. As the trapped electrons are emitted into the conduction band following the termination of the bias pulse, the capacitance of the junction will decay exponentially with a time constant τ_n according to the equation

$$C(t) = C(0)e^{-t/\tau_n} \tag{20}$$

The decaying time constant τ_n of the capacitance transient is related to the emission rate of the trapped electrons according to

$$e_n = \frac{1}{\tau_n} \tag{21}$$

Equation 20 can therefore be written as

$$C(t) = C(0)e^{-e_n t} \tag{22}$$

If the decaying capacitance transient is measured at two different time delays t_1 and t_2 from the termination of the bias pulse, then the difference in the capacitances will be given by

$$S = \Delta C = C(t_1) - C(t_2) = C(0)(e^{-e_n t_1} - e^{-e_n t_2}) \tag{23}$$

The capacitance difference S is the DLTS signal. It can be seen from Figure 4 that as the temperature changes, the time constant of the decaying transient also changes. The strong temperature dependence of trap emission rate and capacitance can be seen in Equations 15 and 20. The time constant at low temperature is longer, reflecting the low emission rate of the trapped electrons due to inadequate thermal energy for the electrons to escape. As the temperature of the sample increases, the time constant becomes shorter, corresponding to a faster rate of emission of the trapped electrons since there is an increasing availability of the thermal energy required for excitation into the conduction band. By monitoring the DLTS signal S, which is the difference in the capacitances at two different delay times, as a function of temperature, it can be seen that at low temperatures ($T \ll T_m$) and at high temperatures ($T \gg T_m$), the signal is very small. However, at an intermediate temperature T_m, the signal will go through a maximum value. At this temperature, T_m, the variation of the DLTS signal S with temperature can be described by

$$\frac{dS}{dT} = \frac{dS}{de_n} \times \frac{de_n}{dT} = 0 \tag{24}$$

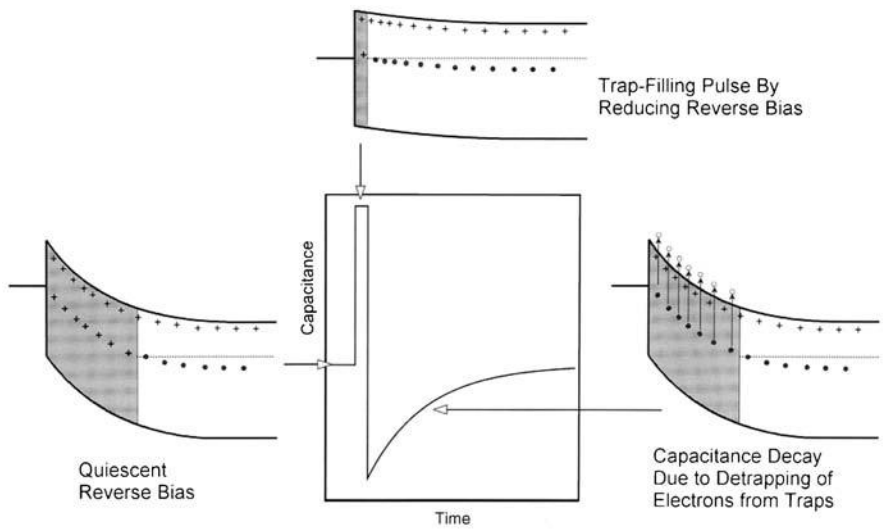


Figure 3. Time variation of junction capacitance during different biasing conditions for a majority-carrier trap.

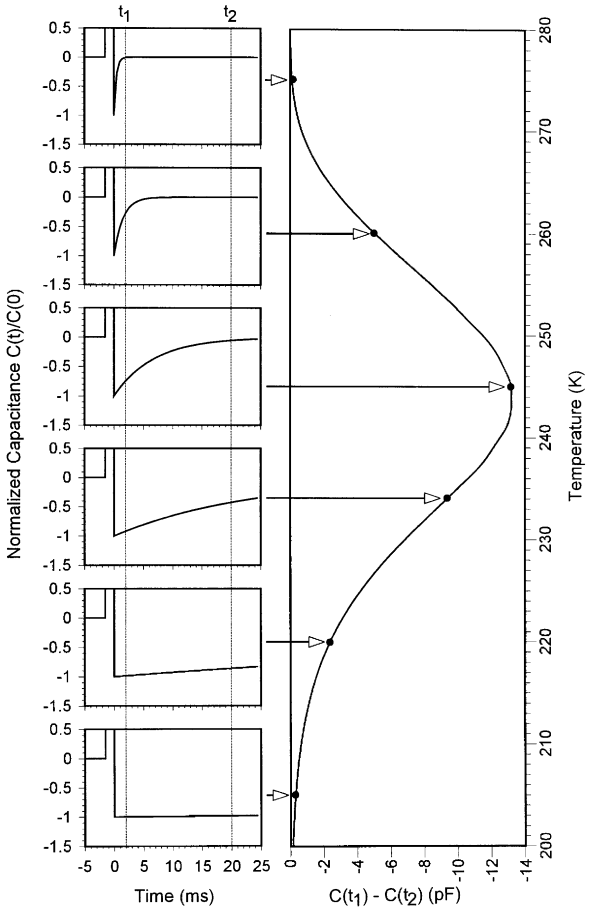


Figure 4. Temperature dependence of the time constant of a capacitance transient.

Therefore,

$$\frac{dS}{de_n} = -t_1 e^{-e_n t_1} + t_2 e^{-e_n t_2} = 0 \tag{25}$$

which gives

$$e_n = \frac{\ln(t_2/t_1)}{t_2 - t_1} \tag{26}$$

This equation shows that if the time delays t_2 and t_1 are known, then at the temperature T_m at which the signal S is a maximum, the emission rate of the trapped electron, valid only for that particular temperature T_m, can be calculated. By varying t_2 and t_1 but keeping the ratio t_2/t_1 constant (thus changing to a different value of e_n) and repeating the temperature scan, a different curve will be obtained, with the signal S peaking at a different temperature T_m. By repeating the procedure for several values of e_n and obtaining the corresponding set of temperatures T_m, an Arrhenius plot of $\ln(e_n/T^2)$ against $1/T$ can give the activation energy and the capture cross-section of the electron traps.

The efficiency of DLTS lies in the way in which the trapping parameters (activation energy and capture cross-section) are obtained experimentally. In the experiment, the capacitance transients are sampled at two different instants of time t_1 and t_2 from the end of the bias pulse. The expression $(t_2 - t_1)/\ln(t_2/t_1)$ can be considered as a time window. The reciprocal of this time window is the rate window. In essence, as the temperature changes, the emission rate of the carriers also varies until a temperature is reached at which the emission rate matches the rate window set by t_2 ant t_1.

Figure 5 shows an example of how an Arrhenius plot is made from the DLTS signals.

PRACTICAL ASPECTS OF THE METHOD AND METHOD AUTOMATION

The first DLTS experimental setup was reported by Lang (1974) and is popularly known as the boxcar technique. A block diagram of a basic DLTS spectrometer using the boxcar technique is shown in Figure 6. The high-frequency capacitance meter measures the capacitance across the junction device and typically has a 1 MHz test frequency. A pulse generator with variable pulse width and pulse height as well as a variable dc offset bias is required to provide a steady-state or quiescent reverse bias as well as the trap-filling pulse. A standard dc power supply can also be used instead of the dc offset bias of the pulse generator. In fact, a standard dc power supply may be able to give a wider range of reverse bias. An analog-to-digital converter is needed to digitize the analog output of the capacitance

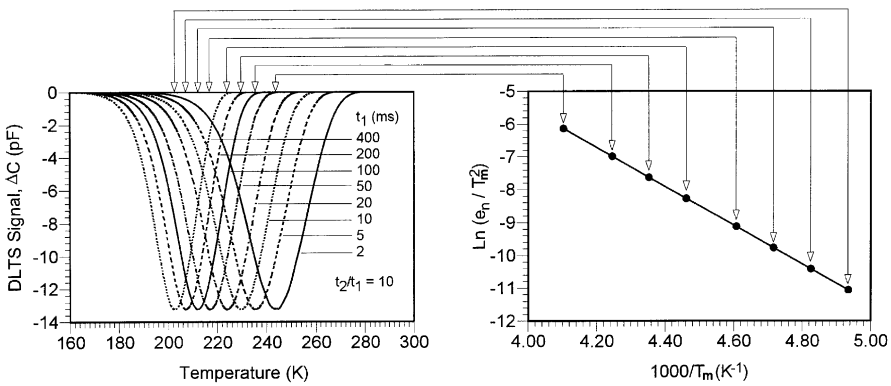


Figure 5. Deriving Arrhenius plots from DLTS signals.

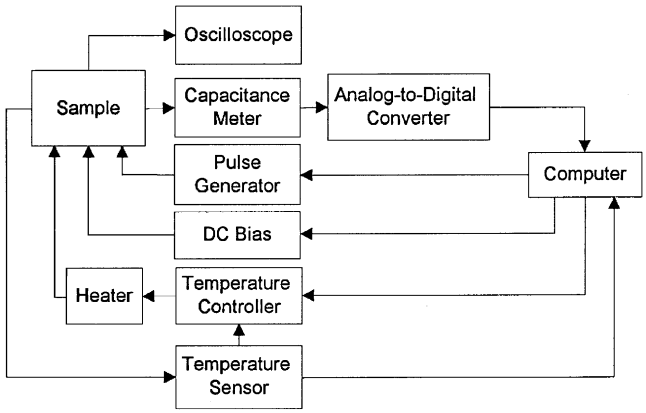


Figure 6. A DLTS spectrometer using the boxcar technique.

meter. A fast digital voltmeter with digital output to a computer can also be used for this purpose. It is also helpful to have an oscilloscope to visually monitor the capacitance transient. Poor or degraded contacts on the sample can give a noisy or spurious signal, and an oscilloscope is a good diagnostic tool for detecting any problem. A computer is necessary for data collection, analysis, and archiving. In the past, an analog *x-y* pen recorder was commonly used, but such recording devices do not allow for data manipulation, analysis, or storage. The temperature of the sample must be variable. A wide temperature range is required to detect the trapping levels from near the energy band edges to the middle of the band gap. A temperature range of about 77 to 380 K would be sufficient to detect trapping levels from about 0.1 eV to ~0.85 eV for most semiconductors. A higher temperature is needed for wide-band-gap semiconductors. Since the sample temperature needs to be as low as 77 K, a mechanical pump and a vacuum sample chamber or cryostat are needed to prevent ambient moisture from condensing on the sample. Although a temperature controller would be helpful for precise temperature control, it is not a necessity if the sample temperature can be slowly varied at a rate small enough to avoid any temperature lag between the temperature sensor and the sample.

The DLTS system shown in Figure 6 is more flexible in the way the data can be processed. The system would allow for the complete decay transient to be digitized, and the rate window technique can then be applied by writing a simple program to analyze the stored data files after the experiment. However, in addition to saving the complete data for later analysis, it is advisable that the DLTS signal be plotted in real time on the computer by simply plotting the difference between two capacitance values taken from the data file at two fixed delay times against the sample temperature at which the data were measured. In this manner, it is possible to see immediately whether there is any peak in the DLTS curve. In the event that there is no peak observed as the temperature is scanned, the operator can choose to terminate the experiment early without wasting much time. The operator should consider different choices of delay times and biasing conditions in order to obtain an optimum DLTS curve. A typical experiment would consist of gradually heating the sample from the low end of the temperature range to the high end and recording the capacitance transient at an appropriate temperature interval. If the complete transient is recorded, then only a single temperature scan is needed. Previously, DLTS temperature scans had to be repeated several times because only a single fixed-rate window can be applied at each scan. This is tedious and sometimes detrimental to the sample because repeated temperature cycling may degrade the sample. The automated approach described here is akin to a multichannel approach in which all the rate windows can be applied simultaneously. Obviously, if the complete transient can be digitized using the system in Figure 6, then it is also possible, especially with the power of modern desktop personal computers, to extract the time constants numerically by nonlinear curve fitting of each transient.

An alternative method to extract the DLTS curves that employs an instrumental approach as opposed to the digital technique described above is shown in Figure 7. In this method, an electronic circuit is used to perform the subtraction of the capacitance values to give the DLTS signal. The circuit essentially has two sample-and-hold amplifiers, each of which is separately triggered to hold the analog voltage corresponding to a capacitance value at a preset time delay from the end of the bias pulse. Since there are two sample-and-hold amplifiers, two analog voltages corresponding to the capacitances at two different parts of the capacitance transient are stored. The outputs from the sample-and-hold amplifiers are then fed into a differential amplifier to give an analog signal that is proportional to the difference between the two capacitances. This analog signal is the DLTS signal and can be further amplified and recorded by a digital voltmeter or analog-to-digital converter and then transferred to a computer. Each circuit can only provide one rate window or a single DLTS curve. The circuit can be built as a module, and by using several modules together, several DLTS curves corresponding to different rate windows can be obtained at the same time in a single temperature scan.

The elegance of the original DLTS lies in the instrumental extraction of the time constant from the capacitance decay without the use of numerical analysis. Other instrumental methods (Auret, 1986; Day et al., 1979; Kimerling, 1976; Miller et al., 1975, 1977) are available to measure the rate window. Various automated DLTS

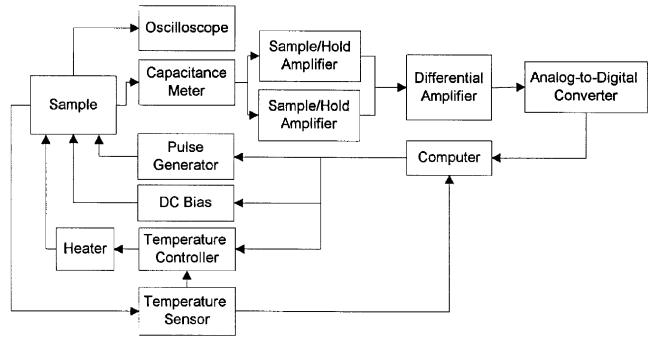


Figure 7. A DLTS spectrometer using the sample-and-hold amplifiers technique.

systems have also been reported (Chang et al., 1984; Holzlein et al., 1986; Kirchner et al., 1981; Okuyama et al., 1983; Wagner et al., 1980). A standard test method for using the DLTS technique has also been published (ASTM, 1993). Due to the spectroscopic character of DLTS and the significance of electrically active defects in devices, this technique can be adapted for industrial semiconductor process control (Jack et al., 1980).

Since the introduction of DLTS, many variations of the technique have been introduced. Examples of such variations are electron beam scanning DLTS (Petroff and Lang, 1977), photo-excited DLTS (Brunwin et al., 1979; Mitonneau et al., 1977; Takikawa and Ikoma, 1980), current or conductance DLTS (Borsuk and Swanson, 1980; Wessels, 1976), constant-capacitance DLTS (Johnson et al., 1979) and double-correlation DLTS (Lafevre and Schultz, 1977).

DATA ANALYSIS AND INITIAL INTERPRETATION

For each value of the emission rate e_n corresponding to a set of t_1 and t_2, a graph of the DLTS signal as a function of temperature is plotted. The emission rate e_n is related to t_1 and t_2 according to Equation 26. The temperature T_m at which the peak occurs is recorded. This graphing procedure is repeated for several emission rates, which are obtained experimentally by changing t_1 and t_2. For each value of e_n, there is a corresponding value of T_m. An Arrhenius plot of e_n/T_m^2 against $1/T_m$ is then produced. From Equation 16, it can be seen that the slope of the linear graph will give the activation energy and the y intercept will give the capture cross-section. The concentration of the defects can be calculated according to Equation 19. The concentration can also be derived from the DLTS peak only if t_1 is chosen to be very near to $t = 0$ and $t_2 \neq t_1$ (or $t_2 \div \infty$). See Principles of the Method for further details.

SAMPLE PREPARATION

A sample for DLTS measurement must contain a voltage-controlled depletion region. Schottky barrier diodes and p-n diodes are suitable samples. A Schottky barrier diode is the most convenient type of sample. This diode is simply a metal placed in contact with a semiconductor to form an energy barrier or Schottky barrier at the interface. The selected metal must have a sufficiently high Schottky energy barrier to produce a rectifying contact with a low leakage current. The surface of the semiconductor must be clean, because surface impurities can reduce the effective barrier height and increase current leakage. The metal film, or gate contact, can be deposited on the semiconductor by vacuum evaporation, sputtering, or chemical vapor deposition. For standard capacitance-voltage measurement, a mercury probe in contact with the semiconductor is adequate to form a Schottky barrier diode. However, this contacting technique is not advisable in DLTS because the temperature of the sample has to be varied over a wide range. Schottky barrier diodes are easily formed on semiconductors of low dopant concentration. They tend to exhibit large leakage current with increasing dopant concentration. Deep-level transient spectroscopy measurement does not work well with leaky diodes.

A p-n diode is made by forming a junction between an n- and a p-type material. Because the interface between the n- and p-type materials must be clean, this type of diode is usually formed by controlled introduction of dopants either by alloying, diffusion, or ion implantation or during epitaxial growth.

Both Schottky and p-n diodes also need low-resistivity ohmic contacts to allow for electrical connections between the semiconductors and the electronic equipment. It is easier to form an ohmic contact on a semiconductor of high dopant concentration ($\sim 10^{18}$ cm^{-3}) than one with a low dopant concentration, because the narrow width of the Schottky barrier formed by a metal on a highly doped semiconductor encourages current conduction by carrier tunneling through the thin barrier. Ohmic contact formation on highly doped material may or may not require alloy formation by thermal annealing. In some semiconductors, it is possible to get some metals to form good ohmic contacts even when the semiconductors have a low carrier concentration of about 10^{14} to 10^{15} cm^{-3}. The formation of ohmic contacts in such cases would invariably include alloy formation through rapid thermal annealing of the metal-semiconductor interface. A common sample configuration is one in which the semiconductor material of interest is grown on a highly doped or low-resistivity substrate of the same material. Schottky or p-n diodes are formed on the top surface while ohmic contact is made at the back of the low-resistivity substrate.

In DLTS experiments using Schottky diodes, it is possible to avoid ohmic contact problems by making the back contact on the substrate larger than the cross-sectional area of the depletion region of the sample. In practice, the size of the gate contact is made smaller than the back contact. In this configuration, the larger back contact will have a larger capacitance than that of the depletion region. The resultant capacitance will be largely that of the depletion region, which is the region of interest.

It is often necessary for the diode cross-sectional area to be known so that it can be used in data analysis. In Schottky diodes, diode structures of the required size can be made by vacuum evaporation or by sputtering of metal through a shadow mask containing holes of the required dimensions. Alternatively, the metal film can be patterned through photolithography and etching to produce Schottky diode structures of different sizes. The p-n junction diodes can only be patterned through photolithography and etching.

Since any sample containing a bias-controlled depletion region can be used for DLTS measurement, this technique can be applied not only to individual discrete diodes but also to junction diodes found in devices such as bipolar and field-effect transistors. It can also be used for measurement of interface state density at an insulator–semiconductor interface in an MIS device. Deep-level transient spectroscopy can therefore be used for analysis of actual transistors. The potential applications of DLTS

as a diagnostic tool for actual devices make it a powerful technique for understanding the role of electrically active defects in device performance and reliability.

PROBLEMS

The shortcoming of DLTS, or any method that requires a depletion region, is that it is not effective for high-resistivity materials due to the problems involved in obtaining suitable junctions on such materials. Furthermore, due to the low concentration of free carriers in such materials, an application of a reverse bias may cause the whole sample to be depleted. This shortcoming led to the introduction of photo-induced current transient spectroscopy (PICTS or PITS; Fairman et al., 1979; Hurtes et al., 1978; Look, 1983; Tin et al., 1987; Yoshie and Kamihara, 1983a,b, 1985), which employs the same gating technique but uses a light pulse instead of a voltage pulse and does not require a depletion region to be formed.

The interpretation of DLTS curves is susceptible to errors. For example, a high circuit impedance (which could be due to high sample resistivity or poor contacts) will cause a disproportionately large time constant that is due to the increase in the reactance-capacitance time constant of the measurement setup and not to a very deep trapping center. In such cases, the DLTS curves lack sharp or well-defined peaks. Such samples are commonly encountered in studies of irradiation effects in semiconductors (Lang, 1977; Tin et al., 1991a,b). A broad peak can also be indicative of a range of closely spaced energy levels producing capacitance transients with multiexponential characteristics. In certain cases, it is possible to separate the closely spaced peaks by using large t_2/t_1 ratios or large values of $t_2 - t_1$. When using DLTS for depth profiling of defects, the effect of fringing capacitance around the gate contact also needs to be considered.

A featureless or flat DLTS curve does not mean an absence of trapping centers. Experiments with different quiescent bias and pulse height settings should be tried to determine the optimum parameters for subsequent high-resolution scans.

Traps with very large or very small capture cross-sections are not easily detected (Miller et al., 1977). Different experimental conditions and samples should be used to provide as much detail of the defect structure as possible.

The steady-state capacitance of the sample changes with temperature. This causes the DC level to shift with temperature. A baseline restoring circuit will resolve this problem (Miller et al., 1975). Another solution to this problem is to reduce the rate of temperature increase to ensure that the sample temperature does not change appreciably while the capacitance is relaxing to its steady-state value after each bias pulse.

During the peak of each bias pulse, the capacitance meter will show an overload. The capacitance meter should have a fast pulse overload recovery in order to measure fast transients.

Data acquisition should only be initiated after the meter has recovered from overload.

LITERATURE CITED

American Society for Testing and Materials (ASTM). 1993. Standard test method for characterizing semiconductor deep levels by transient capacitance techniques, ASTM F978-90. *In* 1993 Annual Book of ASTM Standards, pp. 534–541. ASTM, West Conshohocken, Pa.

Auret, F. D. 1986. Considerations for capacitance DLTS measurements using a lock-in amplifier. *Rev. Sci. Instrum.* 57:1597–1603.

Borsuk, J. A. and Swanson, R. M. 1980. Current transient spectroscopy: A high-sensitivity DLTS system. *IEEE Trans. Electron Devices* ED-27:2217–2225.

Brunwin, R., Hamilton, B., Jordan, P., and Peaker, A. R. 1979. Detection of minority-carrier traps using transient spectroscopy. *Electron. Lett.* 15:349–350.

Chang, C. Y., Hsu, W. C., Uang, C. M., Fang, Y. K., Liu, W. C., and Wu, B. S. 1984. Personal computer-based automatic measurement system applicable to deep-level transient spectroscopy. *Rev. Sci. Instrum.* 55:637–639.

Day, D. S., Tsai, M. Y., Streetman, B. G., and Lang, D. V. 1979. Deep-level transient spectroscopy: System effects and data analysis. *J. Appl. Phys.* 50:5093–5098.

Fairman, R. D., Morin, F. J., and Oliver, J. R. 1979. The influence of semi-insulating substrates on the electrical properties of high-purity GaAs buffer layers grown by vapor-phase epitaxy. *Inst. Phys. Conf. Ser.* 45:134–143.

Holzlein, K., Pensl, G., Schulz, M., and Stolz, P. 1986. Fast computer-controlled deep level transient spectroscopy system for versatile applications in semiconductors. *Rev. Sci. Instrum.* 57: 1373–1377.

Hurtes, Ch., Boulou, M., Mitonneau, A., and Bois, D. 1978. Deep-level spectroscopy in high-resistivity materials. *Appl. Phys. Lett.* 32:821–823.

Jack, M. D., Pack, R. C., and Henriksen, J. 1980. A computer-controlled deep-level transient spectroscopy system for semiconductor process control. *IEEE Trans. Electron Devices* ED-27: 2226–2231.

Johnson, N. M., Bartelink, D. J., Fold, R. B., and Gibbons, J. F. 1979. Constant-capacitance DLTS measurement of defect-density profiles in semiconductors. *J. Appl. Phys.* 50:4828–4833.

Kimerling, L.C. 1976. New developments in defect studies in semiconductors. *IEEE Trans. Nucl. Sci.* NS-23:1497–1505.

Kirchner, P. D., Schaff, W. J., Maracas, G. N., Eastman, L. F., Chappel, T. I., and Ransom, C. M. 1981. The analysis of exponential and nonexponential transients in deep level transient spectroscopy. *J. Appl. Phys.* 52:6462–6470.

Lafevre, H. and Schultz, M. 1977. Double correlation technique (DDLTS) for the analysis of deep level profiles in semiconductors. *Appl. Phys.* 12:45–53.

Lang, D. V. 1974. Deep-level transient spectroscopy: A new method to characterize traps in semiconductors. *J. Appl. Phys.* 45: 3023–3032.

Lang, D. V. 1977. Review of radiation-induced defects in III-V compounds. *Inst. Phys. Conf. Ser.* 31:70–94.

Lang, D. V. 1979. Space-charge spectroscopy in semiconductors. *In* Thermally Stimulated Relaxation in Solids (P. Bräunlich, ed.). pp. 93–131. Springer-Verlag, Heidelberg.

Look, D. C. 1983. The electrical and photoelectronic properties of semi-insulating GaAs. *In* Semiconductors and Semimetals, Vol. 19 (R. K. Willardson and A. C. Beer, eds.). pp. 75–170. Academic Press, New York.

Look, D. C. 1989. Electrical Characterization of GaAs Materials and Devices. John Wiley & Sons, New York.

Miller, G. L., Lang, D. V., and Kimerling, L. C. 1977. Capacitance transient spectroscopy. *Ann. Rev. Mater. Sci.* 7:377–448.

Miller, G. L., Ramirez, J. V., and Robinson, D. A. H. 1975. A correlation method for semiconductor transient signal measurements. *J. Appl. Phys*. 46:2638–2644.

Milnes, A. G. 1973. Deep Impurities in Semiconductors. John Wiley & Sons, New York.

Mitonneau, A., Martin, G. M., and Mircea, A. 1977. Investigation of minority deep levels by a new optical method. *Inst. Phys. Conf. Ser.* 33a:73–83.

Okuyama, M., Takakura, H., and Hamakawa, Y., 1983. Fourier-transformation analysis of deep level transient signal in semiconductors. *Solid-State Electron*. 26:689–694.

Petroff, P.M. and Lang, D. V. 1977. A new spectroscopic technique for imaging the spatial distribution of nonradiative defects in a scanning transmission electron microscope. *Appl. Phys. Lett.* 31:60–62.

Sah, C.T., Forbes, L., Rosier, L. L., and A. F. Tasch, Jr. 1970. Thermal and optical emission and capture rates and cross sections of electrons and holes at imperfection centers in semiconductors from photo and dark junction current and capacitance experiments. *Solid-State Electron*. 13:759–788.

Takikawa, M. and Ikoma, T. 1980. Photo-excited DLTS: Measurement of minority carrier traps. *Jpn. J. Appl. Phys.* 19:L436–438.

Tin, C. C., Barnes, P. A., Bardin, T. T., and Pronko, J. G. 1991a. Near-surface defects associated with 2.0-MeV $^{16}O^{+}$ ion implantation in *n*-GaAs. *J. Appl. Phys.* 70:739–743.

Tin, C. C., Barnes, P. A., Bardin, T. T., and Pronko, J. G. 1991b. Deep-level transient spectroscopy studies of 2.0 MeV $^{16}O^{+}$ ion implanted *n*-InP. *Nucl. Instrum. Methods* B59/60:623–626.

Tin, C. C., Teh, C. K., and Weichman, F. L. 1987. Photoinduced transient spectroscopy and photoluminescence studies of copper contaminated liquid-encapsulated Czochralski-grown semi-insulating GaAs. *J. Appl. Phys.* 62:2329–2336.

Wagner, E. E., Hiller, D., and Mars, D. E. 1980. Fast digital apparatus for capacitance transient analysis. *Rev. Sci. Instrum.* 51:1205–1211.

Wessels, B. W. 1976. Determination of deep levels in Cu-doped GaP using transient-current spectroscopy. *J. Appl. Phys.* 47:1131–1133.

Yoshie, O. and Kamihara, M. 1983a. Photo-induced current transient spectroscopy in high-resistivity bulk material. I. Computer controlled multi-channel PICTS system with high-resolution. *Jpn. J. Appl. Phys.* 22:621–628.

Yoshie, O. and Kamihara, M. 1983b. Photo-induced current transient spectroscopy in high-resistivity bulk material. II. Influence of non-exponential transient on determination of deep trap parameters. *Jpn. J. Appl. Phys.* 22:629–635.

Yoshie, O. and Kamihara, M. 1985. Photo-induced current transient spectroscopy in high-resistivity bulk material. III. Scanning-PICTS system for imaging spatial distributions of deep-traps in semi-insulating GaAs wafer. *Jpn. J. Appl. Phys.* 24: 431–440.

KEY REFERENCES

ASTM, 1993. See above.

Describes a standard procedure for DLTS.

Lang, 1974. See above.

A pioneering paper on DLTS.

Lang, 1979. See above.

A review of the DLTS technique.

Look, 1989. See above.

Provides a comprehensive treatment of various aspects of electrical characterization of semiconductors.

Miller et al., 1977. See above.

Provides an overview of various aspects of DLTS.

CHIN-CHE TIN
Auburn University
Auburn, Alabama

CARRIER LIFETIME: FREE CARRIER ABSORPTION, PHOTOCONDUCTIVITY, AND PHOTOLUMINESCENCE

INTRODUCTION

The charge carrier lifetime is one of the most important parameters characterizing a semiconductor. This is mainly for two reasons: (1) the carrier lifetime determines the performance of devices while, at the same time, (2) it is a sensitive measure of material quality and cleanliness. Carrier lifetime may impact device operation directly, such as by affecting the turn-off speed of a diode or a thyristor operating at high injections or the leakage current of a *pn*-junction. Carrier lifetime may also affect operation indirectly, e.g., by affecting the current gain of a bipolar transistor. Thus, a high carrier lifetime is desirable in many cases, whereas a low lifetime favors specific applications (e.g., high-speed operation). Furthermore, for optically active devices, one is interested in radiative recombination, and other recombination mechanisms must be suppressed to enhance quantum efficiency.

To monitor material quality and cleanliness, the carrier lifetime is one of the most sensitive parameters available, able to detect impurities down to a level of $\sim 10^{10}$ cm^{-1} (Schroder, 1997), i.e., one impurity atom per $\sim 10^{12}$ host atoms! In fact, for silicon, the highest lifetimes ever reported (Yablonovitch et al., 1986; Pang and Rohatgi, 1991) are still believed to be limited by impurity/defect recombination and not by intrinsic recombination mechanisms such as radiative recombination. This makes lifetime characterization an important tool for verifying material quality and also for monitoring contamination during subsequent processing. Thus, lifetime measurements are also performed during CMOS processing even though field-effect transistors are unipolar devices, and thus not directly affected by the carrier lifetime (except for leakage current).

Carrier lifetimes and lifetime-characterization techniques have been studied extensively since the start of semiconductor research. This has resulted in the development of a large number of techniques based on different physical principles (Orton and Blood, 1990; Schroder, 1990). Yet, very few standard techniques exist and no single

technique is dominant. This is in part due to the different requirements for various semiconductor materials, reflecting vastly different lifetime values, but also due to different geometry and probing depths. Finally, different recombination mechanisms or injection ranges are addressed by the different methods as they focus on various applications of the material or the device. As a consequence, different lifetime measurement techniques often yield vastly different results, even for the same sample!

This unit reviews the theory and common concepts in carrier-lifetime measurement methodology and focuses on a few important and general techniques used to characterize lifetime in semiconductor materials. Thus, device lifetime characterization methods have been left out. We do not intend to provide detailed descriptions of the reviewed methods; this information can be gathered from the cited references. Our aim is instead to present basic principles of the methods, methodology, and associated difficulties, such that the reader will be able to critically analyze data provided by the methods and to perform simple measurements using the techniques. Also, a selection guide is provided to direct users to the most appropriate method for their specific needs.

The three methods described in this unit are free carrier absorption (FCA), photoconductivity (PC), and photoluminescence (PL). We have selected these methods based on the following criteria.

Free Carrier Absorption is not of widespread use but is a very general tool applicable to most semiconductor materials and also to devices. It provides an exceptional advantage, as different geometries and very different injection regimes can be realized, and therefore many different materials and recombination mechanisms may be characterized. We use it here to illustrate many common concepts in lifetime measurement methods, and, as a consequence, it is presented first.

Photoconductivity is one of the few standard methods used both for wafer fabrication quality monitoring and for clean-room process monitoring. However, at wafer fabrication the transient conductivity following a laser excitation pulse is detected by electrical contacts on a crystal boule while process monitoring is performed using contactless probing of the conductivity.

Photoluminescence is a particularly useful method for characterizing direct-band-gap semiconductors, but has also been applied to indirect band-gap materials. This relies on the extreme sensitivity of many photodetectors, whereby even single-photon counting is available in most optical laboratories. The PL method has also become the workhorse when small volumes must be analyzed, such as in thin-film studies, for quantum-well characterization, or for the rapidly expanding field of semiconductor nanostructures.

The unit commences with a review (see discussion of Theory, below) of the carrier-lifetime concept and various recombination mechanisms, also illustrating some major difficulties affecting lifetime characterization, such as trapping and surface recombination. The next section (Theory) briefly outlines the different techniques and their fundamental principles of operation. Next, the three different techniques are described in detail. Finally, a selection guide is provided for the different lifetime applications.

THEORY

Carrier-Lifetime Concept

In a semiconductor at equilibrium at a certain temperature, a balance exists between thermal generation of carriers, i.e., electron-hole pairs, and recombination of carriers, such that the electron and hole concentrations, designated by n and p, respectively, remain constant (for a general textbook see Sze, 1981). In this way, the law of mass action, $np = n_i^2$, is fulfilled, where n_i is the intrinsic carrier concentration at this temperature. For any departures of the carrier concentrations from this balance such that the mass action law is no longer valid, the system will try to revert to equilibrium. In a general case this is governed by the continuity equation

$$\frac{dn}{dt} = \frac{1}{q}\frac{\partial J_n}{\partial x} + (G_n - R_n) \tag{1}$$

where q is the elementary charge, J_n the electron current density, and G_n and R_n the generation and recombination rates for electrons in the semiconductor. A similar equation holds for holes. Assuming that no electric field is present, the current density is given only by a diffusion current

$$\frac{dn}{dt} = D_n\frac{\partial^2 n}{\partial x^2} + (G_n - R_n) \tag{2}$$

which is known as the diffusion equation, where D_n is the electron diffusion coefficient, which has been assumed independent of concentration and position. In a **homogeneous** case, the derivative with respect to x vanishes and a very simple equation remains. However, to treat the generation and recombination terms more easily, one may expand carrier densities in terms of excess concentrations Δn and Δp

$$n = n_0 + \Delta n; \qquad p = p_0 + \Delta p \tag{3}$$

Also, one may note that band-to-band recombination must be proportional to the product of n and p: i.e., $R_n = Bnp$, where B is the radiative recombination coefficient. In thermal equilibrium, this recombination must be balanced by the band-to-band generation rate; thus, $G_n = Bn_0p_0$. Inserting into Equation 2, and for the homogeneous case one arrives at

$$\frac{d\Delta n}{dt} = -B(n_0\Delta p + p_0\Delta n + \Delta n\Delta p) \tag{4}$$

From this expression, it is clear that the system will revert to equilibrium both from a case where **excess carriers** have been introduced (i.e., when Δn and Δp are positive) as well as for a case where a **carrier deficit** is present (when Δn and Δp are negative). Examples of these situations are the return to equilibrium, (a) following an optical

generation pulse and, (b) following a reverse biasing pulse forming a depletion layer as in a deep level transient spectroscopy (DLTS) experiment (DEEP-LEVEL TRANSIENT SPECTROSCOPY).

Examination of Equation 4 further reveals that two cases may be distinguished: (1) When Δn and Δp are small with respect to their equilibrium values (n_0,p_0) and (2) when they are much larger than they are at equilibrium. The former case is called low-level injection (ll) case, which is sometimes called **minority carrier injection**, while the latter case is referred to as the **high level injection** (hl) case. We see that recombination is linear in Δn for the ll case and quadratic in Δn for the hl case for band-to-band recombination. As a final simplification, we will consider the case when Δn equals Δp, as in optical generation of carriers or during forward injection in a *pn*-junction diode (in a case of negligible trapping; see discussion of Trapping, below). For minority carrier injection, Equation 4 reduces to

$$\frac{d\Delta n}{dt} = -B(n_0 + p_0)\Delta n \tag{5}$$

The solution to this equation is a simple exponential decay

$$\Delta n(t) = \Delta n(0) \cdot \exp(-t/\tau) \tag{6}$$

with lifetime $\tau = 1/[B(n_0 + p_0)]$ as given by Equation 5. The exponential decay and the lifetime defined in this manner are analogous to the decay of radioactive isotopes. More generally, the lifetime may be defined using a general recombination term R according to

$$\tau = \frac{\Delta n}{\Delta R} = -\frac{\Delta n}{(d\Delta n/dt)} \tag{7}$$

where R is the recombination rate of excess electron-hole pairs. The lifetime may not be a constant as in Equation 5, but may be a complicated function of injected carrier density, i.e., of Δn and Δp, and of other parameters.

Recombination Mechanisms

In the derivation of the rate equation governing recombination leading to Equation 5, radiative band-to-band recombination was assumed to be the single recombination mechanism, dominant in direct-band-gap semiconductors. However, several physically different recombination mechanisms exist, each of them adding a recombination term to Equation 2. Indeed, all of these mechanisms also have their corresponding generation mechanism, as stated in the **principle of detailed balance** (Van Roosbroeck and Schockley, 1954). Thus, a generation term must be added for each of them. This ensures that dn/dt equals zero at equilibrium, such that the carrier concentrations are maintained at the n_0 and p_0 values. Note that all equations are valid both for intrinsic material as well as for doped semiconductors.

In Figure 1, common recombination mechanisms are depicted. The first one, the **Shockley-Reed-Hall** (SRH) or **multiphonon** mechanism (Hall, 1952; Shockley and Reed, 1952) dominates recombination in most indirect bandgap semiconductors, such as Si, and involves the trapping of an electron at a defect level within the bandgap followed by the capture of a hole to the same state (and at the same physical location). The reaction may start with the hole capture, depending on the relative capture strengths for holes and electrons and their relative concentrations, but effectively results in the loss of one pair of carriers upon completion. Since recombination is directly proportional to the trap concentration, the corresponding lifetime is a measure of the defect concentration in the sample. The lifetime may be calculated (Sze, 1981) from the effective SRH recombination rate (actually $G-R$ in Equation 2)

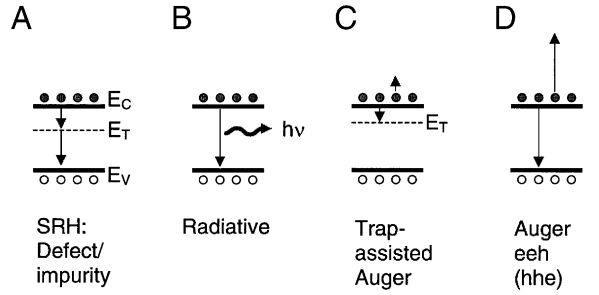


Figure 1. Fundamental recombination mechanisms in semiconductors. From left to right: (**A**) recombination through deep traps (SRH); (**B**) radiative recombination; (**C**) trap-assisted Auger recombination; (**D**) Auger recombination.

$$R_{SRH} = \frac{pn - n_i^2}{\tau_p\{n + n_i\exp[(E_T - E_i)/kT]\} + \tau_n\{p + n_i\exp[(E_i - E_T)/kT)]\}} \tag{8}$$

where E_T and E_i are the trap energy level and the Fermi energy level for intrinsic material, respectively, T is the absolute temperature, and k is the Boltzmann constant. τ_p and τ_n are defined as

$$\tau_p = \frac{1}{\sigma_p v_{th} N_T}; \qquad \tau_n = \frac{1}{\sigma_n v_{th} N_T} \tag{9}$$

where v_{th}, N_T, σ_p, and σ_n denote thermal velocity, trap density, and cross-sections for hole and electron capture by the trap, respectively. The SRH lifetime may be derived using Equations 3 and 7, and is given by the expression

$$\tau_{SRH} = \frac{\tau_p(\Delta n + n_0 + n_i e^{(E_T - E_i)/kT}) + \tau_n(\Delta n + p_0 + n_i e^{(E_i - E_T)/kT})}{p_0 + n_0 + \Delta n} \tag{10}$$

assuming $\Delta n = \Delta p$. Figure 2 displays the functional form of τ_{SRH} versus carrier injection level, and with trap energy level, relative to the edge of the conduction band ($E_C - E_T$). In this example, *n*-type Si with a doping concentration of 1×10^{15} cm^{-3} was selected, and lifetime values were fixed to $\tau_p = 10$ μs and $\tau_n = 44$ μs, respectively. From the figure, it is

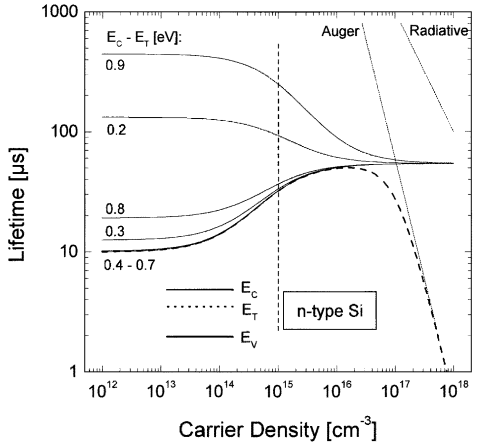


Figure 2. The lifetime injection dependence in n-type Si, doped to 10^{15} cm^{-3}, for recombination through traps according to the SRH formalism (solid curves) with trap energy level as parameter ($\tau_p =$ 10 μs, $\tau_n = 44$ μs). The radiative and Auger recombination rates are shown as dashed lines using generally accepted values for their respective cross-sections. The long-dash line includes SRH ($E_C - E_T = 0.5$ eV) and Auger recombination.

clear that trap energy levels close to the middle of the band gap are most efficient under minority carrier injection, while at high injections all levels are equally effective. The high-injection lifetime, from Equation 10 for infinite Δn, is given by

$$\tau_{hl} = \tau_n + \tau_p \tag{11}$$

The second recombination mechanism depicted in Figure 1 is the **radiative** process, or band-to-band recombination, where an electron and a hole recombine, resulting in the emission of a photon with the characteristic energy of the band gap. From Equation 4, the radiative lifetime is given by

$$\tau_{RAD} = \frac{1}{B(p_0 + n_0 + \Delta n)} \tag{12}$$

which is valid also at high carrier injection. While the intrinsic radiative lifetime dominates recombination in direct-bandgap semiconductors, B is about four orders of magnitude smaller in indirect semiconductors such as Si and Ge (see discussion of Photoluminescence, below, and Table 1).

The third recombination process depicted in Figure 1 is **trap-assisted Auger** recombination (Landsberg, 1987). Here, the capture of an electron (or a hole) at a trap is

Table 1. Measured Carrier Lifetimes and Typical Recombination Coefficients in Some Selected Semiconductors at 300 K[a]

Material	Measured Lifetimes	Radiative Coefficient (cm^3s^{-1})	Auger Coefficients (C_n) (cm^6s^{-1})	Auger Coefficients C_p (cm^6s^{-1})	C_n+C_p (hl)
Indirect					
Si	Typical: 10 to 200 μs Highest: 40 ms[e] Lowest: ~1 ps[f]	1×10^{-14} [b]	2.8×10^{-31} [c]	1×10^{-31} [c]	1.35×10^{-30} [d]
Ge	Typical: 1 to 200 μs	5.2×10^{-14} [g]	8×10^{-32} [g]	2.8×10^{-31} [g]	1.1×10^{-31} [h]
4H-SiC	Typical: 10 to 500 ns Highest: 2 μs[i]	1.5×10^{-12} [j]	$<5 \times 10^{-31}$ [j]		7×10^{-31} [j]
Direct					
GaAs	Typical: 1 ns to 2 μs[k]	2×10^{-10} [g]	1.8×10^{-31} [g]	4×10^{-30} [g]	7×10^{-30} [l]

[a] The resulting effective lifetime can be obtained using Equations 12–16, noting the appropriate injection level, Δn, in a particular measurement. The "measured lifetime" is determined by impurity/defect recombination (SRH) except for clean material of direct-band-gap type semiconductors where the lifetime is set by the radiative lifetime. The last column refers to measurements at high level injection.
[b] Schlangenotto et al. (1974).
[c] Dziewior and Schmid (1977), on highly doped material.
[d] Refers to high-injection Auger coefficient in lowly doped material (Grivickas and Willander, 1999).
[e] Yablonovitch et al. (1986)
[f] Amorphous material (Smit.h et al. , 1981).
[g] Schroder (1997, p. 362).
[h] Auston et al. (1975).
[i] Bergman et al. (1997).
[j] Galeckas et al. (1997a).
[k] Wight (1990).
[l] Strauss et al. (1993).

not accompanied by phonon emission as in the usual SRH case, but its energy is instead transferred to a second electron (or hole). The rate equation in this case predicts a quadratic dependence in carrier concentrations (i.e., proportional to *np*, *nn,* or *pp*) just as in the case of radiative recombination; however, the rate constant is also proportional to the trap density. Therefore, radiative and trap-assisted Auger may be combined into a common radiative coefficient according to

$$B = B_{\mathrm{RAD}} + \beta_{\mathrm{TRAP}} \cdot N_{\mathrm{T}} \tag{13}$$

where B_{RAD} signifies the purely radiative coefficient and β_{TRAP} is a coefficient for the trap-assisted process.

Finally, **Auger recombination**, depicted as the fourth mechanism in Figure 1, is the recombination of an electron and a hole while the energy is transferred to a third carrier. As three carriers are involved, Auger recombination only dominates at very high injections. However, in highly doped semiconductors, as in emitter regions of devices, Auger recombination becomes large and often sets the lifetime in these regions. At high injections, the Auger lifetime is given by

$$\tau_{\mathrm{AUG}} = \frac{1}{(C_{\mathrm{p}} + C_{\mathrm{n}}) \cdot \Delta n^2} = \frac{1}{\gamma \cdot \Delta n^2}; \gamma \equiv C_{\mathrm{p}} + C_{\mathrm{n}} \tag{14}$$

where C_{p} and C_{n} are the Auger coefficients for the *hhe* and the *eeh* processes, respectively. In the minority carrier regime, the Auger lifetime is defined as

$$\tau_{\mathrm{Auger}} = \frac{1}{C_{\mathrm{p}} p_0^2}; \quad \tau_{\mathrm{Auger}} = \frac{1}{C_{\mathrm{n}} n_0^2} \tag{15}$$

representing the Auger lifetime in highly doped p-type and n-type material, respectively. Note that both C and γ are used to denote Auger coefficients and B and β are used for the radiative coefficient.

For calculating the resulting total lifetime, one may note that according to Equation 2, recombination is additive, and when using Equation 7, lifetimes should be summed as inverse numbers

$$\frac{1}{\tau} = \frac{1}{\tau_{\mathrm{SRH}}} + \frac{1}{\tau_{\mathrm{RAD}}} + \frac{1}{\tau_{\mathrm{Auger}}} \tag{16}$$

In Figure 2 the resulting total lifetime injection dependence is shown as a dashed line where Auger and radiative recombination in Si have been included with traditional values for B, C_{p}, and C_{n} (Schlangenotto et al., 1974; Dziewior and Schmid, 1977). The individual contributions of these recombination rates are shown as dotted lines. For the minority carrier **doping dependence** in Si, see, e.g., Häcker and Hangleiter (1994) and Rohatgi et al. (1993).

Finally, Table 1 reports typically measured carrier lifetimes for a few indirect-band-gap semiconductors and for GaAs. Radiative and Auger coefficients have also been included in the table. Coefficients for radiative and Auger recombination for other semiconductors may be found in Schroder (1990).

Generation Lifetime

When a carrier deficit exists, i.e., when both Δn and Δp are negative, recombination according to Equation 8 is negative, meaning that carriers are generated. This carrier generation is mediated through the trap level acting as a stepping stone to cross the band gap and is much more efficient than ordinary band-to-band generation. Such a situation exists, e.g., in the depletion layer in a device, and carrier generation results in additional leakage currents. As a consequence, recombinative defects are also important for majority carrier (unipolar) devices such as MOS transistors.

The generation lifetime may be derived from the SRH expression, Equation 8, in the limit when pn reduces to zero (also $n = p = 0$) and is given by

$$\tau_{\mathrm{g}} = \tau_{\mathrm{p}} \exp[(E_{\mathrm{T}} - E_{\mathrm{i}})/kT] + \tau_{\mathrm{n}} \exp[(E_{\mathrm{i}} - E_{\mathrm{T}})/kT] \tag{17}$$

Analyzing Equation 17, one finds that for $E_{\mathrm{T}} \cong E_{\mathrm{i}}$, the generation lifetime is of the same magnitude as τ_{n} and τ_{p}. However, for trap levels far from the intrinsic Fermi level, E_{i}, the generation lifetime may be orders of magnitude larger than the corresponding recombination lifetime. Thus, we conclude that lifetime measurement techniques probing the generation lifetime, e.g., the **pulsed MOS capacitance** technique, should not be used to measure the recombination lifetime, which is important for bipolar devices as in the base of a transistor, for example. This is extensively discussed in Schroder (1990) and Baliga (1987).

Trapping

In the treatment of recombination so far, it has been assumed that $\Delta n = \Delta p$ and that a single defect center is active with a single trap energy level. If, however, the defect concentration, N_{T}, is high relative to the doping density and injected carrier densities, Δn and Δp, recombination rates may be different for electrons and holes [due to their different capture cross-sections, $\sigma_{n(p)}$]. This leads to unbalanced carrier densities, resulting in nonexponential carrier decays.

A similar situation also occurs when a high concentration of a shallow defect level is present in addition to the main recombination level. The shallow level then acts as a trap which temporarily holds the carrier before releasing it to the band again (the difference with a deep center being that emission to the band of an electron is far more probable than completion of the recombination by the capture of a hole). This results in a variable time delay before recombination through the main recombination level takes place, yielding a nonexponential excess carrier decay transient. According to the most general definition (in Ryvkin, 1964, and in other textbooks) any level in the forbidden gap acts as a recombination level if its position is between the quasi-Fermi levels for electrons and holes. On the contrary, levels that are outside this region (i.e., outside quasi-Fermi levels at the particular injection) act as temporary traps. Examples are porous Si and amorphous Si containing a large density of trap centers of different energy levels, E_T, which in most experiments

exceeds the injected carrier density by several orders. For porous Si the decay transient follows a **stretched exponential** shape according to $I = I_0 \exp(-(t/\tau)^p)$ where τ is a characteristic lifetime and p is a power index describing the curvature in a semilogarithmic plot. We conclude that because the recombination lifetime for high trap densities may not be derived in a simple way, the observed lifetime is multiplied by a factor dependent on the delay time of each trap energy level.

Aspects of Surface Recombination and Diffusion

Inspection of Equation 2 reveals that the local excess carrier decay is not only affected by recombination, which is of primary interest here, but is also affected by carrier gradients. Carrier gradients may result not only from an inhomogeneous excess carrier excitation, but also from differences in the local recombination rate. The effect may be dramatic, such that the diffusion term (all three coordinate directions need to be considered) totally dominates dn/dt and no useful information can be extracted for the recombination term. In fact, diffusion is a limiting factor for most lifetime measurement techniques, and, for an initially homogeneous excess carrier profile, local lifetime information is only obtained within a radius of a **diffusion length**, L_D, given by

$$L_D = \sqrt{D \cdot \tau} \tag{18}$$

For low carrier lifetimes, as occurs in direct-band-gap semiconductors, the diffusion length is relatively short and the spatial resolution is usually set by the specific lifetime measurement technique. In contrast, in indirect-gap semiconductors of good quality material, characterized by high carrier lifetimes, the diffusion length may be several hundred microns. In this case carriers easily reach sample or wafer surfaces and the state of these surfaces becomes critical. Thus, high surface recombination results in a flow of carriers towards the surfaces, leading to a drainage of carriers in the interior bulk and an incorrectly determined value for the carrier lifetime. This is illustrated in the leftmost schematic of Figure 3 and confirmed by the experimental data of Figure 4. In this figure, instant carrier density profiles are shown in a 1-mm-thick Si sample as

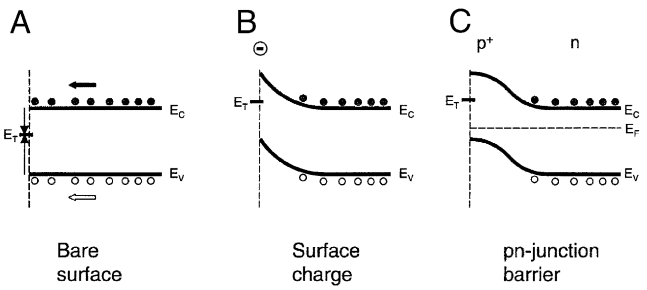


Figure 3. Influence of surface on carrier dynamics. From left to right, schematics illustrate (**A**) high surface recombination at a bare, nonpassivated surface; (**B**) shielding of surface states due to charge-induced band-bending; and, finally, (**C**) barriers introduced by a *pn*-junction (also shielding the surface).

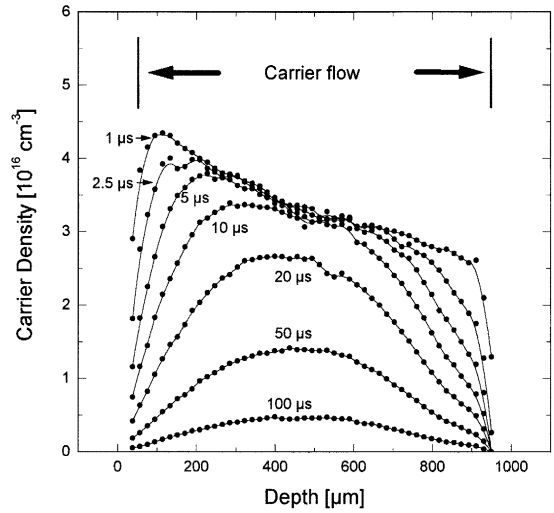


Figure 4. Depth resolved free-carrier-absorption measurements in a 1-mm-thick Si sample illustrating diffusive flow of carriers to surfaces of high surface recombination rate. Data have been recorded at full excitation (top curve) and at successive time moments after the excitation pulse (as given in µs at each respective curve). The optical excitation pulse of wavelength $\lambda = 1.06$ µm impinged onto the sample from the left, yielding the slightly inhomogeneous initial excitation.

measured following a short optical generation pulse. The data clearly illustrate the effect of surface recombination, which, in this case, totally dominates the observed lifetime. Notably, surface recombination is not a different recombination mechanism, but simply SRH recombination occurring at the surface, e.g., at dangling bonds. By the same token, defects **close** to the surface may not be distinguished from defects **at** the surface.

Recombination at or near the surface can be described by the surface recombination velocity, s, in units of cm/s. Its defining equation is

$$D\frac{\partial n}{dx} = -s \cdot n \tag{19}$$

The solutions of the diffusion equation, Equation 2, in a one-dimensional transient case, with boundary values for the surface recombination velocity according to Equation 19, are nonanalytic (Waldmeyer, 1988). For long times after the initial excitation, however, a "steady-state" carrier profile develops, resulting in an exponential excess carrier decay. The corresponding effective lifetime, τ, is modified by the "surface lifetime," τ_s, according to

$$\frac{1}{\tau} = \frac{1}{\tau_b} + \frac{1}{\tau_s} \approx \frac{1}{\tau_b} + \frac{1}{d\left(\frac{1}{2s} + \frac{d}{D\pi^2}\right)} \tag{20}$$

where τ_b is the bulk lifetime, D the diffusivity at the appropriate injection, and d the smallest dimension of the sample (usually the wafer thickness). The last approximate equality was proposed by Grivickas et al. (1989), the uncertainty being less than 4% over the entire range of s values.

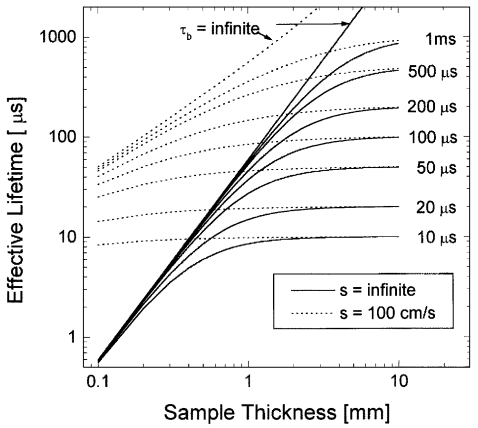


Figure 5. Calculated effective lifetime in silicon, with bulk lifetime as a parameter, versus sample thickness, assuming infinite surface recombination (solid lines) and an ambipolar diffusion coefficient of 17 cm^2/s. The dotted curves were calculated using Equation 20 with a surface recombination velocity of 100 cm^2/s.

Note that, for a single surface, the factor of two in the denominator of Equation 20 should be removed.

Figure 5 shows the results of a calculation according to Equation 20 of the effective lifetime in a Si wafer as a function of thickness. Data for two values of the surface recombination velocity are displayed, and in the calculation an ambipolar diffusivity of $D_a = 17$ cm^2/s was used (Linnros and Grivickas, 1994). The data illustrate well the limitation in measured lifetime imposed by surface recombination, which may only be solved using well-passivated sample surfaces or extremely thick samples ($d^2 \gg D\ \pi\tau_b$).

Methods to reduce high surface-recombination rates are nontrivial, and, in general, such methods may alter the lifetime distribution or may involve etching of the surface. For silicon, a thermal oxide of good quality may reduce surface recombination substantially, although the high formation temperature would probably alter the defect distribution of the sample. Other methods involve different surface passivation schemes, or even measurements with the Si wafer immersed in a passivating HF bath (Yablonovitch et al., 1986). A different approach involves the creation of charges at the surface, resulting in a surface band bending, as illustrated in the middle panel of Figure 3 (Katayama et al., 1991). Finally, the right part of Figure 3 shows the situation when a *pn*-junction is present, as in a highly doped emitter in a device. In this case, the bending of the bands results in barriers, which reflect either electrons or holes (being minority carriers in the highly doped region) such that recombination is inhibited (Linnros, 1998a). Thus, a measurement in the *n* region would reflect the "true" bulk lifetime.

Related Physical Quantities

Although the primary concern of this unit is lifetime measurement techniques, several of these methods may be used to extract other related physical quantities. In fact, some methods do not even measure the lifetime directly, but extract the lifetime from the measurement of other quantities such as the diffusion length. A summary of related physical quantities and a brief comment on how they can be measured is displayed in Table 2.

OVERVIEW OF CARRIER LIFETIME CHARACTERIZATION TECHNIQUES

In this section, a brief overview of the most common methods for carrier lifetime characterization will be given. We do not intend to provide a full list of all methods; examples can be found in the textbook of Schroder (1990) and that of Orton and Blood (1990). Instead, we will focus on fundamental principles of operation, applicability, and strengths and weaknesses. Finally, a comparison of a few techniques from the viewpoint of certain criteria will be provided together with a selection guide.

Table 2. Carrier Lifetime and Related Physical Quantities[a]

Physical Quantity	Symbol	Injection Case	Dominant Regime	Measurement Principle (Examples)
SRH lifetime	τ_{SRH}(ll)	ll	Indirect semiconductor	See techniques of this unit
	τ_{SRH}(hl)	hl		
Radiative lifetime	τ_{RAD}(ll)	ll	Direct semiconductor	PL-transient
	τ_{RAD}(hl)	hl		PL-transient
Auger lifetime	τ_{Auger}(ll); $\gamma_n\gamma_p$	ll	High doping	Measure ll-lifetime
	τ_{Auger}(hl); γ	hl	High injection	Nonexponential transient
Diffusion length	L_D	ll	—	e.g., EBIC/OBIC
Diffusivity	D_nD_p	ll	—	Time-of-flight, transient grating
	D_a (ambipolar)	hl		Transient grating
Mobility	$\mu_n\mu_p$ (majority)	ll		Conductivity/Hall
	$\mu_n\mu_p$ (minority)	ll		Haynes-Shockley
Surface recomb. velocity	s_ns_p	ll, hl	—	See several techniques of this unit

[a] Some of these quantities may be determined in lifetime-type experiments. The fourth column indicates for which type of material a recombination mechanism is dominating while the fifth column indicates suitable measurement methods.

For an overview, the different methods may be sorted in several different ways: either according to physical principles, according to methods used for carrier injection, according to methods used for probing of excess carriers, or according to whether techniques are contactless, nondestructive etc. Unfortunately, there are no unambiguous ways to perform such sorting, partly because variants of the same method would fall into different categories. The approach taken here is to divide the methods into **optical techniques**, **diffusion-length-based methods**, and **device-related techniques**. We note that this distinction is not clear, as optical techniques may contain current probing (as in photoconductivity) and diffusion-length-based methods normally use optical excitation for carrier injection. In addition, several variants exist of techniques where electrical injection is replaced by optical injection, or vice versa. Finally, all techniques may be classified into steady-state methods, modulated methods, or transient/pulsed methods. Several of the methods presented here, indeed, exist in two or all three of these variants.

Optical Techniques

Figure 6 shows the principles of three different techniques based on optical injection of excess carriers. The carriers, i.e., electron-hole pairs, are normally generated by a laser or other light source with photon energy larger than the fundamental energy gap of the semiconductor. The impinging light beam may be continuous, modulated, or pulsed. The resulting excess carrier concentration is either maintained at a higher nonequilibrium value depending on lifetime (if the impinging light is continuous), mimics the imposed modulation with some possible delay (if the light is modulated), or exhibits a transient decay following the initial excitation pulse (if the light is pulsed).

In the PC method, excess carriers represent a modulation of sample conductivity, which is sensed as a current

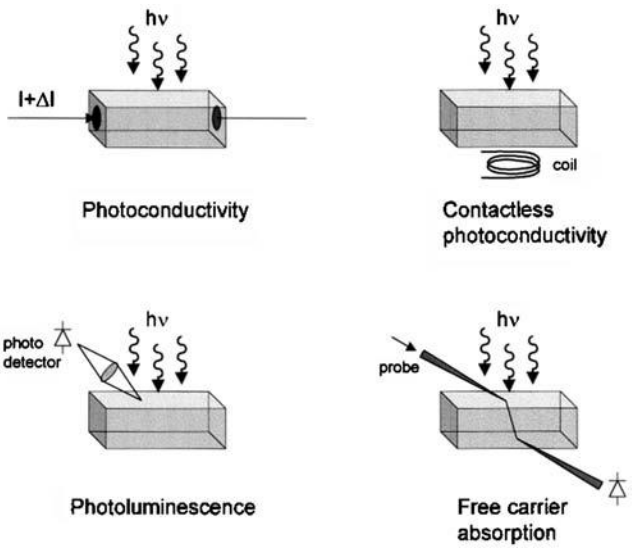


Figure 6. Principles of common optical techniques to measure carrier lifetime.

increase in the external circuit. Contactless variants of this method are the **microwave PC-decay** method and the **radiofrequency PC-decay** method. In the former method, excess carriers are probed by microwaves reflected due to a change in the dielectric constant of the sample resulting from the excess carriers. In the latter technique, excess carriers are detected by a coil reacting on the difference in sample conductivity (see Fig. 6).

The PL method is only applicable for semiconductors exhibiting radiative recombination. This is not a severe limitation, as PL techniques are extremely sensitive and may also be used for semiconductors of indirect bandgap. The excess carrier density is probed by the resulting increase in PL yield. Filtering to a specific wavelength range is a main advantage, reducing background noise and allowing recombination studies of a particular material (e.g., in multiple layer structures). Note that the PL yield tracks the excess carrier density through the radiative process, even in the presence of other possibly dominant recombination mechanisms.

Finally, in the FCA method, excess carriers are sensed by an infrared (IR) beam (or a visible beam, provided that the photon energy is lower than the bandgap energy) transmitted through the sample. The absorption is more or less proportional to the excess carrier density, allowing a calibration of the injected carrier density. The two-beam method can support several different geometries, facilitating depth-resolved studies.

Diffusion-Length-Based Methods

According to Equation 18, the carrier lifetime is proportional to the carrier diffusion length squared. The diffusion coefficient is taken to be constant. This is a good approximation in the low-injection case and for homogeneously doped material where the value of D can be found from mobility-doping data using Einstein's relation, $D = kT\mu/e$. At higher injection levels, however, this is no longer true, as the diffusion coefficient may change substantially when the ambipolar injection regime is approached, and for even higher injection levels, a decrease has been observed due to carrier-carrier scattering (for Si see Linnros and Grivickas, 1994). Thus, we conclude that the carrier lifetime may only be obtained safely from a measurement of the diffusion length in the minority carrier regime. Furthermore, we note that, due to the short radiative lifetimes of direct-bandgap semiconductors, diffusion lengths are extremely short, making diffusion-length-based methods less attractive for these materials.

Figure 7 presents principal schemes of four popular methods to measure diffusion lengths. In the **optical beam induced current** (OBIC) method or in the **electron beam induced current** (EBIC) method, excess carriers are generated locally by a focused laser beam or electron beam, respectively. The carriers diffuse in all directions and the fraction collected by an adjacent *pn*-junction will be detected as a current increase, ΔI, in the outer circuit. By varying the distance of impingement of the generation beam from the *pn*-junction depletion edge, the diffusion length can be measured and the local lifetime extracted.

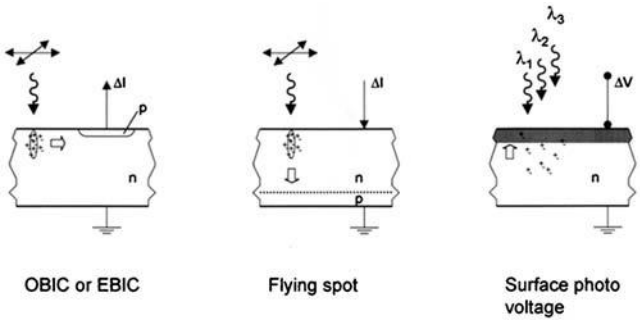


Figure 7. Principles of techniques to measure carrier diffusion lengths for lifetime derivation. Block arrows indicate direction of carrier flow toward *pn*-junction, or, for the surface photovoltage technique, toward the surface (shaded area indicating depletion zone due to band bending).

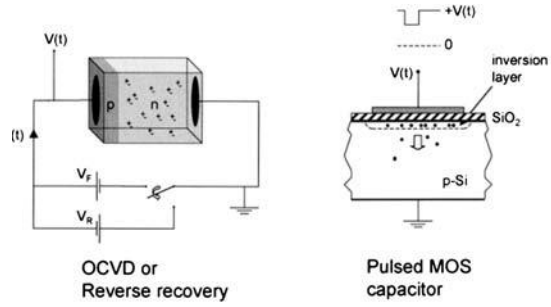


Figure 8. Principles of some device-related techniques.

Using a different geometry, where the *pn*-junction is present at the backside of a sample, lifetime mapping may be performed. Here, the fraction of carriers that reach the *pn*-junction will be collected as illustrated by the block arrow in the center panel of Figure 7. This geometry may be used both in EBIC and OBIC experiments where, for the latter, it is sometimes called the **flying spot** technique (Bleichner et al., 1986). A main disadvantage of the technique is the sensitivity to surface recombination and charging effects on the surface (because in this case only a small fraction of the generated carriers are collected by the *pn*-junction).

Another technique which is suitable for lifetime mapping of bare wafers (without *pn*-junctions) is the **surface photo voltage** technique (SPV), see the right-hand panel of Figure 7. The surface of the probed sample is treated chemically to provide surface charges, inducing surface band bending. Excess carriers, generated optically, reach the surface to induce a voltage shift, ΔV, at the surface, which is measured by a local contact probe. The diffusion length can be extracted by measurements at different wavelengths to provide different penetration depths in the semiconductor (see Schroder, 1990).

Device Related Techniques

Lifetime techniques based on device measurements fall into a separate class of methods, not because of different physical principles, but because they provide **local measurements** of the lifetime in operating devices. This is also a restriction, as for large component areas only an effective lifetime is obtained, averaged over device area and depth. The main advantage is the relatively simple setup, as measurements may be performed directly on the processed wafer, e.g., at a probe station, facilitating automated measurements on a number of devices and with the help of only electrical instruments. In a few cases, measurements may also be performed without the detrimental effects of surface recombination, e.g., for devices covered with a passivating oxide.

The left panel of Figure 8 shows a schematic of lifetime measurements performed on diode structures. In a first period of the measurement, electron-hole pairs are generated continuously during forward conduction by injection from emitters. Then, the circuit is broken by a fast electrical switch, resulting in an open circuit for the rest of the measurement interval, as in the **open circuit voltage decay** (OCVD) technique. If the lifetime of the excess carriers is significantly longer than the switching time, the carriers will remain within the device, their concentration decreasing as a result of recombination. For this time period, the diode will work as a solar cell, providing an open circuit voltage that decays logarithmically with carrier concentration. Thus, from the measured *V(t)* transient, the carrier lifetime can be deduced.

In the alternative technique of **reverse recovery** (RR), the switch is not left open but is switched immediately to a reverse voltage. Under such conditions, the remaining excess carriers will support a large current in the reverse direction until carriers have been depleted at the *pn*-junction such that a depletion layer may form.

Finally, in the **pulsed MOS capacitor** technique, minority carriers are provided by an inversion layer, as in an MOS transistor channel. By subjecting the top electrode of the MOS diode to a voltage pulse, the minority carrier concentration in the inversion layer must change accordingly. However, this cannot happen instantaneously since, for a positive pulse (*p*-type Si), minority carriers must be generated thermally or, for a negative pulse, minority electrons must recombine. In the latter case electrons are pushed into the substrate to recombine with majority holes. Thus, for positive pulses, a transient measurement of the capacitance provides the generation lifetime, whereas a negative pulse yields the recombination lifetime. As described above (see Theory), these lifetimes generally have very different values. The main advantage of the method remains practical, as MOS diodes may be easily provided as test structures to obtain lifetime information on fully processed CMOS wafers.

STEADY-STATE, MODULATED, AND TRANSIENT METHODS

As stated in the beginning of this unit, lifetime determination methods divide naturally into three basic groups: steady-state methods, modulated methods, and pulsed or transient methods (Ryvkin, 1964; Orton and Blood, 1990). Steady-state methods depend on measuring the dc magnitude of physical quantities such as PC,

photoelectromagnetic (PEM) effect, or PL. A number of modulated methods for monitoring carrier lifetime depend on simultaneously monitoring the amplitude and the phase lag between a measured signal and a sine- or square-wave modulated excitation. Pulse methods depend on measuring the decay of the carrier density following an excitation pulse. These three concepts themselves cover many of the somewhat more specialized device-related techniques (e.g., carrier collection in depletion regions of *pn* junction devices) and most of lifetime measurements based on diffusion length determination (Orton and Blood, 1990; Schroder, 1990). Here, a question naturally comes into mind as to whether there is any advantage or disadvantage to the measurement of lifetime by each of these methods.

The answer can follow two distinct paths. A simple answer is that in describing carrier lifetime we have to choose a preferable technique suited for a particular material, an expected range of lifetimes, and an expected accuracy of the measurement. For instance, one might prefer to use the modulated PL technique for characterizing light-emitting, highly doped, direct-band-gap semiconductors where the radiation yield is high and a lifetime is expected in 10^{-7} to 10^{-9} s range. It can be imagined that any dc method, from a sensitivity point of view, or a pulsed method based on PC, will have difficulties providing the necessary short pulses and would pose much greater demands on the detection system. Another answer is related to a more fundamental aspect of the carrier lifetime and its relation to other material parameters affecting recombination, as will be clarified below for modulated and quasi-steady-state-type methods. Transient methods are, therefore, explicitly described in forthcoming sections of the unit.

Principles of the Modulation-Type Method

The principle can be demonstrated by an oversimplified model. Suppose the sample is irradiated uniformly by a harmonic modulated optical source at frequency ω with a generation intensity $I = I_0 + I_1\exp(i\omega t)$, $i^2 = -1$. The concentration of minority carriers throughout the sample can easily be given from the continuity equation (Equation 1) under steady-state conditions, neglecting surface effects and nonlinear recombination terms. A solution can be presented of the form

$$\Delta n = \Delta n_0 + \Delta n_1 \exp(i\omega t) \tag{21}$$

where Δn_1 is complex (an expression of the fact that Δn and I have different phases). The expressions for amplitude and phase are

$$|\Delta n_1| = \frac{\tau I_1}{\sqrt{1+\omega^2\tau^2}} \quad \text{and} \quad \tan\varphi = -\omega\tau \tag{22}$$

The behavior of $|\Delta n_1|$ and φ are illustrated in Figure 9. It is clear that either parameter may be used to obtain τ when the signal is measured as a function of modulation frequency. The sensitivity of the phase lag is high enough in the interval $0.1 < \omega\tau < 10$ since the phase lag is 45° at

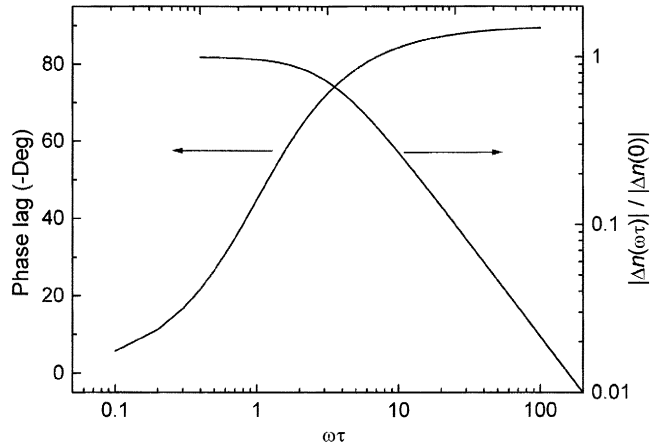


Figure 9. The frequency dependence of $|\Delta n_1|$ and φ as given by Equation 22, showing that both parameters may be used to measure carrier lifetime under harmonic modulation conditions.

$\omega\tau = 1$. To obtain τ from the amplitude versus frequency curve demands measurements up to frequencies of at least a factor of ten higher, i.e., when $\omega\tau \sim 10$. Note that the effective lifetime τ' in a modulated measurement is given by

$$\tau' = \frac{\tau}{\sqrt{1+i\omega\tau}} \tag{23}$$

and the effective carrier diffusion length from the relation $L_D = (D\tau)^{1/2}$ follows as

$$L_D' = \frac{L_D}{\sqrt{1+i\omega\tau}} \tag{24}$$

This clarifies the principle of the modulation methods. The phase shift can be determined very precisely at a single selected frequency, using a narrow-band receiver and lock-in detection, thus separating noise problems, which are usually associated with a direct measurement of a transient response in any pulsed lifetime method. In practice, it is important to study both amplitude and phase shift over a wide frequency range in order to establish an appropriate regime, after which routine characterization may rely on a single measurement of phase shift at a moderately low frequency.

Principles of the Quasi-Steady-State-Type Method

From Figure 9, it also follows that if the modulation frequency varies very slowly compared to the effective lifetime, $\omega\tau \ll 1$, the phase lag approaches zero. In this case a quasi-steady-state condition is reached and the amplitude of the slowly modulated signal offers an expedient method for steady-state lifetime measurement. Under steady-state illumination, a balance exists between the generation and the recombination of electron-hole pairs. Expressing the photogeneration and recombination rates as photoinduced current and recombination current densities (J_{ph} and J_{rec}, respectively) it may be stated as

$$J_{\text{ph}} = J_{\text{rec}} \tag{25}$$

As a consequence of this balance, an excess concentration of electrons and holes is established in the material. The total recombination in a sample of thickness d can conveniently be expressed in terms of an average excess-carrier density Δn_{av} and an average lifetime τ_{av}, leading to:

$$J_{ph} = \Delta n_{av} q d / \tau_{av} \tag{26}$$

which, in essence, is a version of the classic relationship, $\Delta n = \Delta G \tau_{av}$ (Ryvkin, 1964). For clarity, we state the classical analysis of the photoconductance, which, therefore, is a nearly direct way of probing the excess carrier density. Equations 25 and 26 can be iterated to find the average carrier lifetime as

$$\tau_{av} = \Delta\sigma / [J_{ph}(\mu_n + \mu_p)] \tag{27}$$

where μ_n and μ_p are corresponding mobility of electrons and holes. The conductance $\Delta\sigma$ and the incident light intensity can be measured using a calibrated instrument and a reference solar cell or photodiode, respectively. For a given illumination, J_{ph} can be estimated through available computer programs or lookup tables. In a typical 380-μm-thick silicon wafer with low reflection losses, the standard global AM1.5 solar spectrum produces a total photogeneration J_{ph} of ~38 mA/cm^2 (AM1.5 standard reference spectrum defined in ASTME 892). The value of J_{ph} relative to the reference solar cell can be adjusted for a particular sample to take into account different surface reflectivities, sample thicknesses or illumination spectra.

Compared to a transient decay approach, the quasi-steady-state method allows a lifetime measurement without fast electronics or short light pulses. The range of measurable lifetimes is only limited by signal strength. Thus, the measurement is preferred for samples of low dark conductance. To measure very short lifetimes, the light intensity can be increased. Note, however, that the average lifetime determined does not depend on the details of the carrier distribution within the sample. The local excess carrier density is essentially equal to the average Δn_{av} obtained from Equation 26 when the surfaces of the sample are well passivated and the carrier diffusion length is greater than the sample thickness. It is possible to find general steady-state solutions of the one-dimensional continuity equation subjected to boundary conditions for surface recombination in a semiconductor wafer (or in n^+p junctions) as, for example, expressed by Schroder (1990). The short-circuit current versus open-circuit voltage ($I_{sc} - V_{oc}$) characteristics can be equivalent to steady-state photoconductance measurements in solar-cell-type devices (Sinton and Cuevas, 1996).

Data Interpretation Problems

In practice, however, there are many cases where nonuniform carrier generation, surface recombination, nonuniform bulk recombination or nonlinear dependencies of recombination parameters introduce complications for both modulation and steady-state methods. The problem can be further complicated by various sensitivity factors for detecting Δn_1. All these effects imply that carrier diffusion currents at characteristic frequencies of the higher recombination terms (for modulated frequency $\omega\tau_i$) have to be considered. As a consequence, solutions for Δn_1 will be modified substantially. To calculate these expressions for the phase shift between the induced signal and the generation term, some kind of expression can be adopted. However, the resulting general expressions are too complicated to be useful.

A better procedure is to derive approximate expressions for the measured quantity by introducing various simplifying conditions in the steady state. Then, to make the appropriate substitution of τ' for τ (Equation 23) or L' for L (Equation 24), a number of expressions for modulated methods, where the algebra is reasonably tractable, is provided by Orton and Blood (1990). Specific microwave lifetime expressions by a harmonic optical generation method, which allows separation of the bulk lifetime and the surface recombination velocity on two wafer sides, is presented by Otaredian (1993).

An even more complicated situation can appear if carrier bulk lifetime is determined by the distributed recombination levels in the energy band, as is the typical case in amorphous semiconductors. Some guiding expressions for frequency-resolved modulated luminescence methods in amorphous semiconductors are provided by Stachowitz et al. (1995). Therefore, decay measurements can hardly be converted to modulation measurement data and vice versa, since mathematical procedures depend on the interpretation of the particular lifetime distribution function.

Summary of Strengths and Weakness

From the fundamental nature of the carrier lifetime and its relation to other material parameters, the strengths and weakness of different types of lifetime methods can be summarized as follows.

Steady-State and Modulation-Type Methods. The major strength of these methods is that they provide a rapid and less complicated measurement. The lifetime may be obtained either from the amplitude or the phase lag of Δn_1. The equipment is typically inexpensive.

The major weakness of these methods is the complicated interpretation of the recorded data. In particular, this becomes important if a weak feature of the lifetime has to be extracted correctly. Interpretation in terms of injection-dependent recombination parameters is typically not allowed. In steady-state type methods, the probe samples a relatively large volume, determined by the carrier diffusion length, preventing high-resolution measurements.

Pulsed-Type Methods. As explicitly shown by examples in the following sections, the major strength of these methods is an adequate interpretation of the measurement results. Injection dependency of the lifetime and other related material parameters (carrier diffusivity, surface recombination) can be interpreted with the same algorithm. There is a large background of knowledge related to pulsed methods that has been developed over the past

40 years. Some of the equipment operating in narrow bandwidth range can be relatively inexpensive, while other methods can provide very sophisticated measurements with high lateral or depth resolution for the extracted lifetime.

The weakness of pulse-type methods is that measurements typically require specialized equipment for working on different materials and in different bandwidth ranges. Pulsed techniques are sensitive to background noise problems and reference standards must be used for quantitative measurements.

The impact of recombination centers (usually generalized as carrier trapping) on the lifetime measurement type under steady state or transient conditions is, therefore, not clear. This is particularly the case if recombination of carriers in semiconductors takes place via several parallel recombination channels. The analysis of the experimental data, which is intended to unravel the individual contributions of the participating recombination channels is, however, often done on an intuitive basis rather than by a physically motivated formalism. The typical presumption that $\Delta n = \Delta p$ is only valid for the case of perfectly symmetric capture rates of electrons and holes at recombination centers, and holds for high excitation conditions. Some authors claim that while this presumption is a reasonable approximation for steady-state conditions, it may lead to grossly erroneous results for transient conditions even in the limit of a small concentration of recombination centers (Brandt et al., 1996). On the other hand, the shape of the transient measurement, in certain cases, makes possible the visualization of the evolution of carrier trapping quite evidently.

FREE CARRIER ABSORPTION

The FCA method is a noncontact, fully optical pump-probe technique suitable for bulk lifetime measurements. It is based on optical injection of excess carriers, most often by a short laser pulse, and probing of the excess carrier decay by a continuous wave, long-wavelength laser beam transmitted through the sample (Ling et al., 1987; Waldmeyer, 1988). It is less common as compared to other techniques described in this unit and no commercial apparatus has been developed. Yet, it is a quite universal technique, able to accommodate very different sample structures and semiconductor materials. The technique can also be easily set up in a research laboratory that has access to a pulsed laser of appropriate wavelength. So far, the technique has been applied to indirect type semiconductors (Si, Ge, SiC, porous Si, etc.).

Another advantageous feature is a possibility for control of the injection level. This results from a homogeneous excitation volume, enabling a calibration of the carrier density. Thus, lifetime extraction may be performed at different injections probing different recombination mechanisms. This feature has inspired the selection of the FCA method as a demonstration unit to illustrate various concepts, as stated in the previous Theory section. For general reviews of the technique, see Linnros (1998a,b), Grivickas et al. (1992), and Waldmeyer (1988).

Principles of the Method

The absorption of long-wavelength light induced by free carriers in a semiconductor, i.e., for photon energies less than E_g, can be used to derive the carrier density in a sample. The absorption mechanism is related to acceleration of electrons and holes by the electric field of the incident radiation and generally increases with wavelength (Pidgeon, 1980). Elementary theory predicts a dependence of the absorption according to (for free electrons of density n)

$$\alpha_{\mathrm{FCA}} = \frac{\lambda^2 q^3}{4\pi^2 c^3 n^* \varepsilon_0} \cdot \frac{n}{m_{\mathrm{n}}^2 \mu_{\mathrm{n}}} \tag{28}$$

where λ is the wavelength, q the electron charge, c the velocity of light, n^* the real part of the refractive index, ε_0 the permittivity in vacuum, and m_{n} and μ_{n} the electron mass and mobility, respectively. In reality, different absorption mechanisms may dominate and the quadratic dependence on wavelength may be replaced by a more general power law: $\alpha \sim \lambda_{\nu}$ where ν is a power index to be experimentally determined. The most important consequence of Equation 28 is the linear dependence of the absorption on carrier concentration, n.

For lifetime measurements, free carrier absorption may consequently be used to monitor the instant excess carrier density following optical injection of electron-hole pairs. As carriers are generated and recombine in pairs (assuming negligible trapping) it follows that the measured absorption is the sum of the contributions from electrons and holes, even though their individual absorption cross-sections may be different.

The transmitted probe beam intensity through a sample is described by

$$I(t) = I_0 \exp[-\alpha_{\mathrm{probe}}(t) \cdot d]; \quad \alpha_{\mathrm{probe}}(t) = \alpha_0 + \Delta\alpha(t) \tag{29}$$

where I_0 is the incident beam intensity, d the sample thickness, and α_0 a constant absorption coefficient related to, for example, dopant-induced free carriers. $\Delta\alpha(t)$ is the absorption coefficient, which is related to the density of **free excess carriers** within the sample. Usually, a linear dependence as in Equation 28 is assumed

$$\Delta\alpha(t) = \sigma_{\mathrm{FCA}} \cdot \Delta n(t) \tag{30}$$

where $\Delta n(t)$ is the excess carrier density ($\Delta n = \Delta p$). Figure 10 shows measurements of $\Delta\alpha$ versus Δn in silicon for two different probe wavelengths. The data indicate a linear behavior up to $\Delta n = \sim 10^{17}$ cm^{-3}, followed by a slightly increased absorption cross-section at higher carrier concentrations.

In a general case, the carrier density may not be homogenous throughout the probed volume of the sample, and Equation 30 may be replaced by

$$\Delta\alpha(t) = \sigma_{\mathrm{FCA}} \int_0^d \Delta n(z,t) dz \tag{31}$$

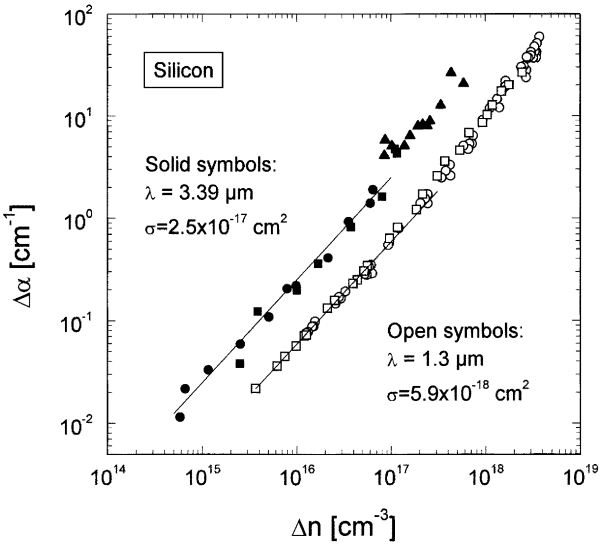


Figure 10. Absorption induced by excess carriers in silicon for two probe wavelengths. The excess carrier concentration was calculated using measured absorbed optical power within a sample and assuming a unity quantum efficiency for the conversion of photons to electron-hole pairs.

where $\Delta n(z,t)$ is the local excess carrier density along the probe beam trajectory z within the sample, and a linear cross-section has been assumed as in Equation 30.

The pump beam photons generate free carriers in the form of electron-hole pairs, provided that the wavelength matches the fundamental absorption edge of the semiconductor sample. In the transient FCA method, the pump beam is represented by a short laser pulse. The carrier density depth profile can be calculated from the generation function described by

$$g(z,t) = g_0(t) \cdot \frac{\alpha_{\text{exc}}(1-R)^2 \exp(-\alpha_{\text{exc}} \cdot z)}{1 - R^2 \cdot \exp(-2\alpha_{\text{exc}} \cdot z)} \qquad (32)$$

where $g_0(t)$ is the incident flux density of photons, R the sample reflection coefficient (assuming equal reflectivity on front and back surfaces), and α_{exc} the absorption coefficient for the pump beam. An inspection of Equation 32 suggests that for an in-depth, homogeneously excited sample, $\alpha_{\text{exc}} \cdot d$ should be less than unity, and therefore, the penetration depth should be large compared to the sample thickness (refer to the experimental data of Figure 4 where this condition is barely fulfilled, $\alpha_{\text{exc}} \sim 10$ cm^{-1} yielding $\alpha_{\text{exc}} \cdot d \sim 1$).

Figure 11 is a schematic of a **transient** FCA measurement setup using a **parallel** geometry, i.e., both excitation and probe beams enter the sample from the same side. The pump beam is pulsed and the excited area should be large compared to typical diffusion lengths in the sample. The probe beam is from a continuous wave laser with a wavelength tuned toward the IR compared to the band-gap absorption. By focusing, the probed sample area is smaller than the excitation area and may, therefore, be regarded as laterally homogenous. The transmitted probe beam is

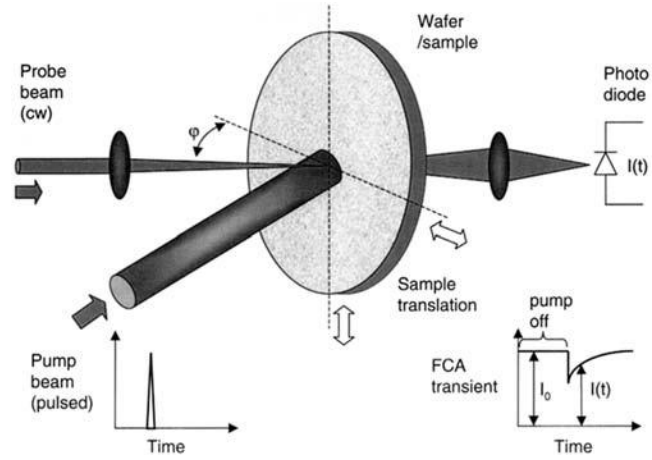


Figure 11. Schematic of a pump-probe experiment to measure carrier lifetime in a parallel geometry. Note that $90^\circ - \varphi$ should be close to the Brewster angle to minimize multiple reflection of the probe beam.

collected by a lens and focused onto a photodetector, usually a *pin*-photodiode. To allow measurements on different areas of the sample, such as for lifetime mapping, the sample may be scanned in the plane of intersection of the two beams (e.g., by motorized translators). The two inserted time diagrams depict the laser-pulse shape and the detected absorption transient. The characteristic time for the FCA signal to revert to equilibrium is a measure of the carrier lifetime. Indeed, the temporal pulse width of the excitation laser must be short compared to the carrier lifetime of the sample in order not to affect the lifetime extraction.

An alternative geometry is the perpendicular setup (Grivickas et al., 1992; Linnros et al., 1993), shown in Figure 12, where the probe beam impinges on a cross-sectioned surface of a sample, i.e., at right angles with

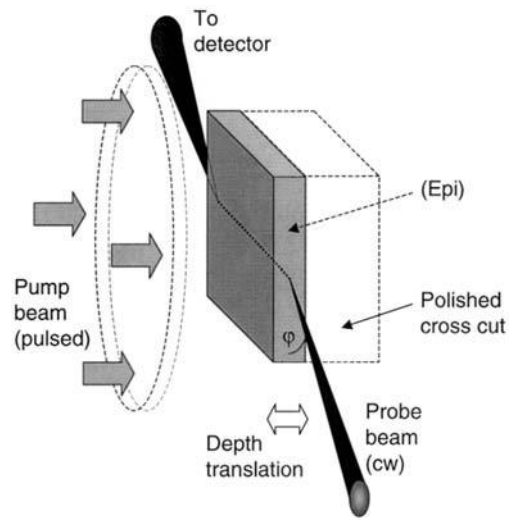


Figure 12. Schematic of a pump-probe experiment to measure carrier lifetime in a perpendicular geometry. Dashed lines show geometry when a surface epilayer is investigated.

respect to the pump beam. This geometry enables depth-resolved measurements by moving the probe beam impingement spot on the cross-sectioned surface, or rather by moving the sample in the depth direction. As indicated by dashed lines, this geometry also allows lifetime characterization of an epilayer on top of a substrate, provided the epilayer is sufficiently thick that the probe beam may be focused and confined in this layer (Galeckas et al., 1997a).

In the selection between the parallel and the perpendicular geometry, one notes that the parallel scheme results in a measurement of the average lifetime throughout a wafer (provided the penetration depth of the pump beam is large with respect to sample thickness). This is advantageous in many cases, such as during lifetime mapping. On the other hand, the perpendicular scheme provides a few other important advantages in addition to depth resolution. First, a uniform excitation is ensured along the beam path, even though excitation may be fairly inhomogeneous in depth. Second, samples may be cut in such a way that the probe beam path is quite extended, yielding a high sensitivity in terms of detected carrier concentration (refer to Equations 29 and 30, using a large d). This is often a prerequisite to reach low-level injection conditions.

Finally, a related method is the modulated free carrier absorption technique (Sanii et al., 1992; Glunz and Warta, 1995) where, instead of a pulsed excitation beam, a modulated (usually sinusoidal) excitation beam generates carriers. The probe beam then monitors the resulting carrier modulation amplitude (or phase lag), which drops to zero (phase lag increases to 180°) at frequencies approaching the inverse lifetime ($f \sim 1/\tau$) as carriers are unable to follow the imposed modulation (see Principle of the Modulation-Type Method).

Practical Aspects of the Method

Pump Laser Selection. For measurement of bulk lifetimes, a homogeneous excess carrier excitation in the sample is preferable to ensure the extraction of an in-depth average value of the bulk lifetime. As stated previously, this demands a relatively low absorption coefficient of the pump light, in turn requiring a photon energy close to the energy band gap. As an example, a yttrium-aluminum-garnet (YAG) laser operating at $\lambda = 1.06$ μm is ideally suited for Si wafers, as the absorption is ~ 10 cm^{-1}, yielding a depth inhomogeneity of the initial carrier concentration less than ~30% for normal 350-μm thick wafers (also see Fig. 4, where absorption at this wavelength is illustrated by the initial carrier concentration, i.e., the top depth profile).

In contrast, when thin layers close to a semiconductor surface must be analyzed (in a parallel geometry as in Fig. 11), a higher absorption coefficient is desirable. An example is the extraction of the carrier lifetime in a thin epitaxial layer on top of a substrate (Galeckas et al., 1997b). In such a geometry, the part of the generated excess carrier profile penetrating into the substrate needs to be suppressed, requiring an excitation wavelength considerably shorter than for the previous case. One should keep in mind, however, that the result of such a measurement must be critically analyzed, since diffusion of carriers towards the surface and into the bulk becomes a more serious limitation as the thickness of the probed layer is reduced, and could seriously affect the measured carrier lifetime. In other words, the useful time domain of the expected lifetimes is pushed toward very short lifetimes, often into the ns or ps regime.

Finally, pulse duration and beam size must be considered. Whereas pulse duration must be kept below the shortest expected lifetime in the sample, minimum beam size should be at least a few carrier diffusion lengths in diameter. From practical aspects and ease of alignment, a diameter of a few millimeters is desirable, often necessitating the use of a beam expander.

Probe Laser Selection. The selection of probe laser wavelength is determined by sensitivity and focusing requirements. Generally, a high sensitivity is the primary requirement in speeding up lifetime measurements, in particular for lifetime mapping. As free carrier absorption increases with wavelength (refer to Equation 28; notice the predicted quadratic dependence), long wavelengths toward the IR range are preferable and the choice is often set by laser availability. For increased lateral or depth resolution, short wavelengths are advantageous. But, optimizing resolution is not a simple task, as the optical resolution is not set simply by the Airy-focusing spot size. Instead, the **effectively probed volume** in the sample determines the resolution as set by sample thickness, focusing lens, and wavelength. For a more detailed treatment, the reader is referred to Linnros (1998a). Finally, carrier diffusion most often sets the practical resolution, at least for indirect-band-gap materials with lifetimes in the μs range.

As probe lasers, HeNe lasers of ~1-mW power were traditionally used at operating wavelengths of 3.39, 1.3, or 0.632 μm, depending on band gap and the above selection criteria. For increased sensitivity, a CO_2 laser with 10.6-μm wavelength would be preferable (Polla, 1983). Recently, semiconductor lasers have become abundant, offering several wavelengths in the visible and near-IR at affordable prices. Also, relatively intense lasers have become available (100 mW) offering increased measurement speed, although care must be taken not to affect the carrier dynamics by heating. A polarized laser is to be preferred (see Geometrical Considerations).

Detection Electronics. The signal from the photodetector is normally amplified and subsequently fed into an oscilloscope. To reduce noise, it is important to limit the bandwidth to a minimum with respect to characteristic times as set by the detected recombination transients. This may most easily be achieved by reducing the oscilloscope bandwidth, an ability provided by most modern equipment. Even though transients might be observed by the eye on the oscilloscope screen, digital oscilloscopes are far superior, providing digital averaging on a large number of transients. In the selection of a digital oscilloscope, two features (besides analog bandwidth and sampling rate, which need to be compatible with expected lifetimes)

are particularly important for FCA measurements: (1) The ability to average detected transients, usually from an 8-bit A/D converter, into a 16-bit memory, providing greatly enhanced sensitivity to small FCA transients; and (2) an arbitrary offset bias on the input channel, offering the possibility to offset the steady-state I_0 value and increase the gain on the transient part of the signal.

Geometrical Considerations. To design the geometry of the setup, one needs to pay particular attention to interference effects of the probe beam within the sample. This results from the high index of refraction of semiconductors, yielding typical reflection coefficients of ~30%. For very rough surfaces, interference effects may be reduced, but in some cases (for normal Si wafers with nonpolished backside) the problem may still persist. The remedy is to lower the reflected part of the beam using a polarized probe beam and a Brewster angle geometry, as defined by $\tan(\theta_i) = 1/n_s$, where n_s is the semiconductor refractive index and θ_i the incidence angle. For Si the Brewster angle is ~74°. Therefore, the sample needs to be inclined with respect to the probe beam as indicated in Figures 11 and 12 [where φ corresponds to $(90^\circ - \Theta)$]. For depth profiling using the perpendicular geometry, the same holds for the probe beam entering the cross-sectioned surface of the sample.

As for the pump beam, interference is not a critical issue, but, the use of a Brewster-angle geometry definitely reduces the pulse-to-pulse noise of absorbed power and the resulting injected carrier density.

Method Automation

Computer Interfacing. The automation of lifetime measurements using the FCA technique is more or less mandatory, as the absorption transient may only be observed directly on an oscilloscope screen under strong carrier injection. The first automation step is the interfacing of a digital oscilloscope with a computer. The interface should have high data transmission capability, usually in the range of ~1 Mb/s, as found in modern instrument bus interfaces. A computer code to handle the reading of decay data may be written in any programming language and should be able to handle storage and presentation of data, preferably also including conversion of raw data to carrier concentration.

Lifetime Depth Profiling. For depth-resolved measurements (Linnros et al., 1993), computer control of a motorized micropositioner is needed. The micropositioner should be able to handle movements of 1 μm with good precision. The actual depth scan is performed by stepwise motion to the next position after the accumulation and readout of an averaged FCA transient. As focusing of the probe beam is crucial, the measurement is preceded by scanning the probe intensity across the depth. The fall-off of the intensity at sample edges is an indication of probe beam focusing. Thus, several such scans using different lens-to-sample distances allows the determination of correct focusing by optimizing for steep edges.

Lifetime Mapping. Lifetime mapping demands a high degree of automation and precision-motorized positioners in the two orthogonal scanning directions (Glunz and Warta, 1995; Linnros et al., 1995; Linnros, 1998b). As in the case of depth profiling, data accumulation and readout is followed by stepwise motion to the next position. The computer code controlling the mapping now becomes relatively sophisticated in order to optimize speed (becoming a primary issue), data handling, selection of positions within the periphery of a wafer, and finally ease of handling for the user. Data storage also becomes critical, and normally only the extracted lifetime is stored for each sample position. Here, the procedure for correct lifetime extraction is essential, as signal-to-noise ratio must be traded against measurement speed.

Data Analysis and Initial Interpretation

Extraction of Carrier Decay Transient. The measured absorption transient, as recorded by a digital oscilloscope and transferred to a computer, may be transformed to a carrier decay transient using Equations 29 and 30 and solving for $\Delta n(t)$

$$\Delta n(t) = \frac{1}{\sigma_{FCA}} \cdot \frac{1}{d} \ln \frac{I_0}{I(t)} \tag{33}$$

where α_0 now has been lumped together with I_0 as it represents a constant absorption during the measurement. To correctly define I_0, a few data points before the actual beginning of the pump pulse are needed, as shown in the inset of Figure 11. Thus, triggering of the oscilloscope starts slightly before the pump pulse. Finally, sample thickness must be measured and an appropriate absorption cross-section must be inserted. However, to extract the lifetime, neither thickness nor cross-section is needed, as is evident from Equation 7.

Lifetime Analysis. The lifetime may be derived straightforwardly by applying Equation 7 with some numerical averaging profile over a few data points. An alternative route is by plotting $\log\Delta[n(t)]$ versus t followed by fitting of a straight line to some limited range of data points. This yields a better control of the lifetime extraction procedure, as different recombination mechanisms may be distinguished in the carrier decay transient. This is illustrated in Figure 13, where a recorded absorption decay transient has been converted to excess carrier density. The measurement was made on a p-type Si wafer doped to 3×10^{16} cm^{-3} with boron, and a $\lambda = 3.39$-μm probe beam was used (excitation by a pulsed $\lambda = 1.06$-μm beam). To increase the signal-to-noise ratio, averaging of several thousand decays was performed at a repetition frequency of 100 Hz. This yielded a detection sensitivity of $<10^{14}$ carriers/cm^3 (signal-to-noise level), as seen in the figure.

Analyzing the decay shape, one may distinguish three distinct regions. (1) At high injections nonexponential recombination is present due to Auger recombination (2), while at slightly lower injection, but still above the doping density, a linear region is observed where a high-injection lifetime may be extracted. Finally, (3) at low injections, the

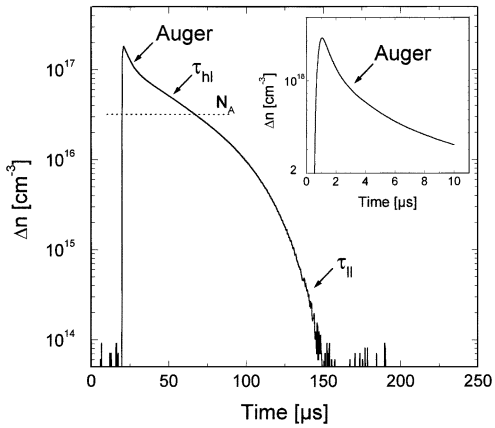


Figure 13. Excess carrier decay transients following an optical generation pulse in Si (p-type, $N_A = 3 \times 10^{16}$ cm^{-3}, thickness: $d = 430$ μm). The excess carrier density has been calculated using a linear cross-section (see Fig. 10), from absorption transients of a 3.39-μm probe beam. Inset demonstrates Auger recombination at high injections.

minority carrier lifetime may be derived. The inset shows a transient recorded at very high injection, where Auger recombination prevails.

In Figure 14, extracted lifetimes from the data of Figure 13 and a few additional recorded transients are displayed. Lifetimes were derived by a stepwise linearizing procedure of the data points in the transients. Solid lines represent a fit to the SRH formula (assuming a mid-gap trap energy level; see Equation 10 and Fig. 2), accomplished by varying the high- and low-injection lifetimes.

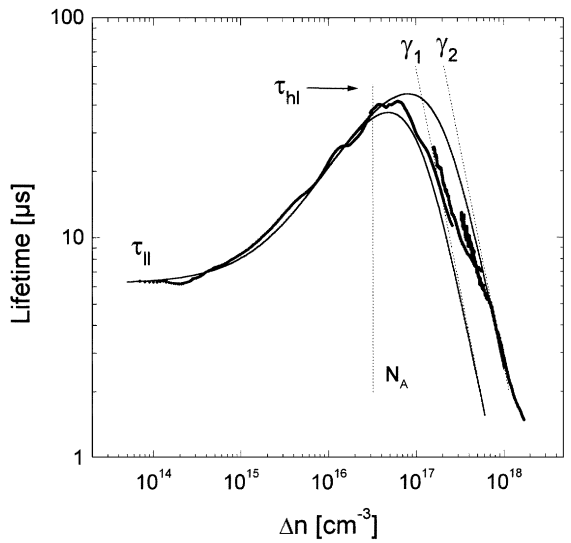


Figure 14. Calculated lifetime-injection dependence (solid circles) from experimental excess carrier decays as displayed in Figure 13. The two solid curves have been calculated according to the SRH theory (using Equations 10, 14, and 16) assuming a minority carrier (electron) lifetime of 6 μs and a high injection lifetime of 60 μs, and for two Auger coefficients. $\gamma_1 = 1.7 \times 10^{-30}$ cm^{-6}/s and $\gamma_2 = 3.8 \times 10^{-31}$ cm^{-6}/s (from Dziewior and Schmid, 1977, and Grivickas and Willander, 1999, respectively).

Auger recombination is also included (see Theory) where γ_1 and γ_2 represent two different Auger coefficients (Grivickas and Willander, 1999). One may note the good correspondence with theory and the broad transition range between the pure high-level (hl), low-level (ll), and Auger-injection case, complicating the interpretation of extracted lifetimes.

From the discussion above, one may conclude that the FCA technique allows a detailed study of carrier recombination at different injection levels, whereas an accurate determination of the minority carrier lifetime and/or the high injection lifetime may only be derived from plots similar to Figure 14 (by extrapolating to the pure ll or hl limit). In most cases, however, the lifetime may be estimated at one particular injection level, allowing comparison of different samples or lifetime mapping. Thus, the extracted lifetime value may not be the "true" minority carrier lifetime but represents a characteristic lifetime describing sample purity. The appropriate injection level for lifetime extraction depends on practical considerations (detection sensitivity) as well as on device applications for the analyzed sample/material (a tentative device may operate under high level injection and, therefore, the hl-lifetime is of primary interest).

Analysis of Lifetime Mapping Data. Figure 15A and B show examples of lifetime maps for two Si wafers. Carrier lifetimes were extracted from decay data at each position after the conversion to carrier concentration. The lifetime fit was performed using a fixed injection interval (as an example, the high injection lifetime could be derived from Figure 13 by a fit to the data in the interval 4×10^{16} to 7×10^{16} cm^{-3}). The mapping data have been displayed in the form of contour plots using the software Surfer for Windows (Golden Software).

Sample Preparation

Sample preparation ranges over the entire spectrum, depending on semiconductor material, surface properties, and measurement geometry. The listing below is structured according to different types of applications.

Virgin Wafers: Surface Passivation. To extract the bulk lifetime in virgin wafers, at least for Si wafers of good quality, is probably the most difficult task for any lifetime measurement technique. The reason is surface recombination in combination with large diffusion lengths of carriers (see Fig. 5). Indeed, the manufacturer specification for lifetimes is usually based on measurements on boules before cutting into wafers.

There are a couple of approaches to circumvent this problem. One technique is based on a passivation of the "dangling bonds" at the surface by hydrogen. This may be performed by a low-temperature chemical process (Linnros, 1998b), in a bath of HF (Yablonovitch et al., 1986), or by other chemical passivation methods (Horanyi et al., 1993). An alternative is to charge the native oxide of a Si wafer with ultraviolet (UV) light (Katayama et al., 1991) in order to provide band bending that will prevent carrier diffusion to surfaces (see center panel of Fig. 3).

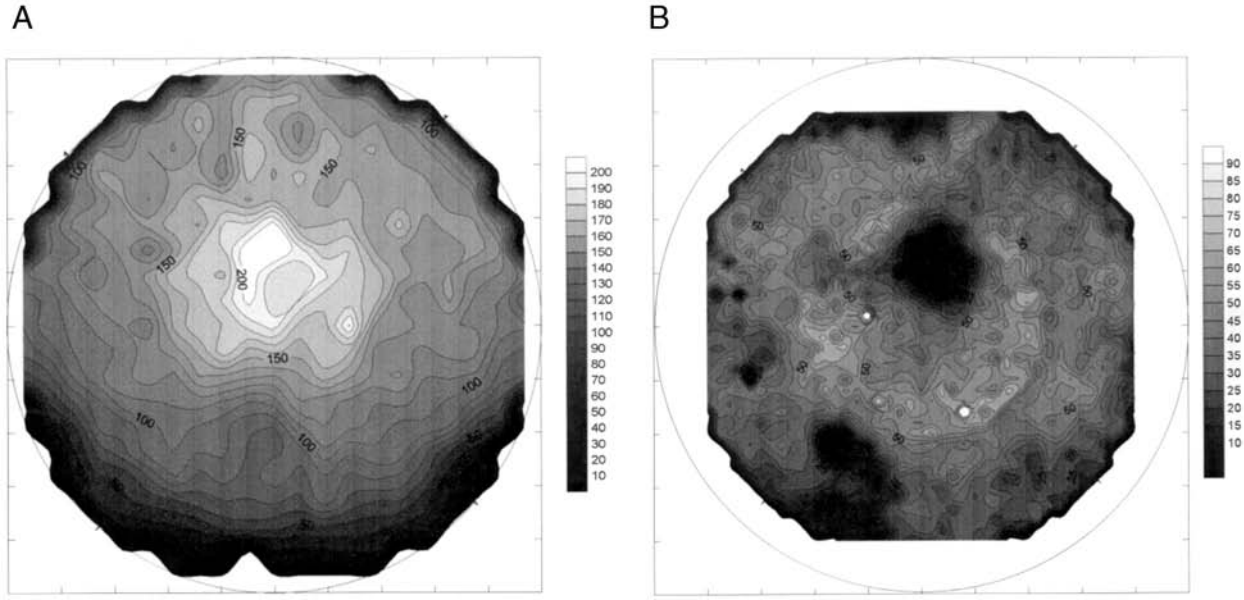


Figure 15. Examples of lifetime maps from processed Si wafers of FZ-NTD material. The data have been gray-scale coded with highest lifetimes in white. Circles at the periphery of the mapped areas indicate 4-in. wafer area. (**A**) Furnace-annealed wafer at 1050°C, 4 h, where low-lifetime area at lower part indicates contamination from the boat. (**B**) Wafer having passed a full thyristor process sequence. Note central high lifetime area corresponding to a ~45-mm diameter device as well as black, low-lifetime areas indicating contaminated regions (Linnros et al., 1995).

Common to all these methods are, however, problems with stability and repeatability.

Naturally, for Si wafers or other semiconductors with short carrier lifetimes, the problem of surface recombination is reduced, but, it still needs to be carefully addressed. Finally, an optical problem may arise for wafers/samples with nonpolished backsides as this may spread the probe beam such that the detector signal vanishes. In such cases it is necessary to polish the backside.

Processed Wafers and Devices: Removal of Metal and Highly doped Layers. In many cases, processed wafers are excellent for measurements of the bulk lifetime using the FCA technique, often without any preparation (Fig. 15). The reason for this is that furnace processing often results in a surface oxide, which may provide a very low surface recombination velocity. Another advantageous case is when dopants are diffused from the surfaces. This results in *pn*-junctions, nn^+- or pp^+-junctions that serve as barriers for carrier diffusion towards surfaces (see right panel of Fig. 3). For highly doped emitters, however, the probe beam may be completely attenuated by the high permanent concentration of carriers in these layers. A solution would then be to either remove this layer partly (by etching or polishing) or by decreasing the absorption using a shorter probe beam wavelength (Fig. 10 and Equation 28). For devices, metal layers must be removed to ensure transparency to pump and probe beams (Linnros et al., 1995).

Depth Profiling: Cross-Sectioning of Samples. For depth-resolved measurements using the perpendicular geometry, a cross-section must be cut from the wafer/sample. Special attention has to be paid to the parallelism of the surfaces and to the thickness of the cross-section. The latter determines the obtainable depth resolution and a graph for simple estimations for Si samples may be found in Linnros (1998a). The parallel surfaces of the cross-section normally have to be polished for the probe beam to enter and exit the sample and still remain collimated. Using a good cutting machine with a diamond blade, however, a mirror-like surface may be obtained. Polishing should be performed using conventional procedures, employing either diamond paste or abrasive paper of decreasing grain size.

Problems

As for most carrier lifetime measurement techniques, surface recombination is one of the main problems. This results from carrier diffusion, and the critical parameter is the diffusion length of carriers in comparison to sample dimensions. Thus, for short carrier lifetimes (e.g., in direct-band-gap materials) the effects of surface recombination may be negligible. On the other hand, one may be interested in carrier lifetimes of thin films or epilayers, and surface or interface recombination may again be a limiting factor. Different passivation methods to overcome surface recombination are briefly addressed above and in Linnros (1998a).

The diffusion length also excludes using the technique for local lifetime determination, as, e.g., in the top surface region of active devices. In such cases, one needs to resort to lifetime techniques based on generation lifetime as in a depleted *pn*-junction.

A specific problem with the FCA technique is the determination of minority carrier lifetime under very low injection conditions, as needed for wafers with low levels of doping. This is due to the limited sensitivity of optical absorption techniques at low carrier concentrations. The sensitivity may be optimized using large samples (in the probe beam direction), long wavelength probing, and extended averaging. In silicon, sensitivities reaching $\sim 10^{11}$ carriers/cm^3 have been reached (Grivickas et al., 1992) although the injection level during lifetime mapping is, for practical reasons, limited to $\Delta n > 10^{15}$ cm^{-3}.

Another problem, although very rarely encountered, is the measurement of extremely long carrier lifetimes (roughly for $\tau > 1$ ms in Si). In such cases the diffusion length may approach several millimeters, and a large excitation spot is needed to reduce effects of radial outdiffusion of carriers from the excited area (Linnros, 1998a).

PHOTOCONDUCTIVITY

The measurement of PC decay in various experimental modifications is widely used as a convenient method for measuring carrier lifetimes in semiconductors. As an example, the archetypal method of PC-decay under a constant current applied by means of ohmic contacts has been set as a standard technique for measurement of the minority-carrier lifetime in bulk germanium and silicon for >30 years (Standard Method for Measuring the Minority-Carrier Lifetime in Bulk Germanium and Silicon, F28 in ASTM, 1987). Numerous noncontact PC methods, such as microwave reflectance and rf bridges, have recently become quite popular because their nondestructive quality and also their suitability for use in a clean-room environment. Most of the considered techniques do not require particularly expensive and bulky instruments. This motivates the wide diffusion of noncontact PC decay techniques into diverse areas of carrier lifetime measurements including on-line processing of silicon wafers.

Principle of PC Decay Methods

The principle of the experiment involves shining a pulse of light into the semiconductor sample to generate excess carriers and monitoring the time decay of the corresponding excess conductivity. Since electrons and holes have opposite charge, the total conductivity is the sum of the partial conductivities, which is always positive. Consequently, conductivity can be written in the form

$$\Delta\sigma(t) = q[\Delta n(t)\mu_n + \Delta_p(t)\mu_p] \cong q\Delta n(t)[(\mu_n + \mu_p)] \quad (34)$$

where q is the electron charge and μ_n, μ_p are the corresponding drift mobilities of electrons and holes that characterize PC for different injection levels.

Before proceeding to details, it is worthwhile to reiterate two general points. First, the approximate equality for the last term of Equation 34 is valid in the absence of asymmetric capture within an impurity recombination model (Blakemore, 1962) or the absence of carrier trapping (Ryvkin, 1964). In general, these conditions are valid at high injections for carrier generation by the electromagnetic radiation in the fundamental band with a photon energy $h\nu \geq E_g$, where E_g is the forbidden gap energy. Then both Δn and Δp are proportional to the optical energy absorbed in that time and in the unit volume. Numerous experiments have shown that recombination misbalance effects can be neglected at any injection for measured lifetimes in the range 1 to 500 μs of crystalline Si and Ge (Graff and Fisher, 1979). However, we note that the approximate equality in Equation 34 does not hold for low injection conditions in some highly irradiated, highly damaged crystalline semiconductors and in polycrystalline or amorphous films (an example is provided by Brüggemann, 1997). The same indeterminate equation may appear in compensated compound semiconductor materials, since some of them may contain a substantial concentration of nonstoichiometric deep centers acting as carrier traps. These cases might require quite specific methods for carrier lifetime extraction from PC measurements. In the following, we shall assume that the approximate equality in Equation 34 is an inherent property of the semiconductor material. This assumption leaves complicated interpretation of PC transients out of the scope of this unit.

Second, it has usually been assumed that the PC decay process senses the carrier concentration decay where mobilities are constant parameters. Such an assumption may be made in the great majority of actual cases. However, this cannot be true at high injection where electron-hole scattering reduces mobility substantially, for example, in high-resistivity Si (Grivickas et al., 1984). A few papers have also reported that carrier mobility was changed during PC transients in high-resistivity polycrystalline and compound semiconductors because of the effective recharging of scattering clusters. We do not attempt to provide an exhaustive accounting for these effects. As explained earlier (see Theory), in an extrinsic semiconductor, under low injection conditions, we shall assume that the decay of excess carrier concentration is always controlled by minority carriers through SRH recombination and either by a direct band-to-band or an Auger-type recombination mechanism.

Practical Aspects of the Standard PC Decay Method

By the standard technique, PC decay is monitored through the change of sample conductance in a constant electric field. Current is passed through a monocrystalline semiconductor specimen by means of ohmic contacts. The experimental arrangement in its simplest form is shown in Figure 16. The sample is bar-shaped with dimensions of typically $l \gg d$, w and has ohmic contacts on its end faces. The light beam is oriented normal to the applied electric field (a case of transverse PC). The intensity decreases with depth in the sample according to Beer's law. Neglecting multiple reflection, the photoconductance of the whole sample can be obtained (Ryvkin, 1964) as

$$\begin{aligned}\Delta G &= (w/l)q(\mu_n + \mu_p)I_0[1 - \exp(-kd)] \\ &= (w/l)q(\mu_n + \mu_p)\int_0^d \Delta n_x dx, \end{aligned} \quad (35)$$

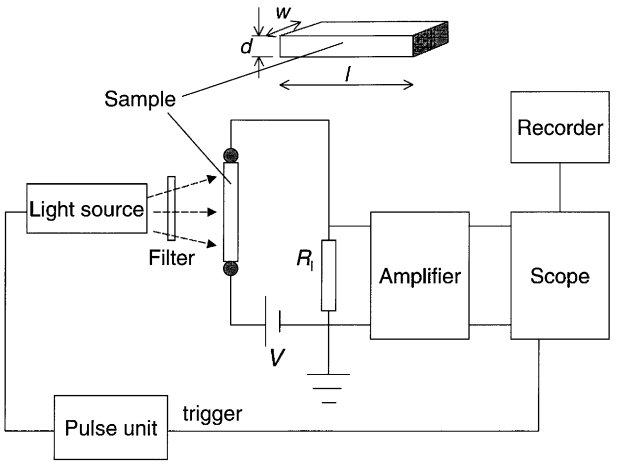


Figure 16. Basic experimental setup for photoconductive decay measurements in a constant electric field. Excess carriers are monitored as an increase in sample conductance.

where I_0 is intensity penetrated from the front surface of the sample and k is the absorption coefficient. For sufficiently thick samples ($kd \gg 1$) all light energy is absorbed in the sample and the photoconductance is independent of the absorption coefficient. Therefore, in this case the carrier distribution in the x direction is highly inhomogeneous.

In addition to the transverse PC (Fig. 16), the longitudinal PC is sometimes investigated. In this type of experiment, a semitransparent electrode is illuminated with the light beam parallel to the field. In general, the longitudinal case has a more complex relationship with fundamental semiconductor parameters. We confine our discussion in this unit to transverse photoconductivity.

As shown in Figure 16, the sample is connected in series with a battery supplying a voltage V and a load resistance R_L. Consequently, interrupting the light, the current in the circuit has a constant component and an alternating one. The voltage drop may be monitored with an oscilloscope or by a signal averager between the ohmic contacts (or between a second pair of potential contacts in a four-terminal measurement). If, in general, voltage drop is monitored on the load resistance R_L, a rather complex relationship between the alternating voltage received at the amplifier input and the variation of ΔG can be expected under the action of light. In some special cases, this relationship is simpler (Ryvkin, 1964). For example, if a small load resistance is used, $R_L \ll (G_0 + \Delta G)^{-1}$, where G_0 is the equilibrium conductance, then the relationship is linear because the illumination does not greatly alter the electric field distribution in the sample and the load resistance. This case sometimes is called the **constant field** regime. The resistor must be nonreactive, with a resistance at least 20 times less that of the sample in the excited state, to provide a condition of essentially constant field. As described by Ryvkin (1964), two other regimes, such as the **constant current** regime, $R_L \gg (G_0 + \Delta G)^{-1}$, or the **maximum-sensitivity** regime, $R_L = (G_0)^{-1}\,[1 + \Delta G/G_0]^{-1/2}$, can be frequently used. While the last two regimes, in general, do not imply proportionality between the signal and ΔG, a linearity is obtained when the relative change of the photoconductance is small ($\Delta G/G_0 \ll 1$). Thus, a constant current regime is recommended for minority carrier lifetime measurements in bulk Si and Ge by the ASTM standard F28 (ASTM, 1987).

The threshold of ΔG measurement sensitivity depends on various types of noise superimposed on the measured signal and the type of recording instrument. The minimum detectable photoconductance is also governed by the value of the equilibrium conductance, G_0, other conditions being equal (Ryvkin, 1964).

PC Method Automation for High-Frequency Range

Ordinary PC transient measurements are quite restricted on a nanosecond time scale because capacitance and/or induction of conventional electrical circuit elements produce failure effects. Recently, PC lifetime measurements in the 1- to 15-GHz frequency range have been developed through strip-line techniques by driving so-called Auston switches (Lee, 1984). Coplanar metal waveguides have been monolithically integrated on the semiconductor sample to provide characteristic impedance $\sim$50 Ω, with a corresponding spacing (photoconductive gap) of a few micrometers wide. Free carriers can be generated in the photoconductive gap by sharp femtosecond laser pulses and accelerated in an applied dc bias field, producing an electrical transient. The output of the circuit can be measured via high-frequency connectors attached to the striplines by a fast sampling oscilloscope. In this case, lifetime resolution down to $\sim$30 ps can be achieved.

Sampling experiments can also be performed without the need for an expensive fast oscilloscope. In this case, two photoconductive switches, one serving as a gate and having a very fast response function, can be connected in parallel and excited with delayed optical pulses. The temporal profile of an incident signal voltage pulse, $v_{\text{sign}}(t)$, can be measured by scanning the time delay, Δt, between the signal pulse and the gate pulse. The sampling yields a dc current given by the signal correlation

$$I(\Delta t) = \int \nu_{\text{sign}}(t) f_{\text{samp}}(t - \Delta t)\, dt \qquad (36)$$

where f_{samp} is the known sampling function corresponding to the optical gate pulse. Relatively inexpensive lock-in amplifiers can measure dc currents, provided that colliding optical pulses are chopped at low frequency. The lifetime can be estimated from mathematical simulations of the correlation function $I(\Delta t)$. The time resolution is limited by the duration of the sampling function of the gate switch. The ideal sampling function is a delta pulse, and, in fact, this has motivated the fabrication of a gate switch made from semiconductor materials with reduced carrier lifetime in a subpicosecond range (Smith et al., 1981). Therefore, the gate sample in this case should be regarded as a distributed circuit element, and the propagation of the wave through the photoconductive gap should be properly treated (Lee, 1984). Four factors have been shown to be of special importance. These are the free carrier lifetime, local dynamic screening, velocity overshoot, and the

carrier transient time from the photoconductive gap (Jacobsen et al., 1996).

High-frequency measurements usually involve high excitation levels and high electric fields. In these cases, carrier recombination is affected by fundamental processes of carrier momentum randomization, carrier thermalization, and energy relaxation in the bands (Othonos, 1998). At very high injections, PC in semiconductors usually obeys a nonlinear dependence as a function of light intensity. The most probable cause of any nonlinearity in a photoconductor is related to carrier lifetime. Pasiskevicius et al. (1993) have proposed to use PC correlation effects for measuring nonlinearities in ultrafast photoconductors. In this way, correlation signals are recorded on a single switch using a double pulse excitation scanned relative to one another. A peak will be observed on the autocorrelation traces when the photocurrent is a superlinear function of optical power and a dip when this function is sublinear. The widths of these extrema will be determined by the duration for which the switch remains in the nonlinear state.

The typical experimental setup used by Jacobsen et al. (1996) is sketched in Figure 17. Colliding pulses (100 fs) of a mode-locked Ti:sapphire laser are divided into two beams, with one beam passing through a delay line. Both beams are mechanically chopped and focused onto a biased photoconductive switch, and the relative arrival time is varied by mechanically moving the delay line. The current flowing in the photoconductive gap is measured by using a current-sensitive preamplifier, and the signal is sent to a lock-in amplifier. The preamplifier operation is slow compared to the time scale of the carrier dynamics, so a time average of the current is measured. In order to suppress noise when performing a photocurrent correlation measurement, the two beams are chopped at different frequencies f_1 and f_2, and the lock-in amplifier is referenced to the difference frequency $|f_1 - f_2|$. In this way, the correlation measurement integrates only the product of the carrier

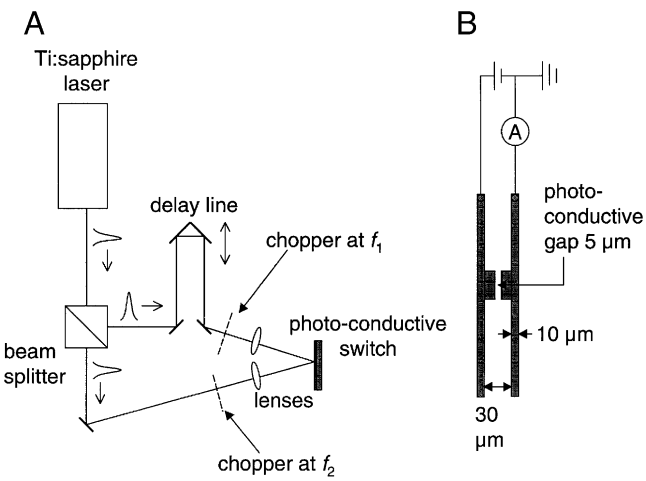


Figure 17. (**A**) The experimental setup for the photocurrent autocorrelation measurements in subpicosecond range. (**B**) Antenna design used as a photoconductive switch. The laser spot is focused in the 5-μm gap in the middle of the transmission line. Bias is applied across the gap and the current is measured.

density and the local electric field that corresponds to both pulses. Therefore, a rapid recharging of the photoconductive switch is required by a short recombination lifetime. Lifetimes as short as 200 fs have been measured using this correlation scheme.

Additional information on PC processes can be obtained from a set of compatible electrooptic techniques based on Pockels, Kerr, Franz-Keldysh effects, as reviewed by Cutolo (1998).

Detailed protocols of minority-carrier lifetime determination in bulk Si and Ge by a classical PC technique are provided in F28 in the Annual Book of ASTM Standards (ASTM, 1987). Other applicable documents of ASTM Standards are E177 (Recommended Practice for Use of the Terms Precision and Accuracy as Applied to Measurement of a Property of a Material (ASTM, 1987, Vol. 14.02) and F43 Test Methods for Resistivity of Semiconductor Materials (ASTM, 1987, Vol. 10.05).

The specimen resistivity should be uniform. The lowest resistivity value should not be <90% of the highest value (see Test Methods F 43; ASTM, 1987). The precision expected when the F28 method is used by competent operators in a number of laboratories is estimated to be better than ±50% for minority-carrier lifetime measurements on Ge or ±135% for measurements on Si, as defined in Recommendation Practice E177.

A similar, but not equivalent method is under the responsibility of Germany DIN Committee NMP 221. This is DIN 50440/1, Measurement of Carrier Lifetime in Silicon Single Crystals by Means of Photoconductive Decay: Measurement on Bar-Shaped Test Samples. This document is available from Beuth Verlag GmbH, Burggrafenstrasse 4-10, D-1000 Berlin 30, Germany.

Data Analysis and Related Effects

For reliable interpretation of the experimental data by the standard PC method it is important that the measuring electric field E be uniform throughout the sample. This implies the use of good ohmic contacts, and it is also necessary for contacts to cover the whole area of the end faces. Also, the sample cross-section and resistivity must remain constant over the sample length. Another requirement is that the magnitude of the electric field be low, such that the injected carrier density is not disturbed by excess carrier sweepout. The appropriate criterion can be obtained applying a drift term to the continuity equation (Orton and Blood, 1990). This leads to the condition

$$(\mu E)^2/4D \ll \tau_{\text{eff}}^{-1} \tag{37}$$

where D is carrier diffusion coefficient. If $\mu = 10^3$ $\text{cm}^2\text{V}^{-1}\text{s}^{-1}$ and $\tau_{\text{eff}} = 1$ μs, and applying the Einstein relation, $D = k_{BT} \cdot \mu/q$, for the diffusivity, this results (at 300 K) in the requirement that E should be much less than 10 V/cm. It is advisable to check experimentally that a change in E has no effect on the measured decay time.

Another possible source of error lies in the generation of photovoltages at the contacts if these are not ohmic. The signal should be observed without applied voltage in order to determine the magnitude of such a photovoltage. The

amplitude of any photovoltaic signal should be less than 1% of the desired photoconductive signal. Also, the same PC amplitude with applied voltage must be observed for both polarities of the constant-voltage source. If the specimen voltages for the two polarities agree to within 5%, the contacts may be considered as nonrectifying. All connections to the contacts should be made with shielded cables, which must be as short as possible. The requirements for the performance of the preamplifier should be considered together at the same time. For working in the small-signal regime, it is advisable to mask contacts and illuminate only the central part of the semiconductor bar (F28, ASTM, 1987).

For the measurement of minority carrier lifetime, a finite injection level is required under low-level injection conditions. This level can be obtained by the standard PC measurement technique from the ratio of the voltage change, ΔV, across the illuminated specimen, to the voltage, V, across the specimen when the specimen is not illuminated. When the ratio $\Delta V/V$ is restricted to <0.01, the corrections to the lifetime value may be ignored. When the modulation of the measured voltage across the specimen exceeds 0.01, a correction is required to find the minority carrier lifetime, τ_{ll}. From the voltage decay time, τ_V, which is the quantity actually measured, a correction to the minority carrier lifetime is recommended as determined by the following equation (F28, ASTM, 1987)

$$\tau_{ll} = \tau_v[-(\Delta V/V)] \tag{38}$$

Sample Preparation

The largest possible specimen should be used, to minimize the contribution of surface recombination to the observed lifetime value. By the method description in ASTM Standard F28 (ASTM, 1987), it is also recommended that surface effects be attenuated by reducing surface recombination. Alternatively, the specimen surface can be sandblasted or ground to produce a reference surface for which a limiting high surface recombination rate may be calculated (see Equation 20 and Fig. 5). It is recommended that the corrections applied to an observed filamentary specimen lifetime never exceed the observed value. That is, the condition for low level-injection lifetime, $\tau_{ll} > \tau_b/2$, should never be violated. Tables 3 and 4 show the maximum values of the bulk lifetime, τ_b, which may, therefore, be determined on three standard small-sized specimens according to the F28 method.

Although many methods may be used for making contact connections (e.g., diffusion or implanted n^{++} or p^{++} layers), it is suggested by F28 that pressure contacts of metal braid or metal wool can be used for measurements at low injection. Thick sheets of Pb or In have also been found to be acceptable for this purpose. The ends of Ge specimens should be plated with either Ni, Rh, or Au. Nickel plating is satisfactory on the ends of n-type Si specimens and rhodium plating on the ends of p-type Si specimens. Copper should be avoided in the plating operation. Electrical contacts formed by alloying with an Au-Sb-Si eutectic can be recommended for PC measurements on n-Si at high-injection levels (Grivickas et al., 1984).

Different technological procedures for minority-carrier lifetime retention in Si wafers—before and after thermal heat treatment in various atmospheres, as well as after chemical treatments, after final deionized water rinsing, etc.—have been presented with respect to minority-carrier lifetime analysis by Blais and Seiler (1980), Graff and Pieper (1980), and Graff and Fisher (1979).

Table 3. Dimensions of the Three Small-Sized Filamentary Specimens for which Limiting Bulk Lifetime Data are Presented in Table 4

Size Designation	Length (cm)	Cross-Section (cm)
A	1.5	0.25 × 0.25
B	2.5	0.5 × 0.5
C	2.5	1.0 × 1.0

Table 4. Maximum Values of Bulk Lifetime, τb, in μs, for the Three Size Designations in Table 3, as Recommended by the F28 Photoconductivity Method for Minority-Carrier Lifetime Measurements at 300 K

Material	Size A (μs)	Size B (μs)	Size C (μs)
p-type Ge	32	125	460
n-type Ge	64	250	950
p-type Si	90	350	1340
n-type Si	240	950	3600

Microwave PC Decay Method

A wide range of alternative PC decay methods, which have the advantage of not requiring electrical contacts to the sample, depend on the measurement of **microwave reflectance** power. Due to the plasma effect, the dielectric constant of the semiconductor is a function of excess conductivity. A variety of experimental arrangements is possible depending on available microwave frequencies: e.g., through strip line systems (1–4 GHz), or through X-band (8–12.4 GHz), or through the Ka-band waveguide apparatus (28.5–40 GHz). The choice can be made between transmission or reflection systems, the use of a resonant cavity, or geometrical configurations within a partially closed system or outside it. Some methods are described by Orton and Blood (1990). Several commercial apparatus for microwave reflectance lifetime measurements are also available. Therefore, we briefly describe the most common experimental system, which illustrates the general features of microwave methods.

A simplified example of such a setup is shown in Figure 18A. Radiation from a frequency-stabilized microwave source is divided at the circulator. Power reflected from the sample arm is again split and is transmitted to a crystal detector in the other arm. Also mounted in the sample arm is a sliding short circuit that allows the sample to be placed in the position of a uniform microwave field. In other representative setups, the sample is simply backed by a metallic plate, which serves to increase the reflected power. The power reflection coefficient R' describes

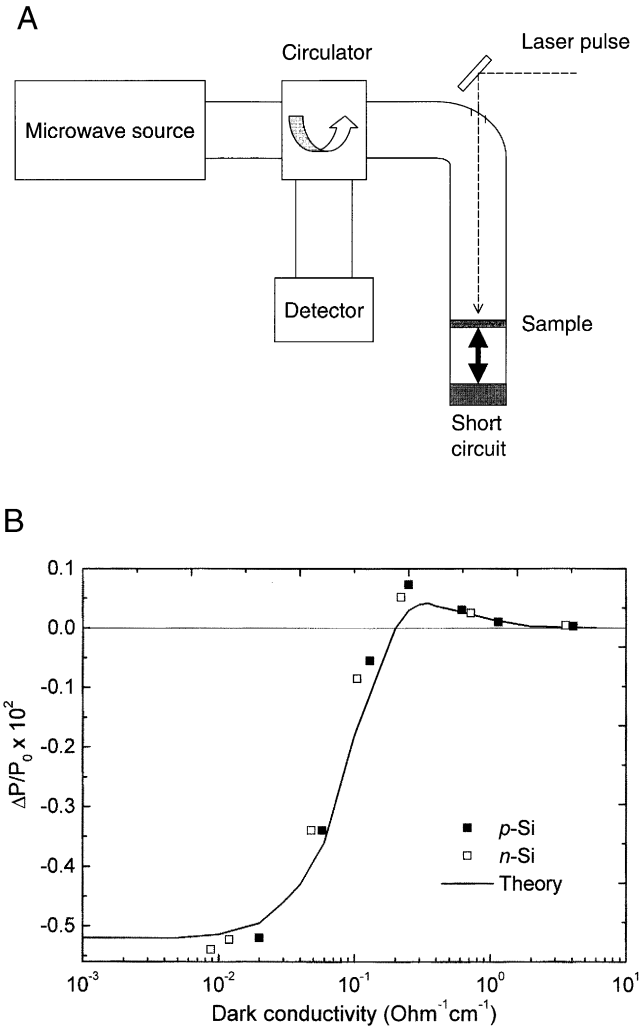


Figure 18. (**A**) Schematic representation of the microwave setup to perform quantitative lifetime measurements by microwave reflectivity. (**B**) Microwave conductivity signal amplitude at 35.5 GHz induced by exciting p-Si and n-Si wafers of 1 mm thickness with 1064-nm YAG light plotted versus the dark conductivity of the samples, as measured by Kunst and Beck (1988).

macroscopically the interaction between the sample and the microwaves. The change of the reflected microwave power from the sample is measured as a function of time when the sample is subjected to pulsed illumination. The time resolution is determined by the sharpness of the pulse, which could easily be in the nanosecond range. An oscilloscope (not shown in Fig. 18A) monitors the detector signal.

A usual approximation to the time-resolved microwave signal change $\Delta P(t)/P_0$ is given by

$$\Delta P(t)/P_0 = A \cdot \Delta\sigma(t) \tag{39}$$

A is called the **sensitivity factor** and is given by

$$A = \left(\frac{1}{R'(\sigma_0)}\right)\left(\frac{\partial R'(\sigma_0)}{\partial \sigma_0}\right) \tag{40}$$

Hence, A can be determined by measuring R' for different values of dark conductivity, σ_0, in a fixed cell geometry. It can also be calculated from electromagnetic theory (Kunst and Beck, 1988; Otaredian, 1993; Swiatkowski et al., 1995). For a first-order perturbation, two effects of an increase in the conductivity can be distinguished. First, an increase in the conductivity leads to an increase of the microwave absorption within the sample (if the microwave field in the sample remains constant) and, second, an increase in the conductivity leads to a decrease in the absorption due to the dark conductivity σ_0. The first effect yields a decrease in R' and the second leads to an increase in R' with the conductivity. However, it is obvious that the second effect is only important if the dark conductivity is large. In a general case, A can be positive or negative depending on the dark conductivity. In the low dark conductivity range, where $R'(\sigma_0)$ decreases linearly with increasing conductivity, higher-order terms can be neglected, and the relation

$$R' = 1 - A\sigma_0 \tag{41}$$

has been derived (Kunst and Beck, 1988).

Figure 18B demonstrates the amplitude of the microwave signal $\Delta P/P_0$ as a function of dark conductivity range for 1-mm-thick Si wafers measured by Kunst and Beck (1988), as determined at a microwave frequency of 35.5 GHz. All data are normalized to a constant excitation density below 10^{14} cm^{-3}. The solid curve represents a theoretical calculation. At higher dark conductivity, the absolute value of the sensitivity factor decreases until it becomes zero at $\sim 10^{-1}$ $(\Omega\text{cm})^{-1}$. Thereafter the sensitivity factor is positive, and, after a short increase, diminishes with increasing conductivity. So, for some combinations of conductivity and sample thickness, a frequency exists at which the sensitivity equals zero. A comprehensive calculation of reflection coefficient and conductivity sensitivity functions [10^{-5} to 10^{1} $(\Omega\ \text{cm})^{-1}$ range] for frequencies $f =$ 0.1 to 100 GHz has been provided by Otaredian (1993).

The sensitivity factor is proportional to the square of the microwave electric-field strength, $A \approx E^2(z)$, where z is a direction along the waveguide. The dependence of $A(z)$ on z is determined exclusively by the nonuniformity of the probing microwave field in the z direction. As $E(z)$ has a sinusoidal behavior, the sample should be placed at a maximum of the standing microwave. Thus, it is recommended that the thickness of the sample should be much less than a quarter-wavelength ($d < 10$ mm for X-band and $d < 2$ mm for Ka-band) even for low dark conductance samples.

Reference to sensitivity resolution leads one to consider the question of the skin depth, which determines the depth of material actually monitored by the microwave probe. At 10 GHz, skin depth is $\sim$500 μm for a conductivity of $1\,(\Omega\text{cm})^{-1}$ and varies approximately as ($\sigma_0 \times$ frequency)$^{-1/2}$. As quantitatively shown by Swiatkowski et al. (1995), the decay process has to be interpreted with care for the intermediate dark conductance range and in cases of higher injection because of the inhomogeneous excess photoconductivity in the z direction. Also, diffusion processes during the transient must be considered. In

general, for higher-conductance samples, data acquisition at lower microwave frequency is preferred.

In order to overcome this problem, the use of a constant-bias light in combination with a small excitation laser pulse has been successfully applied by some researchers for microwave reflectivity decay measurements (Stephens et al., 1994). Consequently, the modulation of the excess carrier concentration due to the laser pulse is small compared to the background injection level from the bias light, also eliminating differences in skin depths (see also Otaredian, 1993). The measured effective lifetime is shown to be a differential quantity, which may differ significantly from the actual effective lifetime (Schuurmans et al., 1997).

A number of researchers extended the microwave method to obtain minority carrier lifetime maps over whole semiconductor slices by mounting the sample on a positioning table. Spatial resolution may be obtained by focusing the light beam to a small spot or by monitoring the local resistivity with a small microwave probe. A lateral resolution of <100 μm is possible. It was shown that the signal could be maximized in terms of $\Delta P/P_0$ if the sample is excited in the center of a microwave cavity resonator, which typically has a higher-quality factor than a simple waveguide system. However, the rather cumbersome handling (instability) and the large time constants of microwave cavity resonators with high-quality factors have to be taken into account.

Radio Frequency PC Decay Method

In the radio frequency (rf) method, rf bridges, operating in the 1 to 900 MHz range, have been implemented for monitoring transient PC of semiconductors. The basic idea of the technique is that a semiconductor sample can be coupled to a rf bridge without the need for actual physical wire contact. Either inductive coupling by a 2 to 5 turn radio coil (Yablonovitch and Gmitter, 1992) or capacitive coupling by two flat electrodes (Tiedje et al., 1983) can be utilized (Fig. 19A and B). In effect, the coil or the capacitor is the primary of the transformer circuit and the semiconductor wafer is a secondary. The rf apparatus monitors the absolute sheet conductivity of the semiconductor, both in static and pulsed modes. The effective resistance in the secondary circuit is boosted by the coil turn ratio squared, when monitored in the primary circuit. Tuning capacitors are included in the circuit. One starts the measurement by adjusting the bridge with the sample mounted but without the light source. This procedure effectively nulls out the sample's dark conductance. The circuits are designed to digitize the reflected rf signal due to the impedance imbalance caused by the sheet conductivity changes in a semiconductor when the light is on. Small conductivity changes in the semiconductor will only perturb the tuning, and the reflected rf wave from the primary circuit will be linearly proportional to sheet conductivity. The output bridge signal is then mixed with an adjustable phase-shifter signal of the same frequency to give sum and difference output components. The difference component is a signal that is detected and recorded using an oscilloscope. Because there is no direct contact with the sample,

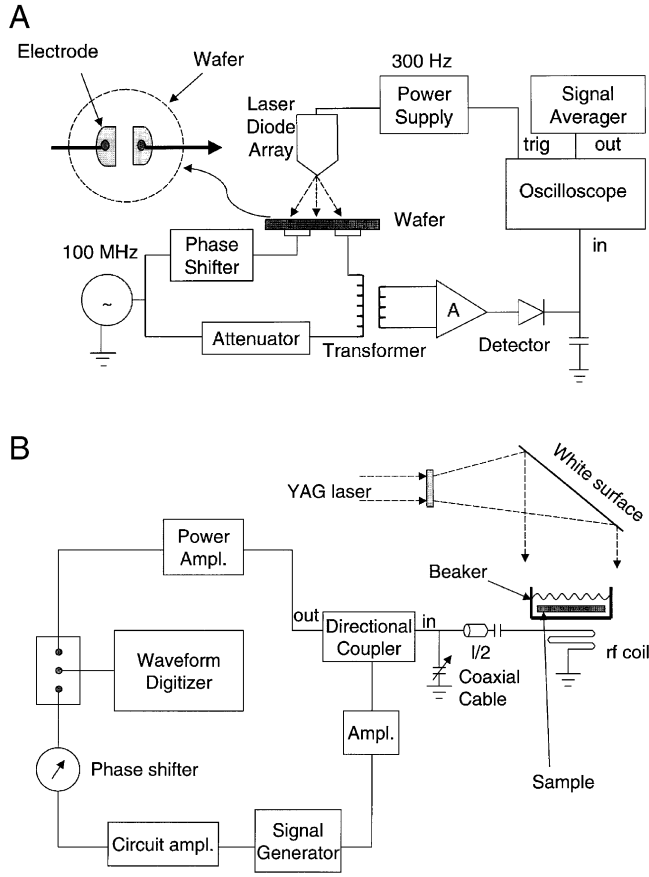


Figure 19. Schematic of the rf bridges for carrier lifetime measurements. (**A**) Setup with capacitator coupling at 100 MHz as used by Tiedje et al. (1983). (**B**) Setup with inductive coupling at 500 MHz as utilized by Yablonovitch and Gmitter (1992).

electronic measurements can be performed both in aqueous media (Fig. 19B) as well as in a vacuum.

The measured quantity is proportional to $\Delta\sigma$ where the constant of proportionality can be found by calibrating the detection circuit with semiconductor wafers of known conductivity (Tiedje et al., 1983). In general, a higher frequency increases the sample signal and also gives the circuit a better time response to monitor changes of the sample's conductance. The setup of Figure 19B, operating at 500 MHz frequency, extends measurements in the nanosecond time range to detect lifetimes on III-V semiconductors. While the skin depth is lower at lower frequencies, at 100 MHz it allows the electromagnetic radiation to penetrate even highly doped samples of 10 $(\Omega\text{cm})^{-1}$ to a thickness of a few hundred microns. Lateral resolution of the lifetime by the rf method is, however, quite limited due to the large area of coupling, and is of the order of several mm^2.

Problems

All PC decay measurements suffer from the following limitations.

1. It is very difficult to obtain separate quantitative information on the electron and the hole PC alone,

since both are equally present in the total PC. Partial conductivities may be obtained only in some specific cases. Such cases appear in damaged (amorphous) semiconductors at low injection levels (see Trapping).

2. Carrier mobilities are always weighting factors for the PC decay. In addition, the sensitivity factor A in all microwave reflectivity techniques, or the proportionality factor in rf techniques, must be properly calibrated for each particular thickness and dark conductance of the sample. Linearity of these constants should be checked and maintained for PC transients.

Except for the microwave technique, where local resolution is possible, only integrated lifetime over a relatively large area of the excited material can be represented by the PC decay.

While PC can be utilized in either a near-surface or bulk excitation mode, the main problems (as in any volume-integrated measurement) are the precise knowledge of the injection level and separation between surface and bulk contributions.

PHOTOLUMINESCENCE

Measurement of minority carrier lifetime in direct band gap semiconductors is usually performed by the PL technique, since emission yield is rather high. This is a nondestructive and noninvasive measurement that is very time-efficient. For process control and evaluation, one can measure lifetime in a simple structure and evaluate its performance. The individual components of multiple epitaxial layers can often be separately probed where only small quantities of material are needed. The lifetime, the diffusion length, and the quantum efficiency could be inferred. The main advantage is based on the fact that PL is an optical spectroscopic technique, i.e., it gives energetically resolved information. In particular, it is sensitive to the chemical species of impurities, which can be detected even at very low densities (see PHOTOLUMINESCENCE SPECTROSCOPY). This motivates the wide diffusion of this technique in the field of materials science and solid-state physics.

Principles of the PL Method

The PL decay method measures the radiation emitted by a material after optical excitation; in other words, it monitors various radiative recombination paths of photoexcited electron-hole pairs. The transient decay of the PL signal is called the PL lifetime. When combined with a model of the transient response, which is derived from the continuity equation, the minority-carrier lifetime and other recombination lifetimes are calculated. The principle of the experiment is very simple: excess carriers are injected by an appropriate light source in their fundamental absorption band, and the corresponding emission range is selected by a suitable monochromator and detected by a photoreceiver. The photoluminescence spectrum can usually be divided into three main energy regions whose width depends on the particular material. The numbers quoted in the following are characteristic of direct gap semiconductor crystals, e.g., GaAs of $E_g = 1.43$ eV.

Near-Band-Gap Emission ($E_g - 2E_{ex} < \hbar\omega < E_g + 2kT$, Where E_{ex} Is The Exciton Binding Energy). In this region, interband transitions, free and bound exciton recombination, and bound electron-free hole transitions are observed. At room temperature and above, exciton and shallow donor-related luminescence is thermally quenched. Band-to-band recombination dominates the spectra, and the PL edge position can even be used to determine the alloy composition of compound semiconductors such as $Al_xGa_{1-x}As$ (Pavesi and Guzzi, 1994).

Shallow Impurity Emission ($E - 200$ meV $< \hbar\omega < E_g - 2E_{ex}$). In this region, acceptor-related recombination is present. The information that one obtains from luminescence in this region also reflects the chemical species of the impurities involved in the transition, their relative densities, compensation ratios, etc.

Deep Level Luminescence ($\hbar\omega < E_g - 200$ meV). In this region, the recombination of deep impurities and defects is observed. The study of low-energy luminescence also gives information on important properties of the crystal such as stoichiometry, nonradiative recombination channels, defect densities, etc.

The radiative band-to-band recombination rate R of free electrons and holes (Fig. 1) is given by

$$R = B \times pn(\mathrm{cm^{-3}s^{-1}}) \tag{42}$$

Here p and n denote total concentrations of holes and electrons. The B-coefficient is specific to a particular semiconductor and is proportional to the dipole matrix elements between the conduction and valence band wave functions. This coefficient can be established by calculation from the integrated absorption spectrum using the detailed balance relationship (Van Roosbroeck and Shockley, 1954). Therefore, experimentally determined B coefficients are much larger for direct than for indirect band gap materials and have quite different temperature dependencies. For example, B is 2×10^{-10} cm^3/s for GaAs and $\sim 10^{-14}$ cm^3/s for Si at 300 K (see Table 1).

The PL decay is bimolecular into high injection conditions. The simplest recombination process, however, consists of direct band-to-band radiative recombination under low injection conditions. From Equation 12, the low-injection radiative lifetime is

$$\tau_{RAD} = 1/(BN) \tag{43}$$

where N is the majority-carrier density. In this case, the total PL output is simply given by summing the excess minority carriers over the active volume and multiplying by the radiative probability. When self-absorption and reflection of the internally generated photon are taken into account, the PL intensity is modified by

$$I_{PL}(t) = \frac{1}{\tau_{RAD}} \int_V A'(\mathbf{r}) \Delta n(\mathbf{r}, t)\, dV \tag{44}$$

Here Δn is excess minority carrier concentration in a p-type semiconductor (Δp in n-type) and $A'(\mathbf{r})$ is an optical interaction function accounting for self-absorption and interfacial reflection of internally generated photons, the extraction factor which accounts for the light-collection geometry, etc. For now, we will take $A'(\mathbf{r}) = 1$. We see from Equation 44 that $I_{PL}(t)$ tracks $\Delta n(t)$ in time. However, the photoluminescence decays with the minority-carrier lifetime only when there is no charge flow out of the active volume. A more accurate measurement of minority-carrier lifetime is therefore obtained from confinement structures which prevents minority-carrier diffusion loss, such as from $n^+/n/n^+$ AlGaAs type heterostructures (Ahrenkiel, 1992).

Practical Aspects of the Method

It is important to recognize that the external PL decay $I_{PL}(t)$ measures the overall minority carrier lifetime, not the radiative lifetime, because, in general, the recombination rates add as given by Equation 16. All nonradiative recombination paths (e.g., via deep recombination centers) compete with the radiative process in the bulk of the semiconductor and also with surface (or interface) recombination (see Theory) in the context of the decay concept. It is convenient to note that the internal efficiency η_i, defined as the ratio of radiative to total recombination rate, can be expressed in terms of Equation 43

$$\eta_i = \frac{\tau_{\text{eff}}}{\tau_{\text{RAD}}} = \tau_{\text{eff}} BN \tag{45}$$

For a steep absorption edge and high emission efficiency, the reabsorption of the photons generated by radiative recombination, termed **photon recycling**, leads to an increase in both the lifetime and the diffusion length of minority carriers. According to the calculations of Asbeck (Asbeck, 1977), this behavior produces an apparent increase in radiative lifetime which, for an active layer thickness d, can be written in terms of a photon recycling factor $\phi(d)$ as

$$\tau_{\text{eff}}^{-1} = \tau_{\text{nonRAD}}^{-1} + [\phi(d)\tau_{\text{RAD}}]^{-1} \tag{46}$$

where τ_{nonRAD} is the bulk minority carrier nonradiative lifetime. The parameter $\phi(d)$ is a complicated function of the internal radiative efficiency, the critical angle for total internal reflection, and the absorption coefficient, and can only be calculated numerically. The Asbeck calculations of $\phi(d)$ for GaAs heterostructures with different double $Al_xGa_{1-x}As$ window layer compositions are shown in Figure 20. One sees that the curves are not linear below 10 μm thickness where the transition to near-linear response appears. The recycling factor is ~10 for 1.2×10^{18} cm^{-3} doped GaAs active layer with a 10 μm thickness.

Method Automation

The experimental setup complexity is proportional to the needs; for characterization, a very simple and cheap apparatus is sufficient. The earlier references describe a number of techniques (Orton and Blood, 1990; Schroder, 1990).

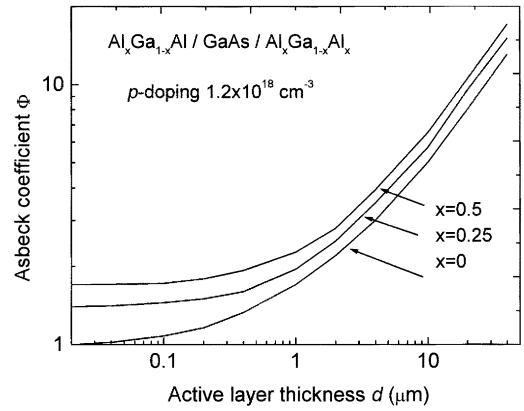


Figure 20. Calculations of the Asbeck coefficient for photon recycling in a double heterostructure $Al_xGa_{1-x}As/GaAs/Al_xGa_{1-x}As$.

Most workers have used short pulses from a laser or light-emitting diode (LED) to generate excess carriers. Alternative methods of excitation are also possible, the most popular being the injection from a metal contact or a pn-junction. The so-called cathodoluminescence decay method differs only in the use of a pulsed electron beam or x ray as the excitation source.

A more advanced technique for measuring photoluminescence decay with a high sensitivity and a fast time resolution is called time-correlated single-photon counting (Ahrenkiel, 1992). In order to use this technique, the photodetector must be able to detect single photons produced by the emitted light. The typical commercial photodetectors with sufficient sensitivity for photon counting are cooled photomultiplier tubes or microchannel plates (MCP). The transient time is greatly reduced in an MCP down into the picosecond range. The photodetector is followed by a high-speed amplifier, which is connected to the photon-counting apparatus. A schematic of the experimental setup is shown in Figure 21A. The photons emitted by the sample are focused on the input slit of a scanning monochromator tuned to the appropriate wavelength. The time-correlated photon counting electronics are identical in the remainder of the detection system. Using a beam splitter and a photodiode, a small fraction of the excitation pulse is deflected to a fast photodiode. The electric output of the photodiode triggers the time-to-amplitude converter (TAC). The pulse-height discriminator is necessary to block electrical pulses produced by thermal and other nonphotonic sources. The first collected PL photon initiates an electrical pulse in the photodetector, which is passed by the amplitude discriminator. The electrical pulse produces a stop message at the TAC. The TAC output is a ramp generator with amplitude proportional to the time delay between the laser pulse and the arrival of the first photon. The TAC signal is fed to a multichannel pulse-height analyzer (MCA), which collects one count per detected photon. A count is stored in a channel appropriate to the time delay and a maximum of one count is recorded for every laser pulse. In this way, a histogram of the PL decay is built in terms of counts versus time.

Many different techniques have been used for obtaining time resolution in luminescence spectroscopy (Fleming,

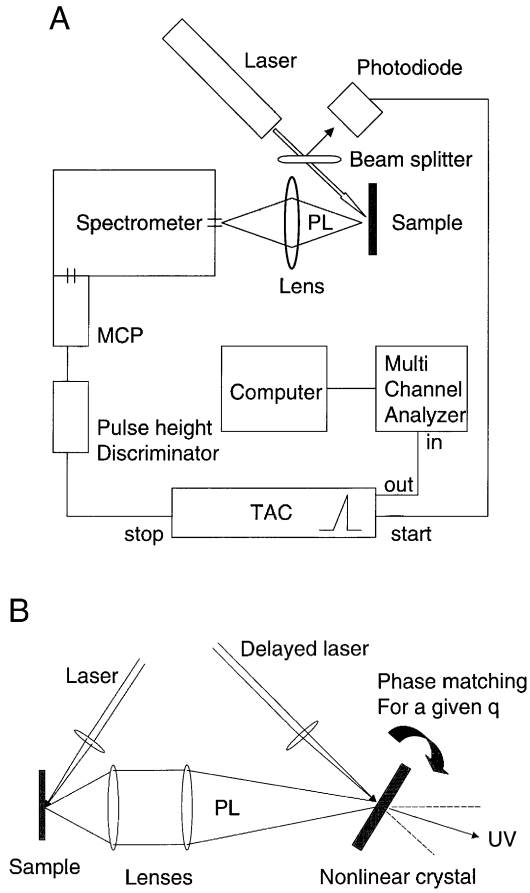


Figure 21. (**A**) Schematic of the PL decay measurement apparatus with time-correlated single-photon counting. (**B**) Schematic arrangement for the collection of luminescence for up-conversion. Sum frequency radiation is generated in a nonlinear crystal.

1986). While some photodetectors can be made with much faster response times than a typical photomultiplier of ~50 ps, their low sensitivity restricts their use in detecting weak signals. Another technique, preferred for obtaining time resolution in the 1 to 100-ps range, has been to use a streak camera. Some single-shot streak cameras offer a time resolution of better than 1 ps in certain cases. However, a drawback is the spectral response of the photocathode, which typically does not extend beyond ~900 nm.

The best hope for achieving luminescence time resolution comparable to today's available lasers of fs pulse width appears to be techniques that use nonlinearity induced by the laser pulse as a gate for recording the luminescence. In this technique, the luminescence excited by an ultrafast laser is mixed with the laser in a nonlinear crystal ($LiIO_3$, β-BaB_2O_4, CS_2) to generate sum or difference frequency radiation (see Fig. 21B). Since the mixing process takes place only during the presence of a delayed laser pulse, this provides time resolution comparable to the laser pulse width, as given by the experimental implementation of Shah (1988). Sum frequency generation, also known as wavelength up-conversion, is widely used because the availability of excellent photomultipliers in the visible and UV spectral regions makes this technique extremely attractive for time-resolved luminescence spectroscopy of infrared luminescence.

Data Analysis and Interpretation

The net effect of PL lifetime has to be analyzed in terms of SRH and surface recombination in the same way as for other methods. A model case of an n-type semiconductor with excess injected carriers ($\Delta n = \Delta p$) exceeding the doping density, N_D, will be analyzed here as an example. In contrast to previous analyses (see description in the Theory section) we assume here that SRH recombination and radiative recombination are the only two main recombination channels that have to be considered in direct-bandgap semiconductors. From Equations 7 and 10, it follows that the instantaneous lifetime becomes (Ahrenkiel, 1992)

$$\frac{\partial I}{\partial t} = \frac{I + I^2}{\tau_p[1 + I] + \tau_n I} \tag{47}$$

where the instantaneous injection level is $I = \Delta p(t)/N_D$ and the SRH electron and hole lifetimes τ_n, τ_p are described by Equation 9. The decay rate is nonexponential over an appreciable range of excess hole densities. The solutions of Equation 47 are

$$\begin{aligned} I(t) &= \exp(-t/\tau_p) \quad (I \ll 1) \\ I(t) &= \exp[-t/(\tau_p + \tau_n)] \quad (I \gg 1) \end{aligned} \tag{48}$$

One can fit the decay with these asymptotic rates and then define the initial injection ratio used in the experiment. The characteristic downturn of the decay curve has been seen in numerous devices and is an indicator of SRH recombination. When surface or interface recombination is the dominant mechanism (in thin devices), the photoluminescence decay may be fit with a function of Equations 47 and 48, replacing electron and hole lifetimes with the corresponding effective lifetimes: $\tau_{sp} = d/(2s_p)$ and $\tau_{sn} = d/(2s_n)$. The quantities s_p and s_n are interface recombination velocities for holes and electrons, respectively (refer to Equations 19 and 20). By varying the active layer thickness d, one can determine if the dominant recombination mechanism is bulk or surface related.

If a substantial concentration of excitons is present in a semiconductor at low temperatures, the PL decay becomes quite complex. The exciton level population should be included in the analysis as governed by the rates of formation and dissociation of excitons. Some description can be found in Orton and Blood (1990) and in Itoh et al. (1996). Usually, a strong coupling condition between the free exciton level and the conduction band can be used as a starting point. In a general case, the PL decay of free carriers can be approximated by two exponentials, and it becomes different from the carrier lifetime associated with PC (Ridley, 1990).

In disordered materials (such as amorphous, polycrystalline, or porous semiconductors) the PL decay and lifetime is usually represented either by a stretched-exponential or a power-law line shape. Physical interpretation of effective lifetimes or a characteristic distribution

Table 5. Application of Selected Transient Lifetime Techniques for Characterization of Different Semiconductor Materials

Lifetime Technique	Need for Contacts	Measured Quantity	Application for semiconductors		
			Indirect Band Gap	Direct Band Gap	Amorphous
PC direct	Yes	Photoconductance	++	+	+++
PC-microwave	No	Reflected microwave power	+++	+	+
PC-rf bridge	No	Eddy current (electromotive force)	++	++	+
FCA	No	Photoinduced absorption	+++	+	+
PL	No	Photoemission yield	+	+++	++

[a] All the considered techniques share the common feature of using a pulsed laser beam for excitation of excess electron hole pairs. Symbols: +, Applicable, but scarcely used; ++, relevant, commonly used; +++, beneficial, frequently used.

of lifetime parameters, however, very much depends on the particular assumptions for the recombination and transport models (Stachowitz et al., 1995; Roman and Pavesi, 1996).

Problems

From the discussion above, PL also has some disadvantages that can be summarized as follows:

1. It is very difficult to obtain quantitative information on the lifetime-doping-density relation or the lifetime-trap-concentration relation by PL alone. Some efforts, however, have been made in this direction, but the proposed methods require very accurate calibrations, checking very carefully the line shape of the spectrum (Schlangenotto et al., 1974; Pavesi and Guzzi, 1994).
2. Photoluminescence provides information on radiative transitions. Only indirectly can one study the nonradiative recombination processes, essentially by analyzing the quantum efficiency.
3. Photoluminescence is not a bulk characterization technique, as briefly explained above. Only a thin, near-surface region can be investigated, defined roughly by the larger of the light penetration depth and minority carrier diffusion length.

SUMMARY AND SELECTION GUIDE

The three lifetime characterization methods described in more detail in this chapter—free carrier absorption (FCA), photoconductivity (PC), and photoluminescence (PL)—have in common optical generation of excess carriers by a pulsed light source. Their common advantages may be summarized as:

1. They offer the possibility of controlling the injection level, Δn, by varying the laser pulse energy (enabling the minority or the high-injection lifetime to be studied).
2. They offer the possibility of localized measurements by focusing the excitation beam.
3. They are versatile techniques that may be used for several different semiconductors.
4. They are contactless techniques (except for the direct PC technique) allowing simple (or no) sample preparation.

Table 6. Comparison between Selected Lifetime Techniques for Limiting Parameters of Measurement Conditions

Lifetime Technique	Sensitivity Range (Δn) (cm^{-3})[a]		Dark Conductivity Range (σ_0)[b] $(\Omega_0 cm)^{-1}$	Measurement Bandwidth[c]	Commercial Apparatus
	Minimal	Maximal			
PC direct	10^{11}	5×10^{19}	<100	<40 THz	Yes
PC-microwave	10^{9}	5×10^{16}	<5	<10 GHz	Yes
PC-RF bridge	10^{13}	3×10^{17}	<10	<1 GHz	No information
FCA	10^{11}	10^{19}	≤200	<100 GHz	No information
PL	10^{9}[d]	10^{19}	≤1000	<500 GHz	Yes

[a] Specified for material of low dark conductivity, $\sigma_0 = 0.01\ (\Omega cm)^{-1}$.
[b] Specified for sample thickness of 500 μm.
[c] The exact value of the bandwidth is also a function of the time duration of the utilized laser pulses. Also, detection electronics may reduce these values considerably.
[d] For Group III-V semiconductors, which are characterized by high radiation yield.

Table 7. Comparison between Selected Lifetime Techniques for Sensitivity, Speed, hl/ll Lifetime, Depth Resolution, and Lateral Resolution

Lifetime Technique	Typical Decay Interval (Decades)[a]	hl/ll Lifetime[b]	Depth Resolution (μm)	Mapping Resolution (μm)	Mapping Speed
PC direct	2	++	—[c]	—	—
PC-microwave	2–3	+	—[c]	<100	Medium
PC-RF bridge	2–3	+	—[c]	—	—
FCA (∥)	3	+ +	—[c]	<10	Low
FCA (⊥)	3–5	+ + +	5	—	—
PL	3–5	+	—[c]	<1	High

[a]The exact sensitivity can be a function of the characteristics of the semiconductor material under test.
[b]Symbols: +, possible, however, causes methodical problems; ++, medium, some methodical problems; + + +, easy, small methodical problems.
[c]Indirectly through the measurements with variable absorption depth.

Their disadvantages may be summarized as:

1. They may yield erroneous lifetime values when carrier trapping is dominant (in this case a modulated or quasi-steady-state variant of the technique may successfully be applied).
2. Carrier diffusion and surface recombination may dominate recombination and, thus, the sought-after material parameter such as minority carrier lifetime or high injection lifetime may not be evaluated (as true for almost any lifetime measurement technique).
3. They require relatively complex experimental setup.

Tables 5–7 compare the different methods reviewed in this unit, and may be used as a selection guide for different semiconductor applications. Thus, Table 5 evaluates the advantages of each method for different types of semiconductors, while Table 6 summarizes the sensitivity and ultimate frequency range of the different methods. Notably, the frequency range indicated represents the ultimate limit of the measurement principle, while standard laboratory equipment such as photodetectors or oscilloscopes often yield a frequency limit much below 1 GHz. An exception is the PL up-conversion technique (see Fig. 21B) where pico- or femtosecond lasers may be used for characterizing ultrashort lifetimes. Finally, Table 7 summarizes some aspects of the techniques which are important for detailed recombination studies (e.g., several decades of the decay must be measured to safely extract the hl/ll-lifetime and the the radiative and Auger recombination rates). Also indicated is the applicability of the techniques for lifetime mapping and for depth-resolved studies.

As a final comment, one may note that even though the principle of each method is relatively simple, some experience must be gained to evaluate aspects of, e.g., surface recombination or carrier diffusion, before the derived lifetimes can be trusted as true characteristics of the semiconductor material bulk properties or as measures of the material's purity. It is, indeed, recommended that alternative methods be used to analyze the same sample. If other methods are not at one's disposal, one may gain further insight by varying the experimental conditions widely, e.g., by using different injection levels, different excitation and probe areas, etc. This may hopefully unravel different recombination mechanisms and separate carrier diffusion effects from true recombination phenomena.

LITERATURE CITED

Ahrenkiel, R. K. 1992. Measurement of minority-carrier lifetime by time-resolved photoluminescence. *Solid State Electron.* 35:239–250.

ASTM. 1987. Annual Book of ASTM Standards. American Society for Testing Materials, Philadelphia.

Asbeck, P. 1977. Self-absorption effects on the radiative lifetime in GaAs-GaAlAs double heterostructures. *J. Appl. Phys.* 48:820–822.

Auston, D., Shank, C., and LeFur, P. 1975. Picosecond optical measurements of band-to-band Auger recombination of high-density plasmas in germanium. *Phys. Rev. Lett.* 35:1022–1025.

Baliga, B. J. 1987. Modern Power Devices. John Wiley & Sons, New York.

Bergman, J. P., Kordina, O., and Janzén, E. 1997. Time resolved spectroscopy of defects in SiC. *Phys. Status Solidi A* 162:65–77.

Blais, P. D. and Seiler, C. F. 1980. Measurements and retention of recombination lifetime. *In* Lifetime Factors in Silicon, ASTM STP 712, pp. 148–158. American Society for Testing and Materials, Philadelphia.

Blakemore, J. S. 1962. Semiconductor Statistics. Pergamon Press, Elmsford, N.Y.

Bleichner, H., Nordlander, E., Fiedler, G., and Tove, P. A. 1986. Flying-spot scanning for the separate mapping of resistivity and minority-carrier lifetime in silicon. *Solid State Electron.* 29:779–786.

Brandt, O., Yang, H., and Ploog, H. 1996. Images of recombination centers on the spontaneous emission of semiconductors under steady-state and transient conditions. *Phys. Rev. B* 54:R5215–R5218.

Brüggemann, R. 1997. Hole response time and the experimental test of the Einstein relation. *Phys. Rev. B*. 56:6408–6411.

Cutolo, A. 1998. Selected contactless optoelectronic measurements for electronic applications. *Rev. Sci. Instum.* 69:1–24.

Dziewior, J. and Schmid, W. 1977. Auger coefficients for highly doped and highly excited silicon. *Appl. Phys. Lett.* 31:346–348.

Fleming, G. R. 1986. Chemical Applications of Ultrafast Spectroscopy. Oxford University Press, New York.

Galeckas, A., Grivickas, V., Linnros, J., Bleichner, H., and Hallin, C. 1997b. Free carrier absorption and lifetime mapping in 4H SiC epilayers. *J. Appl. Phys.* 81:3522–3525.

Galeckas, A., Linnros, J., Grivickas, V., Lindefelt, U., and Hallin, C. 1997a. Auger recombination in 4H-SiC: Unusual temperature behavior. *Appl. Phys. Lett.* 71:3269–3271.

Glunz, S. W. and Warta, W. 1995. High resolution lifetime mapping using modulated free carrier absorption. *J. Appl. Phys.* 77:3243–3247.

Graff, K. and Fisher, H. 1979. Carrier lifetime in silicon and its impact on solar cell characteristics. *Top. Appl. Phys.* 31:173–211.

Graff, K. and Pieper, H. 1980. Carrier lifetime measurements for process monitoring during device production. *In* Lifetime Factors in Silicon, ASTM STP 712, pp. 136–147. American Society for Testing and Materials, Philadelphia.

Grivickas, V., Linnros, J., Galeckas, A., and Bikbajevas V. 1996. Relevance of the exciton effect on ambipolar transport and Auger recombination in silicon at room temperature. *In* Proceedings of the 23rd International Conference on the Physics of Semiconductors, Vol. 1, pp. 91–94. World Scientific, Singapore.

Grivickas, V., Linnros, J., Vigelis, A., Seckus, J., and Tellefsen, J. A. 1992. A study of carrier lifetime in silicon by laser-induced absorption: A perpendicular geometry measurement. *Solid State Electron.* 35:299–310.

Grivickas, V., Noreika, D., and Tellefsen, J. A., 1989. Surface and Auger recombinations in silicon wafers of high carrier density [in Russian]. *Lithuanian Phys. J.* 29:48–53; ASTM STP 712 (Sov. Phys. Collect.) 29:591–597.

Grivickas, V. and Willander, M. 1999. Bulk lifetimes of photogenerated carriers in intrinsic c-Si. *In* EMIS Data Review no. 20, Properties of Crystalline Silicon, pp. 708–711. INSPEC, London.

Grivickas, V., Willander, M., and Vaitkus, J. 1984. The role of intercarrier scattering in excited silicon. *Solid State Electron.* 27:565–572.

Häcker, R. and Hangleiter, A. 1994. Intrinsic upper limits of the carrier lifetime in silicon. *J. Appl. Phys.* 75:7570–7572.

Hall, R. N. 1952. Electron-hole recombination in germanium. *Phys. Rev.* 87:387–387.

Horanyi, T. S., Pavelka, T., and Tuttö, P. 1993. In situ bulk lifetime measurement on silicon with a chemically passivated surface. *Appl. Surf. Sci.* 63:306–311.

Itoh, A., Kimoto, T., and Matsunami, H. 1996. Exciton-related photoluminescence in 4H-SiC grown by step-controlled epitaxy, *Jpn. J. Appl. Phys.* 35:4373–4378.

Jacobsen, R. H., Birkelund, K., Holst, T., Jepsen, U. P., and Keiding, S. R., 1996. Interpretation of photocurrent correlation measurements used for ultrafast photoconductive swich characterization. *J. Appl. Phys.* 79:2649–2657.

Katayama, K., Kirino, Y., Iba, K., and Shimura, F. 1991. Effect of ultraviolet light irradiation on noncontact laser microwave lifetime measurement. *Jpn. J. Appl. Phys.* 30:L1907–L1910.

Kunst, M. and Beck, G. 1988. The study of charge carrier kinetics in semiconductors by microwave conductivity measurements. II. J. *Appl. Phys.* 63:1093–1098.

Landsberg, P. T. 1987. Trap-Auger recombination in silicon of low carrier densities. *Appl. Phys. Lett.* 50:745–747.

Lee, C. H. (ed.) 1984. Picosecond Optoelectronic Devices. Academic Press, San Diego.

Ling, Z. G., Ajmera, P. K., Anselment, M., and DiMauro, L. F. 1987. Lifetime measurements in semiconductors by infrared absorption due to pulsed optical excitation. *Appl. Phys. Lett.* 51:1445–1447.

Linnros, J. 1998a. Carrier lifetime measurements using free carrier absorption transients: I. Principle and injection dependence. *J. Appl. Phys.* 84:275–283.

Linnros, J. 1998b. Carrier lifetime measurements using free carrier absorption transients: II. Lifetime mapping and effects of surface recombination. *J. Appl. Phys.* 84:284–291.

Linnros, J. and Grivickas, V. 1994. Carrier diffusion measurements in silicon with a Fourier-transient-grating method. *Phys. Rev. B* 50:16943–16955.

Linnros, J., Norlin, P., and Hallén, A. 1993. A new technique for depth resolved carrier recombination measurements applied to proton irradiated thyristors. *IEEE Trans. Electron Devices* 40:2065–2073.

Linnros, J., Revsäter, R., Heijkenskjöld, L., and Norlin, P. 1995. Correlation between forward voltage drop and local carrier lifetime for a large area segmented thyristor. *IEEE Trans. Electron Devices* 42:1174–1179.

Orton, J. W. and Blood, P. 1990. The Electric Characterization of Semiconductors: Measurement of Minority Carrier Properties. Academic Press, London.

Otaredian, T. 1993. Separate contactless measurement of bulk lifetime and surface recombination velocity by harmonic optical generation of the excess carriers. *Solid State Electron.* 36:163–172.

Othonos, A., 1998. Probing ultrafast carrier and phonon dynamics in semiconductors. *J. Appl. Phys.* 83:1789–1830.

Pang, S. K. and Rohatgi, A. 1991. Record high recombination lifetime in oxidized magnetic Czochralski silicon. *Appl. Phys. Lett.* 59:195–197.

Pasiskevicius, V., Deringas, A., and Krotkus, A., 1993. Photocurrent nonlinearities in ultrafast optoelectronic switches. *Appl. Phys. Lett.* 63:2237–2239.

Pavesi, L. and Guzzi, M. 1994. Photoluminescence of $Al_xGa_{1-x}As$ alloys. *J. Appl. Phys.* 75:4779–4842.

Pidgeon, C. R. 1980. Free carrier optical properties of semiconductors. *In* Handbook on Semiconductors, Vol. 2 (T.S. Moss, ed.) p. 223. North-Holland Publishing, Amsterdam, The Netherlands.

Polla, D. L. 1983. Determination of carrier lifetime in Si by optical modulation. *IEEE Electron Device Lett.* EDL-4:185–187.

Ridley, B. K. 1990. Kinetics of radiative recombination in quantum wells. *Phys. Rev. B* 41:12190–12196.

Rohatgi, A., Weber, E. R., and Kimerling, L. C. 1993. Opportunities in silicon photovoltaics and defect control in photovoltaic materials. *J. Electron. Mater.* 22:65–72.

Roman, H. E. and Pavesi, L. 1996. Monte-Carlo simulations of the recombination dynamics in porous silicon. *J. Phys. Condens. Matter* 8:5161–5176.

Ryvkin, S. M. 1964. Photoelectric Effects in Semiconductors. Consultants Bureau, New York.

Sanii, F., Giles, F. P., Schwartz, R. J., and Gray, J. L. 1992. Contactless nondestructive measurement of bulk and surface recombination using frequency-modulated free carrier absorption. *Solid State Electron.* 35:311–317.

Schlangenotto, H., Maeder, H., and Gerlach, W. 1974. Temperature dependence of the radiative recombination coefficient in silicon, *Phys. Status Solidi A* 21:357–367.

Schroder, D.K. 1990. Semiconductor Material and Device Characterization. John Wiley & Sons, New York.

Schroder, D. K. 1997. Carrier lifetimes in silicon. *IEEE Trans. Electron Devices* 44:160–170.

Schuurmans, F. M., Schonecker, A., Burgers, A. R., and Sinke W. C., 1997. Simplified evaluation method for light-biased effective lifetime measurements. *Appl. Phys. Lett*. 71:1795–1797.

Shah, J. 1988. Ultrafast luminescence spectroscopy using sum frequency generation. *IEEE J. Quantum Electron*. 24:276–288.

Shockley, W. and Reed, W. T. 1952. Statistics of the recombination of holes and electrons. *Phys. Rev.* 87:835–843.

Sinton, R. and Cuevas, A. 1996. Contactless determination of current-voltage characteristics and minority-carrier lifetimes in semiconductors from quasi-steady-state photoconductance data. *Appl. Phys. Lett*. 69:2510–2512.

Smith, P. R., Auston, D. H., Johnson, A. M., and Augustyniak, W. M. 1981. Picosecond photoconductivity in radiation-damaged silicon-on-sapphire films. *Appl. Phys. Lett*. 38:47–50.

Stachowitz, R., Schubert, M., and Fuhs, W. 1995. Nonradiative recombination and its influence on the lifetime distribution in amorphous silicon (*a*-Si:H). *Phys. Rev. B* 52:10906–10914.

Stephens, A. W., Aberle, A. G., and Green, M. A., 1994. Surface recombination velocity measurements in the silicon-silicon dioxide interface by microwave-detected photoconductivity decay. *J. Appl. Phys*. 76:363–370.

Strauss, U., Ruhle, W. W., and Köhler, K., 1993. Auger recombination in intrinsic GaAs. *Appl. Phys. Lett*. 62:55–57.

Swiatkowski, C., Sanders, A., Buhre, K.-D., and Kunst, M. 1995. Charge-carrier kinetics in semiconductors by microwave conductivity measurements. *J. Appl. Phys*. 78:1763–1775.

Sze, S. M. 1981. Physics of Semiconductor Devices, 2nd ed. John Wiley & Sons, New York.

Tiedje, T., Haberman, J. I., Francis, R. W., and Ghosh, A. K. 1983. An RF bridge technique for contactless measurement of the carrier lifetime in silicon wafers. *J. Appl. Phys*. 54:2499–2503.

Van Roosbroeck, W. and Shockly, W. 1954. Photon-radiative recombination of electrons and holes in germanium. *Phys. Rev*. 94:1558–1560.

Waldmeyer, J. 1988. A contactless method for determination of carrier lifetime, surface recombination velocity, and diffusion constant in semiconductors. *J. Appl. Phys*. 63:1977–1983.

Wight, D. R. 1990. Electron lifetime in *p* -type GaAs: Hole lifetime in n-type GaAs. *In* EMIS Data Review No. 2, Properties of Gallium Arsenide, pp. 95, 107. INSPEC, London.

Yablonovitch, E., Allara, D. L., Chang, C. C., Gmitter, T., and Bright, T. B. 1986. Unusually low surface recombination velocity on silicon and germanium surfaces. *Phys. Rev. Lett*. 57:249–252.

Yablonovitch, E. and Gmitter, T. J. 1992. A contactless minority lifetime probe of heterostructures, surfaces, interfaces and bulk wafers. *Solid State Electron*. 35:261–267.

KEY REFERENCES

Ahrenkiel, 1992. See above.

Measurement theory and PL methods are described in connection with recent developments in the Groups III-V material technology.

Graff and Fisher, 1979. See above.

Comprehensive study of carrier lifetime distributions in as-grown CZ- and FZ-silicon, before and after thermal treatments, by photoconductivity decay-type measurements.

Orton and Blood, 1990. See above.

Unified account of the physical principles of electrical characterization of semiconductors related to measurement of minority carrier properties.

Ryvkin, 1964. See above.

Describes the basis of generation, transport, and recombination processes of nonequilibrium carriers in semiconductors and methods of experimental investigation.

Schroder, 1990. See above.

Provides a comprehensive survey of electrical, optical, and physical characterization techniques of semiconductor materials and devices. See Chapter 8: Carrier lifetime.

Schroder, 1997. See above.

A review paper where various recombination mechanisms are discussed for silicon and the concept of recombination and generation lifetime is presented.

JAN LINNROS
Royal Institute of Technology
Kista-Stockholm, Sweden

VYTAUTAS GRIVICKAS
Royal Institute of Technology
Kista-Stockholm, Sweden and
University Vilnius, Lithuania

CAPACITANCE-VOLTAGE (C-V) CHARACTERIZATION OF SEMICONDUCTORS

INTRODUCTION

Hillibrand and Gold (1960) first described the use of capacitance-voltage (C-V) methods to determine the majority carrier concentration in semiconductors. C-V measurements are capable of yielding quantitative information about the diffusion potential and doping concentration in semiconductor materials. The technique employs *pn* junctions, metal-semiconductor (MS) junctions (Schottky barriers), electrolyte-semiconductor junctions, metal-insulator-semiconductor (MIS) capacitors, and MIS field effect transistors (MISFETSs). The discussions to follow will emphasize *pn* junctions and Schottky barrier techniques.

C-V measurements yield accurate information about doping concentrations of majority carriers as a function of distance (depth) from the junction. The major competing technique to C-V measurements is the Hall effect (HALL EFFECT IN SEMICONDUCTORS), which, while yielding added information about the carrier mobility, requires difficult, time-consuming procedures to determine carrier-depth profiles. In fact, C-V profiling and Hall measurements can be considered complementary techniques.

In concert with deep-level transient spectroscopy (DLTS; DEEP-LEVEL TRANSIENT SPECTROSCOPY), C-V measurements can quantitatively describe the free carrier concentrations together with information about traps. Defects

appearing as traps at energies deep within the forbidden gap of a semiconductor can add or remove free carriers. The same Schottky diode can be used for both C-V measurements and DLTS measurements. The C-V apparatus and the DLTS apparatus are often a "bundled" setup.

C-V profiling to determine the doping profile of the mobile majority carriers in a semiconductor is a powerful quantitative technique. Some precautions must be taken, especially in dealing with multiple layered structures such as high-to-low doping levels, quantum wells, and the presence of traps. However, most of the problems encountered have been addressed in the literature. Major considerations and precautions to be taken in dealing with C-V profiling data acquisition and reduction are also covered in the ASTM standard (ASTM, 1985). The cost of assembling a C-V profiling apparatus will range between six thousand dollars for a simple, single temperature, manual apparatus to sixty thousand dollars for a fully automated apparatus. C-V profiling used in concert with deep-level transient spectroscopy (DEEP-LEVEL TRANSIENT SPECTROSCOPY) to identify traps, as well as with Hall effect (HALL EFFECT IN SEMICONDUCTORS), to address transport mechanisms, provides a powerful set of tools at reasonable cost.

PRINCIPLES OF THE METHOD

The capacitance at a *pn* or MS junction depends upon the properties of the charge-depletion layer formed at the junction. The depletion region is in the vicinity of the *pn* junction and is "depleted" of free carriers (Fig. 1A) due to the drift field (Fig. 1B) required to maintain charge neutrality. These properties are well known and are discussed in detail in standard texts such as Grove (1967), Sze (1981), McKelvey (1966), and Pierret (1996). This latter reference includes examples of computer simulations and programs for duplication by the reader.

Following Sze (1981) we show an abrupt *pn* junction in Figure 1. In this figure, the band gap of the semiconductor $E_G = (E_C - E_V)$ is defined by the difference between the conduction band energy, E_C, and the valance band energy, E_V. The Fermi energy E_F defines the equilibrium condition for charge neutrality. The difference in energy between the conduction bands as one crosses the *pn*-junction is called the diffusion potential, V_{bi} (or built-in potential). In Figure 2, we show the equivalent case for an MS junction. Here, the barrier potential at the MS junction is ϕ_{bn} in *n*-type material and ϕ_{bp} in *p*-type material. Although MS contacts for both *p*- and *n*-type materials are shown, the discussion will be restricted to *n*-type material only. Consider first the *pn* junction. The regions denoted by $\oplus$ and $\ominus$ indicate the junction region depleted of free carriers, leaving behind ionized donors and acceptors. In this region, we solve Poisson's equation

$$-\frac{\partial^2 V}{\partial x^2} = \frac{\partial E}{\partial x} = \frac{\rho(x)}{\varepsilon} = \frac{q}{\varepsilon}[p(x) - n(x) + N_D^+(x) - N_A^-(x)] \quad (1)$$

or for regions predominately doped *p*-type

$$-\frac{\partial^2 V}{\partial x^2} \approx \frac{q}{\varepsilon} N_D^+ \qquad \text{for} \quad 0 < x \le x_n \quad (2)$$

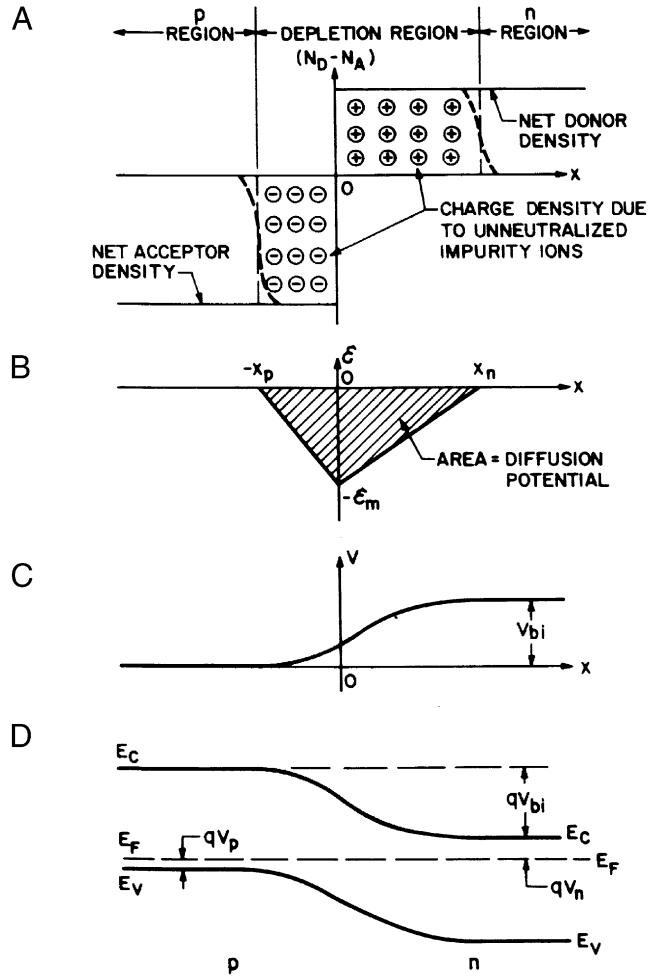


Figure 1. Abrupt *pn* junction in thermal equilibrium (no bias). After Sze (1981). (**A**) Space charge distribution in the depletion approximation. The dashed lines indicate the majority carrier distribution tails. (**B**) Electric field across the depletion region. (**C**) Potential distribution due to the electric field where V_{bi} is the (built-in) diffusion potential. (**D**) Energy band diagram.

or for *n*-type

$$-\frac{\partial^2 V}{\partial x^2} \approx \frac{q}{\varepsilon} N_A^- \qquad \text{for} \quad -x_p \le x < 0 \quad (3)$$

In these equations, V is the voltage, E the electric field, q the electronic charge, $p(x)$ and $n(x)$ the hole and electron concentrations (electric potential) comprising the mobile carriers, $N_D(x)$ and $N_A(x)$ the donor and acceptor doping concentrations, and $\varepsilon = K_S \varepsilon_0$, the permittivity with dielectric coefficient K_S. The spatial dependence, x, is measured relative to the physical location of the *pn*-junction. The solution of these equations in a form useful for C-V measurements is

$$V(x) = V_{bi}\left[2\left(\frac{x}{W}\right) - \left(\frac{x}{W}\right)^2\right] \quad (4)$$

where

$$V_{bi} = \frac{kT}{q} \cdot \ln\left(\frac{N_A N_D}{n_i^2}\right) \quad (5)$$

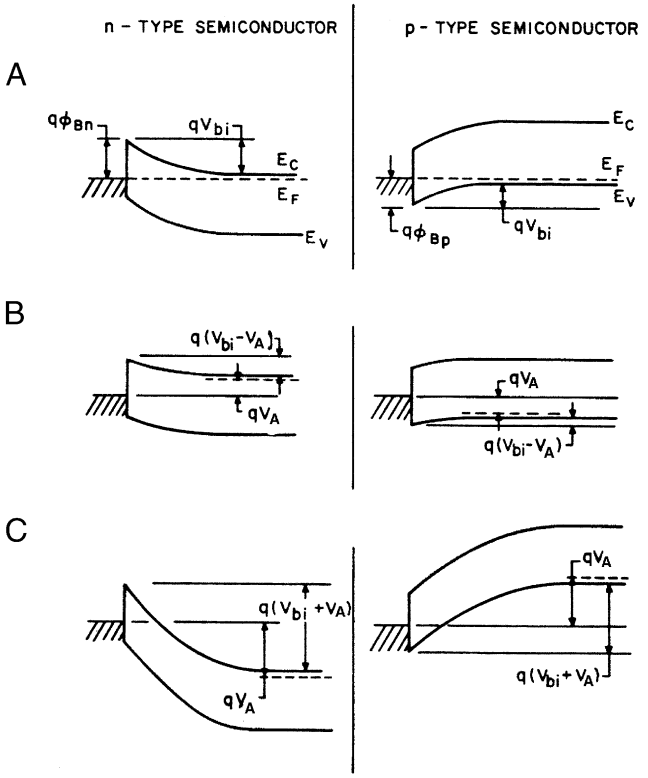


Figure 2. Energy band diagram of metal semiconductor (MS) contacts: ϕ_{Bn} (n-type) and ϕ_{Bp} (p-type) are the barrier heights at the MS interface required to establish equilibrium between the two materials. E_C, E_V, and E_F denote the conduction band energy, the valance band energy, and the equilibrium Fermi energy, respectively. Note that the diffusion potential is defined as $qV_{bi} = \phi B - (E_C - E_F)$. Diagrams after Sze (1981). (**A**) Thermal equilibrium for n- and p-type material. (**B**) Forward bias, $V = +V_A$. (**C**) Reverse bias $V = -V_A$.

is the diffusion potential, previously defined, and

$$n_i^2 = N_C N_V e^{E_G/kT} = 2.332 \times 10^{31} \left(\frac{m_{dc} M_{dv}}{m_0^2}\right)^{3/2} T^3 e^{E_G/kT} \tag{6}$$

in units of cm^{-6}. The intrinsic carrier concentration, n_i, contains the density of available states in the conduction (valence) band, $N_C(N_V)$, and m_{dc} and m_{dv} are the density of states effective masses in the conduction band and valence band respectively, while m_0 is the free electron mass. The band gap is $E_G = (E_C - E_V)$ and T is the absolute temperature. All these physical properties are tabulated in the literature. The depletion width $W = x_n + x_p$ is given by

$$x_n = \left[\frac{2K_S\varepsilon_0}{q} \cdot \frac{N_A}{N_D(N_A + N_D)} \cdot V_{bi}\right]^{1/2} \tag{7}$$

and

$$x_n = \left[\frac{2K_S\varepsilon_0}{q} \cdot \frac{N_D}{N_A(N_A + N_D)} \cdot V_{bi}\right]^{1/2} \tag{8}$$

Thus,

$$W = x_n + x_p = \left[\frac{2K_S\varepsilon_0}{q} \cdot \frac{(N_A + N_D)}{N_A N_D} \cdot V_{bi}\right]^{1/2} \tag{9}$$

See Figure 1A regarding the locations of W, $x_{n,}$ *and* x_p. The depletion width will widen under applied reverse bias, $-V_A$, and V_{bi} must be replaced by $(V_{bi} - V_A) > V_{bi}$. We are now able to calculate the capacitance of the depletion region, assumed to be a parallel-plate capacitor of plate separation W and area A,

$$C = \frac{K_S\varepsilon_0 A}{W} = \frac{K_S\varepsilon_0 A}{\left[\frac{2K_S\varepsilon_0}{q} \cdot \frac{(N_A+N_D)}{N_A N_D} \cdot (V_{bi} - V_A)\right]^{1/2}} \tag{10}$$

for a pn junction. Since C-V measurements cannot separately determine N_A and N_D in pn junctions, the usefulness of C-V profiling is quantitative for p^+n and n^+p junctions only. p^+ and n^+ designate very heavily doped regions reducing x_p and $x_{n,}$ respectively, to negligible thickness in the depletion layer. In this case, Equation 10 reduces to

$$C = \frac{K_S\varepsilon_0 A}{W} = \frac{K_S\varepsilon_0 A}{\left[\frac{2K_S\varepsilon_0}{qN_B} \cdot (V_{bi} - V_A)\right]^{1/2}} = \left[\frac{2K_S\varepsilon_0}{2(V_{bi} - V_A)}\right]^{1/2} A \tag{11}$$

where $N_B = N_D$ for the lightly doped side of p^+n junctions and $N_B = N_A$ for the lightly doped side of n^+p junctions. This same equation obtains for an MS junction, because the depletion layer in the metal is infinitesimally small.

For MS junctions on n-type semiconductors the diffusion potential is given by

$$V_{bi} = \phi_B - (E_C - E_F) = \phi_B - \frac{kT}{q}\ln\left(\frac{N_C}{N_D}\right) \tag{12}$$

where ϕ_B is the barrier height blocking the flow of carriers at the MS interface, E_C is the conduction band energy, and E_F the Fermi energy as previously defined and shown in Figure 2. The density of available states in the conduction band is given by

$$N_C = 2\left(\frac{2\pi m_{dc} kT}{h^2}\right)^{3/2} \tag{13}$$

where h is Planck's constant. For n-GaAs, $m_{dc}/m_o = 0.067$, so that for $T = 300$ K, $N_C = 4.35 \times 10^{17}/cm^3$.

PRACTICAL ASPECTS OF THE METHOD

In Table 1 are listed calculated depletion widths and capacitance for p^+n junctions in GaAs. The p^+ side is heavily doped to 2×10^{18} holes/cm^3, while three representative device-level donor doping levels are given for the n-side. The diodes are assumed to have typical areas of 1 mm^2.

Table 1. Depletion Widths and Capacitance for a GaAs p^+n Junctions[a]

		W (μm)			C (pF)		
N_D (cm^{-3})	V_{bi}(V)	$V_A = 0$	$V_A = -2$	$V_A = -10$	$V_A = 0$	$V_A = -2$	$V_A = -10$
1×10^{15}	1.228	1.30	2.11	3.94	84.9	52.4	28.1
1×10^{16}	1.288	0.42	0.68	1.25	262	164	88.4
1×10^{17}	1.348	0.14	0.22	0.41	792	502	273

[a] p^+ doping; $N_A = 2 \times 10^{18}$ cm^{-3}; $T = 300$ K; dielectric constant, $K_S = 13.2$; junction area, $A = 1$ mm^2.

Table 2 lists calculated values for a metal-GaAs Schottky barrier. Note that the significant difference is the diffusion potential V_{bi}.

The equations thus far assume that the doping concentration is uniform and that all of the dopants are ionized (i.e., $n = N_D$ in n-type material). In fact, a test of uniformity is to plot $1/C^2$ versus V_A which according to Equation 11 should be a straight line with slope proportional to the inverse of the doping concentration and the intercept on the positive abscissa equal to V_{bi}. This is illustrated in Figure 3 for capacitance normalized to its value at zero bias plotted versus V_A/V_{bi}.

C-V measurements are most useful if the differential form of the equations given above can be used. Rearranging and differentiating Equation 11 with respect to V we find

$$\frac{1}{C^2} = \frac{2(V_{bi} - V_A)}{qK_S\varepsilon_0 N_B A^2} \tag{14}$$

which leads to

$$N_B(W) = \frac{2}{qK_S\varepsilon_0 A^2}\left[-\frac{1}{d(1/C^2)/dV}\right] \tag{15}$$

The depth dependence of N_B can be determined from the measured capacitance, C

$$W = \frac{K_S\varepsilon_0 A}{C} \tag{16}$$

Equations 15 and 16 imply that it is possible to probe the semiconductor locally as a function of depth from the junction by merely increasing the reverse bias. We now wish to determine if our equations remain valid for the more general case of nonuniform doping. In the discussion below we follow the excellent description by Palmer (1990). In Figure 4 we depict a nonuniform doping profile, where for an n-type semiconductor $N_D(x)$ varies with position x. From Equations 1 and 2, we see that the electric field will vary according to the charge density $qN_D(x)$. Thus the areas under the $E(x)$ versus x curves are the voltages $V_1 = (V_{bi} + V_{A1})$ and $V_2 = (V_{bi} + V_{A2})$. If $V_{A2} = V_{A1} + \Delta V_A \ll V_{A1}$ then a small change in the depletion width will result, i.e., $\Delta W = W(V_{A2}) - W(V_{A1}) \ll W(V_{A1})$. Now, from Figure 4 we see that

$$V_2 - V_1 = V_{A2} - V_{A1} = E(W)W \tag{17}$$

And because we assumed ΔV_A to be small, we may write

$$E(W) = [dE(x)/dx] \cdot \Delta W = qN_D(x)\Delta W/(K_S\varepsilon_0) \tag{18}$$

But $qN_D(x)\Delta W$ is the charge per unit junction area, i.e., $qN_D(x)\ \Delta W = \Delta Q/A$. Thus, combining this result with Equations 17 and 18 we find

$$\Delta V_A = \Delta QW/(AK_S\varepsilon_0) \tag{19}$$

Since $C \equiv \Delta Q(W)/\Delta V_A$, we recover the relationship derived in Equations 10 and 11.

All of the equations developed above included the doping concentrations N_D (or N_A for p-type semiconductors). However, it is the mobile majority carriers, electrons of concentration n, or holes of concentration p, which respond to the varying applied bias. The donors N_D (or acceptors, N_A) are tightly bound in the lattice. Thus, the condition under which $N_D(W) = n(W)$ requires that the donor ionization energy be small compared to kT (at room temperature, $kT = 0.026$ eV, while the ionization energy for most shallow dopants is 0.010 eV or less). This condition is called

Table 2. Depletion Widths and Capacitance for a Metal n-GaAs Schottky Barrier[a]

		W (μm)			C (pF)		
N_D (cm^{-3})	V_{bi}(V)	$V_A = 0$	$V_A = -2$	$V_A = -10$	$V_A = 0$	$V_A = -2$	$V_A = -10$
1×10^{15}	0.744	1.01	1.95	3.85	109	56.8	28.7
1×10^{16}	0.803	0.33	0.62	1.22	331	177	90.4
1×10^{17}	0.863	0.11	0.20	0.40	989	543	279

[a] Barrier height, $\varphi_{Bn} = 0.9$ eV; $T = 300$ K, dielectric constant, $K_S = 13.2$, junction area, $A = 1$ mm^2.

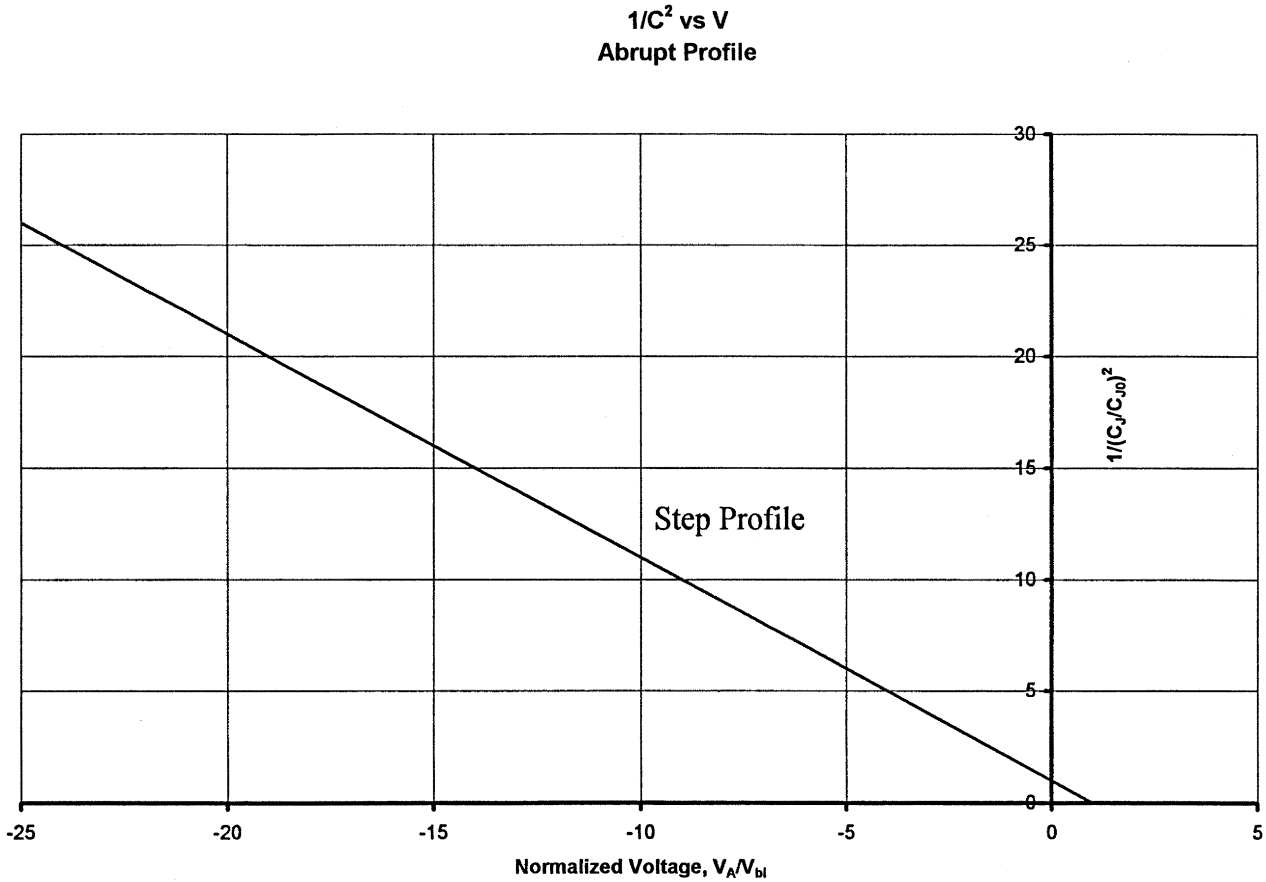


Figure 3. Normalized plot of C^{-2} versus V_A/V_{bi} for a uniform doping distribution indicating the expected straight line plot.

the extrinsic range of operation. Note that if the temperature is too high, however, the semiconductor becomes intrinsic and the free carrier concentration is determined by Equation 6, independently of the dopant concentration. We will show later that if traps are present deep within the forbidden energy gap, Equations 14 and 15 are not strictly true and do not represent either n or N_D. Since the depletion depth is determined by the number of immobile positive charges exposed (in an n-type semiconductor), then a more correct way to write Equation 15 is

$$N^+(W) = \frac{2}{qK_S\varepsilon_0 A^2}\left[-\frac{1}{d(1/C^2)/dV}\right] \tag{20}$$

and depth below the junction is written as

$$x = W(V_A) = \frac{K_S\varepsilon_0 A}{C} = \left[\frac{2K_s\varepsilon_0}{qN_B}(V_{bi} - V_A)\right]^{1/2} \tag{21}$$

Note that for shallow donors (i.e., with donor ionization energy $<kT$ as discussed above) and no traps, $N^+ = N_D{}^+ = n$. However, in general, Equation 20 and Equation 21 are the "working" equations to be used in C-V profiling.

C-V PROFILING EQUIPMENT

In order to apply Equations 20 and 21, the apparatus shown schematically in Figure 5 is employed. The C-V meter superimposes a small ac signal (typically 10 to 15 mV) onto the dc bias so that the voltage drop across the junction is $V_A + v_{ac}$. During the positive half of the ac signal, the reverse bias is reduced slightly, causing a small reduction in the depletion width ($W = W_A - |w_{ac}|$). The depletion width increases slightly during the negative portion of the cycle. Typical of capacitance meters used is the Boonton 72B meter. It operates at a fixed frequency of 1 MHz and has an amplitude of 15 mV. Its capacitance range covers 1 pf to 3000 pf, sufficient to cover the range of capacitance found in Schottky barriers and *pn* junctions (see Tables 1 and 2). Other meters used for C-V measurements include the variable frequency Hewlett-Packard models HP-4271B and HP-44297A LCR meters and the Keithley −590 capacitance meter. The Keithley meter permits measurements at 100 kHz as well as 1 mHz. Most of the apparatuses in use are also interfaced to a computer for automation of the measurement plus real-time data reduction and plotting. Several manufacturers offer bundled C-V profiling packages with turn-key operation.

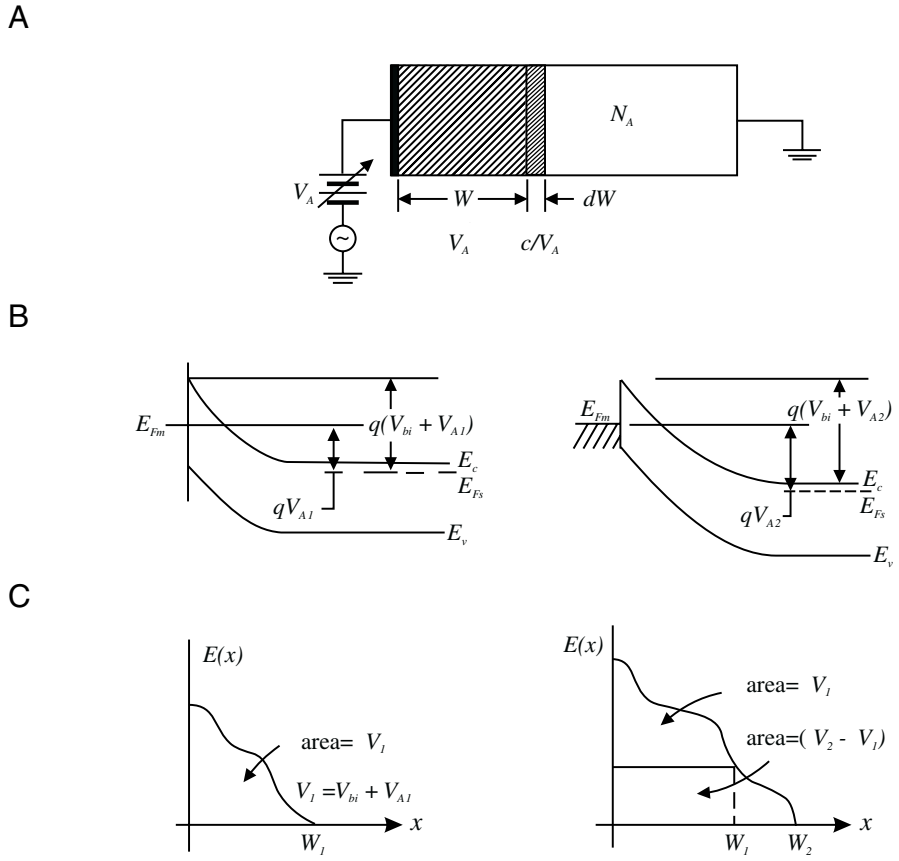


Figure 4. A plot of the electric field versus distance for a nonuniform doping profile. Compare the electrical field here with the electrical field in Figure 1 where the doping on either side of the junction was uniform. $V_{A2} > V_{A1}$. (**A**) Biasing circuit. (**B**) MS junction for applied potentials V_{A1} and V_{A2} with $|V_{A2}| > |V_{A1}|$. (**C**) Electric field versus position for a nonuniform carrier distribution corresponding to V_{A1} and V_{A2} in (**B**).

OTHER C-V PROBING PROCEDURES

Mercury Probe Contacts

The disadvantage of all carrier concentration measurement techniques is the need to establish contacts to the material under investigation. For Schottky barriers, a metal of precise area is deposited onto the wafer using photolithographic techniques. A low-resistance ohmic contact must also be formed in order to complete the circuit. These procedures are destructive, and the material is lost to further use. In an attempt to eliminate these Schottky barrier fabrication procedures, mercury probes have been developed to make temporary contact to the semiconductor surface. A schematic representation of an Hg probe is shown in Figure 6 (Schroder, 1990). The Hg is forced against the semiconductor by pulling it down against a precisely defined area using a vacuum. Simultaneously the Hg is siphoned up against the semiconductor making intimate contact (Jones and Corbett, 1989). The larger-area contact in the figure is the second contact to complete

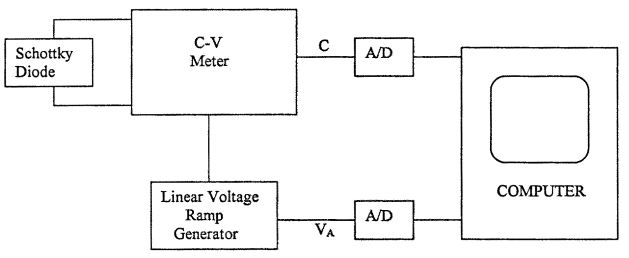


Figure 5. Schematic diagram of a typical automated C-V apparatus. Linear ramp generator provides the reverse bias to the Schottky diode. A/D, analog/digital converters.

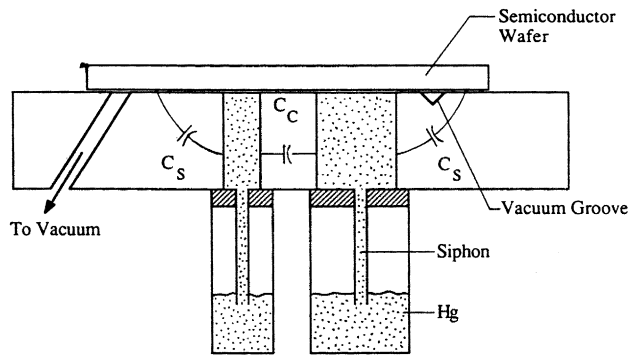


Figure 6. Typical mercury probe. The smaller contacting area is the reverse-biased Schottky barrier. (After Schroder, 1990.)

the circuit. Its area is many times that of the profiling contact, so that the two capacitances in series essentially look like the small reverse biased probing capacitance acting alone. Unfortunately, reproducible results are difficult to obtain, requiring careful surface preparation, very clean Hg, and careful cleaning of the semiconductor surface after use. Furthermore, Hg is toxic and safe procedures must be observed. Good discussions of the use of Hg probes may be found in Jones and Corbett (1989) and Schaffer and Lally (1983).

Electrochemical Profiling

This is another attempt to simplify C-V profiling, in which the Schottky barrier is formed at an electrolyte-semiconductor interface [see SEMICONDUCTOR PHOTOELECTROCHEMISTRY]. Ambridge and Faktor (1975) have described the description of the electrochemical profiler in detail. The measurements can be performed at a constant dc voltage (no modulating signal) and the depth is changed by electrolytically etching the semiconductor between capacitance measurements. The cell cross-section shown in Figure 7 is taken from the review article by Blood (1986). The Schottky barrier is formed by the electrolyte, and the back contact by springs or a plunger, which press the semiconductor against a sealing ring that also serves to define the area of the Schottky contact. The light source is used to generate holes in n-type semiconductors so that etching can take place under reverse bias. When the sample is illuminated, the etch rate is proportional to the photo-induced current flowing through the cell. For p-type material, illumination is not needed because etching takes place under forward bias. Thus, for p-type material the procedure involves forward biasing to etch away material, followed by reverse biasing to make the C-V measurements. Because in principle it is possible to etch to any depth, electrolytic profiling places no limit on the depth that can be probed. Complete details for electrolytic C-V depth profiling can be found in the excellent review by Blood (1986).

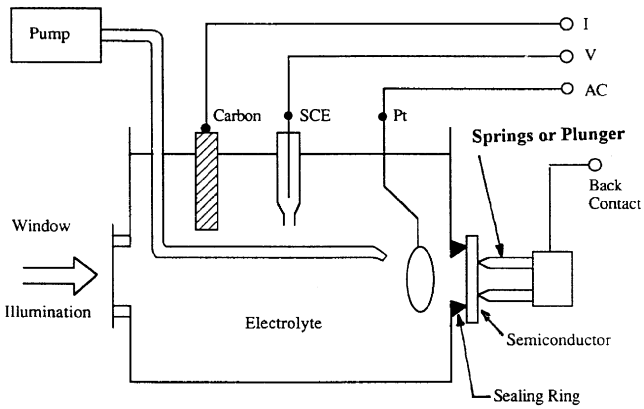


Figure 7. Schematic diagram of an electrochemical cell. The pump is used to agitate the electrolyte to maintain a uniform concentration at the semiconductor surface and to disperse bubbles from the surface. SCE is a saturated calomel electrode, used to maintain the correct overpotential in the cell. Diagram after Blood (1986).

DATA ANALYSIS: TYPICAL C-V RESULTS

An example of a uniformly doped n-GaAs layer grown in the authors' laboratory, using a halide vapor phase epitaxy (VPE) reactor, is shown in Figure 8A. The Schottky diode was fabricated by evaporating a 1.13-mm diameter gold guard ring structure ($A = 1$ mm^2) using photolithographic lift-off. Precise definition of the diode diameter is essential because it enters Equation 20 as the fourth power of d. The measurement of the area in Equation 20 is the largest source of error in determining N^+. This example is chosen to illustrate several points. The doping concentration is $\sim 3.6 \times 10^{15}$ cm^{-3}. The following observations may be made.

1. The profile depth starts at about $W = 0.55$ μm, corresponding to the zero bias depletion depth. C-V profiling is limited in how close to the MS junction data can be obtained by this zero bias distance. This minimum depth is calculated using Equation 21 with $V_A = 0$ and Equation 12 to calculate V_{bi}.
2. The profile is uniform, with no indication of carrier freeze-out at 80 K, indicating the ionization energy of the Te-donor atoms (~ 0.003 eV) is $< kT = 0.007$ eV.

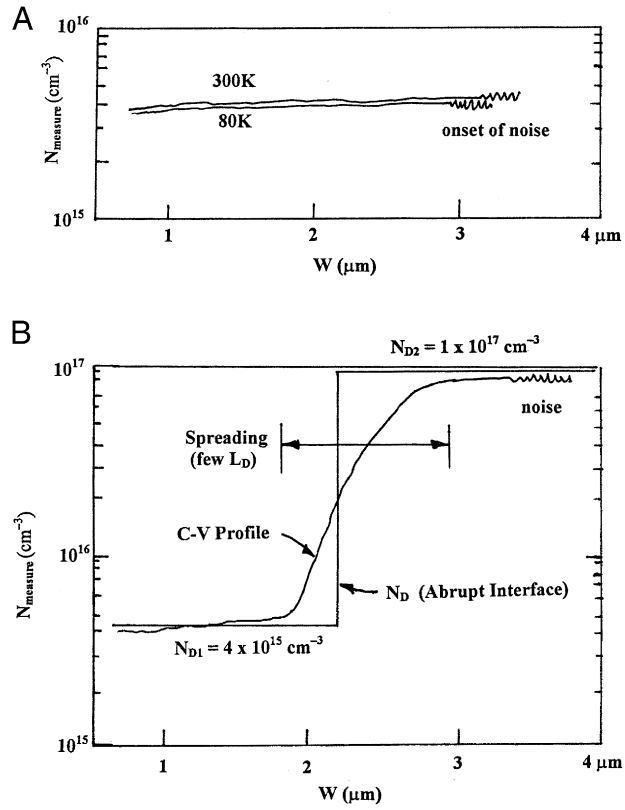


Figure 8. Carrier profiles determined in the authors' laboratory on VPE grown n-GaAs. (**A**) A single layer uniformly doped to $\sim 4 \times 10^{15}$ cm^{-3}. Note the onset of noise limiting the depth of the profile. (**B**) A double layer growth with a heavily doped layer grown first ($N_D = 1 \times 10^{17}$ cm^{-3}) followed by a less heavily doped cap layer ($N_D = 4 \times 10^{15}$ cm^{-3}). The solid straight vertical line indicates the actual abrupt growth interface. Note the gradual transition in the C-V profile due to spreading over several Debye lengths, L_D.

3. The maximum depth depends upon the thickness of the epitaxial layer on the underlying semi-insulating substrate or upon the breakdown voltage of the reverse-biased junction. The breakdown of reverse-biased metal-compound semiconductor Schottky barriers tends to be much lower than predicted from the maximum possible field strength. In this case, the onset of noisy signals at $V_A = -40$ V terminated the measurement. The predicted breakdown voltage for this doping concentration is $V_R = -190$ V. For more heavily doped material, voltage breakdown of the junction limits the maximum depth probed.

The second example in Figure 8B indicates the profile measured for a GaAs structure in which the two layers were grown with the first layer doped to a higher donor concentration of $N_D = 1 \times 10^{17}$ cm^{-3}, followed by a layer doped to $N_D = 4 \times 10^{15}$ cm^{-3}, the top surface layer onto which the Schottky barrier is fabricated. This allows us to observe another limitation of C-V profiling. Although the transition from the lightly doped to the heavily doped layer was abrupt, the C-V data indicate a gradual transition from the first to second layer. This deviation of the majority carriers, $n(x)$, from the dopant profile, N_D, is explained by considering the local distance required to establish equilibrium (see Problems, below).

The data reported in Figure 8 can be recorded and plotted as shown at a rate of ~10 min per figure with automated equipment. The time required to add data acquired at several temperatures is determined by (a) the time to electrically and thermally mount the test structure, and (b) the time required to reset each temperature in the cryostat.

SAMPLE PREPARATION

There is no standard sample preparation procedure for fabricating Schottky barrier devices to measure C-V profiles. In fact, Schottky barrier fabrication is material-specific for each semiconductor. An ideal Schottky barrier structure is one in which the semiconductor material of interest is grown on a heavily doped, low-resistivity substrate of the same material. This reduces the series resistance, R_S, and makes formation of the ohmic back-contact straightforward. After performing the cleaning procedure to be discussed below, the ohmic contact is formed first. This can be as simple as depositing a metal or metal combination followed by a low temperature sinter to "form" the ohmic contact. In a worst case scenario, the metal must be "alloyed" at a higher temperature to react it with the heavily doped substrate. After formation of the ohmic contact, the wafer must be recleaned to prepare the opposite surface for the Schottky barrier contact. Although cleaning procedures are also material-specific, there are certain generic steps common to all semiconductors. It is essential that the surface of the semiconductor be spotlessly clean. To accomplish this, the surface must be stripped of all organic residue and native oxides. Organic materials are removed by agitating the wafer in acetone followed by methanol and rinsing with deionized (18 MΩ) water. Native oxides formed in the water rinse and/or by exposure to laboratory air are usually removed using buffered hydrofluoric acid (buffered HF), a commercially available acid. However, not all semiconductors are etched this way. For example, the standard oxide etch for GaAs is one part hydrochloric acid in one part water (1:1 HCl:H_2O). By comparison, this same HCl:H_2O etch will destroy an InP sample in seconds. The study of newly evolving semiconductors will require developing procedures compatible with the material under study, often by trial and error. If the cleaning and etching procedure is successful, the sample surface will be completely dry as it is withdrawn from the acid etch (i.e., will be hydrophobic). If the surface has wet areas or spots, this means that the cleaning procedure has failed and must be repeated. The importance of the cleaning step prior to metallization cannot be overemphasized.

The sample should be loaded directly into the metallization station immediately after the etching step to minimize the formation of native oxide. A typical metallization for Si and the group III-V compound semiconductors is the deposition, usually by thermal evaporation in vacuum, of a thin (~40 nm) "adhesion" layer of a reactive metal such as chromium followed by a thicker (~400 nm) layer of a noble metal (gold) or a soft metal such as aluminum. This thickness of metal is adequate for wire bonding for cryostat mounting of the sample, or for probing using a standard probing station for single temperature measurements. The only requirement is that the metal in contact with the semiconductor form a low-leakage Schottky barrier. Often the equipment available governs the metal chosen and the procedures adopted in a given laboratory.

If the semiconductor of interest is grown on a high-resistivity substrate, ohmic contact formation may not be possible. In this case, a second large-area MS contact is fabricated simultaneously with the precisely defined Schottky barrier of known area. In a typical 5 × 5-mm test sample, this second contact will be at least 15 times larger than the Schottky gate, and since the capacitances are in series, will introduce an error of no more than 6%. To improve accuracy in this case, the diameter of the Schottky barrier contact is reduced (by a factor of two) to reduce the error to ~1%.

PROBLEMS

Instrument Limitations

The capacitance meter is the ultimate resolution-defining instrument in determining the $n(x)$ versus W profile. The meter should be of sufficient accuracy so that the Debye length, L_D, is the limiting factor for the spatial resolution. Amron (1967) has discussed in detail the limits to accuracy in C-V profiling. It would appear that large signals are desirable and that ΔC should be made as large as possible. But large ΔC deteriorates the spatial localization of the measurement. This can be seen from Equation 10 and Equation 11, where the differential changes in C and W are related through

$$\beta = \frac{\Delta C}{C} = -\frac{\Delta W}{W} \tag{22}$$

where β is a precision parameter introduced by Amron (1967) which sets the error limits in terms of an acceptable standard deviation in $n(x)$ and W. Amron shows that it is desirable to keep $\beta < 0.1$. For a uniform profile, and constant modulation voltage ΔV, ΔC decreases because W increases with reverse bias and C decreases (see Tables 1 and 2). Thus, as ΔC decreases, the profiles become noisier, limiting the depth to which the profiles can be measured. See the increased noise for increasing W in the profiles for GaAs in Figure 8.

We now address the problem of measuring real devices versus ideal structures discussed to this point. The capacitance meter sees the diode as an admittance of the form

$$Y = G + j\omega C \tag{23}$$

and measures the quadrature component, $j\omega C$. However, it is assumed in the measurement that any series resistance, R_S, is small. Series resistance, R_S, includes the bulk resistance of the device under test and any lead resistance. The circuit model is shown in Figure 9. Wiley and Miller (1975) have considered the effects of R_S on the measurement. Calculation of the admittance with R_S in the circuit yields the following equations for C and G

$$C_M = \frac{C}{(1 + R_S G)^2 + (\omega R_S C)^2} \tag{24}$$

and

$$G_M = \frac{G(1 + R_S G) + R_S \omega^2 C^2}{(1 + R_S G)^2 + (\omega R_S C)^2} \tag{25}$$

Only the measured capacitance, C_M, is of interest. For C_M to closely approximate the true capacitance, C, we require $R_S G \ll 1$ and $(\omega R_s C)^2 \ll 1$. Provided that careful fabrication procedures are followed, the $R_S G \ll 1$ specification is easily met, because in good diodes $G \to 0$. This condition requires that leakage currents be minimized, a condition easily met with careful fabrication procedures. The condition requiring $(\omega R_s C)^2 \ll 1$ is more difficult and must be addressed for individual diodes. For example, using a typical capacitance meter operating at $f = \omega/2\pi =$ 1 MHz, the condition on R_S to guarantee 1% accuracy is

$$R_S < 1.6 \times 10^4 / C(pf) \tag{26}$$

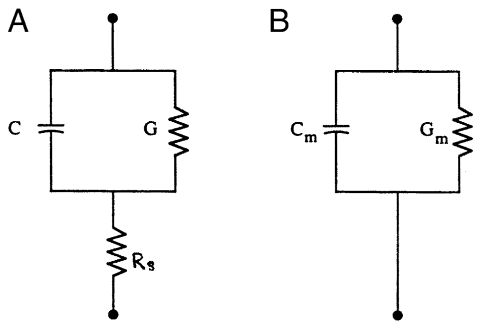


Figure 9. (**A**) Realistic circuit and (**B**) the ideal circuit describing a *pn* junction or Schottky barrier. Resistance, R_S, accounts for the bulk resistance of the semiconductor being measured, and the lead resistance. R_S must be minimized.

with the results being obtained in Ω. For the range of capacitance in Tables 1 and 2 this sets limits of $R_S < 16$ Ω for the largest capacitance (heaviest doping) and $R_S < 571$ Ω for the smallest capacitance (lightest doping). A rule of thumb is that the series resistance must always be <100 Ω. Thus, careful sample preparation is of great importance (see Sample Preparation).

In the discussion above of the data presented in Figure 8B for two layers of significant doping difference, we noted a gradual change in the profile as measured by the C-V apparatus. The reason for this apparent gradual transition in a region where the doping profile changes abruptly is as follows. This distance is the region over which the combination of the diffusion of electrons away from the concentration gradient at the junction is balanced by the drift due to the electric field, and is described by the Debye shielding length given by

$$L_D = \left[\frac{kTK_S\varepsilon_0}{q^2 N_D}\right]^{1/2} \tag{27}$$

L_D is a measure of the distance over which the mobile majority carriers eliminate any charge imbalance through diffusion and drift. Thus, at the abrupt interface, the spatial resolution of the $n(x)$ measurement is limited to a few Debye lengths. Johnson and Panous (1971) have presented detailed calculations of this effect. For further details of profiling errors and instrumental limitations see the excellent paper by Blood (1986).

TRAPS AND THEIR EFFECT ON C-V PROFILING

Finally, we present a short discussion of carrier trapping centers and their effect on carrier profiling using the C-V method. The effects of traps on C-V measurements can be profound, and only an outline of their effects will be presented here. Traps (defects) in semiconductors are a complementary area of study, and are discussed in DEEP-LEVEL TRANSIENT SPECTROSCOPY. Although traps can change the interpretation of C-V profiling, most workers ignore them. Detailed consideration of the effects of traps on C-V profiling can be found in the paper by Kimerling (1974).

We first define a trap as an energy level in the forbidden gap several kT below the shallow donor level (or above the acceptor level). Following the definitions in DEEP-LEVEL TRANSIENT SPECTROSCOPY they are designated acceptor traps if they have negative charge after trapping an electron, and as neutral when empty (i.e., the electron has vacated the trap). Similarly, donor traps are neutral when they possess a trapped electron and are positive after the electron has been emitted.

For discussion purposes we will consider only donors in n-type semiconductors and acceptor-like traps. Further, we assume the concentration of traps, N_{tr}, is $<N_D$. In Figure 10 we show two different trap levels. One is far from the conduction band and is not influenced by the band bending at the MS interface or the effects of reverse bias and the AC modulation of the capacitance measurement. The other trap is closer to the donor level and is

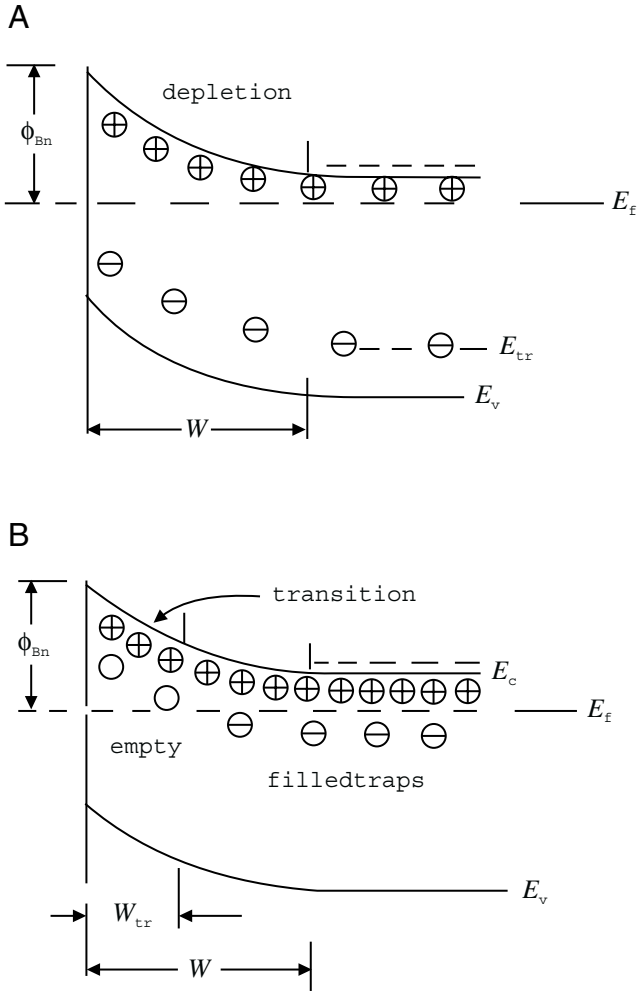


Figure 10. Schematic diagram of traps. (**A**) "Deep" traps, which do not contribute directly to the profile except to compensate some of the donors. (**B**) "Shallow" traps, which may contribute to the C-V profile due to rising above the Fermi level in the transition region. Note: "deep" and "shallow" refer to energy level within the forbidden gap, and not position.

influenced by both the band bending at the MS interface and the modulated signal used to measure the capacitance.

The deep trap, Figure 10A, behaves much like compensating acceptors in a Hall measurement (see HALL EFFECT IN SEMICONDUCTORS). In this case, no acceptor-like trap is ever above the Fermi level, so that in both the depletion region and the bulk semiconductor the measured carrier concentration for all values of reverse bias is

$$n(x) = N^+(x) = N_D(x) - N_{tr}(x) \tag{28}$$

where

$$(E_C - E_{tr}) \gg kT \tag{29}$$

If the trap energy rises above the Fermi level in the depletion region, it may influence the measured donor concentration if the traps can be "uncovered" by the modulation voltage of the capacitance meter. This is the case depicted in Figure 10B. There are now two distances into the depletion region of interest. The depletion region is still designated as W, but the region where the traps are above the Fermi level, and hence empty, is designated $W_{tr} < W$. The region between W and W_{tr} is called the transition region. In this region, if the emission rate, e_n from the trap is small compared to the modulation frequency of the capacitance meter, i.e., $e_n \ll \omega/2\pi$, then only those traps in the region $x < W_{tr}$ contribute to the measured carrier density. However, if $e_n \geq \omega/2\pi$, then the experimenter must account for the electrons emitted from the traps at W_{tr}. These "shallow" traps can confuse the interpretation of the carrier profile determined from a C-V measurement. The interested reader should consult the paper by Kimerling (1974). Fortunately, for most materials trap densities are much less than the doping densities (often N_{tr} is <1% of N_D) and can in many cases be neglected.

LITERATURE CITED

Ambridge, T. and Faktor, M. M. 1975. An automatic carrier concentration profile plotter using an electrochemical technique. *J. Appl. Electrochem.* 5:319–328.

Amron, I. 1967. Errors in dopant concentration profiles determined by differential capacitance measurements. *Electrochem. Technol.* 5:94–97.

ASTM. 1985. Standard test method for net carrier density in silicon epitaxial layers by voltage-capacitance of gated and ungated diodes: ASTM Standard F419. *In* 1985 Annual Book of ASTM Standards. American Society for Testing of Materials, Philadelphia.

Blood, P. 1986. Capacitance-voltage profiling and characterization of III-V semiconductors using electrolyte barriers. *Semicond. Sci. Technol.* 1:7–27.

Grove, A. G. 1967. Physics and Technology of Semiconductor Devices. John Wiley & Sons, New York.

Hillibrand, J. and Gold, R. D. 1960. Determination of the impurity distribution in junction diodes from capacitance voltage measurements. *RCA Rev.* 21:245–252.

Johnson, W. C. and Panousis, P. T. 1971. The influence of Debye length on the C-V measurement of doping profiles. *IEEE Trans. Electron. Dev.* Ed-18:965–973.

Jones, P. L. and Corbett, J. W. 1989. Investigation of the electrical degradation of silicon Schottky contacts due to mercury contamination. *Appl. Phys. Lett.* 55:2331–2332.

Kimerling, L.C. 1974. Influence of deep traps on the measurement of free carrier distributions in semiconductors by junction capacitance techniques. *J. Appl. Phys*. 45:1839–1845.

McKelvey, J. P. 1966. Solid State and Semiconductor Physics. Harper and Row, New York.

Palmer, D. W. 1990. Characterization of semiconductors by capacitance techniques. *In* Growth and Characterization of Semiconductors. (R. A. Stradling and P. C. Klipstein, eds.). Adam Hilger, Bristol, U.K.

Pierret, R. F. 1996. Semiconductor Device Fundamentals. Addison-Wesley, New York.

Schaffer, P. S. and Lally, T. R. 1983. Silicon epitaxial wafer profiling using the mercury-silicon Schottky diode differential capacitance method. *Solid State Technol.* 26:229–233.

Schroder, D. K., 1990 Semiconductor Material and Device Characterization. John Wiley & Sons, New York.

Sze, S. M. 1981. Physics of Semiconductor Devices, 2nd ed. John Wiley & Sons, New York.

Wiley, J. D. and Miller, G. L. 1975. Series resistance effects in semiconductor C-V profiling. *IEEE Trans. Electron. Dev.* Ed-22:265–272

KEY REFERENCES

ASTM, 1985. See above.

Describes the standard, accepted testing procedure.

Blood, 1986. See above.

A review of C-V profiling with special attention paid to electrolytic techniques.

Hillibrand and Gold, 1960. See above.

The original C-V profiling paper. Still relevant today.

Schroder, 1990. See above.

A good collection of materials characterization techniques including a chapter on C-V profiling.

Sze, 1981. See above.

The standard text on device physics. Good discussion of the depletion layers, depletion capacitance and Schottky barriers.

PETER BARNES
Auburn University
Auburn, Alabama

CHARACTERIZATION OF *pn* JUNCTIONS

INTRODUCTION

The *pn* junction of a semiconductor material forms part of many commercially important devices such as bipolar junction transistors (BJTs), heterojunction bipolar transistors (HBTs), junction field-effect transistors (JFETs), and various kinds of diodes (e.g., Zener, rectifying, laser, light emitting). Semiconductor engineers routinely use the current-voltage (*I-V*) measurement technique both for the design of semiconductor devices and for yield or process checks during manufacture. The *I-V* measurement technique is also considered to be a method in materials research because it is used to characterize semiconductor materials and processes used to create these materials. The *pn* junction *I-V* technique is useful in materials research because complementary techniques often do not definitively answer whether the material quality is sufficiently good to operate as required in a semiconductor device. The technique can definitely be used on *pn* junctions present in state-of-the-art semiconductor devices, but in most materials research laboratories, diodes with large feature sizes and considerably simplified sample preparation are preferred for rapid turnaround and lower equipment cost. The technique is most often used when the purpose of characterizing the material is for fabrication of *pn* junction devices such as BJTs, HBTs, JFETs, and diode structures for a number of applications, including semiconductor lasers and light-emitting diodes. In that case, the method can be used primarily as a measure of material quality in a device application or as routine intermediate process characterization during device fabrication.

Measurements of *pn* junctions do not measure material properties per se but rather are used to extract parameters from the ideal current equations presented below (see Principles of the Method). These parameters can be related to physical properties of the *pn* junction material (through theoretical analysis; see, e.g., Sze, 1981), but in practice, this is rarely done because the extracted parameters are more relevant as a qualitative evaluation of the semiconductor material. Some parameters commonly measured for *pn* junctions are the built-in voltage V_{bi} of the *pn* junction, the ideality factor n of current for a forward-biased *pn* junction, and the reverse-biased breakdown voltage V_{br}. Different semiconductors will deviate from the ideal equation in predictable ways. Practical examples of different materials characterized with the *pn* junction *I-V* technique will be presented later (see Practical Aspects of the Method).

Competitive and Complementary Techniques

A number of techniques can be used to study defects directly in semiconductors. They are often used when more insight is needed as to the specific nature of the recombination centers or breakdown characteristics of a *pn* junction structure. As such, they are more complementary than competitive and are used more often for fundamental studies, whereas the *pn* junction *I-V* method is an appropriate method for establishing the actual characteristics of the device of interest. Alhough many more techniques than those presented below can be used to characterize semiconductor material with *pn* junctions, the techniques summarized below are the main complements to *I-V*.

Cathodoluminescence (CL) is a technique in which an electron beam is used to excite electron-hole pairs in the sample and detect luminescence of radiative recombination pathways. By scanning the electron beam using methods similar to scanning electron microscopy (SEM; see TRANSMISSION ELECTRON MICROSCOPY), a map of the luminescence is obtained and is correlated with the spatial variation of defects in the material. The luminescence is analyzed for wavelength of the emission and can be related to different types of defects. Analysis and identification of the defects are empirical in nature, and correlation to theoretical modeling of the defects is often used. Cathodoluminescence is a powerful technique, but it is costly, time consuming, and not able to image defects with nonradiative transitions.

The electron-beam-induced current (EBIC) technique is another way of mapping the defects in a *pn* junction sample. An electron beam is scanned across a sample, and the current through the *pn* junction is measured at an electrode on the other side of the junction (commonly the backside of a doped semiconductor substrate). When the electron beam is scanned across a region with defects, the current through the sample will be reduced as the defects capture electrons and thereby reduce the current to the bottom electrode. As with CL, a defect map can be obtained by this method with equipment, ease of use,

and cost similar to CL. The EBIC method can be used to map defects that recombine by both radiative and nonradiative transitions and is a useful complement to CL. Both CL and the EBIC technique provide information not available by the *pn* junction *I-V* method, such as minority-carrier diffusion lengths (e.g., from a spatial measurement of the dark regions in the EBIC method). The EBIC technique can also yield information regarding the carrier diffusion lengths.

A newer defect mapping technique with less widespread use is near-field scanning optical microscopy (NSOM). It is similar to CL but with optical excitation in the place of electron excitation. Either the excitation or the detected signal (or both if one can accept the small signals) will be through an optical fiber that can be scanned to provide a defect map.

Deep-level transient spectroscopy (DLTS; see DEEP-LEVEL TRANSIENT SPECTROSCOPY) is another technique for studying defects in semiconductors. Its main advantage is that it yields information about defects deep within the forbidden gap. The method can be used for *pn* junctions as well as for Schottky barriers as long as a one-sided abrupt junction is used (abrupt separation of carrier dopants with one side of the junction heavily doped and the other side lightly doped). A number of different measurements are possible with many nuances. One type uses a constant capacitance as the temperature of the sample is scanned. The capacitance C changes abruptly as defect centers are populated or depopulated. These data are analyzed and compared to models to gain fundamental understanding of defects in *pn* junctions. Deep-level transient spectroscopy uses a very similar experimental setup as *pn* junction *I-V* or capacitance-voltage *C-V* methods (with the exception of temperature scanning of the sample) and gives complementary information. Its main drawback is that the data analysis and interpretation can be rather complicated.

Transmission electron microscopy (TEM; see SCANNING ELECTRON MICROSCOPY) is a powerful technique that images the structure of semiconductor materials at the atomic level. Defects are imaged directly. Alhough the power of TEM is undisputed, the technique is costly, sample preparation is difficult and time consuming, and perhaps most importantly, the fraction of the sample analyzed is very small and one cannot be sure if the defects imaged are representative of the entire sample. Its use is as a complementary technique for some of the most difficult problems.

Secondary ion mass spectroscopy (SIMS) can give unrelated but useful information. A beam of ions is used to sputter a sample, and the sputtered ions are mass analyzed to reveal the elemental analysis as a function of depth in the sample. Secondary ion mass spectroscopy can detect many elements with a sensitivity of better than one part in 10^6 and is useful for detecting many impurities and the diffusion broadening of the *pn* junction. This information is useful for distinguishing effects due to impurities from those due to crystalline defects. The SIMS equipment is large and costly to operate, but the commercial availability of reasonable cost analyses makes it an attractive option and a widely used complementary method.

Finally, the *pn* junction *C-V* technique (see CAPACITANCE-VOLTAGE (C-V) CHARACTERIZATION OF SEMICONDUCTORS) provides complementary information to the *I-V* technique. The spatial dependence of doping density extracted from the *C-V* technique is a useful complement to the breakdown voltage in determining material quality.

The *pn* junction *I-V* method is one of the simplest methods available to characterize the quality of semiconductor material. From the discussion above, it is apparent that many complementary techniques are available when the *I-V* data show indications of lesser material quality requiring a deeper understanding of the defects or impurities involved. These complementary techniques are also more generally applicable to any doped (as well as undoped for some techniques) semiconductor material, not just those involving *pn* junctions.

PRINCIPLES OF THE METHOD

The ideal diode equation has been derived by Shockley (1949, 1950) with the following assumptions: (1) there is an abrupt junction where the depletion approximation is valid, i.e., the built-in potential and applied voltages are modeled by a dipole layer with abrupt boundaries and the semiconductor is neutral outside of the boundaries; (2) no generation or recombination takes place in the depletion region; (3) injection currents are low such that injected minority-carrier densities are small compared with the majority-carrier densities; and (4) the Boltzmann approximation applies (used for deriving equations to calculate free carriers in each of the regions; generally assumes the semiconductor to be nondegenerate, i.e., doping levels are such that the Fermi level is not within $3kT$ of the conduction or valence band edges). The reader should consult introductory texts for derivation of the ideal diode equation and for a better understanding of the assumptions and their implications. In that case, the equations of state are solved for the current density,

$$J = q\left(\frac{D_{\mathrm{N}}}{L_{\mathrm{N}}}n_{\mathrm{P0}} + \frac{D_{\mathrm{P}}}{L_{\mathrm{P}}}p_{\mathrm{n0}}\right)\left[\exp\left(\frac{qV_j}{nkT}\right) - 1\right] \qquad (1)$$

where D_{N} and D_{P} are the electron and hole diffusion constants, n_{p0} and p_{n0} are the equilibrium electron and hole concentrations in the n and p regions, L_{N} and L_{P} are the Debye lengths for electrons and holes, V_{j} is the junction voltage, q is the magnitude of the electron charge, k is the Boltzmann constant, and T is the temperature. More generically,

$$J = J_0\left[\exp\left(\frac{qV_j}{nkT}\right) - 1\right] \qquad (2)$$

$$J = J_0\left[\exp\left(\frac{q(V_{\mathrm{a}} - IR)}{nkT}\right) - 1\right] \qquad (3)$$

The applied voltage $V_{\mathrm{a}} = V_{\mathrm{j}} + IR \approx V_{\mathrm{j}}$ for low currents, where R is the extrinsic diode resistance. The first term dominates for positive applied voltages (forward-biased junction) several times greater that kT, and the current

increases exponentially. The second term dominates for negative applied voltages (reverse-biased junction) several times greater than kT, and the reverse current approaches a constant value $-J_0$. In practice, this equation applies fairly well to Ge but less so to Si and GaAs semiconductors. The departures from ideality are mainly due to (1) surface-related currents, (2) the generation and recombination of carriers between states in the band gap, (3) the tunneling of carriers between states in the band gap, (4) the high injection condition, and (5) the series resistance in the diode.

Surface effects are complicated to model and are often neglected in diode measurements (although they are manifested more plainly in bipolar transistors). When recombination in the depletion region is taken into account, the current density can be given by the following expression for a p^+n junction, where p^+ refers to degenerately doped p-type material:

$$J = J_0\left[\exp\left(\frac{qV_j}{nkT}\right) - 1\right] + \frac{qn_iW}{2\tau_0}\left[\exp\left(\frac{qV_j}{nkT}\right) - 1\right] \quad (4)$$

where W is the depletion width, τ_0 is the effective lifetime, n_i is the intrinsic carrier concentration, and the other parameters are as defined previously. In forward-bias conditions, the exponential terms dominate. At very low positive V_j, the second term can dominate for semiconductors with small n_i (band gap $E_g \geq 1$ eV at room temperature as for Si and GaAs). This comes about because J_0 is proportional to $(n_i)^2$. This is illustrated in Figure 1, where the logarithm of current is plotted vs. voltage and region (a) shows an inverse logarithmic slope of $2kT/q$. At higher V_j, the inverse slope approaches the ideal kT/q value in region (b) but again deviates at higher applied bias in region (c). The current in region (c) has been attributed to effects of high injection (conditions where the injected minority carriers approach the same concentration as the majority carriers). The net result is an increase in recombination with an exponential slope of $2kT/q$. Region (b) has been described as ideal, but in practice, recombination can be significant over the entire range of bias and the exponential slope can have a value between kT/q and $2kT/q$. This is accounted for by writing the equation for a practical diode,

$$J = J_0\left[\exp\left(\frac{qV_a}{nKT}\right) - 1\right] \quad (5)$$

where the exponential slope is defined by the ideality factor n, which typically varies as $1 \leq n \leq 2$. A perfect semiconductor crystal will have some amount of recombination, illustrated by the different regions of Figure 1, but will usually have a region that is near ideal. Imperfect semiconductor materials will have defects (traps) or impurities that will cause excess recombination. Measurement of the minimum ideality factor is an important indicator for material characterization that comes out of the pn junction I-V method.

In region (d) of Figure 1, the exponential increase in current stops as the diode current becomes limited by the resistances of the undepleted regions of the diode. This increase follows the equation $V_a = V_j + IR$, with the IR component becoming the dominant term. For the ideal diode of Equations 1 to 3, the reverse current approaches a constant value for applied voltages more negative than $3kT$, but as in the forward-bias case, some assumptions of the ideal diode are not valid. In the depletion region, very little recombination takes place because there are so few carriers there (very few carriers are injected into this region in reverse bias) and generation of carriers is the dominant deviation from ideality. The generation current is given by the expression

$$J_{gen} = \frac{qn_iW}{\tau_e} \quad (6)$$

where τ_e is the electron recombination time. The reverse current is then given by

$$J_R = \frac{qn_i^2\sqrt{D_P}}{N_D\sqrt{\tau_p}} + \frac{qniW}{\tau_e} \quad (7)$$

where N_D is the donor impurity concentration, τ_p is the hole recombination time, and the other terms are as defined previously.

The first term is the diffusion current and the second term is the generation current. Since the depletion layer width is weakly dependent on the reverse bias according to

$$W \propto (V_{bi} + V_j)^{\gamma} \quad (8)$$

where γ is an empirical constant with a range $\frac{1}{3} \leq \gamma \leq \frac{2}{3}$. It is seen that J_R is given by a constant term and a second term weakly dependent on reverse bias.

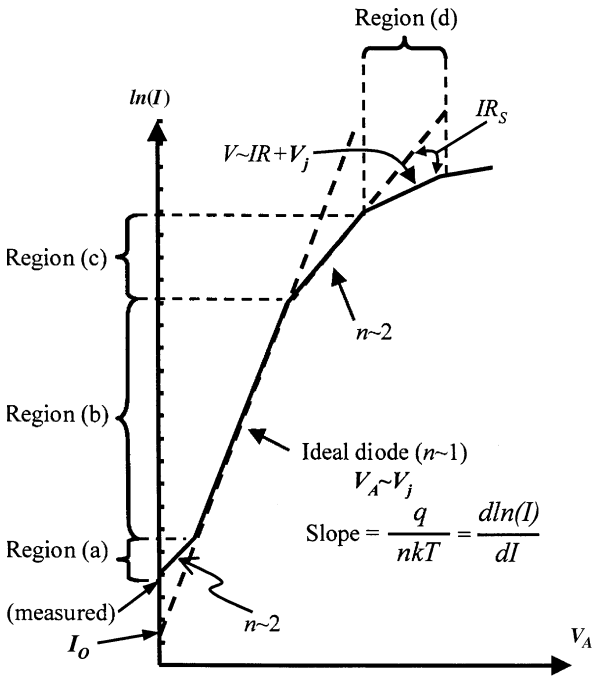


Figure 1. Deviations from the ideal forward-biased pn junction I-V characteristic.

The last subject of interest for this unit is the breakdown behavior of the diodes. There are two types of breakdown: tunneling and avalanche. Tunneling can occur with a narrow barrier (depletion region), which arises from high doping levels for most semiconductors. Applying the appropriate reverse bias matches filled valence band states with unfilled conduction band states, which satisfies the conditions required for tunneling. Tunneling is important for voltage regulator applications (Zener diodes) but is not of any real use for most materials characterization. Avalanche breakdown occurs from the creation of extra free carriers due to the electric field in the depletion region. As the depletion region is reversed biased with larger voltage, the electric field increases to the point at which some of the carriers accelerate so as to gain enough energy to create additional free carriers (termed impact ionization) when they collide with the lattice (or a defect in the lattice). With additional reverse bias, enough additional carriers can be created and gain enough energy accelerating in the field to participate in impact ionization. The term "avalanche" is used to describe this increase in current. The equation describing the breakdown voltage due to the avalanche process is given by

$$V_{\mathrm{br}} = 60\left(\frac{E_{\mathrm{g}}}{1.1}\right)^{3/2}\left(\frac{N_{\mathrm{B}}}{10^{16}}\right)^{-3/4} V \qquad (9)$$

where N_{B} is the background doping concentration and E_{g} is the band gap (Sze, 1981). For a one-sided abrupt junction, V_{br} is proportional to the inverse of the dopant concentration of the lightly doped side of the junction. Reverse breakdown voltages are a commonly measured parameter for *pn* junction semiconductor materials because breakdown voltages can be much smaller than expected if the material contains an excess of defects or if surface breakdown effects are possible.

PRACTICAL ASPECTS OF THE METHOD

The experimental aspects of *pn* junction *I-V* measurements are straightforward to implement and use routinely. The equipment required consists of a semiconductor parameter analyzer (or equivalent; refer to the Appendix), optional data collection and storage capability (typically a computer or workstation), and a probe station. The probe contacting the *n*-type electrode of the diode is grounded, and the other probe contacts the *p*-type electrode. The parameter analyzer is set up to scan the desired bias voltage range, and the data collection software is properly initialized. Unless the interest is in response to light (e.g., a photodetector), measures should be taken to exclude light or ascertain insensitivity to light. High-quality semiconductors with a small enough diode area often will have reverse currents that approach the electrical noise level of the instrumentation; therefore, some care is required in eliminating extraneous levels of electrical noise through proper grounding and proper cables.

As a first example of these measurements, an *I-V* plot for a commercial Si-rectifying diode based on a diffusion process is shown in Figure 2. Silicon *pn* junction *I-V* data

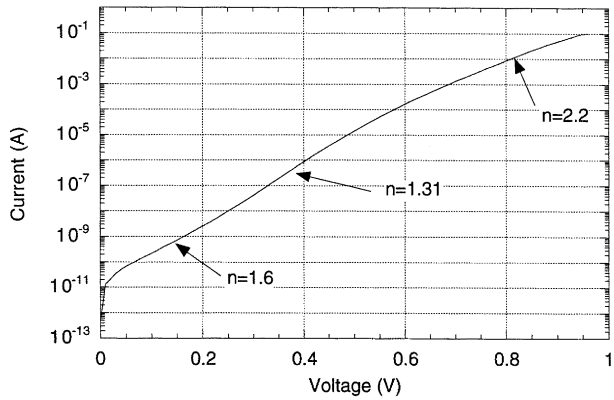


Figure 2. Current-voltage characteristics for a silicon *pn* junction diode.

are well represented in all of the reference texts, and not surprisingly, the data of the rectifying diode resembles that of Figure 1. Three different ideality factors are extracted for the regions of different logarithmic slope. For the "ideal" region, the ideality factor of 1.31 could be modeled for the particular diffusion process, if desired. This is a significant departure from 1.0 and is probably due to significant recombination from a relatively high doping level.

An example of another material system, *pn* diodes for GaAs devices, is shown in Figure 3 for both an HBT and a JFET. The JFET is ion implanted with a p^+ doping of $>10^{19}$ cm^{-3} and an *n*-channel doping of $\sim 3 \times 10^{17}$ cm^{-3}. The ideality factor $n = 1.37$. The JFET is contrasted with a base-collector diode of an HBT whose p^+n junction is grown by metal organic chemical vapor deposition (MOCVD) with a p^+ doping of 2×10^{19} cm^{-3} and an *n* doping of 8×10^{15} cm^{-3}. The ideality factor *n* is measured to be 1.01. The absolute current depends mainly on the area of the sample, which is larger for the HBT. This example illustrates the effect of different materials preparation and doping levels (higher for the JFET) in GaAs *pn* junctions. A large difference in the *n* doping level will introduce recombination, especially with the nonabrupt junctions that can be expected from ion implantation. Other differences are due to intrinsic defect levels from the two methods. The ion implantation process introduces considerable lattice damage that is incompletely removed by the implant activation process (e.g., annealing). The epitaxial growth of the HBT can produce higher quality GaAs, and the ideality factor can be used as a feedback mechanism to reduce the incorporation of point defects during the growth. Figure 3B presents data for the reverse-biased diodes. According to Equation 9, the differences in the breakdown voltages, 30 V for the HBT vs. 8 V for the JFET, can be expected due to the different doping levels in the *n* region. What is less expected is the large difference in the magnitude of the reverse current prior to breakdown. Because the current is weakly dependent on bias, as in Equations 7 and 8, the reverse current is probably due to generation in the depletion region, according to Equation 7.

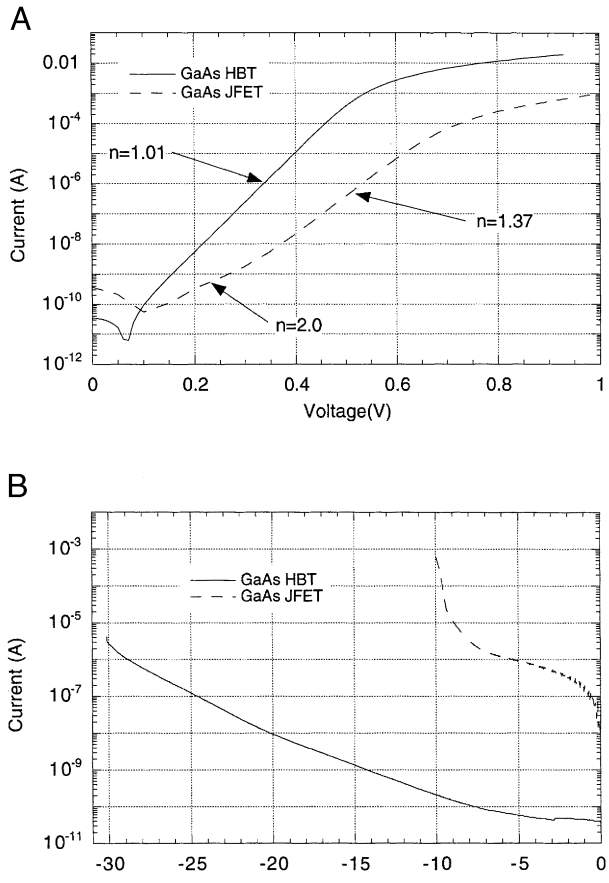


Figure 3. Current-voltage characteristics for GaAs *pn* junction in JFETs and HBTs in (**A**) forward bias and (**B**) reverse bias.

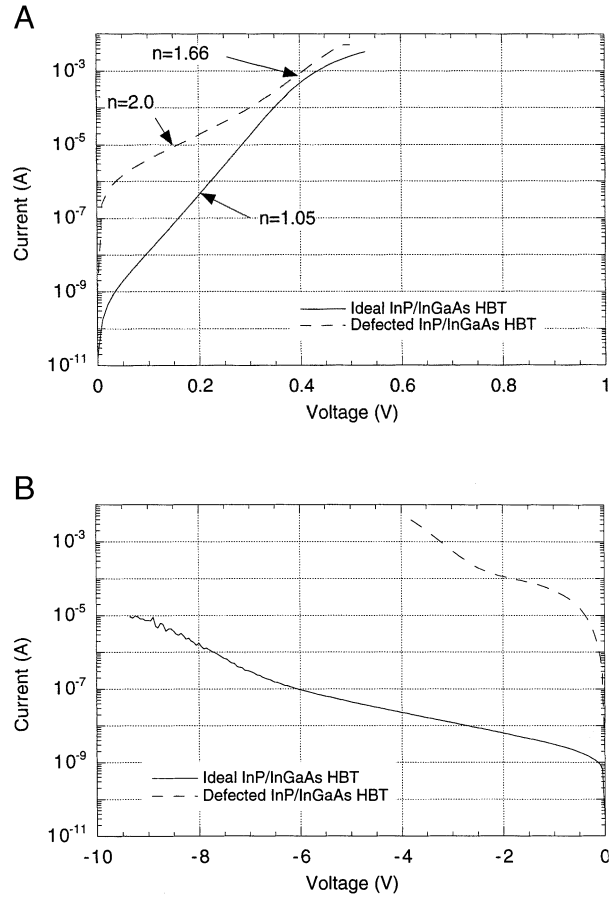


Figure 4. Current-voltage characteristics for the base-collector *pn* junction in near-ideal and InP/InGaAs HBTs in (**A**) forward bias and (**B**) reverse bias.

A third example illustrates the InGaAs material system lattice matched to an InP substrate ($E_g = 0.7$ V). Base-collector diodes are compared for near-ideal and defected HBT material. The base is Zn doped to 3×10^{18} cm^{-3}, whereas the collector is Si doped to 3×10^{16} cm^{-3}. The near-ideal sample shown in Figure 4A has a minimum ideality factor of 1.05, whereas the defected sample shows a minimum ideality factor of 1.66. The defected sample also shows very large recombination currents (second term of Equation 4) at low forward biases with an ideality factor of 2.0. Both samples were grown by MOCVD, but the defected sample contains a large concentration of point defects that are generated during the earlier growth of a degeneratively doped subcollector ($n > 1 \times 10^{19}$ cm^{-3}). Only a slight process change (to anneal out the point defects prior to the growth of the p^+ InGaAs to prevent rapid Zn diffusion) to the MOCVD growth can produce such drastic differences (Kobayashi et al., 1992). As in the GaAs case, a large increase in the reverse current of Figure 4B also accompanies the defected sample. Also, it should be noted that the lower band gap E_g of InGaAs plays a role in the lower breakdown voltages compared to GaAs.

From the examples it is seen how the ideality factor and the reverse currents can give qualitative information about the presence of defects in a *pn* junction. One can get a sense of how the equations presented above (see Principles of the Method) agree qualitatively with some aspects of a given sample. The presentation given here represents the needs of a majority of users of this method, including those without a need for detailed understanding of defects in semiconductors. Another approach, that of detailed theoretical modeling of *I-V* data, also has a role for certain well-characterized material systems but does not lend itself to a convenient summary in this unit.

SAMPLE PREPARATION

For ease of sample preparation, the *pn* junction should be made on a doped (conducting) substrate. A top contact can be made by shadow mask evaporation with a contact metal appropriate to the material (generally one that will make a good ohmic contact), and a back contact can be made with a second evaporation (no shadow mask).

The metal deposition takes place in a conventional metal evaporator. Briefly, this consists of a high-vacuum chamber (base pressure $<10^{-6}$ Torr) with the metal evaporated from a crucible with either an electron beam as the energy source or by thermal heating of the metal by

controllers appropriate for that purpose. The sample in close proximity to a shadow mask is placed in the vacuum chamber and pumped down. A shadow mask is typically a thin sheet of metal that contains openings (usually circular) through which evaporated metal can deposit onto the sample. Away from the openings, the evaporated metal deposits on the mask. After the evaporation, the sample is removed and separated from the shadow mask. A second evaporation to the backside does not need a pattern since it will serve as the sample ground during the electrical test. Temescal and CVC are two companies that market laboratory electron beam evaporators.

Both front and backside contacts should be ohmic, i.e., follow Ohm's law, $V = IR$, where R is the resistance of the contact. The contact resistance should also be sufficiently low so that the junction voltage V_j closely approximates the applied voltage. Aluminum is a common ohmic contact to heavily doped Si, and a GeAu eutectic capped with Ni forms a common ohmic contact material to GaAs after alloying to 400°C (Braslau et al., 1967).

In some cases, small-area diodes are desired or both junction contacts are required on the frontside of the sample. In such cases, photolithography will be required to form the topside contacts. Ohmic contact formation for certain material systems may then require an ohmic contact anneal. If a doped substrate is not feasible or if photolithography is desired, then a *pn* junction can be fabricated by conventional microelectronic processes for diode or transistor fabrication (Ghandi, 1983).

PROBLEMS

In general, the problems associated with this method are minor and manageable, and the method can be mastered with little trouble. Nevertheless, care must be taken to assure repeatable measurements. Poor probe contact can be one reason for poor repeatability. Probe contact problems can arise from, e.g., poor probes, improper force on the probe pad, and vibrations in the probe station. These problems must be carefully addressed and avoided if automated probe stations are to be used. Test algorithms with special open and shorted probe pads can be implemented. The effect of low light levels must be ascertained if sample measurement in room light is desired. Electrical noise can overwhelm low current levels, especially for high-quality reverse-biased samples. Setting the proper integration time of a parameter analyzer can minimize noise. For extremely low current levels (<1 pA), a shielded environment for the samples is desirable. This can be as simple as constructing a metal box around the probes with electrical feedthroughs for the cables.

It is possible for the measurement to degrade the sample due to high voltage or high current levels, and these problems are especially severe in low-band-gap materials because electrons can be sufficiently energetic to create lattice defects. Prior to setting test parameters, data should be verified by repeating measurements for candidate test parameters. In general, one can scan forward bias first. If a sample is sensitive to reverse-bias breakdown measurements, one can scan the reverse bias in the negative direction with a conservative compliance (current limit) level.

LITERATURE CITED

Braslau, N., Gunn, J. B., and Staples, J. L. 1967. Metal-semiconductor contacts for GaAs bulk effect devices. *Solid State Electron.* 10:381–383.

Ghandi, S. K. 1983. VLSI Fabrication Principles. John Wiley & Sons, New York.

Kobayashi, T., Kurishima, K., and Gösele, U. 1992. Suppression of abnormal Zn diffusion in InP/InGaAs heterojunction bipolar transistor structures. *Appl. Phys. Lett.* 62:284–285.

Shockley, W. 1949. The theory of *p-n* junctions in semiconductors and *p-n* junction transistors. *Bell Syst. Technol. J.* 28:435–489.

Shockley, W. 1950. Electrons and Holes in Semiconductors. Van Nostrand, Princeton, N.J.

Sze, S. M. 1981. Physics of Semiconductor Devices. John Wiley & Sons, New York.

KEY REFERENCES

Neudek, G. W. 1983. Modular Series on Solid State Devices, Vol. II: The PN Junction Diode. Addison-Wesley, Reading, Mass.

The modular series on solid-state devices is a set of self-contained topics suitable for the advanced undergraduate student. The pn junction diode, one of the six topics in this series, is covered thoroughly, with emphasis on clear presentation of assumptions and derivations. Where derivations are simplified, an outline of the more rigorous approach is also presented.

Shur, M. 1990. Physics of Semiconductor Devices. Prentice-Hall, Englewood Cliffs N.J.

Covers a wide range of topics similar in scope to Sze (1981) but with a more general representation suitable for the undergraduate level.

Sze, 1981. See above.

A widely used text in graduate and advanced undergraduate courses in semiconductor devices. The derivation of pn junction diodes are faily rigorous, which makes the text of most use to serious students of semiconductors.

APPENDIX: SOURCES AND CONSIDERATIONS FOR MEASUREMENT EQUIPMENT

Agilent Technologies (Palo Alto, CA) is a source for semiconductor parameter analyzers. One can also use a separate voltage source and an ammeter, such as those supplied by Keithley Instruments (Cleveland, OH). One should consider the sensitivity of the current measurement, ease of use, and interface to data collection computers and software in making a choice. A supplier of probe stations is The Micromanipulator Co. (Carson City, NV). Features to consider include the ease and accuracy of translation, automation capability, stability, and repeatability for auto probing, and interface to data collection software.

ALBERT G. BACA
Sandia National Labs
Albuquerque, New Mexico

ELECTRICAL MEASUREMENTS ON SUPERCONDUCTORS BY TRANSPORT

INTRODUCTION

In this unit, we will focus on techniques for measuring the electrical behavior of technologically important superconducting wires, tapes, and other conductors. Depending on the conditions of measurement, electrical behavior can be interpreted to obtain many technologically important quantities. This unit will also discuss the interpretation for some of the most common of these quantities.

When characterizing the electrical behavior of superconductors, it is primarily the voltage-current relationships that are measured. All other quantities are derived from these. Background conditions for the measurements, including temperature, magnetic field, and stress-strain state, vary with the specific measurement and type of information being sought.

The most common parameter to be obtained from a transport current-voltage measurement is the transport critical current (I_c), which is the maximum current that can be carried by the superconductor in the superconducting state. Often this is determined as a function of temperature and magnetic field. Critical current and the corresponding critical current density (J_c) are the most important parameters in determining the utility of a superconducting material for applications. The value of J_c equals I_c normalized by the cross-sectional area perpendicular to the direction of the current flow. This area can be defined by a variety of methods that are discussed below. Closely related to I_c is the "n value" of the superconducting-to-normal (S/N) transition. The n value is an empirical quantity without physical meaning. It is very important, however, in the design of a superconducting magnet. The n value quantifies the sharpness of the superconducting transition and thus determines the critical current margin required for operating a superconducting magnet (i.e., it determines the required ratio of operating current to critical current). Measurements of I_c, J_c, and the n value can be performed on samples that are subjected to tensile, compressive, or bending loads to ascertain the effects of stress and strain on the electrical properties, which is important for magnet and other applications.

Electrical measurements of transport current are used to determine the critical temperature (T_c) of the superconductor. This is the temperature below which the material enters the superconducting state and above which it is in the normal state. Typically, T_c is determined by monitoring the voltage at constant current while decreasing and/or increasing the temperature. By measuring T_c as a function of magnetic field using such an electrical measurement, one can also learn about the magnetic behavior of a high-temperature superconductor.

In metallic low-temperature superconductors (LTSs), electrical measurements can be used to determine the upper critical field H_{c2}, and in high-temperature superconductors (HTSs), they are used to determine the irreversibility field H_{irr}. In both cases, the magnetic field at which $I_c(T,H)$ goes to zero is the desired quantity.

Electrical measurements are used to determine the properties of superconductors beyond the superconducting state. Measurements of the resistivity versus temperature (i.e., the current-voltage relationship as a function of temperature in the normal state) are important for the design of superconducting current leads and magnets. For cables and coils, electrical measurements provide insight into the stability and quench propagation (sudden reversion to the normal state from the superconducting state) behavior. Again, this information is key for the design of a superconducting magnet.

Lastly, electrical measurements can be used to estimate the AC losses (i.e., energy dissipation resulting from an alternating current) in a superconducting wire or cable. In this case, the hysteretic behavior is used to estimate the energy dissipation.

This unit concentrates on techniques for measuring the current-voltage (I-V) relationship in superconducting wires, tapes, and cables. Important differences between LTSs and HTSs are discussed. Emphasis is placed on the I-V measurement and on the interpretation of the results in the context of the technical information sought from the particular experiment.

Competitive and Related Techniques

From the perspective of developing superconducting applications, there are no techniques that compete with electrical measurements for providing key design information. There are, however, magnetic techniques for measuring the same or similar quantities as those measured electrically. Magnetic measurements on superconductors take advantage of the diamagnetism of the superconducting state. By measuring the diamagnetic signal from a superconductor, the current flow in the superconductor is deduced. Magnetization measurements are most commonly performed in a superconducting quantum interference device (SQUID) magnetometer (see MAGNETOMETRY and THERMOMAGNETIC ANALYSIS) but are also made using more novel techniques such as a cantilever beam magnetometer (Brooks et al., 1987: Naughton et al., 1997; Rossel et al., 1998) or Hall probes (Bentzon and Vase, 1999).

Magnetization measurements of J_c are fundamentally different from electrical measurements in that magnetization is the result of a combination of intragranular and intergranular current flow, while an electrical measurement is strictly an intergranular measurement (although it is influenced by intragranular effects). Unlike transport measurements, magnetization measurements are based upon the diamagnetic behavior of type II superconductors (see definition of type I and type II superconductors in Appendix C).

While electrical measurements directly measure I_c and derive J_c from I_c divided by area, magnetization is used to determine J_c by measuring the magnetization hysteresis with applied magnetic field. The difference between the magnetization when increasing the background field and the magnetization when decreasing the magnetic field (the width of the magnetization hysteresis, ΔM) is then proportional to the intragranular J_c and the characteristic scale length of the current loops. This can be the grain size,

the sample width, or a value in between. The specific relationship between ΔM, J_c, and the characteristic scale length also depends upon the sample shape and is determined by the Bean critical state model (Bean, 1962). Details of these calculations are beyond the scope of this unit. As a result of the uncertainty in actual current flow path and the characteristic scale length, it is difficult to obtain an accurate value for J_c from a magnetization hysteresis measurement. Magnetization hysteresis does not provide the key information for evaluating a conductor for superconducting applications, but rather provides a measure of the magnetic flux pinning within the superconducting grains.

In HTS materials, magnetization hysteresis is a common method for obtaining the irreversibility line $H_{irr}(T)$. In this case, the irreversibility field at the measurement temperature is the magnetic field for which $\Delta M(H)$ approaches zero or falls below a predetermined limiting value. Due to subtleties of the technique, this tends to give a low estimate of $H_{irr}(T)$. Similarly, the upper critical field can be obtained by magnetization. In this case, when the superconductor is no long diamagnetic, the upper critical field H_{c2} has been exceeded (see definition of type II superconductor in Appendix C for a discussion of the H_{c2}). Similar to electrical transport measurements, the range of background magnetic fields available for the experiment limits the effectiveness of magnetization for measuring $H_{irr}(T)$ and $H_{c2}(T)$. In general, higher magnetic fields are available for electrical transport because a smaller measurement space is required.

Magnetization is an effective method for measuring T_c of superconductors. Because the measurement is detecting the onset of diamagnetism, magnetization must be performed in the presence of a small magnetic field. Critical temperature T_c can be measured by monitoring the magnetization during either cooling or warming. Differences due to magnetic effects can be seen when the sample is first cooled without versus with an applied H field (zero-field cooling and field cooling) in magnetization measurements during warming. In general, however, magnetization is an effective method for measuring T_c independent of grain-to-grain coupling within the material. Thus, a material can have a sharp superconducting transition by magnetization but not carry any electrical supercurrent by transport.

The alternative to electrical transport measurements for determining the AC losses in a superconductor is calorimetry. In this case, direct thermal measurements (of the deposited energy) are made and the superconducting properties are not determined.

Included below are recent advances in the measurement of the electrical properties of metallic and oxide superconducting materials. The fundamental theory underlying the measurements to be discussed has not changed dramatically over the past 25 years; the technology used to implement the measurements, however, has changed, resulting in more rapid measurement and data analysis with increased accuracy. The application of electrical measurements to the determination of stress and strain effects on superconductors is not covered in this unit. Such a measurement requires the marriage of the techniques discussed here with the application of stress and strain and is thus beyond the scope of this unit. Note, however, that for such an experiment, the principles discussed in this unit remain applicable (also see MECHANICAL TESTING). We do not discuss the measurement and control of the temperature of the sample during measurement. Historically, most measurements have been made on samples that are bath cooled in liquid He or N_2. Recently, interest in variable-temperature measurements using conduction cooling and He gas cooling has increased dramatically. Issues for measurements on samples that are not bath cooled are discussed under Problems. We also do not discuss the measurement of transport *I-V* relationships in superconducting single crystals, superconducting thin films for electronics applications, or multistrand cables. While the general principles of the measurement do not change, the practical aspects of sample preparation and contacts are significantly different.

Techniques for applying transport current ranging from 1 mA to 1 kA for the purpose of measuring a resulting voltage ranging from <1 μV to tens of millivolts are presented in this unit. The application of *I-V* measurements to the determination of the superconducting properties of wires and tape conductors are also discussed, as are some aspects of measurements in the presence of a background magnetic field from the point of view of sample mounting and support, but not from the point of view of generating the background field (see GENERATION AND MEASUREMENT OF MAGNETIC FIELDS). Lastly, some of the primary problems and pitfalls associated with electrical measurements in superconductors are discussed at some length.

The authors refer readers to document IEC 60050-815, International Electrotechnical Vocabulary, Part 815: Superconductivity produced by VAMAS (*www.iec.ch*) for the definitions of important terminology (see Internet Resources). Some key definitions from this document are repeated in Appendix C.

PRINCIPLES OF THE METHOD

The method is based on passing a known current through the sample, typically increasing stepwise or with a smooth ramp, while measuring the voltage across a section of that sample, until the voltage reaches a predetermined level—the critical voltage. The corresponding current level is considered the critical current. Alternatively, a critical resistance level can be chosen to determine the critical current. In the past, voltage-current characteristics were often recorded using *x-y* plotters, resulting in a continuous curve. Currently, mostly digital voltmeters are used, resulting in discrete pairs of voltage and current values that describe the *I-V* characteristics.

Essential Theory

Four-Point Measurement. In a four-point measurement, the voltage drop is measured using two contact points on an object, while the current through the object is introduced using two separate connections that are placed on either side of the voltage contacts. The minimum separation between the current and voltage terminals is

determined by the extent of the voltages associated with the current introduction. Four-point measurements are used to determine the voltage across a section of superconductor, as well as to determine the voltage drop across a precision resistor, commonly referred to as a shunt, that is in series with the superconductor. With the known resistance of the shunt, the current through the superconductor is calculated using Ohm's law. Despite the discretization, the data are referred to as an *I-V* curve.

Ohm's Law. A sample that obeys Ohm's law (Equation 1) is characterized by a voltage drop that is proportional to the current it carries. The ratio of voltage and current is the resistance R:

$$V = IR \tag{1}$$

Superconductors obey Ohm's law when in the normal, non-superconducting state. In the superconducting state, the voltage is a non-linear function of current. Accordingly, resistance is defined as the ratio of voltage and current at a specific current level. Resistivity (ρ) is the intrinsic material property that is related to the resistance through the geometric parameters of length (l) and cross-sectional area (A), as in Equation 2:

$$R = (\rho l)/A \tag{2}$$

Current Sharing between Superconductor and Other Material. In most practical superconductors, the superconducting material is deposited on a substrate or surrounded by a sheath of conductive, non-superconducting material, thereby creating a composite structure. In the case of a LTS multifilamentary composite, the material surrounding the superconducting filaments is referred to as the matrix. For HTS, it is referred to as either the matrix or the sheath. When the transport current approaches the critical current, the superconductor is no longer a perfect conductor, and an electric field (E) and voltage develop. That causes a small redistribution of the transport current from the superconductor to the other element(s) of the composite, and the superconducting composite is said to be in the current-sharing regime. This regime extends to transport currents above the critical current until the point where the superconductor no longer carries a significant fraction of the transport current.

Power Law Transitions. The relationship between transport current and voltage for currents in the current-sharing regime is often approximated by an empirical power law. There are several embodiments of this law, describing electric field or voltage, as in Equation 3, or resistivity as proportional to the transport current, or current density (J), to the power n:

$$E/E_0 = (J/J_0)^n \qquad V/V_0 = (I/I_0)^n \tag{3}$$

where E_0 and V_0 are arbitrary reference electric field and voltage values and J_0 and I_0 are the corresponding current density and current. When E_0 is set to the critical current criterion (E_c), J_0 corresponds to the critical current density. Similarly, if V_0 is set to the criterion voltage (V_c), I_0 corresponds to the critical current.

Properties of Materials Used

Contact Materials. Solders are used to ensure a low-resistance connection to the current leads as well as to hold the sample in place. Rosin-core (60% Sn, 40% Pb) noneutetic solder, commonly used in electronics, is easy to use and provides good bonding to Cu or Ag surfaces, as encountered on LTSs and HTSs. It is generally used to solder to superconductors.

As some HTSs can be damaged by thermal shock and prolonged heating, solders with a melting point lower than rosin-core solder (which melts around 185°C) may be preferred. Several materials such as pure indium, indium alloys, and bismuth alloys are available. However, these materials often bond only moderately and risk poor mechanical and electrical contact.

Tin-silver solders (2% to 6% Ag) have higher melting point than rosin-core solder (e.g., 240°C for $SnAg_3$). They require the use of a separate flux, but their lower resistivity at cryogenic temperatures often justifies their use with high-current cabled conductors.

For low-current measurements (e.g., T_c measurements), solid indium can also be used. In this case, small indium "pads" are pressed onto the sample and the voltage and current wires are pressed into the pads. This is not a mechanically robust contact, but it is the least invasive to the sample and can be effective for small samples or samples of random geometry. This approach is not effective for large currents.

Empirical, Phenomenological, and Heuristic Aspects

Definition of Zero Voltage. From magnetization and trapped flux measurements, it is known that superconductors have a zero-resistance state, a state in which there is a DC current, but no dissipation and therefore no electric field or voltage. When measuring the voltage across a section of superconductor, however, a suitable voltmeter will rarely, if ever, indicate "0 V" when used at its maximum precision. Both random noise and unwanted temperature-dependent DC voltages keep the meter from indicating zero, even if the superconductor is in a zero-resistance state. Direct-current offsets on the order of 1 μV are common. The random noise level determines the lowest voltage rise that can be measured. Therefore, the current value corresponding to the onset of a voltage rise and the upper limit of the zero-voltage regime are dependent on the sensitivity of the instrumentation (see Practical Aspects of the Method).

Critical Current Criteria Options. The critical current of a composite superconductor is an essential parameter to describe the current-carrying capacity. It does not correspond to a single physical quantity, however, nor is it the maximum current that can flow through the conductor. Since conductors can be operated isothermally at current levels significantly above the onset of a voltage rise, a

criterion is required to determine I_c. The minimum detectable voltage is not a suitable criterion to characterize the current-carrying capability because it is instrumentation dependent. More suitable and commonly used are electric field criteria of 10^{-4} V/m (1 μV/cm) and 10^{-5} V/m (0.1 μV/cm). In these cases, the electric field is defined as the voltage divided by the distance between the voltage taps. The dimensions of a sample are better taken into account by a resistivity criterion, which enables the comparison of wires with different cross-sections. Critical current criteria are usually at resistivity levels between 10^{-11} and 10^{-14} Ω-m. Critical current can also be quoted at a specific power density value, i.e., the product of electric field and current density. Although not widely used, this is useful for applications that are sensitive to the heat balance. Other criteria exist, such as the current at which a quench to the normal state occurs or criteria based on the shape of the *I-V* curve. These are used infrequently as they are either difficult to determine with sufficient precision or dependent on circumstances during measurement. Regardless of which criterion is used, it must be clearly described when I_c is reported.

H_{c2} by Extrapolation and H_{irr}. The upper critical field, especially for temperatures that are a small fraction of T_c, is often higher than the maximum magnetic field available to the experimentalist. Thus, H_{c2} often cannot be measured directly. In LTSs, however, J_c, H, and H_{c2} are related through the magnetic flux pinning force density $F_p = J_c \times H$. According to Kramer's scaling law, this is proportional to $h^p(1-h)^q$, where $h = (H_{applied}/H_{c2})$ and p and q are constants (Kramer, 1973). When the $J_c(H)$ data are plotted as $J_c^{1/2}(\mu_0 H)^{1/4}$ versus $\mu_0 H$, known as the Kramer plot, a linear extrapolation toward higher fields intercepts the horizontal axis at H_{c2} (μ_0 is the vacuum permeability). In HTSs, the upper critical field can similarly be determined by electrical measurement of I_c to find the boundary between zero and finite I_c at a given temperature. This quantity is associated with the irreversibility field H_{irr}, which is more commonly obtained. Note that H_{c2} and H_{irr} are fundamentally different parameters, representing different physics: H_{c2} is the boundary between the superconducting and normal states; H_{irr} is the boundary between lossless and lossy current flow due to vortex motion. Crossing the H_{irr} boundary does not take the material out of the superconducting state. The field H_{irr} is also defined as the onset of a finite I_c but is derived from either DC magnetization, AC magnetization, or resistivity measurements. The latter is an electrical measurement. The quantity H_{irr} is determined from a set of *I-V* curves obtained at constant temperature for a range of magnetic fields and is defined as the point where the curvature changes from positive to negative. However, magnetic measurements are preferred for H_{irr}, as it is an intragranular property and thus more suitably measured with an intragranular technique. The electrical measurement (resistivity curvature) is suitable primarily when grain boundary effects are not a concern, i.e., in measurements on single crystals. An example of this is shown in Figure 1, where resistivity measurements as a function of magnetic field and temperature result in a phase

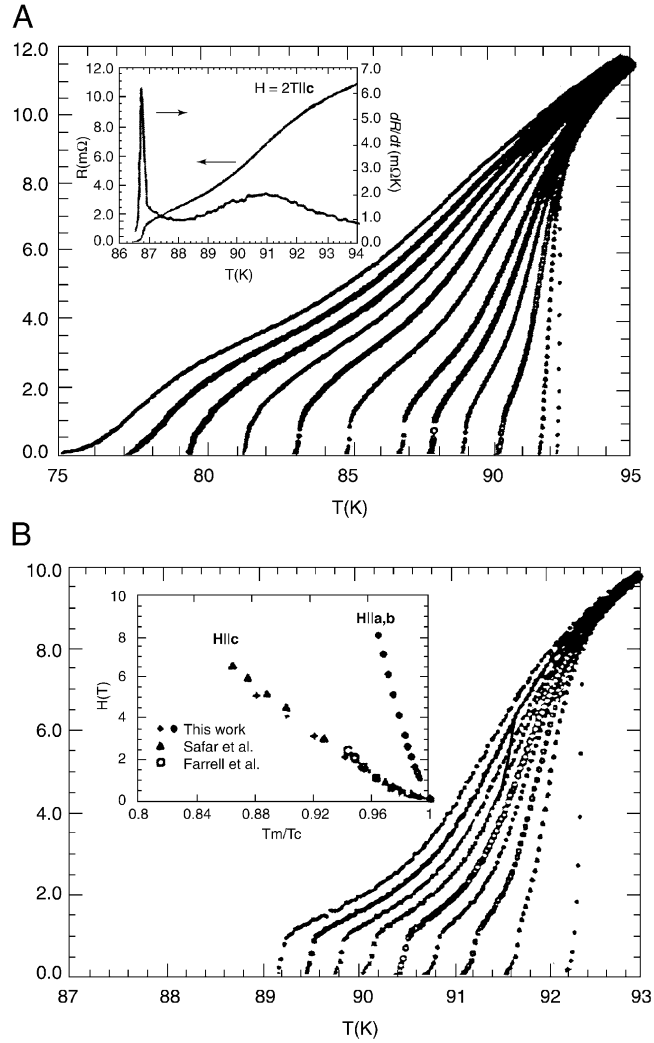


Figure 1. (**A**) Resistive transition in magnetic fields from 0 to 8 T for **H||c** in an untwined $YBa_2Cu_3O_{7-\delta}$ crystal. Inset: Determination of T_m from the inflection peak of dR/dT for $H = 2$ T. (**B**) Resistive transition in magnetic fields of 0 to 8 T for **H||(a,b)**. Inset: Phase diagram of the melting transition for **H||c** and **H||(a,b)** (from Kwok et al., 1992). Note: **a**, **b**, and **c** refer to the crystallographic directions in the $YBa_2Ca_3O_{7-\delta}$ structure.

diagram of the magnetic behavior (Kwok et al., 1992). Similarly, these three methods are used to determine the irreversibility temperature T_{irr}. The line of irreversibility (zero current in a transport measurement or "zero" width of the magnetization hysteresis loop in a magnetization curve) points in an *H-T* plot is known as the irreversibility line. Note also that for HTS materials it is usually not possible to measure H_{c2} at low temperature because the values are higher than the highest magnetic fields that can presently be generated. Thus it is only estimated by extrapolation.

Subtle Assumptions and Approximations

Current Transfer and Transfer Length. When the current is introduced into a superconducting composite, it is typically introduced into the non-superconducting material,

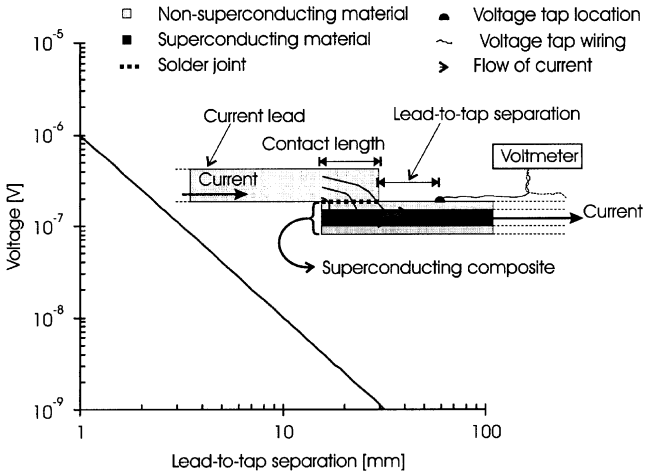


Figure 2. Cross-section of the connection between a current lead and a composite superconductor; all current transfers from the lead to the composite superconductor, and most of the current to the superconducting material within the composite, within the contact angle. A small fraction of the current will transfer from the non-superconducting part of the composite to the superconducting material outside the contact area. The graph represents an example of the unwanted transfer voltage measured versus the lead-to-tap separation for an arbitrary but realistic case.

since that material usually forms the outer surface of the conductor and is easier to make contacts (e.g., solder joints) to. The transfer of current to the superconductor extends beyond the contact area with the current leads (see Fig. 2). The non-superconductor current and its associated resistive voltage decrease quickly with distance from the contact area. It is assumed that the transfer is completed within the distance between the current contact and the voltage contact to the extent that the contact and current transfer area have no measurable effect on the current-voltage characteristic that is measured. See Practical Aspects of the Method for further discussion and references.

Current-carrying Area for Converting I_c to J_c, J_{ct}, J_e, and J_{cm}. The conversion from critical current to critical current density (J_c) is often a source of confusion and inaccuracy. One important source of confusion arises because the area used to calculate the current density can be merely the superconductor (which is still commonplace in HTSs), the nonstabilizer part of the conductor (most common in LTSs), or the entire cross-section. These conventions are all found in the literature. The latter is often, but not exclusively, referred to as the engineering current density J_e. For more detailed discussion regarding the sample cross-sectional area, see Data Analysis and Initial Interpretation.

The transport critical current density J_{ct} is sometimes used to denote J_c obtained by a resistivity or voltage measurement, as opposed to the magnetization J_c (J_{cm}) when magnetization measurements in combination with theoretical models are used. This distinction is particularly important when both electrical and magnetic measurements are used. A lack of distinction in the literature can be a source of confusion, as J_{ct} is the significant parameter for most applications, while J_{cm} is almost always the greater value. Measurement techniques for determining J_{cm} are discussed at length in MAGNETOMETRY.

PRACTICAL ASPECTS OF THE METHOD

Properties of Instruments to Generate Data/Signals

Current Supply. A generic schematic of equipment for electrical measurements is shown in Figure 3. A single-quadrant DC supply is required with a maximum current of at least the critical current. Concerning current stability, it is desirable to have a maximum periodic and random deviation of less than 2% within the bandwidth of 10 Hz to 10 MHz. Control of the current can be implemented in a number of ways, depending on the power supply design. Often supplies offer several options for control of the current. These options include software control, through an instrumentation bus or serial connection with a computer, and manual control using potentiometers or numeric keypads usually on the front panel of the supply. A third category entails external programming with a separately controlled resistance, voltage, or current source, usually to be wired to the back of the supply. This is also known as remote programming.

The current is changed as a function of time using either the step-and-hold approach or a constant ramp rate. A maximum rate corresponding to a 30-s rise time from 0 A to I_c has been suggested as a standard (VAMAS). A higher rate may be required if heat dissipation in the sample or sample support is an overriding issue, which may be the case for HTS-coated technologies. In the worst

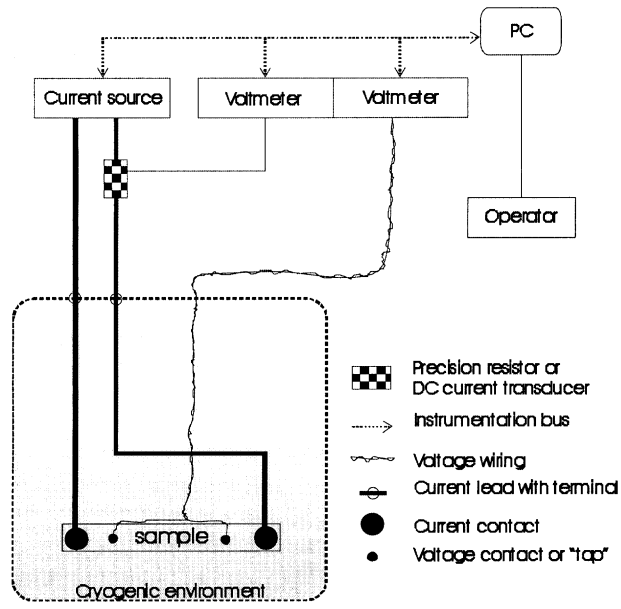


Figure 3. Equipment for electrical characterization of superconductors. No sample support structure (sample holder) is shown, which will generally support the sample, current leads, and voltage wiring in the cryogenic environment. Many variations on this schematic are possible.

cases, measurement times on the order of milliseconds may be required.

Voltmeter. The required accuracy of the voltage measurement depends on the sample length between the voltage connections ("taps") and the I_c criterion used. Typically, tap lengths are on the order of 0.01 to 0.05 m, and a common criterion is 10^{-4} V/m, often expressed as 1 μV/cm. Other criteria tend to lead to a lower critical voltage and lower I_c. Since a minimum precision of ~10% of the critical voltage is desirable, a high-precision voltmeter is required. A 6 1/2-digit voltmeter might have 10^{-7} V resolution but is likely to require extensive averaging to obtain a similar accuracy, resulting in a slow measurement. More accurate single-unit voltmeters or a combination of filters, preamplifiers, and voltmeters is recommended. Either system will be referred to, in the remainder of this unit, as a voltmeter.

An analog-to-digital-converter (ADC) forms the core of every voltage measurement system. The larger the number of bits specified for the ADC, the larger the number of discrete voltage levels that can be detected. This, together with the range of preamplifications available, determines the range and resolution of the voltmeter. The last step in the data acquisition process is to transfer the voltage value as an ASCII string to the software recording or controlling the measurement.

These elements can be implemented as a single-unit voltmeter connected to the computer and controlled through a driver (see Method Automation). Single-unit meters tend to offer convenience, a large range, and high sensitivity. Another approach is to separate the amplifier from the ADC. The latter can be a stand-alone unit or mounted on a card inside a computer. Either approach can be used to measure multiple voltages through the use of a multiplexer, which can be a separate unit or incorporated in the ADC unit or single-unit voltmeter. When each signal is amplified before the multiplexer, the option tends to offer the highest data acquisition speed, since the time required for analog-to-digital conversion with a given relative accuracy decreases significantly as the value of the voltage increases. Voltages below 1 μV may take on the order of one to multiple power-line cycles (16.7 or 20 ms per cycle) to convert, whereas one or a few milliseconds may suffice for signals in the millivolt range. A computer-mounted ADC is likely to offer the fastest transfer of measured voltage values to the software and further improve the data acquisition speed. Data acquisition speed is not the limiting factor in most measurements, however, and considerations of accuracy and cost determine the choice of voltmeter.

Sample Shape

The ideal shapes for electrical transport current measurements are a uniform, homogeneous, cylindrical wire or a rectangle. However, the geometry is beyond the control of the measurement technology and dictated by the processing factors that produce the sample. In electronics applications, patterning of the thin film is typical. This is not standard, however, on wire and tape-based conductors for large-scale applications.

Contacts

Current contact length. The measurement current is passed in and out of the conductor using two solder connections to current leads that are a permanent part of the measurement setup (see Figs. 2 and 3). As a guideline, the length of the contact area for tapes should be at least as long as the conductor width. A longer length will reduce the resistance, and therefore heating, in the contact. Clamping the conductor to the current leads is possible too, but this results in less reliable contacts with higher contact resistance and possible sample damage. For twisted conductors (either conductors with twisted filaments or cables with twisted or braided strands), a contact length of one twist pitch is recommended if the current is not introduced homogeneously around the circumference. Integer multiples of the pitch length are necessary if longer joints are used. This minimizes unevenness in the current distribution within the conductor due to unequal resistances to the leads.

Current Transfer Length. When the current is introduced to the outer surface of the conductor, it will flow through the sheath and/or matrix to the superconductor. The resulting distribution of current extends beyond the contact area itself, resulting in a measurable voltage drop near the contact area. These resistive voltages decrease quickly with the distance to the contact area and are dependent on the current density, matrix resistivity, and conductor geometry. Models have been developed and verified to calculate these voltages (Ekin, 1978; Ekin et al., 1978; Wilson, 1983b; Jones et al., 1989). As a general guideline for tape conductors, a minimum separation should be used between current lead and voltage tap equivalent to the tape width. The minimum separation for round wire is a multiple of the diameter. The required separation strongly depends on the measurement sensitivity and the resistivity of the non-superconducting material(s). It can be as high as two orders of magnitude larger than the sample diameter. A dedicated experiment may be required to determine the minimum separation to avoid measuring the current transfer voltage. Current transfer within a conductor (between filaments) can occur when the temperature, field strength, or field orientation changes abruptly over a short length of sample. When the voltage taps are too close to such a transition, extra noise or spurious, unstable voltages can result.

Voltage Tap Placement. The optimum length between voltage taps is a balance between the increased sensitivity that accompanies a long tap length and the need to have sufficient separation between the voltage taps and the current leads to avoid measuring the voltages associated with the current transfer.The assumption that iso-voltage planes are perpendicular to the long axis of the conductor is valid. Therefore, for tape conductors, the voltage taps can be placed either somewhere on the flat side or on the edge. The voltage tap length is measured along the axis of the conductor between the centers of the wires where they exit the solder.

Sample Handling and Soldering/Making Contacts. Many superconductors are sensitive to strain and, for some conductors, even small deformations can irreversibly affect the superconducting properties. All samples should be handled carefully, and measurement procedures should minimize handling. For ceramic superconductors, thermal shock and the length of time taken to solder leads and voltage taps must be minimized. The use of solders with a low melting point (and an appropriate soldering iron) or preheating the sample and the current leads can help prevent degradation when normal soldering procedures affect the superconducting properties. Often, especially for strain-sensitive conductors, it is not possible to unsolder a sample without damaging it.

Silver paste is also used to make electrical connections and avoids the heating issues associated with soldering. Drawbacks include a limited mechanical strength and creating a risk of unreliable connections. Electrical and mechanical properties vary strongly, depending on the manufacturer and type of paste.

Sample Holder. Typically, the sample is mounted on a sample holder that incorporates the current leads. Sample holders are usually custom designed and show broad variations driven by sample characteristics and experimental conditions. The connections to the leads keep the sample in place. As a result of the difference in thermal expansion between the superconductor and the holder, the sample is subjected to strain upon cooling from the temperature at which the connections are made to the cryogenic temperature at which the superconducting properties are measured. Thus, the holder material should be chosen based in part on the tolerable strain state of the superconductor at the measurement temperature. When a conductive sample holder is used, the current running in the holder should be subtracted from the transport current to determine the critical current. This calculation requires knowledge of the voltage-current characteristics of the section of the sample holder between the voltage tap positions, typically obtained from a dedicated measurement without a sample mounted.

Cooling Options and Procedures

Bath Cooling. The most thermally stable measurement environment is provided by a bath of a cryogenic liquid, such as helium, hydrogen, neon, oxygen, or nitrogen. Heat developed in submerged parts of the experiments will cause evaporation but no significant temperature variations as long as the heat flux is below that which would result in boiling ($\sim$0.35 W/cm^2 for liquid helium at atmospheric pressure). There are significant safety aspects associated with cryogenic liquids with which each user should be thoroughly familiarized. Care should be taken to avoid contamination of the cryogen, which will affect its boiling point. Both the liquid levels above the sample and the ambient pressure may have a significant effect on temperature, and thereby on the properties of the superconductor. To the extent that the properties of the cryogen and the experimental conditions allow, deliberate variations in pressure can be used to vary the bath temperature.

Liquid cryogens, especially liquid helium, are usually contained in a cryostat. These are typically double-walled stainless steel vessels where a vacuum and often radiation shields provide insulation between the inner and outer walls. Cryostats are usually equipped with safety and access valves and pressure and level gauges. Simpler double-walled containers made of glass are often referred to as dewars. Containers that rely on foam as insulator are used for short-term storage of liquid nitrogen. The generation or purchase of cryogenic liquids is beyond the scope of this unit.

Gas Flow. A larger range of temperatures is available when using a gas flow.There is a heat balance between the cooling effect of the gas flow and the heat conduction through the sample support plus thermal radiation from surrounding surfaces. Heaters in the gas flow or on the sample support and/or control of the flow rate is used to maintain a constant temperature and minimize the thermal gradients across the sample. As the heat transfer to the cooling medium is more limited than in the case of bath cooling, the tolerance for heat dissipation is reduced. The gas can be obtained from a liquid bath through a combination of transfer lines or valves and heaters or directly from a liquifier.

Conduction. Conduction cooling has gained in popularity with improvements in cryocooler technology and increased demand for "cryogen-free" systems. The sample, leads, and sample holder are placed in a vacuum, where thermal radiation and heat dissipation are balanced by conduction cooling through a connection to the cryocooler. Care is required to limit thermal gradients inherent in conduction-cooled systems, but electrical isolation should be maintained. Cryocoolers, for example of the Gifford-McMahon type or pulse tubes, can provide the cooling power. Both rely on compression and expansion of a gas. Both are commercially available.

Cooling Procedures. The sample can be cooled by mounting it on the holder, placing it in a cryostat, followed by a cooldown of the entire setup. Alternatively, it can be introduced to a cold cryostat. In either case, thermal shock and contamination of the cryostat are to be avoided. Depending on their porosity, ceramic superconductors tend to be more sensitive to thermal shock than metallic superconductors. For the former, unless specific cooling rate specifications are available, a cooling rate of no more than 1 K/s is a useful guideline. When a cryostat will be cooled down after positioning of the sample, pumping and purging with gas or flushing with gas will remove the air and the water vapor. If the cryostat is already cold and the sample support structure is cylindrical, sliding through an O-ring seal will allow slow introduction of the sample to cryogenic temperatures. The use of a load-lock, a chamber between two vacuum-tight valves and connected to a pump, allows complete separation between the cryogenic and ambient atmosphere. Contamination of the cryogenic liquid may affect its boiling point as mentioned

above, and frozen gases or water may block passages and mechanisms. The tolerance for contamination strongly depends on the properties of the equipment used and may or may not be a significant factor.

Current Ramp Rate and Shape

This section describes three common ways to vary the current as a function of time. Other approaches are often a combination or variation of the methods described here. Upon reaching the desired maximum current, the current is usually reduced to 0 A with an abrupt change, a linear ramp, or the same current-time profile used to increase it. The three approaches are illustrated in Figure 4.

Staircase Step. The most accurate measurements can be performed when the current is changed stepwise, allowing the current through and voltage across the sample to stabilize before each measurement of voltage and current. The time for each measurement is determined by the desired accuracy combined with the properties of the equipment used. This approach can be time consuming, resulting in a relatively high ohmic dissipation and possibly complicating temperature stability. In those cases, a compromise between accuracy and measurement time is required. The staircase step approach can alternately be optimized by tailoring the size of the current step such that it is small during the transition and larger where the derivative of the *I-V* curve changes little. The ramp rate between levels should be limited to minimize transient effects such as current overshoot and AC loss effects in the superconductor. Often, especially with relatively small steps, the maximum rate practically possible is low enough.

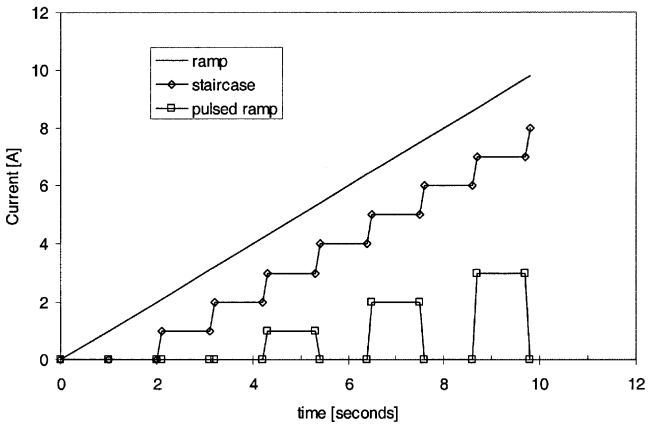


Figure 4. Three common current ramp options in a current-vs.time graph for increasing current and with a limited number of steps: (1) a 1-A/s continuous ramp; (2) a staircase pattern with 1-A steps, 1-s dwell per step, and 10-A/s ramp rate between steps; and (3) a pulsed ramp with 1-s dwell and 10-A/s ramp rate between steps. It is noted that there usually is a delay between a command to the current source and a change in output. This delay can range from less than 0.1 s to on the order of 1 s, depending on current source model and configuration. In reality, changes in ramp rate are not as sharp as depicted but either are more gradual or result in overshoot.

Pulsed Step. To reduce the average power dissipation but maintain the advantages of a staircase step approach, the pulsed-step method is used. In this case, the current is reduced to zero between each step in the pulsed-step approach. Since larger steps are taken, the issue of ramp rate between levels requires more careful consideration.

Continuous-current Ramp Measurements. Another approach is the continuous-current ramp measurement, where the current is changed using a linear ramp and the voltage and current measurements are synchronized. This approach should be taken if control of the sample temperature is poor or if the contact resistance is high, requiring quick measurements to minimize heating, or if the stabilization time is excessive. The ramp rate can vary from 0.1 to 100 A/s. A discussion of problems that result from this approach is given under Problems.

Electrical and Magnetic Phenomena

Superconducting properties vary with applied magnetic fields and may depend on the history of the applied field and, to a lesser extent, the rate of field change. This does not affect the individual *I-V* characterization, but the sequence of measurements and field changes should be carefully considered. The same argument holds for the temperature dependence of the superconducting properties. For HTSs, in particular, the magnetization state and electrical properties strongly depend on whether the superconducting state was reached through in-field cooling or zero-field cooling.

A stabilization period after a field change may be required to allow induced eddy currents to dissipate. As a general guideline, the spatial variation of the applied magnetic field should be below 1% over the length of the sample.

Determining Maximum Measurement Current

Factors that determine the maximum current in a measurement, other than the current supply limit, include the voltage range required for curve fitting in the data analysis, maximum tolerable power dissipation in the sample to maintain thermal control and/or avoid sample damage, and maximum tolerable Lorentz force from either the self-field or an applied magnetic field. It should also be noted that the value of the maximum current affects the magnetization state of the superconductor and may have a small effect on the critical current in subsequent repetitions of the *I-V* measurement.

Instrumentation and Data Acquisition

Usually a personal computer (PC) is the central hub in the measurement system, interfacing with the user, the current source, and the voltmeter. As a minimum, the sample and shunt voltages are read through software on the PC, the latter converted to current, and stored in a file. Additional elements can involve instruments that control or monitor the sample temperature, applied magnetic fields, mechanical loads, and other relevant parameters.

Multiplexers, also known as scanners, can be used to measure different signals with a single meter. Computer and instruments are connected to a bus system. Currently, the General Purpose Instrument(ation) Bus (GPIB) operating under the IEEE 488.2 standard is most commonly used, although serial connections under the RS 232 or RS 442 standard are not uncommon, particularly for power supplies. Converters between the two standards are commercially available.

Instrumentation and data acquisition are areas that are likely to see significant changes as new PC-based communications standards, equipment, and network protocols develop.

Troubleshooting

A quick means to short the current leads is convenient when troubleshooting or testing the operation of the current supply, as it avoids the risk of damaging the sample by overcurrent. When the voltmeter is connected to the sample at zero current, it should read a constant voltage of no more than $\sim 1\ \mu V$. A drifting voltage on the order of ~ 1 mV or more indicates an open circuit; drifting voltages well below 1 mV indicate the presence of parasitic thermocouples in the circuit. A more elaborate discussion of thermocouples and other potential problems follows under Problems. As it can be difficult to distinguish between the zero-voltage state of a superconductor carrying current and no signal at all, one must be concerned with the risk of damaging the sample by overcurrent. A low-current *I-V* measurement at a temperature above T_c is recommended (e.g., at room temperature) such that results can be used to verify proper operation.

Generic Protocol

A generic protocol for performing an electrical characterization of a superconductor follows:

1. Verify cryogenic liquid level and condition of cryogenic equipment.
2. Cut sample to size; prepare current lead and voltage tap connection areas. Measure dimensions. If the sample is easily damaged, the determination of dimensions can be postponed until after the measurement.
3. Verify wiring and leads of sample support by checking for top-to-bottom continuity and absence of shorts.
4. Mount the sample by placing on the support structure and connecting to current leads and voltage tap wires. Verify connections to the sample by resistance measurements from the warm end of the support structure or performing a room temperature *I-V* characterization.
5. Complete all connections.
6. Cool the sample at a controlled rate and bring the sample in the desired condition (e.g., magnetic field magnitude and orientation, stress-strain state).
7. Perform the *I-V* characterization, display, and save the data. At least a preliminary analysis is performed to verify that the data are valid and noise levels are acceptable.
8. Warm the sample at a controlled rate.
9. Dismount the sample.

When performing multiple measurements at varying magnetic fields, temperatures, and other conditions, history effects are important and usually determine the order of temperature changes, field changes (or any other condition), and measurements.

METHOD AUTOMATION

At the most basic level, only the data acquisition and storage are automated. The current is controlled through separate means, and the data are first analyzed after the measurement. In a more fully automated system, the control of the current source and the logic for how and when to change the current is incorporated in the software. This applies specifically to the determination of the maximum current, i.e., when to start reducing the current to zero. Real-time analysis is therefore a key element to automation. Data analysis during or immediately after a measurement will help assess if the obtained data are adequate. It is often pertinent to make that assessment immediately, as there may not be another opportunity to repeat the measurement.

Software to Control the Current Source, Read the Voltmeters, and File and Display the Raw Data

Drivers (software that handles the interaction with specific instruments) are often supplied with the instrument or are available through the National Instruments (NI; *http://www.ni.com*) website, thereby reducing the need for extensive programming. Additional software, usually developed in-house, can provide a graphic interface with the user for control of the instruments as well as for data storage, presentation, manipulation, and interpretation. A wide range of programming languages can be used.

Quench Detection

Quench detection is ideally performed by a circuit independent of the measurement system. When this circuitry determines that the voltage across the sample exceeds a threshold value, it assumes control of the current supply and reduces the transport current to a safe level (typically 0 A) before excessive heating occurs. Options for quench protection are reducing the output of the current supply, circuit breakers to interrupt the current, or a passive circuit using diodes and dump resistors. A shunt of appropriate dimensions and resistance can also be mounted in parallel with and close to the sample and connected to the same current leads. The latter solution may require a correction for the current through the shunt and limit the range of voltages for which the sample current can be accurately determined.

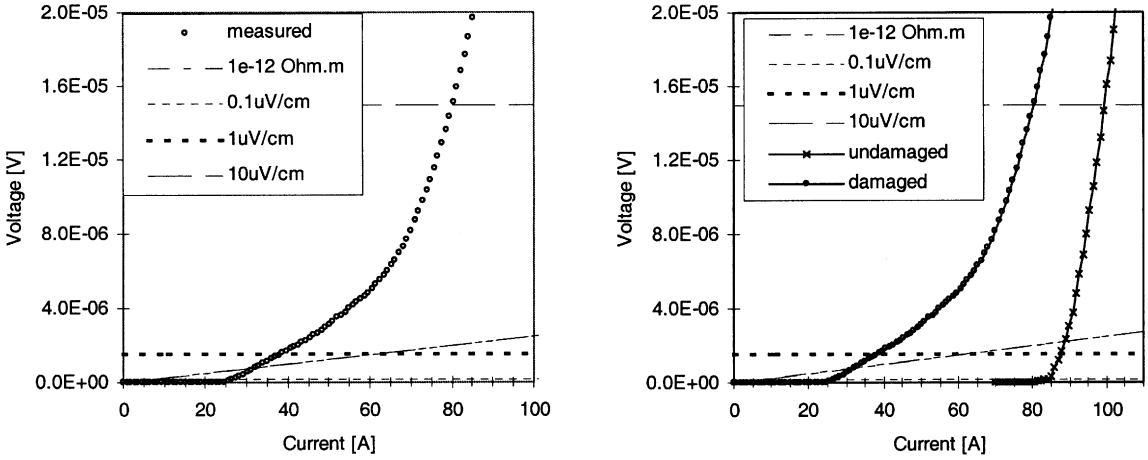


Figure 5. *I-V* characteristic of a damaged sample and a similar undamaged sample. Various critical current criteria are shown.

DATA ANALYSIS AND INITIAL INTERPRETATION

Determining T_c

The value of T_c can be determined electrically either indirectly by extrapolating $I_c(T)$ or directly from measuring the voltage (typically reported as resistance) versus temperature, although the latter is more typical, easier, and more accurate. Measurement of T_c and interpretation of the voltage-temperature data are discussed at length in the VAMAS document on Superconductivity, Part X. In particular, it specifies that T_c is the temperature at which the specified resistivity is 1×10^{-8} Ω-m in the self-field.

In the literature one often reads of the onset temperature and the zero-resistance temperature. In this case, the former refers to a sharp change in the slope of the voltage-temperature curve (observed while the temperature is decreasing), and the latter refers to a resistance or voltage criterion such as the 1×10^{-8} Ω-m established by VAMAS.

Determining I_c and J_c

Selecting the Appropriate Criterion. The appropriate criterion depends on the intended application of the conductor or the intended use of the data (see Principles of the Method). All criteria are somewhat arbitrary since none corresponds to an intrinsic physical quantity or parameter. All provide values between zero voltage and the completely resistive state. Thus, one can identify clear lower and upper bounds (see Figs. 5 and 6). The choice of criterion will affect the value of I_c significantly, especially for conductors with low n values. Comparing measurements on different conductors requires adoption of a single criterion. For example, VAMAS used 10^{-4} V/m. Multiple criteria can be applied to the same data, and the spread in the values, or the absence thereof, is indicative of the sharpness of the transition and implicitly of the quality of the conductor.

The Bseline. The starting point of *I-V* data analysis is the determination of the offset voltage. This offset is a measurement artifact and is subtracted from the *I-V* data before applying the I_c criteria or other data analysis. The simplest approach is to use the voltage corresponding to the lowest current value, usually near 0 A. A more accurate value is the intercept of a line fitting the *I-V* data up to the onset of a voltage rise. This segment of the *I-V* curve is the baseline. In case of a true zero-voltage regime in

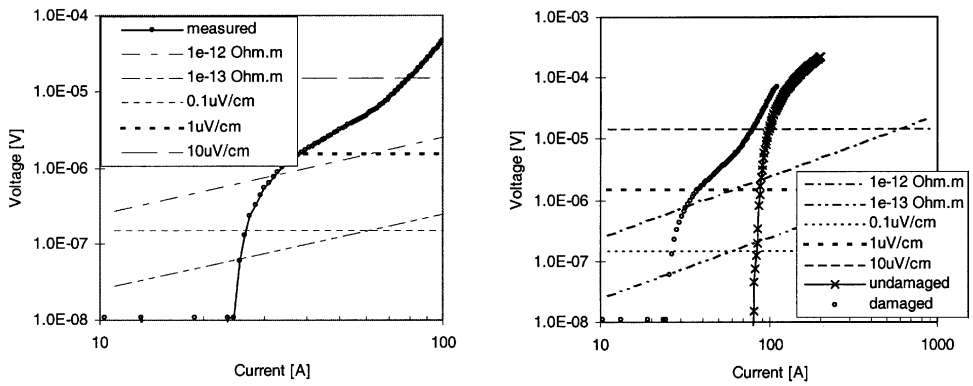


Figure 6. *I-V* characteristic of a damaged sample and a similar undamaged sample in a log-log representation. Various critical current criteria are shown.

the *I-V* characteristic, this value corresponds to the average of the voltage value over this range. There are several reasons why the baseline may have a slope beyond the result of random noise in the data. These reasons include resistive behavior from minor damage or other imperfections in the conductor and resistive behavior from current transfer voltages near the current lead connections. The latter may be subtracted from the data, since it does not represent the properties of the superconductor within the voltage taps, but this is somewhat controversial as it may be difficult to prove the source of the resistive behavior. A suggested guideline is to only subtract baseline slopes that correspond to less than half the critical voltage at I_c and otherwise reject the measurement.

If the *I-V* data were obtained using the current ramp method, the inductive voltage component will automatically be corrected for when using a baseline fit. The first current values might be measured before the ramp rate, and therefore the inductive voltage reaches a stable value and should in that case not be included in the baseline. (Note that the inductive voltage is the result of the self-inductance of the sample.)

Defining the Cross-sectional Area. When reporting J_c, the area used and how the area was measured need to be described in detail. For HTSs, the areas involved are often hard to determine with high accuracy and may vary along the length of the conductor. Tape conductors, for example, tend to have features similar to rounded, but not circular, edges, and in HTS tapes, often the "flat" side is not truly flat after the final heat treatment. Maximum width and maximum thickness are relatively easy to measure, but their product is an overestimation of the cross-sectional area. In addition, the internal geometry of stabilizer and nonstabilizer materials tends to be complex. Digital analysis of cross-sectional pictures in combination with repeated measurements using calipers and micrometers minimizes the errors and uncertainties in areas and therefore J_c and J_e.

Software (Algorithms) to Analyze the Data

A typical electrical measurement on a superconductor results in a number of pairs of corresponding voltage and current values. The main superconducting parameter, I_c, is defined by the intersection of curves describing electrical properties and criteria. Therefore, the pairs of *I-V* data have to be converted, preferably by software, to a curve by either a curve-fit or linear interpolation. Linear interpolation may be sufficiently accurate if the interval between current values is small, the noise level low, or the transition gradual. Linear interpolation in a log-log graph provides a straight *I-V* curve and accurate intersections if the *I-V* data follow the power law over the range of interest. The slope of the graph is the n value. When the data do not follow the power law (described by Equation 3) or if the noise level requires some smoothing, a polynomial fit can be used. Ninth- or tenth-order polynomials often provide a fit that follows the details of the characteristics while filtering out random noise, assuming that more than ~15 data points are available.

Determining *n* Value

Selecting an Algorithm to Determine n. Perhaps the most common method to obtain an n value of Equation 3 is through determination of the slope of the log(V)-log(I) graph in the region of interest. Alternatively, the n value can be determined from a direct fit to the *I-V* data. In the latter case, minimizing the sum of the absolute errors between measurement and fit may yield better results than optimizing for the sum of relative errors. Methods using numerical differentiation exist as well. The power law can be differentiated and rewritten to yield $n = (I/V)(dV/dI)$. The n value is thus interpreted as the ratio of resistance and differential resistance at a given current level. Discrete differentials based on measured data are used to calculate the derivative. Other formulations for the n value exist using first and second derivatives of the *I-V* characteristic (Warnes and Larbalestier, 1986). All calculations require an accurate determination of the voltage offset.

Selecting the Appropriate Data Range. Typically one or two orders of magnitude in voltage around I_c, and corresponding current values, are used to determine the n value when fits are used. If derivatives are involved, an n value is determined for each current value using neighboring data points. The range is then determined by the span of the derivative and the amount of filtering or smoothing required.

Estimating $H_{c2}(T)$ and $H_{irr}(T)$

Magnetic fields $H_{c2}(T)$ and $H_{irr}(T)$ can be estimated by measuring the values for which I_c approaches zero. This was discussed at length under Principles of the Method.

SAMPLE PREPARATION

Sample Geometry Options and Issues

Possible sample geometries include straight, circular, spiral and U-shaped samples. Experimental conditions, especially the volume in which the desired conditions exist, and the desired sample length often determine the sample shape. Some conductors (e.g., NbTi or very thin tapes) can be shaped and tolerate a significant range of bending diameters, but more often a heat treatment of the sample in the desired shape is necessary. If the sample is already heat treated, it tends to be straight or stored on a large-diameter spool. Measuring a straight section may be the only option.

Sample Support Issues

Sample support is mostly a sample holder issue and is discussed in Appendix A and Appendix B. Two aspects also affect sample preparation.

The interface between sample and holder should be uniform when the sample support is required to support the sample against Lorentz forces. However smooth the sample holder surface, additional measures are required when

the sample surface is not perfectly flat. One option is to glue the sample to the holder using an epoxy. This bond is strong and can be thermally conductive, depending on the type of epoxy. Removing the sample without major damage is usually not possible, and this method requires mechanical cleaning of the sample holder or replacement of part of the holder. Different types of grease have also shown good results for ensuring an even contact between the sample and holder. Thermal conductivity is low and excess grease should be removed. The elegance lies in the freezing of the grease at cryogenic temperatures while it is soft at room temperature. Thus, it is easily applied and removed.

A second preparation issue involving the sample holder is associated with thermal prestrain. The strain state of the sample is determined by the integrated thermal expansion of the sample relative to the holder from the temperature at which the solder in the joints solidifies to the temperature after cooling down. The sample and sample holder may not be at the same temperature when the solder solidifies, and a gradient may exist within one or both. Additional measures to ensure that sample and holder are at the correct temperature, such as preheating before soldering, may be required if the margin on thermal precompression is narrow.

Sample Contact Issues

The metallic surface of a sample should be clean before soldering to prevent unreliable contacts. For NbTi, this often means thorough removal of varnish-type insulation. Using an oxygen-free atmosphere as well as zirconium as an oxygen getter will prevent contamination of the surface of Nb_3Sn conductors during heat treatment. Pickling with a 10% to 30% nitric acid solution can remove surface oxidization and should be followed by a rinsing step. Sanding or scraping can clean metallic surfaces on all conductors mechanically, but this involves a significant risk of sample damage. Often solvents are of little use, but flux will improve a solder connection.

It is often difficult to make good connections to ceramic surfaces. Some HTS conductor production techniques result in ceramic outer surfaces, such as dip coating for BSCCO 2212 and most YBCO coated-conductor processing approaches. Application of silver paste followed by a heat treatment can result in contact areas that can carry a modest current density after soldering to a current lead. The heat treatment depends on the brand of Ag paste and properties of the superconductor, so each user should develop his or her own approach. On some conductors, the paste can be applied before the heat treatment that forms the superconductor. Note that large contact areas may be required to prevent high-current density contacts and thus excessive heating. Soldering with indium can give similar results but has the disadvantage that it is very hard to see if the solder wets the surface reasonably and is therefore less reliable. When in doubt about the contact resistance, it is advisable to place an additional set of voltage taps on the current leads just outside the contacts to monitor the heat dissipation during a measurement and compare the contact resistance between samples. The electrical characterization may have to be performed relatively quickly to limit heating, and active quench protection may be necessary to prevent sample burn-out when thermal runaway occurs.

The above also holds for coated ceramic YBCO conductors. The ceramic is typically deposited on a metallic tape substrate leaving one of the flat sides free, but insulating buffer layers between the metal substrate and superconductor make contacts on the metallic side ineffective.

SPECIMEN MODIFICATION

Thermal cycling can affect a sample in several ways. The most dramatic effect occurs in HTSs with pinholes, allowing liquid cryogen to creep into the porous ceramic core slowly. When the conductor is warmed after the measurement or during a quench, the cryogen warms, vaporizes, and expands more quickly than it can escape via the pinholes. The conductor develops localized blisters, and tape conductors may even balloon to near-circular cross-section.

A second sample modification has occurred when an initially straight sample is in the shape of an arc after measurement and warming up. This can occur with samples that are not completely fixed to the holder. Usually this implies that the sample was cooled much quicker than the holder, i.e., a large temperature difference between sample and holder occurred. The sample will contract faster than the support upon cooling and yield plastically when its contraction significantly exceeds that of the support. Thus, the sample has been stretched relative to the holder, stands as an arc upon reaching a stable cryogenic temperature, and remains in that shape. A second possible reason for arced samples is a Lorentz force that is outwardly normal to the sample holder surface, i.e., in the wrong direction.

A third, more subtle effect is elastic-plastic interaction between the superconductor and the matrix due to differences in thermal contraction upon first-time cooling. This is known to occur in ceramic superconductors with ductile pure silver in the matrix, where the stress-strain properties (measured at room temperature) are changed by a thermal cycle to cryogenic temperatures. The plastic yielding mostly occurs during cooling such that most of the change in mechanical properties takes place before electrical characterization. Many HTSs show a small decrease in critical current, on the order of 1%, when cooled down and measured a second time, implying that some degradation occurred. Continued cycling usually has no further effect.

PROBLEMS

Parameters Affecting the Signal-to-Noise Ratio

Pick-up Loops. Pick-up loops, the area between the voltage wires and the sample or the area between sections of the voltage wiring that are not in close contact, work as antennas and pick up electromagnetic noise (see Fig. 7). To minimize this effect, the voltage wiring should be

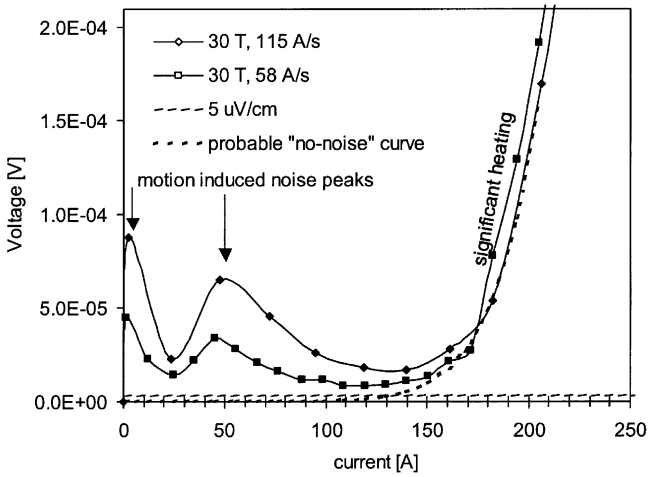


Figure 7. Measured *I-V* data at different current ramp rates of a sample in a magnetic field. The proportionality of the parasitic signal to the ramp rate (shown) and magnetic field (not shown) points to Lorentz force–induced motion and corresponding induced voltages as the source of one problem in this measurement. A second problem is sample heating.

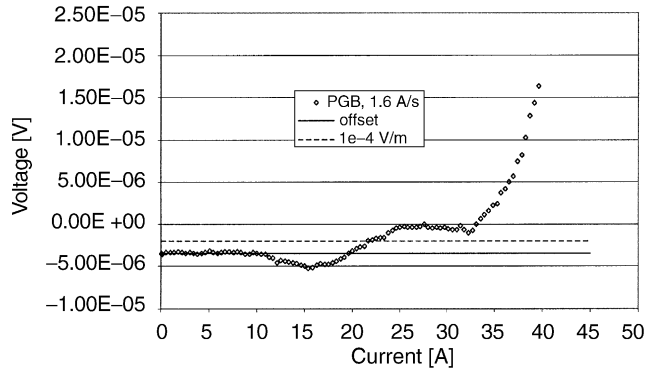


Figure 8. Measurement in which thermal emfs are suspected, given the unusual *I-V* characteristic. See Figure 7 for the standard *I-V* curve.

twisted until very close to the conductor and then follow the surface of the conductor closely until the voltage tap location. Possible sources of noise are the measurement current in the current leads, ambient electromagnetic fields, and applied background magnetic fields. The results can be an unusually high random noise level, voltage spikes when nearby equipment is switched on/off, and, when a background magnetic field is present, repeatable "bumps" that are most easily recognized in the superconducting part of the *I-V* curve. A possible worst-case scenario occurs when measuring electrical properties in a background magnetic field on a sample holder that is not sufficiently rigid. Not only will a loop pick up the ripple in the magnetic field, but also the slightest motion of the sample and holder under the Lorentz force will cause significant induced voltages.

Thermal Electromotive Forces. Thermal electromotive force (emf), or thermocouple voltage, is caused by temperature variations of junctions between dissimilar materials in the voltage circuit, such as solder connections and connectors. Ideally, the voltage wiring should consist of a pair of twisted pure copper wires (insulated), using only copper-to-copper clamped terminals, from the sample to the printboard of the meter. Parasitic DC voltages can originate from junctions that are not in a cryogen bath or submerged junctions exposed to gas bubbles. A draft on a connector at room temperature can easily cause drift of a few microvolts in 30 s. If the junction causing the thermal emf is on the sample itself, then the sample itself may not be at a constant temperature as well. Thermal emfs are not dependent on the polarity of the current, a feature that helps diagnose them when they are caused by the magnitude of the current and associated heating, as is shown in the following example.

Figure 8 shows a measurement in which thermal emfs are suspected to affect the *I-V* response of the superconductor. After switching polarity of both the current and voltage, a second *I-V* curve is measured. Both curves are displayed in Figure 9. Since the voltmeter polarity is changed, the thermal emfs now have the opposite effect on the measurement. The average between those curves approximates the superconductor properties, but caution is required when interpreting such manipulated data, as will be shown below. The above measurements are performed at a constant current ramp rate. When the ramp is interrupted by keeping the current constant for a few seconds at a time, the characteristic of Figure 10 is obtained. The measured voltage drifts at constant current, indicating that the temperature at the junctions causing the thermal emfs varies even at constant currents. Since no unusual temperature variations outside the cryostat are present in this case, it is suspected that the junctions formed by the solder connection between the sample and voltage wiring are exposed to varying temperatures. This implies severe boiling at the sample, since this sample is below the liquid helium surface. To assess the possibility that the sample is heated to temperatures significantly above 4.2 K, measurements at different ramp rates are performed. The temperature variations will decrease with current ramp rate as the total amount of dissipated

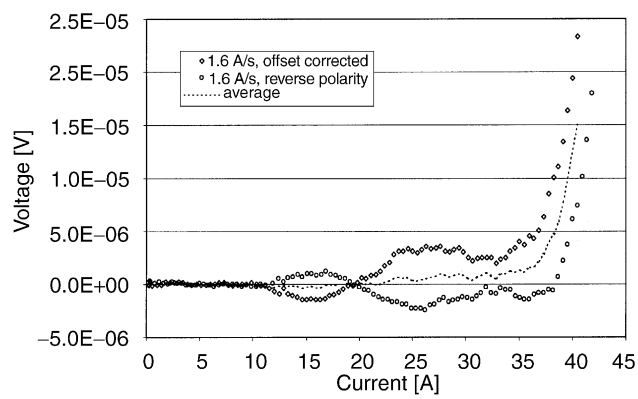


Figure 9. Two *I-V* measurements on the same conductor as in Figure 8, one with reversed polarities. Note the symmetry in the deviation from a standard *I-V* curve.

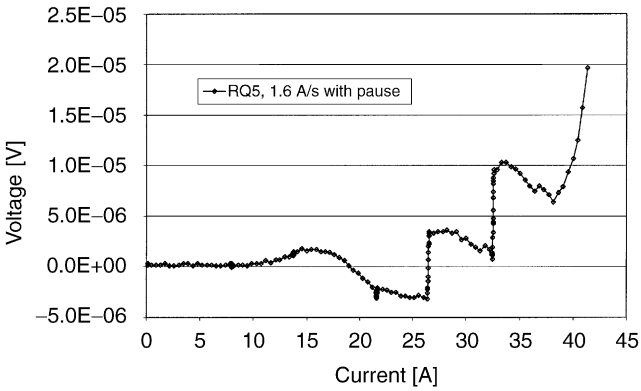


Figure 10. Measurement on the same conductor as in Figure 8, with interruptions of the otherwise constant current ramp. The voltage shows significant drift due to thermal emfs.

heat will decrease. As Figure 11 shows, the measured data strongly depend on ramp rate. It is now apparent that the temperature of the sample itself rises significantly during the measurement, even during the 23-A/s run, and that none of the curves in Figures 8 to 11 represent the properties of the superconductor at 4.2 K. The culprit in this case is a poor solder joint between the sample and the current leads.

Grounding. Only a single point in the circuits of the current and voltage wiring should be grounded. Ground loops should be avoided in this wiring and also in all other connections of the equipment involved. It is usually worthwhile to verify the grounding, or absence thereof, of all circuits, equipment, and wiring. It should be noted that, unlike oscilloscopes, many current sources and digital voltmeters have floating outputs and differential inputs and thus do not provide a ground by default. Voltmeters tend to perform optimally if a point close to the sample, or even on the sample between the voltage taps, is grounded with a dedicated wire to the case of the meter.

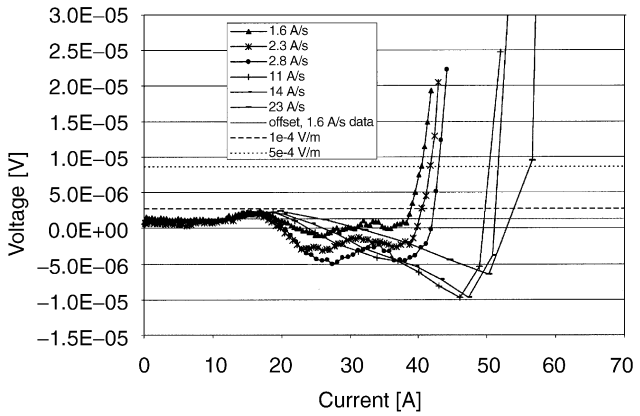


Figure 11. Measurements on the same conductor as in Figure 8 at different current ramp rates. Thermal emfs affect all measurements. The increase in transition current with ramp rate indicates that the temperature of the sample itself is not stable.

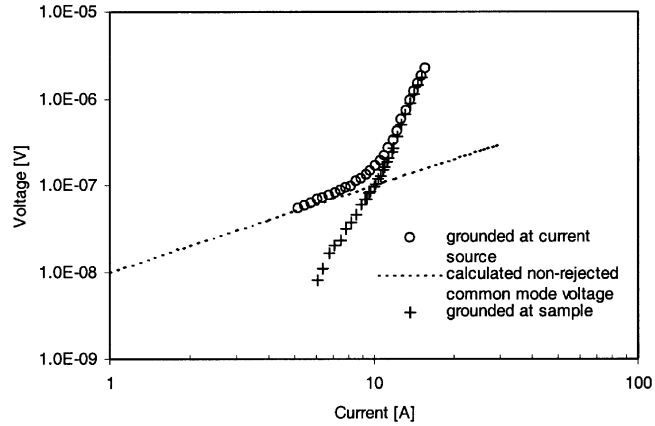


Figure 12. Measured data on one sample with either the current source output terminal or the sample grounded. The calculated nonrejected common-mode voltage with 140 dB common-mode rejection ratio and 0.1 Ω lead resistance explains the error when grounding at the current source.

This is related to the common-mode rejection ratio of the voltmeter (see Appendix A for more detailed discussion) and the resistance of the current leads. Here we compare two measurements on the same conductor in identical sample conditions, one measurement where the output terminal of the power supply is grounded closest to ground and one where the sample is grounded to the case of the voltmeter. The voltmeter has a common-mode rejection ratio of 140 dB (see Appendix A) and the current lead resistance between supply and sample is taken as 0.1 Ω. The increasing voltage drop across the leads with increasing current will cause the sample (common-mode) voltage to increase, unless the circuit is grounded on the sample. Not all of the common-mode voltage is rejected with power supply grounding, and the resulting voltage and its effect on the measured *I-V* curve are presented in Figure 12. If improper grounding is suspected, a multimeter can be used to verify what is grounded to what and where. The ground loops should be removed and the grounding varied, while maintaining safety, and its effect on the measurement assessed. When possible, sensitive equipment should be plugged into a different power outlet group than power equipment or known sources of electrical noise. This process can be time consuming but is a worthwhile investment.

Random Noise and Voltage Spikes. High-power or high-current equipment in the vicinity, especially when switched on or off, can cause voltage spikes and other random noise. Possible sources include overhead cranes, variable-speed motors controlled by pulse-width modulation, and pumps. Check for proper grounding and pick-up loops. Spikes in an *I-V* curve can cause software routines that calculate I_c to yield improper results or can trip quench detection circuits.

In measurements of metallic superconductors (NbTi and Nb_3Sn), reproducible voltage spikes in the low- or intermediate-current part of the superconductive baseline may not be a measurement problem per se but the result of

low-current instabilities in the internal distribution of the self-field, i.e., flux jumps. For more information on flux jumps, see Wilson (1983a). However, this is not a measurement problem that requires a solution, unless it is improperly interpreted by the controlling software.

Current Supply Noise and Related Effects. Noise originating in the current supply can affect the voltage measurement through several routes. First, AC components in the output current cause AC voltages across the sample, which are never perfectly rejected by the voltmeter. Those components may vary as a function of the output current. Switched power supplies also tend to have more high-frequency components in their output current and voltage compared to heavyweight current sources with electronics that are more traditional. Second, a path for noise is provided if the current leads are inductively coupled with the voltage wiring. The following third scenario also occurs, albeit rarely. Voltmeters operate best if the line voltage is stable, although a range of voltages is usually acceptable. If the meter is connected to a heavily loaded or overloaded group of voltage outlets, its input voltage and thereby performance may be affected. The varying load that a current source during a measurement places on the outlet group may then cause systematic errors in the measurement. Connecting sensitive equipment such as voltmeters to voltage outlet groups separate from high-current equipment may help avoid this.

Integration Period. Much of the AC noise has the frequency of the line voltage or a multiple thereof. Its effects are minimized by choosing the integration period of the voltage measurement as a power line cycle. Multiple readings can be averaged to reduce the effect of random variations.

Shorting of Leads

When current leads are shorted, part or most of the measurement current will bypass the sample but often not the shunt, leading to incorrect characterization of the conductor that may not be immediately obvious. In case of a short that allows a significant current through the sample, the transition will be less sharp, the measured I_c value will be too high, and repeatability of the measurement will likely be poor. Repeated or prolonged measurements may warm up parts of the current leads, increasing their resistance and leading to a change in the fraction of the current passing through the sample. If the ramped-current approach is used, a decrease of the inductive voltage without change in ramp rate may signal a short. Shorts may be detected early when the resistance between the leads is measured with the current source disconnected, before the sample is mounted.

Thermal Contraction of the Probe

Thermal contraction in itself is not a problem, but differences in thermal contraction can have very frustrating results. When a rigid current lead is immovably connected at two locations to a structure with a smaller thermal expansion, a thermal gradient will place the current lead in tension. Upon cooldown, the lead may fracture, interrupting the current path. Warming to room temperature for inspection will close the gap, and the circuit will seem intact again. Visual inspection rather than resistance measurements will facilitate finding the defect in the lead. Potential problems are avoided by mounting only one end of the current lead immovably or using flexible elements.

Sample Handling/Damage

Many superconducting materials are brittle and can be damaged by improper handling techniques. In particular, damage can occur during the process of mounting the sample to the probe and soldering/unsoldering the current leads and voltage taps. In addition, some HTSs that are processed with the superconducting oxide layer exposed to the atmosphere can suffer chemical attack from exposure to the air. Thus, these samples must be kept in proper storage before and after measurement, and exposure time to the atmosphere during mounting/dismounting must be kept to a minimum. For an example of the electrical properties of a damaged sample, see Figures 5 and 6.

Sample Heating and Continuous-Current Measurements

Continuous current-ramp measurements are often required when sample heating problems arise. Sample heating can be a significant problem, leading to inaccurate temperature determination and possibly sample burnout. It is most likely to occur when the samples are not bath cooled or when the sample/current lead contacts are problematic. Current ramp rate is one variable that can be used to avoid these problems. With samples that are gas flow or conduction cooled, typically a large ramp rate is used. When current stability after a step is an issue, a slow ramp may prove more effective. Several additional factors complicate continuous, rapid measurements compared to the pulsed or staircase step approach. The measured sample voltage includes an inductive component proportional to the ramp rate and sample inductance that has to be subtracted. The measured properties are no longer steady-state properties. The critical current, even after correction for inductive voltages, will decrease with increasing ramp rates. This effect will be negligible for a significant range of rates. At least initially, however, a number of rates should be tried to determine the effect the chosen ramp rates have. Finally, it should be noted that each pair of voltage and current values represents a range of current values and an effective average voltage over that range. This will tend to "smear out" features of the *I-V* characteristic where it is nonlinear.

Bath Temperature Fluctuations

The bath temperature of a liquid cryogen can fluctuate in time and result in reduced repeatability of *I-V* measurements. Although the effects on I_c are relatively small, they are significant in an accurate measurement. The boiling point of a cryogenic bath and therefore the sample temperature if submerged are affected by contamination of the cryogen bath and pressure variations. Pressure variations

can be caused by either variations of the ambient pressure or changes in the level of cryogen above the sample. The former is most relevant in liquid helium, the latter in liquid nitrogen, as it has a much higher density. The use of overpressure valves and recovery systems for helium gas will also affect the boiling point of the cryogenic liquid. The bath temperature of saturated liquid helium at atmospheric pressure varies 0.011 K/mbar (0.072 K/psi); the equivalent number for liquid nitrogen at 77.4 K is 0.084 K/mbar (0.58 K/psi). With knowledge of the relation between the critical current and temperature, one can calculate the relation between I_c and pressure in the bath at the sample. That slope, and therefore the relative error, will be highest for materials with a low T_c and/or when the bath temperature is close to T_c.

Self-Field Effects

The measurement of I_c of a superconducting sample in a zero background field is often viewed as a zero-field measurement. In fact, the sample self-field can be significant and may need to be factored into the interpretation of the measurement results. Otherwise, two samples that may have identical electrical behavior will appear quite different at "zero field." This can be particularly important for HTS materials, where dI_c/dB can be quite large at very low magnetic fields. As a result, it is often best to compare conductors in the presence of a background field that is sufficiently large that self-field effects become negligible.

Thermal Cycling

Thermal cycling of the sample can affect the sample, particularly if the heating and/or cooling are performed too quickly. This is the result of a combination of thermal expansion mismatch between the sample and the sample holder and fatigue. Even if temperature changes are effected properly (slowly), some property changes can occur in HTS samples during the first thermal cycle (see Specimen Modification).

Sample Quality Issues (Particularly for Powder-in-Tube HTSs)

One problem that can occur in powder-in-tube HTS samples is the penetration and subsequent expansion of the cryogen within the conductor, resulting in the destruction of the sample. This is discussed at length in Specimen Modification.

ACKNOWLEDGMENTS

The authors thank Hans van Eck, Ilkay Cubukcu, S. Hill Thompson, Youri Viouchkov, David Hilton, Yusuf Hascicek, Victor Miller, Bennie ten Haken, and Steven Van Sciver for assistance in developing electrical measurement techniques.

LITERATURE CITED

Bean, P. 1962. Magnetization on hard superconductors. *Phys. Rev. Lett.* 8:250–253.

Bentzon, M. D. and Vase, P. 1999. Critical current measurements on long BSCCO tapes using a contact-free method. *IEEE Trans. Appl. Supercond.* 9:1594–1597.

Brooks, J. S., Naughton, M. J., Ma, Y. P., Chaikin, P. M., and Chamberlin, R. V. 1987. Small sample magnetometers for simultaneous magnetic and resistive measurements at low temperatures and high magnetic fields. *Rev. Sci. Instrum.* 58(1):117–121.

Ekin, J. W. 1978. Current transfer in multifilamentary superconductors. I. Theory. *J. Appl. Phys.* 49:3406–3409.

Ekin, J. W., Clark, A. F., and Ho, J. C. 1978. Current transfer in multifilamentary superconductors. II. Experimental results. *J. Appl. Phys.* 49:3410–3412.

Jones, H., Cowey, L., and Dew-Hughes, D. 1989. Contact problems in J_c measurements on high T_c superconductors. *Cryogenics* 29:795–799.

Kramer, E. J. 1973. Strain scaling laws for fluxpinning in hard superconductors. *J. Appl. Phys.* 44:3:1360–1370.

Kwok, W. K., Fleshler, S., Welp, U., Vinokur, V. M., Downey, J., and Crabtree, G. W. 1992. Vortex lattice melting in untwined and twined single crystals of $YBa_2Cu_3O_{7-\delta}$. *Phys. Rev. Lett.* 69:23:3370–3373.

Naughton, M. J., Ulmet, J. P., Narjis, A., Askenazy, S., Chaparala, M. V., and Hope, A. P. 1997. Cantilever magnetometry in pulsed magnetic fields. *Rev. Sci. Instrum.* 68:11:4061–4065.

Rossel, C., Willemin, M., Gasser, A., Bothuizen, H., Meijer, G. I., and Keller, H. 1998. Torsion cantilever as magnetic torque sensor. *Rev. Sci. Instrum.* 69:9:3199–3203.

Warnes, W. H. and Larbalestier, D. C. 1986. Critical current distributions in superconducting composites. *Cryogenics* 26:643–653.

Wilson, M. N. 1983a. Flux jumping. *In* Superconducting Magnets, Chapter 7. Oxford University Press, Oxford.

Wilson, M. N. 1983b. Measurement techniques. *In* Superconducting Magnets, Chapter 10. Oxford University Press, Oxford.

KEY REFERENCES

Ekin, J. W. 2002. Superconductor contacts. *In* Handbook of Superconducting Materials (D. Cardwell and D. Ginley, eds.). Institute of Physics Press, New York.

Provides an in-depth discussion of making contacts between superconductors and other materials, an important aspect of any situation involving transport current in superconductors.

Jones, H. and Jenkins, R. G. 1994. Transport critical currents *In* High Temperature Superconducting Materials Science and Engineering (D. Shi, ed.) pp. 259–304. Pergamon Press, Elmsford, N.Y.

Discusses superconducting terminology, transport current theory, and practical aspects of measurements for both HTSs and LTSs. Discussed aspects include sample shape, current contact problems, examples and modeling of measurements as function of B and T.

INTERNET RESOURCES

http://www.iec.ch

The International Electrotechnical Commission is the international standards and conformity assessment body for all fields of electrotechnology. The IEC is closely linked to activities within the VAMAS project (Versailles project on Advanced Materials and Standards). It developed document IEC 60050-815,

International Electrotechnical Vocabulary VAMAS Part 815: Superconductivity (1540 Kb, 194 pages, bilingual in French and English, November 2000, CHF 222.00) and related documents regarding measurement standards.

http://www.iec.ch/cgi-bin/procgi.pl/www.iecwww.p?wwwlang=E&wwwprog=dirdet.p&committee=TC&number=90

Technical Committee 90 on superconductivity produces documents and standards on terms and definitions and has provided a critical current measurement method of oxide superconductors; a test method for the residual resistivity ratio of Cu/Nb-Ti and Nb_3Sn composite superconductors; room temperature tensile tests of Cu/Nb-Ti composite; the matrix composition ratio of Cu/Nb-Ti composite superconductors; a critical current measurement method of Nb_3Sn composite superconductors; electronic characteristic measurements; a measurement method for AC losses in superconducting wires; measurement for bulk high-temperature superconductors, i.e., trapped flux density in large grain oxide superconductors; critical temperature measurement, i.e., the critical temperature of composite superconductors; and matrix-to-superconductor-volume ratio measurement, i.e., the ratio of copper to noncopper volume of Nb_3Sn composite superconducting wires.

http://zone.ni.com/idnet97.nsf/browse?openform

National Instruments provides a library of drivers for equipment built by a large range of manufacturers. One can search for drivers.

APPENDIX A: INSTRUMENTS

Current Supply

A single-quadrant DC power supply or current source is required with a maximum current of at least the critical current. A power supply offers the option of user control of either the output current or the output voltage, whereas only the current is user controlled in current sources. Both types can be used.

Ripple. It is desirable to have a maximum periodic and random current deviation of less than 2% of the output current within the bandwidth of 10 Hz to 10 MHz.

Resolution. A minimum current supply output resolution of 1% of the maximum current is suggested. A better resolution is required if the expected I_c is a small fraction of the maximum current, with the desired absolute precision in I_c as the lower limit.

Maximum Output Voltage. A 5-V output rating is generally sufficient with properly sized current leads from the current supply to the sample support and as part of the sample support.

Direct or Inductive with Feedback Loop. This issue is applicable only for currents greater than the kiloampere range and is thus outside the scope of this unit.

Voltmeter

A dedicated high-resolution DC voltmeter, a multimeter with adequate DC voltage specifications, or a multicomponent DC voltage measurement system is required. The latter consists of precision amplifiers, filters, and ADCs. Pen-and-paper X-Y recorders are considered outdated.

Common-mode Rejection. Common-mode rejection is rejection of undesirable AC or DC signal between the low (or minus) input and ground (earth). A suggested minimum ratio [Common-mode rejection ratio (CMMR)] is 120 dB when using a single power line cycle integration period. Every 20 dB represents a factor of 10 in reduction of effective noise.

AC Suppression. The suggested minimum AC normal-mode rejection ratio, i.e., rejection of undesirable AC signals across the DC voltage input terminals, of 60 dB is suggested when using a single power line cycle integration period without filtering.

Precision. A minimum accuracy of 10% of the critical voltage is desirable, where the critical voltage is determined by the I_c criterion and the voltage tap length. The reading resolution or precision of a single reading is always smaller than the accuracy, often by an order of magnitude or more. With typical tap lengths of 0.01 to 0.05 m and a 10^{-4}-V/m criterion, this corresponds to a minimum accuracy of 10^{-7} V and suggests a 10^{-8}-V minimum resolution, i.e., a 7 1/2 digit voltmeter, for this example.

Integration Period. It is desirable to have the option of an integration period of one power line cycle (PLC), since this allows relatively easy suppression of AC noise, compared to shorter integration periods. Increasing the measurement accuracy beyond that obtainable with one PLC is often best achieved by averaging multiple readings of one PLC, rather than increasing the integration period.

Sample Probe/Support

The sample support, also referred to as probe or insert, forms the connection between the sample during a measurement, the room temperature environment of the instrumentation, and the current source. It facilitates introduction of current to a sample on the cryogenic end without excessive heating and is equipped with terminals for current leads to the current source on the room temperature end. Similarly, it contains instrumentation wiring, including voltage tap wiring. In addition, the support provides mechanical support for the sample as needed to prevent motion and mechanical damage. This is particularly important for measurements in large background magnetic field. More details on the sample holder are discussed in Appendix B.

Software Selection and I/O Compatibility

Many software packages are suited for control, graphic representation, and analysis of an electrical characterization. Examples include Labview, C-based languages, and Visual Basic. Computer and instruments require input-output (I/O) capabilities for automation of a measurement and parallel or serial connections. Currently, the GPIB operating under the IEEE 488.2 standard is most commonly used, although serial connections under the RS

232 or RS 442 standard are not uncommon, particularly for power supplies. Converters between the two standards are commercially available. Instrumentation and data acquisition are areas that are likely to see significant changes as new communications standards, equipment, and network protocols develop. Instrument drivers for the software environments listed above are available for most instruments.

APPENDIX B: SAMPLE HOLDER

Sample holders can be made in a large variety of shapes, with different functions, and constructed of different materials. Most sample holders share the following functions:

1. Position the sample at the proper location.
2. Support the forces on the sample, Lorentz force, or otherwise.
3. Maintain a constant temperature or allow effective bath/gas cooling.
4. Separate cryogenic and ambient atmosphere.
5. Connect the sample to the current source and voltmeter and supply wiring for other instrumentation (e.g., temperature-, field-, strain-, and liquid-level sensors).
6. Protect the sample from impact when the sample holder is put in or removed from the cryostat.

Although no standard sample holder exists, commonly used materials include the following:

1. Stainless steel, for structural elements. Note that 304 or A2 grade stainless can be slightly magnetic or become magnetic due to welding or thermal cycling.
2. Copper, and to a lesser extend brass, for current leads. Current leads can be hollow and vapor cooled by cryogenic boil-off to offset or reduce heat input by thermal conduction and ohmic heating. The inlet of the vapor is just above the maximum helium level. Superconductors in parallel to the leads can be used to reduce heating below the vapor-cooled leads.
3. G-10 (glass cloth–epoxy composite) as a non conductive structural element. The thermal expansion of G-10 is anisotropic, with properties in the warp and fill directions comparable to those of many commercial HTS composites.
4. Ti6V4Al alloy to mount Nb_3Sn samples, usually spiral samples on a grooved alloy cylinder with copper end pieces as current connections.
5. Epoxies, in particular filled epoxies like Stycast 2850 FT and similar products.

APPENDIX C: DEFINITIONS

The following definitions are taken directly from VAMAS document IEC 60050-815.

Composite Superconductor. Conductor in the form of a composite of normal and superconducting materials.

Critical Current (I_c). The maximum electrical current that can be regarded as flowing without resistance through a superconductor.

Critical Current Density (J_c). A critical current per unit cross-sectional area of either the whole conductor (overall) or the nonstabilizer part of the conductor.

Critical Temperature (T_c). The temperature below which a superconductor exhibits superconductivity at zero magnetic field and current.

Flux Jump. The cooperative and transitional movements of pinned flux bundles or large number of flux lines as a result of a magnetic instability initiated by a mechanical, thermal, magnetic, or electrical disturbance within a type II superconductive material.

Matrix (of a Composite Superconductor). Continuous longitudinal component of metal, alloy, or other material that is in the normal state at the operating conditions in the composite superconductor.

Meissner State. Superconducting state in a superconductor characterized by perfect diamagnetism.

Mixed State. Superconducting state in which magnetic flux penetrates the type II superconductor in the form of fluxons.

Quench. Uncontrollable transition of a superconducting conductor or device (usually magnet) from the superconducting to the normal conducting state.

Superconducting Transition. The change between the normal and superconducting states.

Superconductivity. A property of many elements, alloys, and compounds in which the electrical resistivity to DC currents vanishes and they become strongly diamagnetic under appropriate conditions.

Type I Superconductor. Superconductor in which superconductivity appears with perfect diamagnetism below the critical magnetic field strength H_c but disappears above H_c when the demagnetization factor is zero.

Type II Superconductor. Superconductor that exists in the Meissner state when the magnetic field strength is below the lower critical magnetic field strength H_{c1}, exists in the mixed state for the field between H_{c1} and the upper critical field strength H_{c2}, and exists in the normal state above H_{c2} when the demagnetization factor is zero.

In addition, the following definition is presented here by the authors:

Current Introduction Length. When a current is introduced into a (composite) superconductor, there is a zone that extends beyond the joint where the current transfers from the matrix to the superconducting core/filaments. The length of this current transfer zone up to the point where the ohmic voltage associated with the remaining current in the matrix is below the sensitivity of the voltmeter determines the current introduction length.

JUSTIN SCHWARTZ
HUBERTUS W. WEIJERS
National High Magnetic Field Laboratory
Florida State University
Tallahassee, Florida

This page intentionally left blank

MAGNETISM AND MAGNETIC MEASUREMENTS

INTRODUCTION

Magnetism and magnetic materials have a long and illustrious history. Lodestone (a variety of magnetite) was the first magnetic material of technological importance—which often means military importance as well—due to its seemingly mystical power of seeking direction, a feat which never fails to fascinate a first-time observer. Lodestones were known to the ancient Egyptian, Greek, and Chinese civilizations, and possibly others. About 2600 B.C., lodestone was instrumental in winning a decisive battle for the first emperor of China, fought in a dense fog against the invading barbarians.

The subject that we generally call magnetism began with the observation that some minerals, such as magnetite, could attract each other and certain other minerals. The name magnetism probably derives from Magnesia, a location in ancient Thessaly in Asia Minor, presumably where such minerals were found. Despite the early, albeit accidental, discovery of magnetic materials, interest in magnetism during the subsequent centuries was largely limited to the construction of better compasses, more in the realm of witchcraft than science. William Gilbert made the first serious scientific study of magnetism during the 16th century with the publication of *De Magnete*, only to be ignored for more than a century. It was not until the 18th century that scientific interest in electricity and magnetism was renewed with great intensity, culminating with the realization of the Maxwell equations in the 19th century. However, a fundamental understanding of magnetism had to await the advent of quantum mechanics and special relativity in the 20th century.

Today, magnetism enters many aspects of our lives, far beyond just compasses and refrigerator magnets. Magnetic materials and magnetic phenomena inconspicuously appear in many complex devices, for example, theft-prevention systems in department stores, magnetic recording disks and read heads in the ubiquitous computers and cellular phones, and in magnetic resonance imaging apparatus. Some of the more exciting new developments are the spintronic devices, which are on the verge of becoming realities. These new devices are based on the spin properties of electrons, as opposed to charge alone, as in conventional electronics.

The subject of magnetism is inseparable from electrical charges, their translational motion and their angular momenta. Magnetic fields can be generated by current-carrying wires, through which the electrical charges (most commonly the electrons) are transported. Magnetic fields can also be realized in the vicinity of magnetic materials with large magnetization, such as permanent magnets. There are various methods by which one generates magnetic fields, ranging from generation of magnetic fields by electrical circuits and permanent magnets to the use of hybrid superconducting magnets, Bitter magnets, and pulsed magnets, necessary for achieving very large magnetic fields. Roughly speaking, magnetic fields up to about 2 tesla (20 kOe) and 10 tesla (100 kOe) can be generated by laboratory electromagnets and superconducting magnets, respectively. There are also national facilities, such as the National High Magnetic Field Laboratory at Florida State University, which accommodate measurements at much higher magnetic fields.

All constituent atoms in a material respond to an external magnetic field, at least weakly due to diamagnetism. The direction of the small diamagnetically induced magnetization is opposite to that of the external magnetic field. However, virtually all magnetic materials of scientific or technological importance are those with sizable atomic moments, which respond strongly to external magnetic fields. The atomic magnetic moments are the results of orbital and spin angular momenta of the electrons in the partially filled atomic electronic shells. These atomic magnetic moments, with magnitude of the order of Bohr magneton (μ_B), are much larger than those induced by diamagnetism. Most of the elements in the periodic table do not have atomic magnetic moments, except some of the 3*d* transition-metals (e.g., Ni, Fe, Co, and Mn) and 4*f* rare-earth metals (e.g., Gd, Nd, etc.), which are the major constituents of all the important magnetic materials.

The atomic magnetic moments can interact with each other via the quantum mechanical exchange interactions. At high temperatures, the atomic moments are not ordered and are in the paramagnetic state. However, at sufficiently low temperatures, the atomic moments can spontaneously order. Those materials in which the atomic moments align parallel to each other are the ferromagnets, of which lodestones, refrigerator magnets, and other more powerful permanent magnets are examples. Those materials that exhibit nonaligned spin structures (e.g., antiferromagnetic and various other types of spin structures) are technologically not useful but scientifically fascinating. Each type of magnetic ordering has a unique order parameter, which, for example, is the magnetization for ferromagnets and sublattice magnetization for antiferromagnets. Various methods of magnetometry can be used to measure the magnetization of materials and the atomic magnetic moments. Neutron diffraction is particularly valuable for measuring the spin structure of magnetic ordering, particularly antiferromagnets and other more complex magnetic orderings, as described in NEUTRON TECHNIQUES.

Most people recognize that a piece of iron can easily be attracted by a permanent magnet, but iron itself cannot attract another piece of iron. This is because in iron there are magnetic domains, a general phenomenon in ferromagnets. Under a sufficiently large magnetic field, a

ferromagnetic material can always be swept into a single-domain state, where all the moments are aligned in one direction. In a small or zero magnetic field, many magnetic domains, separated by domain walls, may form in order to reduce the magnetostatic energy. Within each domain, all the moments remain aligned with a unique magnetization, but the magnetization axes among different domains are not aligned. The size and shape of the domains depend on many factors, most importantly the magnetic anisotropy. In a soft magnet, such as pure iron, domains can be readily formed and swept by a small field because of the small magnetic anisotropy. In a hard magnet, such as the refrigerator magnets and other more powerful permanent magnets, the magnetic anisotropy is so strong that it preserves the magnetization within the material. The magnetic domains, as small as a few tens of nanometers in size, have fascinating shapes and patterns. A variety of techniques have been developed to image magnetic domains and study the domain wall motion.

Electrons have both spin and charge. Many of the magnetic materials are electrical conductors. When electrons traverse magnetic materials, the external field and the magnetization **M** of the magnetic medium affect the conductivity. These are magneto-transport properties, which include a variety of interesting topics, such as magnetoresistance and the Hall effect, and the latter is described in ELECTRICAL AND ELECTRONIC MEASUREMENT. New magneto-transport properties due to spin-dependent scattering as well as spin-polarized conduction have recently been discovered in suitable magnetic heterostructures. These new effects form the basis for some of the new spintronic devices.

Most of the magnetic materials of technological importance are ferromagnets. These include permanent magnets (e.g., ferrite, $SmCo_5$, $Nd_2Fe_{14}B$) that can generate strong magnetic fields, soft magnetic materials (e.g., permalloy, amorphous ferromagnets) that can respond to small magnetic fields, materials that have large magneto-transport properties, and materials that can store magnetization, as in magnetic recording media (e.g., γ-Fe_2O_3, CoTa alloys).

The smallest magnetic entities are magnetic dipoles. Under a uniform magnetic field, a magnetic dipole of dipole moment $\boldsymbol{\mu}$ can only rotate. Under a nonuniform magnetic field **H**, the magnetic dipole can also have translational motion due to the force arising from the field gradient, $\mathbf{F}_\mu = (\mu \cdot \nabla)\mathbf{H}$. For a material of magnetization **M**, the total force per unit volume is $\mathbf{F}_v = (\mathbf{M} \cdot \nabla)\mathbf{H}$. The force on a ferromagnet is attractive and substantial, as evidenced by the refrigerator magnet. For nonmagnetic materials, the force is repulsive due to diamagnetism, and therefore usually weak. However, in an intense magnetic field gradient, the repulsive force can be quite substantial, so much so that it can actually lift or levitate an object against gravity.

Thus far, we have mentioned the magnetic properties due to the electrons in the atoms and the materials. We have ignored the nuclei and other particles because their magnetic moments are three orders of magnitude smaller. However, in certain phenomena, such as nuclear magnetic resonance, Mössbauer effect, muon spin resonance, and neutron scattering, these small magnetic moments are the only ones of importance as described in RESONANCE METHODS.

Units

The descriptions of magnetic properties must include the discussion of units. We are all supposed to be, and indeed have been, using the SI units (Systéme Internationale d'Unités). However, one notable exception is in magnetism, where the systems of units have been confusing and challenging even for seasoned practitioners. Although all the more recent physics textbooks have used exclusively the SI units to describe magnetic quantities as they should, most scientific literature in magnetism, both old and new, continues to use the cgs units, or a mixture of both cgs and SI units. The cgs units also persist on the instrument side. Most commercial magnetometers display magnetic moment in emu (electromagnetic unit) and most gaussmeters measure magnetic field in Oe (Oersted). A practitioner in magnetism therefore must know both the SI and the cgs units.

In the cgs units, the relation among magnetic induction or magnetic flux density (**B**), magnetic field intensity (**H**), and magnetization (**M**) is uniquely defined as

$$\mathbf{B} = \mathbf{H} + 4\pi\mathbf{M} \tag{1}$$

This expression is universally used without ambiguity. Since the units for B (gauss = G), H (oersted = Oe), and M (emu/cm^3) have the same dimension, sloppiness in describing the physical quantities is therefore tolerated (lightface symbols represent the scalar magnitude of the corresponding vector quantity). For example, one can describe the earth's magnetic field as about 0.5 G or 0.5 Oe, and the magnetization of a piece of iron can be specified as $M = 1700$ emu/cm^3 or $4\pi M = 21.7$ kG.

In the SI units, the corresponding relation is

$$\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M}) \tag{2}$$

where μ_o is a fundamental constant of $4\pi \times 10^{-7}$ N/A^2. The units for H and M are in A/m, whereas B is in tesla, or Wb/m^2. Most unfortunately, other forms, such as $\mathbf{B} = \mu_o\mathbf{H} + \mathbf{M}$, have also been used, creating endless confusion.

In Table 1, various common quantities used in magnetism are shown both in the cgs units and the SI units, as well as the conversion factors.

MAGNETIC SUSCEPTIBILITY AND PERMEABILITY

As described earlier, the applied external magnetic field, **H**, the magnetization, **M**, and the magnetic flux density, **B**, inside a medium are related by Equation 1 in cgs units and Equation 2 in SI units. An external field, **H**, induces certain magnetization, **M**, in a sample, and as a result, generates an internal flux density, **B**. We will use the cgs units in the discussion below for brevity. One can easily convert values into SI units by referring to Table 1.

The magnetic response of a medium to an external field is often characterized by two related parameters of

Table 1. Conversion Between cgs Units and SI Units for Selected Magnetic Quantities

Quantity	Symbol	cgs Units ($\mathbf{B} = \mathrm{H} + 4\pi\mathbf{M}$)	Conversion Factor (C) [(SI) = C(cgs)]	SI Units [$\mathbf{B} = \mu 0(\mathbf{H} + \mathbf{M})$] [$\mu 0/4\pi = 10-7\mathrm{N/A2}$]
Magnetic induction	B	Gauss (G)	10^{-4}	tesla(T) $\mathrm{Wb/m^2}$
Magnetic flux	Φ	G-$\mathrm{cm^2}$ Maxwell(Mx)	10^{-8}	weber(Wb)
Magnetic field intensity	H	Oersted(Oe)	$10^3/4\pi$	A/m
Magnetic moment	m	emu, erg/Oe	10^{-3}	A-$\mathrm{m^2}$, J/T
Magnetization	M	emu/$\mathrm{cm^3}$	10^3	A/m
	4πM	G	$10^3/4\pi$	A/m
Mass magnetization	M_g	emu/g	1	A-$\mathrm{m^2}$/kg
			$4\pi \times 10^{-7}$	Wb-m/kg
Volume susceptibility	χ	Dimensionless	4π	dimensionless
Mass susceptibility	χ_g	$\mathrm{cm^3/g}$	$4\pi \times 10^{-3}$	$\mathrm{m^3/kg}$
Permeability	μ	Dimensionless	$4\pi \times 10^{-7}$	Wb/A-m

magnetic susceptibility (χ) and permeability (μ), which are defined as

$$\chi = \mathrm{M}/\mathrm{H} \tag{3}$$

$$\mu = \mathrm{B}/\mathrm{H} \tag{4}$$

and $\mu = 1 + 4\pi\chi$. If the magnetization of a medium is proportional to H, χ is then a constant and becomes the most appropriate parameter to characterize the magnetic response. On the other hand, if M is not linear in H, χ and μ become functions of H, i.e., $\chi = \chi(\mathrm{H})$, $\mu = \mu(\mathrm{H})$. In some situations, such as the case of a ferromagnet, the behavior of M is hysteretic and, therefore, not uniquely determined by a given H. The concepts of χ and μ are no longer very useful, although still adopted, particularly in soft magnetic materials.

PARAMAGNETISM

A paramagnet consists of an assembly of magnetic dipole moments—μ—the atomic magnetic moments. Thermal disturbance randomizes the orientations of these moments. An external field, on the other hand, creates a tendency for a slight alignment of the moments along the field direction. Under the condition of $\mu\mathrm{H} \ll kT$, the induced **M** is proportional to **H**. If the interaction between the moments is negligible, the susceptibility is given by the Curie Law

$$\chi = C/T = N\mu_\mathrm{B}^2 p_\mathrm{eff}^2/3kT \tag{5}$$

where C is the Curie constant, N is the number of magnetic moments per unit volume, $\mu_\mathrm{B} = e\hbar/2mc$ is the Bohr magneton, and p_eff is the effective moment number in μ_B. Curie's Law offers an effective means to determine the magnitude of atomic or molecular magnetic moment in a system. Or, if the moment of interest is known, one can also determine the concentration of the magnetic component. In experiments, one often plots $1/\chi$ versus T. The slope of the straight line provides the information needed.

In many cases, there may exist some sort of interaction between the individual moments. The temperature dependence of χ is then described by the Curie-Weiss Law

$$\chi = C/(T - \theta) \tag{6}$$

where the constant θ is positive (negative) for interaction of ferromagnetic (antiferromagnetic) origin. Equation 6 can be used to determine θ, which is close to the ferromagnetic ordering temperature T_c.

The Curie or the Curie-Weiss Law becomes invalid when the condition of $\mu\mathrm{H} \ll kT$ is not met. This could happen when H is appreciable, or the magnetic moment (μ) of interest is large. The latter is particular relevant for a superparamagnetic system, which consists of a collection of ferromagnetic or ferrimagnetic particles with moments much larger than constituent atomic moments. Under these situations, M becomes the Brillouin function of H, and χ is field dependent.

A nonmagnetic metal or alloy can exhibit a positive and essentially temperature-independent susceptibility when subjected to a field. This phenomenon, called Pauli paramagnetism, is caused by the Zeeman splitting of the spin-up and spin-down conduction bands induced by the field. A thermal equilibrium process leads to the redistribution of the two types of electrons, leaving the system with a net magnetization. Pauli paramagnetic susceptibility is given by

$$\chi = \mu_\mathrm{B}^2 D(E_\mathrm{F}) = 3N\mu_\mathrm{B}^2/2E_\mathrm{F} \tag{7}$$

where $D(E_\mathrm{F})$ is the density of states at the Fermi level (E_F). It should be pointed out that Pauli susceptibility is at least one order of magnitude smaller than the paramagnetic susceptibility of magnetic ions. Discrepancy is often found between experiments and the predicted values according to Equation 7, which is derived from the free electron gas model. In reality, electron-electron or other types of interaction may not be omitted.

Another weak and temperature-independent susceptibility can be obtained from the second-order Zeeman effect, as was first proposed by van Vleck. This susceptibility depends on the energies of the low-lying exited states

$$\chi = 2N \sum |\langle \alpha_i | \mu_z | 0 \rangle|^2 / (E_{\alpha i} - E_0) \qquad (8)$$

where $\langle \alpha_i |$ and $|0\rangle$ represent the excited and the ground states.

The total susceptibility of a material is the sum of Equation 6, Equation 7, and Equation 8, plus the contribution from a negative susceptibility due to diamagnetism.

DIAMAGNETISM

According to Lenz's law of electromagnetism, application of an external field on an electron system induces an electrical current that will generate a field opposing the change of the external field. The microscopic current within an atom or molecule is dissipationless, and will create a magnetic moment vector opposite and proportional to the external field. This so-called diamagnetic susceptibility is a negative constant, given by

$$\chi = -(Ne^2/6mc^2) \sum \langle 0 | r_i^2 | 0 \rangle \qquad (9)$$

where r_i is the electron radius. Diamagnetism exists in all materials. Its susceptibility is small in comparison with Curie paramagnetism. It does not depend on temperature, and is proportional to the total number of electrons in an ion or a molecule. The temperature dependence of various paramagnetic and diamagnetic susceptibilities is shown schematically in Figure 1.

A superconductor displays a unique diamagnetism. The induced supercurrent on the surface of a type I superconductor suppresses completely the magnetic flux density (i.e., $\mathbf{B} = 0$) within the body of the superconductor. This perfect diamagnetism is called the Meissner effect. The susceptibility is, according to Equation 1, $\chi = -(1/4\pi)$ emu/cm^3, which is many orders of magnitude larger than that of ordinary diamagnets. In a type II superconductor, a magnetic field can partially penetrate into the superconductor. Therefore, the susceptibility is smaller than $1/4\pi$ emu/cm^3 in magnitude, and is dependent on the field strength.

Because of the negative susceptibility, there is a repulsive force between any diamagnetic material and the source of the external nonuniform field. A levitating superconductor on a magnet is a famous example for such repulsion. Researchers have also used a magnetic field to levitate liquid helium droplets (Weilert et al., 1997) and biological materials (Valles et al., 1997).

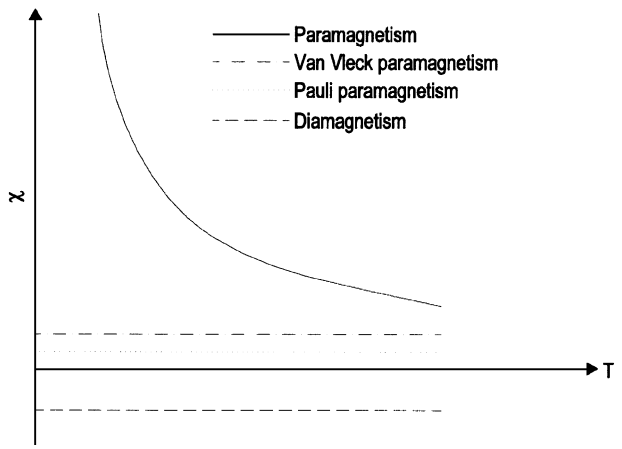


Figure 1. Temperature dependence of susceptibility for paramagnetic and diamagnetic systems.

ANTIFERROMAGNETISM

In an antiferromagnet, magnetic moments at adjacent atomic sites or planes develop a spontaneous antiparallel alignment below a critical phase transition temperature, T_N. The physical mechanism of antiferromagnetism is due to the exchange interaction, which originates from the many-body Coulomb interactions among electrons. Above T_N, the long-range order is destroyed, and the susceptibility is given by the Curie-Weiss Law (Equation 6) with a negative θ. Below T_N, the net magnetization is zero in the absence of an external field, because of the antiparallel configurations of the moments. An external field will induce spin canting and generate a net magnetization, which is proportional to H. In general, there are two susceptibilities, $\chi_\perp$ and $\chi_{||}$, corresponding to the external field perpendicular and parallel to the axis of the spins (easy-axis) respectively.

Figure 2 shows schematically the temperature dependence of susceptibility of an antiferromagnetic crystal. Below T_N, $\chi(T)$ changes abruptly from the Curie-Weiss behavior and branches into $\chi_\perp(T)$ and $\chi_{||}(T)$. As T approaches 0 K, $\chi_{||}$ tends to vanish rapidly. The quantity $\chi_\perp$, on the other hand, remains constant or weakly dependent on T. For polycrystalline or powder materials, $\chi(T)$ falls between $\chi_\perp(T)$ and $\chi_{||}(T)$ as a result of averaging.

FERROMAGNETISM

Ferromagnetic solids are an important class of materials used in a wide range of applications. Many magnetometry

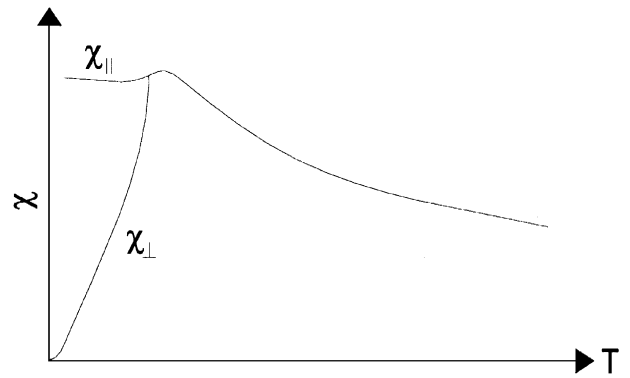


Figure 2. Temperature dependence of susceptibility for an antiferromagnetic system.

techniques have been developed to study the properties of these solids. With exchange interaction as its physical origin, ferromagnetism exhibits a long-range ordering of parallel magnetic moments below a second-order phase transition temperature T_c. Above T_c, the magnetic behavior can be described by the Curie-Weiss Law (Equation 6). From the slope [i.e., the constant C in Equation 5 and Equation 6 of the straight line $1/\chi(T)$ versus T], one can deduce the size of the atomic magnetic moment.

Below T_c, the magnetization is not uniquely dependent on external field H, but rather is determined by the magnetic history of the sample. Experimentally, one either plots M or B versus a cycling H. These plots are called the magnetic hysteresis loops. Since in a ferromagnet M tends to be much larger than H, the two loops (M versus H and B versus H) look similar. Figure 3 shows a schematic hysteresis loop of M versus H and the initial magnetization curve. In the unmagnetized state, the net M is zero because of the magnetic domain structure. As H is increased, M follows the initial magnetization curve until magnetic saturation is reached. Afterwards, the cycling of magnetic field between a positive and negative value causes the M to form a closed loop.

There are a few parameters to characterize the hysteresis loop. The saturation magnetization M_s corresponds to a complete parallel alignment of magnetic moments. The remnant magnetization M_r is the nonvanishing value when H is reduced to zero. The coercivity H_c is the reversing field needed to bring M to zero. The area enclosed by the hysteresis loop is related to the so-called energy product, which is a measure of the prowess of a permanent magnet.

For a permanent magnet, the larger the M_s, M_r, and H_c, the better. In other words, a loop with a large area and a large energy product is preferred. Because of the hysteretic nature of the magnetization process, the definitions of χ and μ according to Equation 3 and Equation 4 are not useful.

For a soft magnet, the H_c is in general very small ($\sim 10^{-2}$ Oe), and so is the loop area. Because of their low energy loss, soft magnetic materials are extensively used in applications involving alternating fields, such as transformers and RF inductors. One often uses modified definitions of susceptibility and permeability to characterize a soft magnetic material. Efforts have been made to enhance permeability for superior performance.

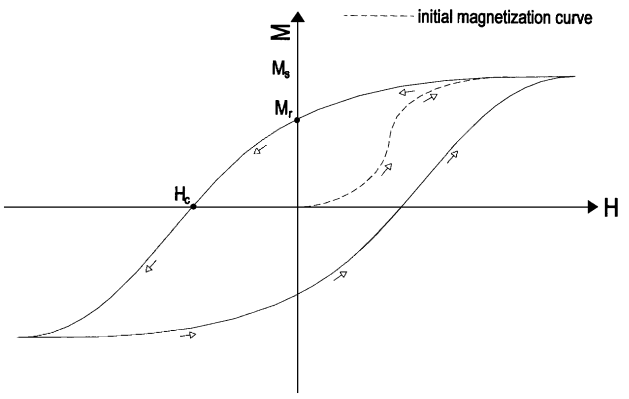


Figure 3. Hysteresis loop for a ferromagnetic system.

In some applications, such as in RF systems, the magnetic field scale is small. The relevant parameter is the initial permeability, defined as

$$\mu = (dB/dH)_{H=0} \tag{10}$$

As **H** increases, $\mu = B/H$ will increase and reach a certain maximum value, which is called maximum permeability μ_M. In some pure and well annealed soft magnetic materials, μ_M can be as large as 10^6. Often permeability data given in the literature refers to μ_M. At each given field, one can also specify a differential permeability $\mu = (dB/dH)_H$.

If a magnet is operated under a bias field ($\mathbf{H}_0$) plus a small alternating field ($\Delta\mathbf{H}$) around $\mathbf{H}_0$, **M**, and **B** will form a minor loop. In this case, one can also define an incremental permeability $\mu_D = (\Delta B/\Delta H)_H$.

When μ is large, as in a typical soft magnetic solid, χ is approximately proportional to $\mu(= 1 + 4\pi\chi \approx 4\pi\chi)$. The aforementioned definitions of various permeabilities can be extended to the corresponding susceptibilities.

LITERATURE CITED

Valles, J. M., Jr., Lin, K., Denegre, J. M., and Mowry, K. L.. 1997. Stable magnetic field gradient levitation of *Xenopus laevis*: Toward low gravity simulation. *Biophys. J.* 73:1130–1133.

Weilert, M. A., Whitaker, D. L., Maris, H. J., and Seidel, G. M. 1997. Magnetic levitation of liquid helium. *J. Low Temp. Phys.* 106:101.

KEY REFERENCES

Anderson, H. L. (ed.). 1989. A Physicist's Desk Reference, pp. 7–10, American Institute of Physics, New York.

Ashcroft, N. W., and Mermin, N. D. 1976. Solid State Physics. W.B. Saunders, Philadelphia.

Chikazumi, S., and Charap, S. H. 1978. Physics of Magnetism. Krieger, Melbourne, Fla.

Cullity, B. D. 1972. Introduction to Magnetic Materials. Addison-Wesley, Reading, Mass.

Kittel, C. 1996. Introduction to Solid State Physics, 7th Ed., John Wiley & Sons, New York.

Morrish, A. H. 1965. The Physical Principles of Magnetism. John Wiley & Sons, New York.

C. L. CHIEN
G. XIAO

GENERATION AND MEASUREMENT OF MAGNETIC FIELDS

INTRODUCTION

In basic science areas, such as semiconductor physics, superconductivity, structural biology, and chemistry, magnetic fields are used as a powerful tool for analysis of matter because they can change the physical state of a

system. In the presence of a magnetic field, the angular momenta (spins), the orbital motion of charged particles, and therefore the energy of the investigated physical system, are altered. We must realize, however, that the size of the induced effects are very modest compared to other physical quantities. One Bohr magneton (typical range of atomic magnetic moments) in a magnetic field of 10 tesla (T) (a 10 T field is ~200,000 times the strength of the earth's magnetic field) has an energy of 0.58 meV, corresponding to a temperature of 6.7 K, a small value in comparison to room temperature. On the other hand, the generation of a magnetic field of 10 T already requires a noticeable engineering effort, with continuous fields of 50 T being the upper limit of what can be achieved today. Pulsed magnetic fields can be generated up to 100 T for durations in the millisecond range, and up to 1000 T on the microsecond time scale. Therefore, experiments must be chosen judiciously. Very often, a low-noise sample environment, such as low mechanical vibrations, small field fluctuations (in the case of dc fields), and low temperatures are decisive to the success of scientific measurements in magnetic fields.

EFFECTS AND APPLICATIONS OF MAGNETIC FIELDS

The use of magnetic fields and their effects on matter are based on a few micro- and macroscopic principles. These principles are utilized not only in basic science areas but also for significant industrial applications (Schneider-Muntau, 1997), and most of them are also employed to measure magnetic fields. These include the following.

1. Magnetization of nuclei and its decay. This effect is the basis for nuclear magnetic resonance (NMR) and magnetic resonance imaging (MRI; NUCLEAR MAGNETIC RESONANCE IMAGING). A considerable industry has developed with numerous applications: NMR spectroscopy, imaging ("scanner") with relevance in medical science, structural biology, and pharmaceutical development, but also for the determination of material properties and nondestructive process and quality control. NMR allows the most accurate measurement of magnetic fields.
2. Induced currents. The electromagnetic induction, i.e., the interaction between conductors and magnetic fields, gives rise to many effects, the most significant applications of which are the generator and electromotor. Magnetic levitation of trains, magnetic propulsion, and magnetic damping of conductive liquids for crystal growth are based on this same effect. Rotating coil gauss meters and extraction flux integrators represent easy ways to measure magnetic fields. Pick-up coils are a standard method for the measurement of pulsed fields.
3. Force on moving charged particles ("Lorentz force"). Accelerators and detectors for particle physics, plasma confinement in fusion devices, isotope separation, ion cyclotron resonance mass spectrometers, and electron beam deflections in TV sets and computer monitors are the most significant applications of this effect. The Lorentz force is, as well, the primary limiting factor in the generation of high magnetic fields, since this force can easily exceed the mechanical strength of the conductors of an electromagnet. The Hall effect (see DEEP-LEVEL TRANSIENT SPECTROSCOPY) is widely implemented for sensors in industry, and it is also the most common device for the measurement of magnetic fields.
4. Magnetization. In a magnetic field, matter is polarized and its magnetic anisotropy results in a force. This effect is employed for separation, alignment of crystals, and levitation. Today's high-field magnets are so powerful that most diamagnetic matter, for instance water and therefore small animals and plants, can be suspended.
5. Energy content. The energy stored in a magnetic field can be effectively utilized in industrial applications. Small systems are available for improvement of the quality of electrical power. Several demonstration projects in the 10-kJ/10-kW range have proven the feasibility of magnets as energy storage systems.

GENERATION OF MAGNETIC FIELDS

Magnetic fields can be generated by magnets created from hard magnetic materials and by electromagnets. Permanent magnets offer a simple, fast, and cost-effective way to provide small-volume fields. *Permanent magnets* are being employed for MRI applications, accelerator storage rings, and wiggler magnets. Fields up to 4 T have been achieved. Devices exist that can produce variable fields or fields with changing orientation. The basic principles of permanent magnets and some applications are described below.

Permanent magnets can also be made from bulk high-temperature superconductors, such as $YBa_2Cu_3O_7$ single crystals. In recent developments it has been shown that a magnetic field of 10 T can be maintained in a mini-magnet of 2-cm diameter, cooled to 42 K (Weinstein et al., 1997).

High and very high fields can only be generated with *electromagnets*. First, some basic principles common to all electromagnets are described, followed by a more detailed treatment of *superconducting magnets*, *resistive magnets*, and *pulse magnets*.

Generation of the highest fields possible is limited by the strength of materials and, for resistive magnets, by the available energy or power source. Because of the resulting magnet size and therefore cost, as well as the required human resources and technical infrastructure, on one hand, and the need for high magnetic fields for research and technology developments, on the other hand, governments have established national facilities as central institutions to serve the scientific community. We provide a short overview of the most important *magnet laboratories* around the world. Each of these laboratories entertains an intensive user program, and access is usually free of charge.

Permanent Magnets

Modern permanent magnets are ideally suited to generate magnetic fields of magnitudes comparable to their spontaneous polarization J_s. Polarization (J), in tesla, is related to magnetization (M) in A/m by the equation $J = \mu_0 M$, where $\mu_0 = 4\pi \cdot 10^{-7}$ Tm/A. In the centimeter-gram-second (cgs) system, M is measured in emu/cm^3 (1 emu/cm^3 = 1 kA/m), $\mu_0 = 1$, and the polarization $4\pi M$ is measured in gauss (1 G = 10^{-4} T); the cgs energy product is measured in megagauss-oersteds (1 MGOe $\approx$ 8 kJ/m^3). The fields can be static or variable, uniform or nonuniform. They are generated by compact permanent magnet structures requiring no continuing expenditure of energy for their operation. The main classes of hard magnets are based on hexagonal ferrites, such as $BaFe_{12}O_{19}$, for which the intrinsic polarization of the pure phase is $J_s = 0.5$ T, and rare earth intermetallic compounds such as $Nd_2Fe_{14}B$, for which $J_s = 1.6$ T. Permanent magnets are now fully competitive with electromagnets for generating fields of up to $\sim$2 T; indeed when a field with a rapid spatial variation is required, they may offer the only practical solution.

Ferromagnetic materials normally have a multidomain structure below their Curie point. The atomic moments within each domain are aligned along some easy direction; the electric currents associated with the quantized spin and orbital moments of the electrons add to produce a net circulating current density of order 1 MA/m, which corresponds to the spontaneous magnetization M_s (1 MA/m $\equiv$ 1.26 T). In the unpolarized state, the moments of the different domains are aligned in different directions and the net magnetization is zero. The domain structure is eliminated by applying an external field, H, which is sufficiently large to align the magnetization of each domain and saturate polarization (Fig. 1). What happens when the field is removed depends on the nature of the material. In a soft magnetic material, the domain structure reforms as soon as the applied field is less than the local demagnetizing field, and hysteresis is negligible; in a hard magnetic material, the polarization remains close to its saturation value well into the second quadrant of the hysteresis loop, and falls to zero at a negative value known as the coercive field, H_c. The hysteresis loop is broad, and for a true permanent magnet the coercive field should be $>M_s/2$.

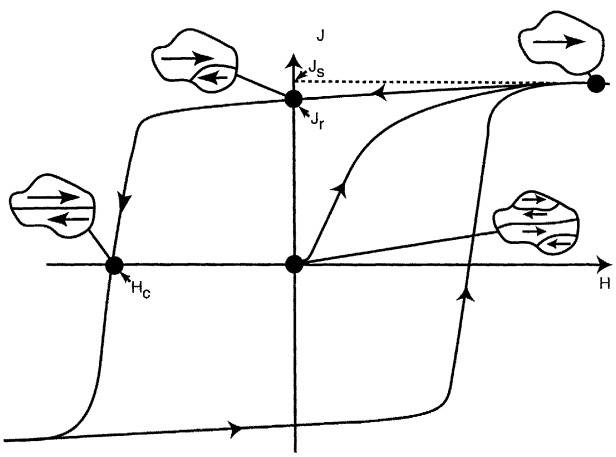


Figure 1. The hysteresis loop of a permanent magnet.

The fully magnetized state is inherently metastable and tends to revert to the equilibrium, multidomain configuration—a process that must be indefinitely delayed in practical magnets. There are two main strategies for achieving this. Permanent magnets are generally constituted of a main phase which has a uniaxial crystal structure (rhombohedral, tetragonal, hexagonal) and strong easy-axis anisotropy so that the magnetization can lie only along the c axis. The anisotropy energy is often represented by an equivalent anisotropy field, B_a. There is, therefore, an anisotropy energy barrier to magnetization reversal. Where the barrier is lowest, a reverse domain may nucleate, and eventually the domain wall may sweep across a particle, completely reversing its magnetization. One approach is to inhibit nucleation of reverse domains by producing a microstructure of tiny ($\sim$1 to 5 μm) crystallites with smooth surfaces, possibly separated by a nonmagnetic secondary phase. The hexagonal ferrites, $SmCo_5$ and $Nd_2Fe_{14}B$ are nucleation-type magnets. The other is to pin any domain walls by local defects associated with nanoscale inhomogeneity in the magnetic material. Sm_2Co_{17} is a pinning-type magnet. The decay of the permanent moment with the logarithm of time is an effect known as magnetic viscosity. It may be of the order of 1% in the first year after magnetizing the material, and cannot be avoided. The medium-term irreversible losses can, however, be exhausted by thermal aging.

The *energy product* $(BH)_{max}$ is the usual figure of merit for a permanent magnet; it corresponds to twice the maximum energy that can be stored in the magnetic field produced around an optimally shaped magnet. In order to maximize the polarization and energy product, the individual crystallites should be aligned with their c axes in a common direction. Details of the intrinsic magnetic properties (Curie temperature, T_C, spontaneous polarization, J_s, and anisotropy field, B_a) of a range of modern permanent magnet materials are listed in Table 1, together with the properties of typical magnets made from them—coercive field, H_c, remanent polarization, J_r, and energy product, $(BH)_{max}$. The polarization of the magnet is normally less than that of the hard phase because of imperfect crystallite alignment, or the presence of nonmagnetic secondary phases needed to achieve coercivity.

All of these magnets can be regarded as transparent, to a first approximation. This means that the polarization of any block of magnetic material is unaffected by the presence of other magnet blocks in its vicinity. Transparency greatly simplifies the design of magnetic circuits. The flux produced by an arrangement of permanent magnets at a point $\mathbf{r}$ is then the product of a geometric factor $f(\mathbf{r})$ and the polarization of the magnets.

The flux density, B, produced by a small magnet of moment $m = MV$ Am^2 is nonuniform and anisotropic. Here $M = J/\mu_0$ is the magnetization, μ_0 is the free space permeability, and V is the magnet volume. The two perpendicular components of the field produced in the plane of a small dipole are

$$B_r = (\mu_o m/4\pi r^3)2\cos\theta \quad \text{and} \quad B_\theta = (\mu_o m/4\pi r^3)\sin\theta \quad (1)$$

Table 1. Properties of Permanent Magnet Materials

Compound	Structure	T_C (K)	J_s (T)	B_a (T)	H_c (MA/m)	J_r (T)	$(BH)_{max}$ (kJ/m^3)
$BaFe_{12}O_{19}$	Hexagonal	723	0.48	1.65	0.28	0.39	28
$SmCo_5$	Hexagonal	993	1.05	40	1.40	0.88	150
Sm_2Co_{17}	Rhombohedral	1100	1.30	6.5	0.90	1.08	220
$Nd_2Fe_{14}B$	Tetragonal	585	1.61	7.6	1.00	1.25	300

In order to generate a uniform field over some region of space, segments magnetized in different directions need to be assembled. An effective way of doing this is to build the structure from long segments with a moment l Am. In this case, the field is given by

$$B_r = (\mu_o l/4\pi r^2)2\cos\theta \quad \text{and} \quad B_\theta = (\mu_o l/4\pi r^2)2\sin\theta \quad (2)$$

so that the magnitude

$$|B| = \sqrt{B_r^2 + B_\theta^2} \quad (3)$$

is now independent of θ, the angle between the radius vector and the magnetization. The dipole ring shown in Figure 2A is an ideal structure where the direction of magnetization in the magnet is given by $\alpha = 2\phi$, with $\phi = \pi - \theta$. The field generated is uniform within the ring and zero outside, provided the ring is very long. The value B_i of the field in the bore for a material with an ideal square hysteresis loop (J_S is the spontaneous magnetization) is

$$B_i = J_s \ln(r_2/r_1) \quad (4)$$

The device illustrates nicely the principle of flux concentration. For example the field generated at the end of a long bar magnet is just J_r, but larger values may be achieved in a magnetic circuit containing segments magnetized perpendicular to the field direction in the airgap. To obtain a field equal to the polarization, we need $r_2/r_1 = 2.7$, but a field of twice J_s needs a radius ratio of 7.4 and 20 times as much magnetic material. Even if materials existed with sufficient coercivity to withstand the large reverse fields acting on some regions of the structure, permanent magnets required to produce multi-tesla fields soon become impractically large.

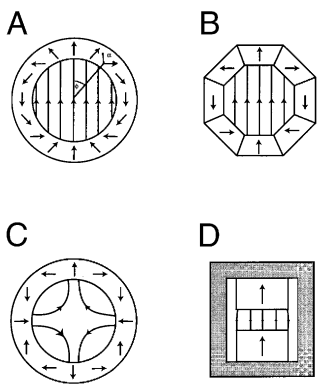


Figure 2. Permanent magnet structures used to generate (**A**) a uniform field, and (**C**) a quadrupole field. The structure shown in (**B**) is a simplified version of (**A**), and (**D**) is a different design, using a soft iron yoke (shaded).

The ring structures can be used to generate higher multiple fields with a different ratio of α and ϕ; for a quadrupole field, for example (Fig. 2C), $\alpha = 3\phi$.

The rings of Figure 2A and C are difficult to realize in practice, and simplified versions with their length comparable to their diameter can be constructed from trapezoidal segments (Fig. 2B). Other designs composed of prismatic segments generate uniform fields in rectangular or triangular bores. It may be advantageous to include soft material in the design to provide a return path for stray flux, as shown in the yoked design of Figure 2D.

These permanent magnet structures are also useful when a rotating magnetic field is needed, or when it is necessary to toggle a field between two directions differing by a fixed angle (such as 180°), in order to reverse the field.

The task of generating a uniform variable field may be undertaken with permanent magnets, provided some movement is permitted. A useful design (Leupold, 1996) consists of two nested dipole rings. The one illustrated in Figure 3A uses cylinders of the type shown in Figure 2B. Dimensions are chosen so that each cylinder produces

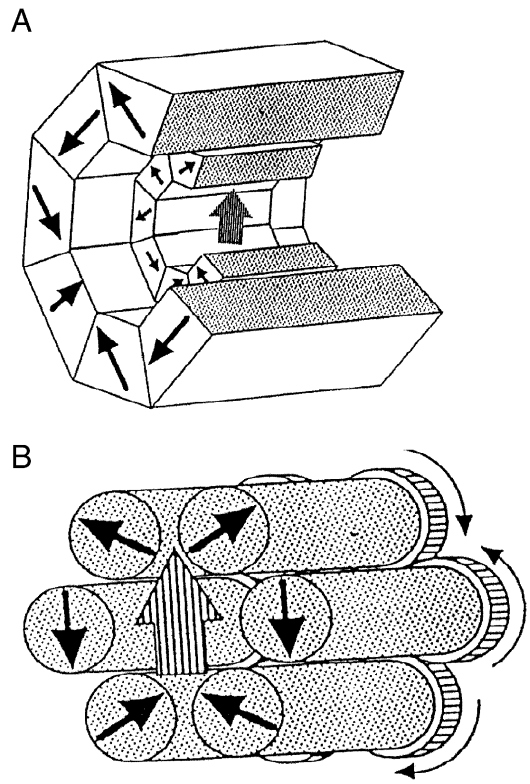


Figure 3. Practical permanent-magnet structures for generating variable fields. (**A**) A double cylinder ("magic cylinder"), and (**B**) a 4-rod design ("magic mangle").

the same field in its bore, and they can be independently rotated about their common axis. In this way, a field in the range $-2B_i < B < 2B_i$ can be produced in any direction in the plane perpendicular to the common axis of the cylinders. The same device can also generate a rotating field of any magnitude from zero up to $2B_i$. When access parallel to the field is desired, a different design, based on rotating cylindrical rods, is preferred. One with six rods is shown in Figure 3B. With four rods, the maximum field available is just J_s (Coey and Cugat, 1994).

There are innumerable applications of permanent magnets as flux sources. They are extensively used in small dc motors and actuators. One example, which accounts for a significant fraction of the global production of $Nd_2Fe_{14}B$, is the voice coil actuator that moves the read head across the surface of a hard disc. Static fields produced by permanent magnets are also used for magnetic separation, magnetic annealing, charged-particle beam control (e.g., wiggler magnets for synchrotron light sources), MRI, NMR spectroscopy, and oil and water treatment, among other applications. In the laboratory, permanent magnet flux sources of this type are the most versatile and convenient way of generating variable fields in the range 1 mT to 2 T. They are compact, and need no cooling water or high-current power supply. However, the limitation on the polarization of existing (Table 1) and foreseeable hard magnetic materials to less than ~2 T will mean that permanent magnet flux sources will only be economically justified for fields <3 T. For fields >3 T, superconducting solenoids with cryogenic or cryogen-free refrigeration will be needed. Fields >20 T call for Bitter or hybrid magnets. For most purposes, however, permanent magnet flux sources are destined to displace the iron-yoked electromagnet from the dominant position it has occupied over the past 150 years.

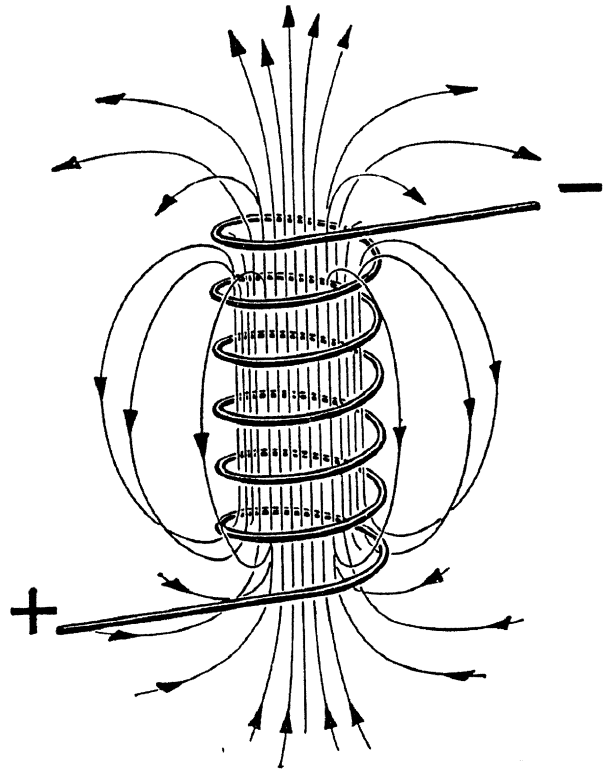


Figure 4. The magnetic field of a coil. (Courtesy of Gerthsen, 1966.)

Electromagnets

In this section, we consider a few basic relationships between current density and magnetic field for a solenoid coil geometry. In the second part, we briefly deal with electromagnetic forces ("Lorentz force") and a simplified description of the resulting stresses.

Figure 4 shows a single-layer coil, wound from a piece of conductor, and the resulting magnetic field as described by the magnetic field lines. Solenoids are coils with many turns and layers, and the current can be described as a current density in a volume element. The law of Biot and Savart gives the field contribution at the center of the coil, i.e., at $r = 0$ and $z = 0$, of a small volume element dV carrying the current density j [as in the literature, j is used here for current density (in units of A/m^2), to be distinguished from polarization J (in units of T) in permanent magnets] in circumferential direction (Fig. 5).

$$\mathbf{H} = \frac{1}{4\pi}\int \frac{\mathbf{r}' \times \mathbf{j}}{r'^3}\, dV \qquad (5)$$

From symmetry consideration it follows that on axis we only have an axial field component, B_z, which we can calculate by integrating over a magnet with inner radius a_1, outer radius a_2, total length $2b$, as shown in Figure 5. From here on, we use the magnetic induction $B = \mu_0 H$ for the magnetic field

$$\mathbf{B} = \frac{\mu_0}{4\pi}\int_0^{2\pi}\int_{a_1}^{a_2} 2\int_0^b \frac{\mathbf{r}' \times \mathbf{j}}{r_3'} r\, dr\, d\varphi\, dz \qquad (6)$$

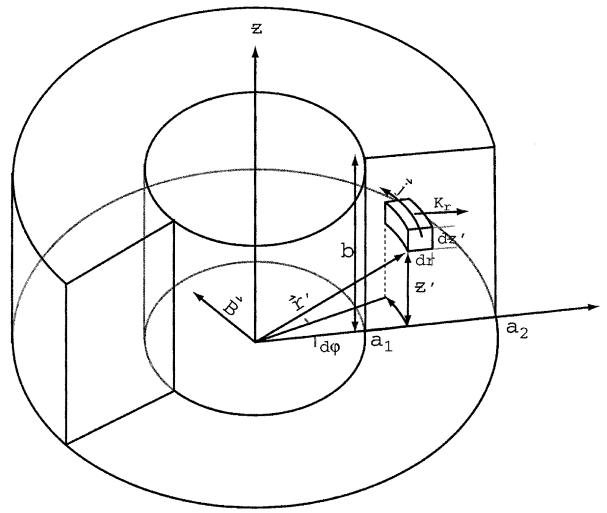


Figure 5. Field contribution of a conductor element.

We obtain for constant current density j over the coil volume the following expression for the magnetic field at the center of the solenoid with $\alpha = a_2/a_1$, and $\beta = b/a_1$.

$$B_z = \mu_0 j \ln \frac{\alpha + \sqrt{\alpha^2 + \beta^2}}{1 + \sqrt{1 + \beta^2}} \quad \text{or} \quad B_z = \mu_0 j a_1 F(\alpha, \beta) \qquad (7)$$

where $F(\alpha,\beta)$ is the so-called field factor that is entirely geometry dependent, i.e., the solenoidal magnet shape can be optimized (Montgomery, 1969). The cross-section of the coil with the overall current density j can be replaced by n turns of wire carrying a current of I occupying only a fraction λ of the total volume. This "space factor," λ, i.e., the ratio between the active section of the winding and its total section, is of great significance, since insulation, reinforcement, and wire shape tend to dilute the wire current density $j_{\text{wire}} = j/\lambda$, and, therefore, the achievable fields. Optimized designs, with the goal of either achieving minimum volume and, therefore minimizing wire cost, or of achieving the highest possible field, will deviate from the above assumption of an overall constant current density and subdivide the winding into regions of different constant current densities to match conductor characteristics with respect to critical current or stress/strain limitations.

Electromagnetic Forces. The design of magnets, especially of high-field magnets, is limited by the Lorentz force, i.e., the stress and strain levels induced by this force, via the coil geometry, on the conductor and the support structure. The precise description of the forces and resulting mechanical loading is extremely complicated and requires a detailed knowledge of the electromagnetic field distribution and the mechanical structure in all three dimensions. Extensive computer programs have been written to solve this issue satisfactorily (e.g., Markiewicz et al., 1999). We present here a simple approach to calculate, in an approximate way, stress levels in a magnet.

The Lorentz force K is a force per volume V (Fig. 5)

$$\frac{d\mathbf{K}}{dV} = \mathbf{j} \times \mathbf{B} \qquad (8)$$

Considering only the z component of the field B_z, a current density in circumferential direction j_t, and no mechanical coupling to other parts of the coil, one obtains a radial force component K_r of

$$dK_r = j_t B_z r \, dr \, dz \, d\varphi \qquad (9)$$

and the hoop stress σ_t of an unsupported current sheet results in

$$\sigma_t = j_t B_z r \qquad (10)$$

For instance, a 10-T solenoid with a current density of 4×10^8 A/m^2 has to support a hoop stress of 200 MPa in its innermost wire layer of 100-mm diameter. Since the magnetic field is proportional to the current density, the Lorentz force increases with the square of the magnetic field. The performance of most magnets and the achievable fields are limited by the materials available to support the induced stress or strain.

Superconducting Magnets

While all metal conductors show a gradual decrease in resistivity with lower temperature, superconductors show zero resistivity very abruptly below a certain temperature, the critical temperature T_c. Another typical property of superconductors is that they have a critical field, B_c, and a critical current density j_c, beyond which superconductivity is destroyed. Superconducting magnets have to be operated within the described B-j-T volume (A-C-E in Fig. 6) since the superconductor transitions into normal resistivity outside this volume. The prevailing conductors are alloys of niobium-titanium and niobium-tin with T_cs of ~12 K and 27 K, respectively (C in Fig. 6). Therefore, the magnets are operated at the temperature of liquid helium (4.2 K) or superfluid helium (1.9 K) (point F). At that temperature the conductor has a lower maximum current density (point D) at zero field, which is again considerably lower in a magnetic field (point G). A magnet is therefore operated along the load line F-G. The new development of high-temperature superconductors (HTS) will extend the useful region toward 100 K; however, at present they are competitive only at low temperatures. Figure 7 compares promising HTS conductors with commercially available low-temperature superconductors.

The generation of magnetic fields through the application of superconductivity can now be considered as a well established technology. Many difficult technical problems have been resolved in the last few years. Thousands of small research magnets are now in routine operation worldwide, producing fields in the range from 5 to 20 T. Both NMR and MRI are the key industrial applications (Killoran, 1997). These applications require extraordinarily high homogeneity of the magnetic field in space and time—typically 10^{-9}—which is achievable solely with superconducting technology. When the magnet is charged to full current, a superconducting switch is closed across the terminals, x, and the power supply is disconnected.

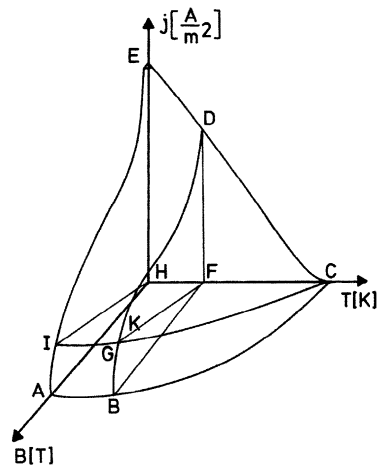


Figure 6. Critical B-j-T volume of a superconductor.

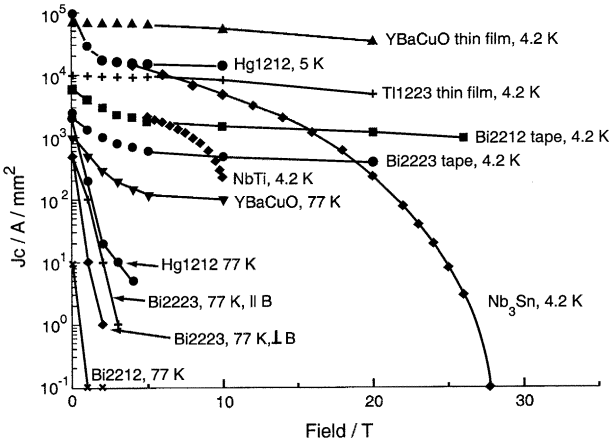


Figure 7. Critical current density within the superconductor of high-temperature superconductors as a function of field for 4.2 K and 77 K. NbTi and Nb_3Sn are plotted for comparison.

The current continues to flow in a closed loop ("persistent mode"), and because of its extremely low resistance—with a decay time of the order of 100,000 years—the magnet is almost a *perpetuum mobile*.

Quench and Training. Superconducting magnets quite often show a performance that is lower than the expectation, which is the field value computed from the short conductor sample characteristics; i.e., the magnets quench prematurely. This is indicated by point K on Figure 6 compared to G. The word "quench" has been coined to describe a process wherein the conductor, starting from a point in a magnet and expanding over the entire volume, passes from the superconducting into the normal state. The stored magnetic energy is dissipated as thermal energy within the coil volume. The heating is not uniform and "hot spots" are generated. The temperature peaks at the origin of the quench and can even result in the melting of the conductor. Modern designs, especially for magnets with high energy content, provide a quench-protection system and even heaters to assure uniform energy dissipation and temperature. Many magnets show a progressive improvement up to a certain level with repeated quenching. This effect is called "training." There is common agreement today that the problem is created by some sudden energy release as the current, and with it the mechanical loading through the Lorentz force, is ramped up. This may originate from some weak spots in the winding structure, cracking of the impregnation under the impact of the induced strain, and small rearranging of the conductor with friction. Because of the extremely small heat capacity of the materials at low temperatures, small energy releases create substantial temperature increases that bring the conductor temperature beyond the critical temperature, T_c. Heat conduction within the winding pack is poor and slow; therefore, the assumption of adiabatic heating is justified. The results necessitate that disturbances must be limited to extremely low values of the order of microjoules per cubic millimeter, corresponding to a maximum sudden wire movement of only a few micrometers. A sound mechanical structure, a precise winding pattern, crack-free epoxy impregnation, a design without slip planes and shear force releases, and a mindful attention to the many mechanical details, have proven to prevent training.

Protection. During a quench, a magnet may be exposed to excessive voltages or temperatures, and the magnet has to be protected against any damage. There are active and passive techniques for doing this. The active technique involves the discharge of the magnet energy into an external dump resistor. In the interest of maximum energy extraction, a small time constant $\tau = L/R_e$ (L is the inductance of the magnet, R_e the external resistor) is preferable. The discharge voltage across the terminals, $V_e = I_m R_e$ (I_m is the magnet current), however, increases and may damage components by arcing. The magnet designer's job is to find a compromise between maximum quench temperature and discharge voltage. This involves the appropriate selection of the operating current (defined by the energy content; $E = 0.5LI_m^2$), the conductor cross-section, and the inductance. A significant feature is a completely reliable quench detection and switch operation system.

Passive techniques are, per definition, reliable. The disadvantage, however, is that all of the energy is dissipated into the magnet volume and the coolant, i.e., liquid helium. The preferred solution is subdivision of the winding into a sufficiently large number of sections that are shunted by a resistor. The resistors can also be used to heat the windings to initiate a distributed quench for a more uniform temperature distribution. Another possibility exists in integrating shorted coupled resistive windings that take up a fraction of the energy. An optimized design will divide the magnet into many subcoils, grade the conductor to suit the local field, construct the coil reinforcement as a wire wound and short circuited overbending for quenchback and heating, and combine it with a quench detection system that dissipates the coil energy in quench heaters. Computer codes have been written that predict quench propagation and coil currents quite reliably by including ac losses (Eyssa et al., 1997). The passive protection has the disadvantage that there is also an energy loss, and therefore increased helium losses, when the magnet is energized or the current setting is altered.

Stability and Losses Due to Changing Fields. The critical-state model describes the penetration of a magnetic field into a superconductor. Because of the superconductor's zero resistivity, screening currents are induced. If the critical current density is surpassed, the screening currents decay resistively to the critical current, allowing the magnetic field to penetrate. All regions of a superconductor are, therefore, carrying either currents at the level of the critical current or no current at all. At full penetration, the current reaches the critical value, I_c. This state may become unstable because the critical current density falls with increasing temperature, and flux motion in the superconductor generates heat. The so-called "adiabatic stability criterion" can be derived from this, with the result that sufficiently thin conductors are stable against flux jumps. In a composite conductor, and above its critical current, the excess current is carried by the stabilizer. A voltage drop is set up which is determined by the stabilizer

resistivity. Copper is usually selected as matrix material as it provides good electrical and thermal conductivity and is essential to protect the conductor against burnout in case of a quench. However, a copper matrix retains the handicap that the filaments are coupled in a changing field via the resistance of the matrix. The solution to this problem consists of twisting the conductors, so that the induced currents cancel over one twist pitch. Twisting is effective against external fields; however, it is not effective against the self-field, induced by the transport current within the conductor. This then leads to the "transport current stability criterion," predicting that wires up to diameters of ~0.5 mm will be stable against flux jumps for the adiabatic case. Further improvements in the theory of stability have been achieved by including the time dependence of heat, current, and magnetic flux, resulting in the so-called "dynamic stability conditions." Obviously, heat conduction and surface cooling are most important for dynamic stability. Experience has shown, nevertheless, that filamentary conductors perform better than theory predicts. As a consequence, modern superconductors are the result of an ongoing development effort, and are tailored to their applications. Figure 8 shows, as an example of the highly advanced wire technology, a bronze route conductor developed especially for high-field NMR magnets.

A detailed knowledge of the induced losses in a changing magnetic field is important for determining operating conditions and the load for the cryogenic system. In superconductors, as in all conductors, changing magnetic fields induce voltages that generate a current flow through the resistance of the current loop. The difference to resistive conductors arises from the fact that the changing magnetic field generates a varying magnetic flux in the filaments. This variable flux results in induced voltages that act on the shielding currents. Heat is dissipated as described above, i.e., the loss mechanism is of a resistive nature. The resistance is highly nonlinear and the loss per cycle is independent of the cycle time. These losses are called hysteresis losses. Clearly, the induced currents, and with it the losses, depend to a high degree on the design of the conductor, the resistivity of the matrix, the resistance between filaments and the matrix ("coupling losses"), the direction of the pulsed magnetic field, and the twist pitch. In addition, we must deal with self-field losses if the transport current changes. Collectively, these losses are known as "ac losses," and their calculation is highly complicated. Many industrial applications (e.g., particle accelerator magnets, superconducting motors and generators, levitated trains, or energy-storage systems) involve changing magnetic fields. In general, hysteresis losses can be reduced by making the filaments as fine as is feasible. Coupling losses can be minimized by shortening the twist pitch, but only if the changing field is perpendicular to the conductor. Self-field losses can be controlled only by reducing the wire diameter.

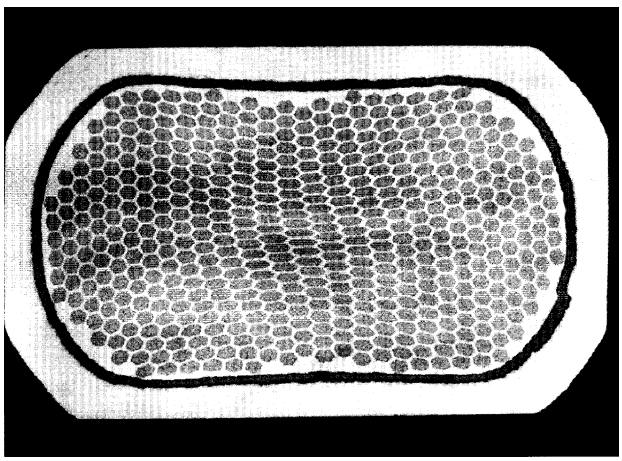


Figure 8. Modern large cross-section (3.17 × 1.97 mm^2) Nb_3Sn conductor with outer stabilizer (25% Cu) and Ta barrier. The conductor contains 123,000 filaments of 4-μm diameter. The critical current density at 20 T and 4.2 K is ~100 A/mm^2, and 200 A/mm^2 at 1.8 K. (Courtesy of Vacuumschmelze, Hanau, Germany.)

Resistive Magnets

Resistive magnets, unlike superconducting magnets, have no constraints, such as critical current density or temperature, imposed on their achievable performance. Resistive magnets require electrical power and, therefore, efficient cooling, but are not limited by any physical effect as superconductors—i.e., in principle it is possible to generate an infinitely high magnetic field. We will see later that power requirements become excessive very quickly, therefore, the generation of very high magnetic fields is only a question of economics.

Resistive magnets are optimized to generate the highest magnetic field possible with the available power. We will present the basic relationships between field, power, and different current density distributions to define the criteria for an optimum design.

Power-Field Relations. A small conductor element dV as shown in Figure 5 with resistivity ρ and current density j will dissipate the following amount of power

$$dW = j^2 \rho \lambda \, dV \tag{11}$$

If we assume a radial dependence of the current density, $j(r)$, but constant resistivity, ρ, and space factor λ, we obtain for a rectangular coil cross-section

$$W = 4\pi b \rho \lambda \int_{a_1}^{a_2} j(r)^2 r \, dr \tag{12}$$

From Equation 6, the same assumptions yield the following:

$$B_z = \mu_0 \int_{a_1}^{a_2} \frac{b}{\sqrt{b^2 + r^2}} j(r) dr \tag{13}$$

Equations 12 and 13 can be combined to the following general form:

$$B_z = \mu_0 G(\alpha, \beta) \sqrt{\frac{\lambda W}{\rho a_1}} \tag{14}$$

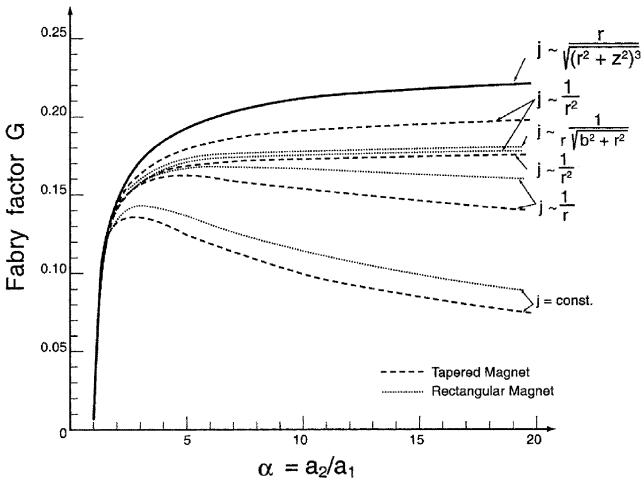


Figure 9. The Fabry factor for different radial current densities. β is optimized for each α value.

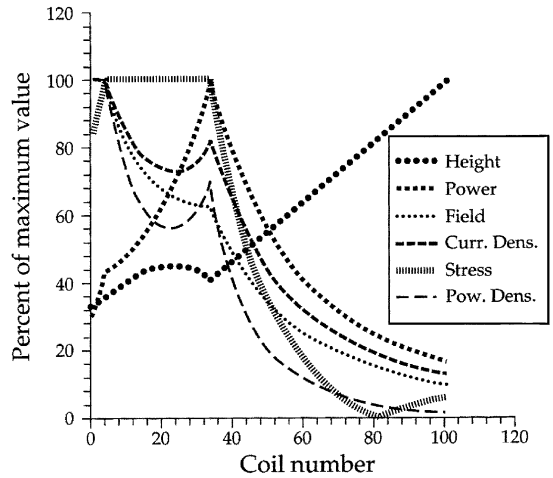


Figure 10. Optimization of a resistive magnet with the constraints of maximum power density and hoop stress.

where $G(\alpha,\beta)$ is the so-called Fabry factor, which is a geometry factor, and B_z is the magnetic field generated at the origin, i.e., in the midplane of the solenoid at $r = z = 0$. Since Equation 14 relates the generated field to the consumed power, it describes the efficiency of the chosen radial current density distribution and solenoid geometry. We can now compare the G-factor of different current density distributions and geometry. Figure 9 shows that, for constant current density and for $j \sim 1/r$, G has a maximum, but that for the other current density distributions, the efficiency can be increased by making the magnet bigger and bigger. Of special interest is the so-called Kelvin distribution (top line in Fig. 9), as it shows the highest efficiency. It is plotted here for a solution that allows the current density to change radially and axially. It is based on the concept that each volume element of the magnet should produce the same amount of field per watt of power input. Practical solutions approximate the optimum value quite closely by subdividing the magnet into several coils and properly selecting the power distribution.

The relationships shown here hold true for magnets with low power levels. High-field magnets have many additional constraints. The power levels are so extremely high that evacuation of the dissipated power through cooling, i.e., the heat flux, is limited by the available cooling surface. Additional cooling requires increased space, no longer available for current conduction, i.e., the space factor λ becomes a function of the power density. The stress levels are so intense that the use of copper alloys is necessitated, i.e., the resistivity ρ also becomes a function of the power level. Additionally, the required reinforcement further increases the space factor. The equations become so complicated that numerical methods are required to solve them. Figure 10 displays the result of a calculation made to illustrate the various constraints. The current density was illustrated only to vary radially. The magnet was split into 100 coils to obtain a quasi-continuous picture. Three regions can be distinguished: (1) the outer region where the optimization has no constraints and finds $j \sim 1/r^2$ and $b = r$, which is the most efficient geometry and power distribution (see Fig. 9); (2) the center region, where the current density is reduced because of the imposed hoop stress limitation; and (3) the inner region where the current density is reduced because the power density is limited for reasons of insufficient cooling surface. An optimum design will attempt to meet both limitations simultaneously. Due to these constraints, resistive magnets become rather inefficient with higher fields, i.e., their power requirements increase exponentially with field.

As result of a newly developed technology, ("Florida Bitter Magnet"), resistive magnets now generate up to 33 T in a 32-mm room temperature bore (Eyssa et al., 1998). These high-field magnets have a power dissipation of up to 20 MW; therefore, efficient and reliable cooling is a prerequisite. Due to the high power densities, the coolant has to be in direct contact with the conductor. Therefore, deionized water of high electrical resistivity is used. Efficient cooling requires high water velocities and a highly turbulent flow to reduce the laminar boundary layer and the temperature drop across this layer as much as possible. Typical values are 20 m/s for the water velocity as upper limit to avoid cavitation. A laminar layer of about only 5-μm results, which accounts for a temperature gradient of up to 40 K. The roughness and its structure determine heat transfer and friction factor, and its optimization is the aim of extended numerical efforts.

Hybrid Magnets. Hybrid magnets combine the advantages of superconducting and resistive magnets. They consist of an outer superconducting coil that provides the background or booster field for the resistive insert magnet. The outsert requires only negligible power compared to a resistive magnet, and power savings of 10 to 20 MW can be realized, depending on bore diameter and field level. Because of the necessary large bore, however, it represents a major capital investment. For the insert, normal, resistive conductors are utilized, as their current-carrying capability is unlimited in comparison to a superconductor. The design of hybrid magnets is a challenging task. The outer magnet has to generate a high field in a large bore

with the consequence of high hoop stresses (as follows from Equation 10). The insert has to withstand not only its own Lorentz forces but also those from the field contribution of the outsert. For economic reasons, the bore size of the outsert is kept as small as possible; therefore, the insert must dissipate its power in a rather small volume. In the case of a misalignment of the magnetic centers of the two systems, forces are generated that are transmitted through the cryostat of the superconducting magnet. In case of a partial burn-out or short-circuit of one of the magnets, these forces can reach the alarming magnitude of several MN. Because of the magnetic coupling, the outsert must also support induced currents arising from sudden field changes from the insert, for instance, in case of an emergency switch-off. Stability requirements and the Lorentz forces lead to a special conductor design for the outsert, the so-called cable-in-conduit conductor, where the conductor is in direct contact with superfluid helium. The stainless steel shell handles the mechanical forces.

There are several hybrid magnets in service worldwide. As of 1999, the highest field (44 T) was generated in the U.S. at the National High Magnetic Field Laboratory (NHMFL), followed by the magnet laboratory in Tsukuba, Japan (37 T). The NHMFL is working on a new insert design that will push the limit even higher, with a goal of generating 50 T in a 32-mm bore.

Pulse Magnets

The generation of pulsed magnetic fields represents an economic avenue to high fields at a small fraction of the cost of continuous magnets. Pulsed magnets are the only way to produce fields above the level of continuous magnets. Disadvantages are the induced voltages and eddy currents in the sample through the changing field. Pulse lengths are typically in the micro- to millisecond range. The use of pulsed fields for research purposes is, therefore, limited to systems where the above effects can be avoided, for instance, in samples of low conductivity. Modern electronics offer fast acquisition time and high resolution.

Devices to generate pulsed fields consist of an energy-storage system that is charged slowly and then discharged rapidly on the magnet. This may be a capacitor bank (typically in the range from 1 to 10 MJ) or a generator in which the energy is stored in the rotating mass of the winding, sometimes complimented by a flywheel. Generators can provide stored energies in the 100-MJ to 3-GJ range. There also exists the option of procuring energy pulses out of the power mains, if the locally available power grid is sufficiently stable.

Discharges from capacitor banks are uncontrolled in the sense that the resulting pulse form is determined by the circuit, i.e., a damped sine wave. Its pulse width is set by the inductance L of the magnet and the capacitance C of the bank, and the damping by the coil and circuit resistance. In first approximation, i.e., for negligible damping, we can write the following equation for the peak current (where V_0 is the voltage of the capacitor bank)

$$I_{\max} = V_0\sqrt{\frac{C}{L}} \tag{15}$$

Similarly, the following equations can be written for the peak power

$$N = V_0^2\sqrt{\frac{C}{L}} \tag{16}$$

and for the pulse-rise time

$$t_0 = \frac{\pi}{2}\sqrt{LC} \tag{17}$$

For the generation of high fields, a high peak current, and equally importantly a high power transfer, from the bank to the coil is needed. This requires a high-voltage capacitor bank and a low-inductance coil. For example, a 10-kV, 1-MJ bank would deliver in excess of 50 kA and 500 MW into an 80-T magnet (Li et al., 1998). Assuming an inner coil diameter of 10 mm, the stress on the conductor will be 2 GPa (Equation 10), i.e., in the range of the strongest materials available today. Often a "crowbar" is added to the discharge circuit to increase the pulse length. At peak current, the capacitor bank is disconnected and replaced by a resistor R, so that the current decays in the LR circuit with a time constant τ of

$$\tau = \frac{L}{R+r} \tag{18}$$

where r is the (time dependent) resistance of the magnet. The total pulse time is then $t \approx t_0 + 3\tau$.

Generators and power supplies fed from the power mains require a unit that controls the energy flow to the magnet and also rectifies the current. The control unit can be used to achieve various pulse forms. Figure 11 gives examples of the pulse shapes that can be generated with the unique 60-T magnet at the user facility of the NHMFL at Los Alamos National Laboratory (LANL).

The upper limit for the above-described technique is considered to be 100 T. There are no known materials that can withstand the Lorentz forces beyond this field level, and magnets inevitably become destructive. To

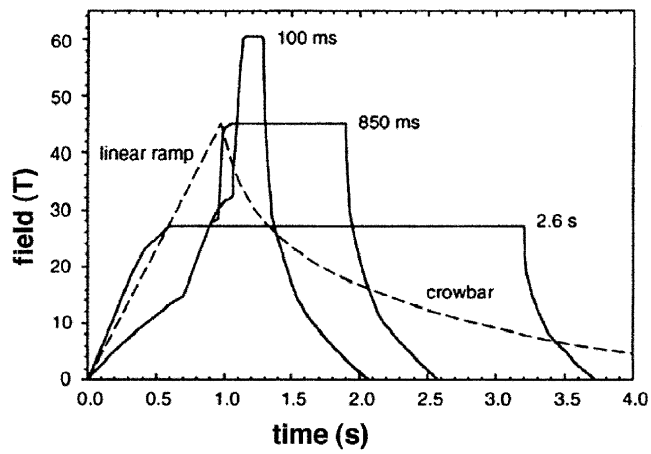


Figure 11. Some of the pulse forms obtainable with a controlled discharge at NHMFL user facility at LANL.

generate even higher fields, two concepts are applied. The first concept is delivering power into a coil faster than the time the coil takes to explode. Fields up to 300 T during microseconds can be created this way by the discharge of a low-inductance capacitor bank into a single turn coil of a few millimeters diameter. Although the coil explodes, because of the symmetry of the explosion, cryostat and sample remain intact. The second concept is to compress a large volume magnetic field into a small volume. Fields up to 500 T can be achieved through electromagnetic flux compression, and well above 1000 T through flux compression with explosives. In both cases, the peak field is achieved in a few microseconds, and the sample is destroyed at this moment.

MAGNET LABORATORIES FOR USERS

It may have become clear from the above that design, construction and operation of high-field magnets is a very challenging and also very costly task, requiring considerable efforts in manpower and installations. It is therefore obvious that a few specialized laboratories of significant size have developed with the goal of serving a broad user community. In the following we give a short overview of the most important laboratories. A detailed review can be found in Herlach and Perenboom (1995).

Continuous Magnetic Fields

It is interesting to realize that the number of laboratories where magnetic fields are used as a research tool has grown considerably over the last several years, accompanied by a more than proportional growth of publications in this area. The strong trend toward higher fields, made possible by new investments (Tallahassee, Tsukuba, Grenoble, and Nijmegen), and the development and evolution of new technologies and materials, will continue. A large and growing user community takes advantage of these central facilities, which are well-equipped with magnets in the 20- to 45-T range, along with the necessary scientific instrumentation (Crow et al., 1996). At present, the development goal is 50 T as the highest continuous field.

Superconducting magnets are now available up to 20 T, and many laboratories employ such a system. We list here only laboratories which can offer continuous fields above 25 T with resistive and hybrid magnets. These are: High Magnetic Field Laboratory, Grenoble, France; Nijmegen High Field Magnet Laboratory, Nijmegen, The Netherlands; High Field Laboratory for Superconducting Materials, Sendai, Japan; National High Magnetic Field Laboratory, Tallahassee, Florida, U.S.A.; and Tsukuba Magnet Laboratories, Tsukuba, Japan.

Pulsed Magnetic Fields

Pulsed magnetic fields represent an economical access to high fields in the 50- to 80-T range. Worldwide, many laboratories provide assistance and support a user program. Experimentation has become easier through the development of new techniques, such as digital data recording and miniaturization. Ultralow temperatures and extreme high pressures have been combined successfully with pulsed fields (Herlach, 1985). The most remarkable development is a 100-T system that will generate pulses in the 10- to 20-ms range, a factor of over 1000 above existing systems (Baca et al., 1999).

There are ~30 laboratories worldwide that do experimental work in pulsed magnetic fields, and which provide assistance to users. We list here only a subgroup, selected from Herlach and Perenboom (1995), which provide fields above 60 T. These are: Fachbereich Physik, Humboldt Universität, Berlin, Germany; H.H. Wills Physics Laboratory, University of Bristol, Bristol, U.K.; Katholieke Universiteit Leuven, Leuven, Belgium; National High Magnetic Field Laboratory, Los Alamos, New Mexico, U.S.A.; Research Center for Extreme Materials, Osaka University, Osaka, Japan; Princeton Plasma Physics Laboratory, Princeton University, Princeton, New Jersey, U.S.A.; National Pulsed Magnet Laboratory, University of New South Wales, Sydney, Australia; Megagauss Laboratory, Institute for Solid State Physics, Tokyo, Japan; Service National des Champs Magnétiques Pulsés, CNRS, Toulouse, France; and Tsukuba Magnet Laboratories, NRIM, Tsukuba, Japan.

A few laboratories provide fields above 100 T. These are: Fachbereich Physik, Humboldt Universität, Berlin, Germany; National High Magnetic Field Laboratory, Los Alamos, New Mexico, U.S.A.; Scientific Research Institute of Experimental Physics, VNIIEF, Sarov, Russia; Pulsed Magnetic Field Installation, Russian Academy of Science, St. Petersburg, Russia; and Megagauss Laboratory, Institute for Solid State Physics, Tokyo, Japan.

MEASUREMENTS OF MAGNETIC FIELDS

The knowledge about the strength and direction of a magnetic field is essential for the following three main purposes.

1. Calibration of a magnet as function of another parameter, typically the magnet current, so that it can be used for its application. The result also tells the magnet engineer how much the constructed magnet corresponds to its design values.
2. Determination of the magnetic environment of a sample to be investigated. Magnetic fields are used as a tool to measure the dependence of some parameters as a function of the applied magnetic field. Therefore, the investigator has to verify, through measurements, the properties of the desired field, e.g., strength, direction, homogeneity, and stability. Equally important for reference measurements is the knowledge of the zero field environment when the magnet is switched off. Magnetization of ferromagnetic parts, for instance, as housing or support structure, can produce noticeable low-field contributions.
3. Fringe fields of magnets have an impact on electronic devices and machinery. The knowledge of their existence is required to initiate and determine the necessary amount of countermeasures, such as

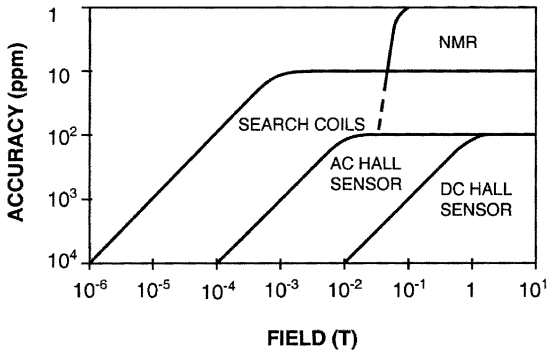


Figure 12. The main measurement methods for magnetic fields and their accuracy and field range. (After Henrichsen, 1992.)

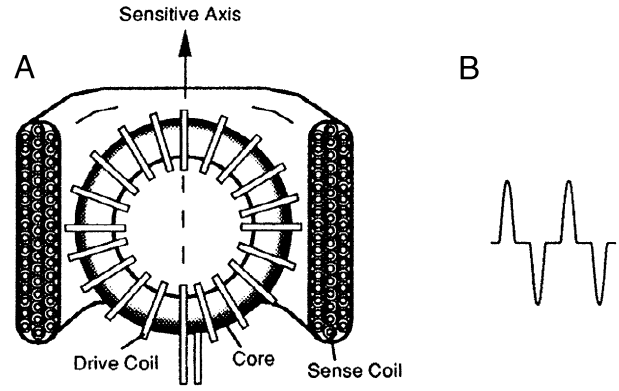


Figure 13. (**A**) Magnetometer with a toroidal core and (**B**) output pulses. (After Fraden, 1993.)

shielding or orientation with respect to the field lines. Fringe fields of high-field magnets can be important, especially with magnets with large bore diameters. Safety measures must be considered for health reasons (pacemakers, for instance), and the force on ferromagnetic parts, like chairs, racks, tools, or gas bottles. The space around a magnet, therefore, has to be cleared carefully of any apparatus and must be cordoned off as a restricted area for the protection of personnel.

The following describes a selection of important measuring devices of magnetic fields. We start the discussion with low-field devices: the compass and the flux-gate magnetometer. We then concentrate on medium-to-high-field measuring techniques: search coils, Hall sensors, and NMR devices. The Hall sensor is the singularly most important, and the most commonly used, device in the measurement of magnetic fields; as such we provide a more detailed description of a new development, the silicon Hall sensor. Figure 12 is a comparison of the most important measurement methods in view of their accuracy and effective useful field range (Henrichsen, 1992).

A large variety of instruments are commercially available for the measurement of magnetic fields.

Low-Field Sensors

Fringe Field Sensors. The existence of low magnetic fields, such as fringe fields of magnets or the fields of magnetized parts, can easily be demonstrated by a simple device: the magnetic compass. The magnetic compass was invented in China about 2500 B.C. and played a role in navigation as early as the 10th century A.D. For a simple demonstration of fringe fields, a screwdriver suspended on a thin thread works well. This can also serve as an indicator of the direction of the field lines and be used for a visual field mapping.

Another simple method to detect low magnetic fields is the monitor of a computer. Black and white cathode ray tube (CRT) monitors change their pattern at ~1 mT; color CRT monitors change their colors at about half that value and therefore, are also a very simple way to check for fringe fields.

Flux-Gate Magnetometer. The flux-gate magnetometer can be utilized for accurate measurements of weak magnetic fields. It is based on a ferromagnetic core with a rectangular hysteresis curve on which detection and excitation coils are wound. Whenever magnetic flux lines are drawn inside the ferromagnetic core, a pulse is induced in the detection coil. A pulse of opposite sign occurs when flux lines are pulled out. The amplitudes of these pulses are directly proportional to the flux intensity; their polarity indicates the direction of the flux lines. The alternate excitation current drives the core into and out of saturation. Figure 13 shows a magnetometer with a toroidal core. It has the advantage that the excitation flux cannot induce current in the sense coil. Flux-gate magnetometers offer linear measurement of the field and its direction, and are especially well suited for the detection of weak stray fields around magnets. Sensitivity can be as low as 20 pT. Commercial devices offer ranges of ~1 mT with a resolution of 1 nT.

High-Field Sensors

Hall Effect Sensors. The Hall effect is a galvanomagnetic effect that occurs by the interrelationship between the electrical currents flowing in a material that is being subjected to a magnetic field (see DEEP-LEVEL TRANSIENT SPECTROSCOPY). It can best be exploited in semiconductor materials in which a high carrier mobility and a low free carrier concentration is achievable. The force F on a free charge carrier with charge e in a semiconductor is driven by the electric field F_E and deflected by the Lorentz force F_L, which acts perpendicular to the direction of the motion and the magnetic flux density

$$\mathbf{F} = \mathbf{F}_E + \mathbf{F}_L = e\mathbf{E} + e[\mathbf{v} \times \mathbf{B}] \tag{19}$$

If a plate-shape piece of semiconductor material in the xy plane carries a current I in the x direction, then an applied flux density B in the z direction produces a transverse electric field in the y direction, the Hall field (Fig. 14).

The integral of this Hall field over the plate can be measured at its outer edges as the Hall voltage, V_H, and

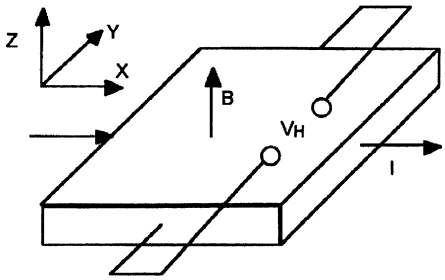


Figure 14. The Lorentz force acting on moving charge carriers in a semiconductor plate causes them to build up an electric field which can be measured by the Hall voltage.

is approximately proportional to the product of the flux density B and the current I

$$V_H = S_H IB \tag{20}$$

Magnetometers/Compound Semiconductor Sensors. A typical Hall effect magnetometer, often called a gaussmeter, incorporates several circuits and controls: a current supply for the sensor, an amplifier with sufficient gain to raise the level of the sensor output to drive an analog/digital converter, zero and calibration adjustments, and analog and/or digital display and voltage output, and either analog or digital signal processing hardware, as well as software to compensate for probe-temperature effects, nonlinearities, and voltage offsets. To date, the most commonly used Hall probes have been based on Hall generators made of compound semiconductors such as InSb and especially InAs. Such generators have a sensitivity S_H (as defined by Equation 20) ranging from 50 mV/AT to 3 V/AT, and enable the gaussmeter to offer full-scale ranges from 30 μT to 30 T, with an accuracy in the 0.1% to 1% range and a precision as high as 10 ppm full scale. Several versions of these Hall elements have been used to measure magnetic fields as high as ~34 T and at temperatures as low as 1.5 K (Rubin et al., 1975, 1986).

Theoretically, a high carrier mobility is crucial for achieving high sensitivity, low offset, and low noise. However, the application of the compound semiconductor Hall elements also implies some drawbacks, such as long-term instability and high-temperature cross-sensitivity. To overcome these difficulties, nowadays high-precision magnetometers are increasingly based on silicon Hall generators.

Silicon Hall Sensors. Although silicon is not a high-carrier-mobility material, it offers many other advantages that make up for this drawback. The doping is usually n-type, with a density around $10^{15}/cm^3$, so that at room temperature the device can be considered as strongly extrinsic. Large-series fabrication of these devices and the application of highly sophisticated and mature silicon technology result in narrow tolerances for the parameters. Moreover, the parasitic effects in these devices are well understood and can be described by simple equations. These characteristics allow efficient means to develop structural and analog electronic compensation of all parasitic effects. A typical value for the proportionality factor (equivalent to sensitivity) S_H, in a modern silicon device is 400 V/AT.

The Integrated Vertical Hall Sensor. Whereas early devices were made as bulk or thin-film discrete components, modern Hall sensors are made using highly advanced semiconductor integration technology, making large numbers of high-quality devices available at low cost. In addition, the sensors can be manufactured as buried devices featuring low noise and very good long-term stability. Their extremely small dimensions make them especially suitable for measurements with high spatial resolution, even in highly inhomogeneous fields.

Besides integrated Hall plates with conventional geometry, the above-described technology also made it possible to implement a sensor measuring the magnetic flux density parallel to the chip surface. In the vertical Hall device (Popovic, 1995), the general plate shape is chosen in such a way that all electrical contacts become available at the surface of the chip. Its geometry evolves from the conventional shape, applying a conformal mapping method where all angles are conserved along with the galvanomagnetic effects mentioned before (Fig. 15). Such devices are adapted to fabrication by integrated technology in an almost perfect way, since the main sources for noise and offset can be reduced to a minimum, yielding highly accurate magnetic sensors. These structures have also been realized to measure two or three components of the magnetic field synchronously in the same spot (Schott and Popovic, 1999), making them suitable for highly demanding tasks like scanning or controlling fields in rather complex magnetic systems.

Even in a modern vertical Hall magnetic sensor, the proportionality between the electrical output signal and the field applied on the device is not ideal. This is due to physical effects within the material of the active zone (charge carrier transport phenomena) as well as the structural composition of the device (depletion zones, surface effects) and the tolerances of the fabrication process itself (mask misalignment, gradients in doping profile). When working at high accuracy, all of these effects have to be taken into account. This can be achieved by applying two different compensation strategies: internal and external. Internal compensation eliminates the influences of offset and of voltage-dependent modulation of the active zone. It is based on mutual cancellation of parasitic effects by adding them with opposite sign.

External compensation eliminates the remaining quadratic temperature cross-sensitivity and the problem of nonlinear behavior in stronger fields. The Hall device follows simple equations with these parameters, so that the analog electronic compensation function can easily be synthesized. Operating the Hall devices in constant

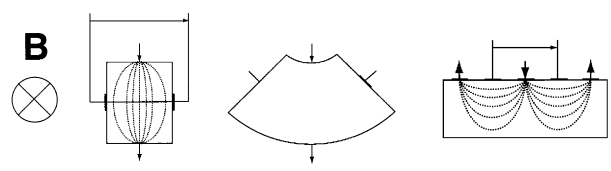


Figure 15. Derivation of the vertical Hall device by conformal mapping of Figure 14.

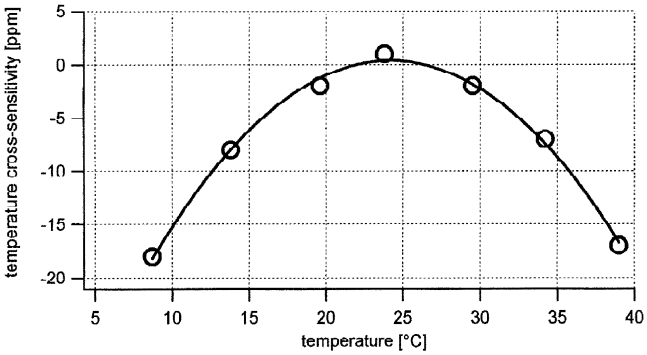


Figure 16. Dependence of sensitivity on temperature. A typical vertical Hall device varies $<5 \times 10^{-5}$ within ± 10 K around room temperature.

current mode allows for implementing all external compensation through bias-current modulation.

Temperature Cross-sensitivity. The Hall factor for electrons in n-type silicon around room temperature increases $\sim 0.1\%$/K. The device's input resistance also rises linearly with temperature. Profiting from this fact, the Hall device itself can be used as a temperature sensor. A simple resistor, put in parallel with the device's input resistance, compensates perfectly for first-order temperature dependence. The less critical remaining parabolic drop of sensitivity with temperature is eliminated by adding a corresponding supplementary current to the biasing current. As a result, sensitivity stays constant over a wide temperature range (Fig. 16).

Offset. Offset in integrated sensors is mainly due to doping nonuniformity and mask misalignment during the fabrication process. According to the Hall device circuit model (Popovic, 1995), offset can be put down to an asymmetry of the resistances between the current contacts and the Hall contacts. Therefore, it is directly correlated to the material resistivity, causing it to rise with temperature and, due to the magnetoresistance effect, also with magnetic field. Since it cannot be distinguished from the output signal resulting from a dc or slowly varying magnetic field, it has to be compensated internally. A first, coarse compensation is provided by the probe's structure (Fig. 17). Two

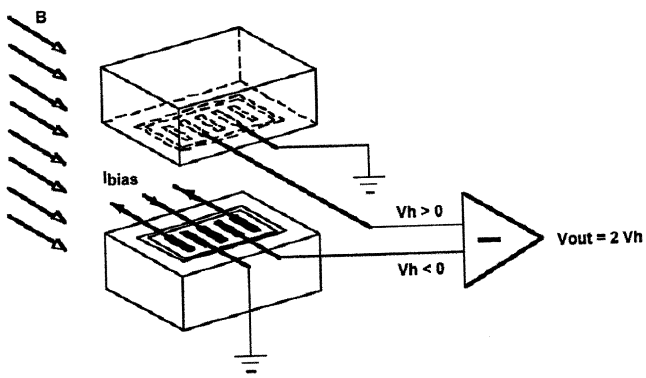


Figure 17. Two Hall devices, oriented in the field so that the Hall voltages are opposed, allow for direct compensation of offset and active zone modulation.

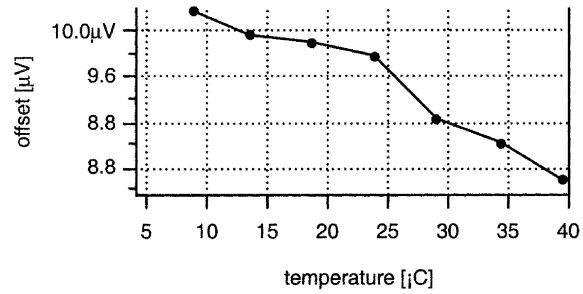


Figure 18. Dependence of the offset on temperature. The equivalent offset field in the specified temperature range from 15 to 35°C is <10 µT.

Hall devices, preferably neighboring elements from the wafer, are subjected to the same parameters during the fabrication process, and therefore, have similar offset. When mounted in such a way that they see an applied field in opposite directions, the value of the measured field can be extracted by subtracting their Hall voltages. This value contains only a very small residual offset, which is the difference in offset between both devices. Compensation of this residual offset makes use of the good correlation between offset drift and input voltage drift with temperature. It consists of subtracting a fraction of the input voltage from the output Hall voltage by electronic means. After the application of these methods, the remaining offset and its drift with temperature are limited to a minimum (Fig. 18).

Nonlinearity. Nonlinearity of Hall devices is described by the variation of sensitivity with the magnetic field H

$$\mathrm{NL} = (S_\mathrm{H} - S_{\mathrm{H}_0})/S_{\mathrm{H}_0}, \qquad \text{where} \quad S_\mathrm{H} = \Delta V_\mathrm{H}/\Delta B \quad (21)$$

Three different sources of nonlinearity can be distinguished. Two of them are technology-independent and described as material nonlinearity and geometry-related nonlinearity. These two depend on the same physical parameters in a similar way, so that they can be used for mutual cancellation in an adapted design. Residual nonlinearity is compensated by modulating the bias current through the Hall elements in a polynomial manner with the applied field, so that field transducers with linearity as good as 0.01% in fields up to 2 T can be produced. The third effect is of specific interest in buried devices like the vertical Hall device. In order to decrease leakage current and border effects, the active zone of the vertical Hall device is surrounded by a depletion layer, formed by a reverse-polarized pn-diode. Its thickness depends on the potential difference between the diode's p and n contacts. This causes the active volume to change in shape and the device to change in sensitivity with a rising Hall or input voltage, just as in a junction-field-effect transistor. However, these effects are significantly reduced when appropriate means for biasing are applied (Schott and Popovic, 1997) or by using two devices seeing the magnetic field in opposite direction.

Magnetoresistive Sensors. There are several materials that change their resistivity in a magnetic field. At low

temperatures, some materials like bismuth change their resistance by a factor of 10^6. The effect is caused by the Lorentz force, which forces the electrons to move on curved paths between collisions and thereby increases the resistivity. In contrast to the Hall effect sensor, the magnetoresistor, when made as a III-V semiconductor (InSb) can be divided into small, long connected strips. This results in a high initial resistance of 50 to 1000 Ω. Only small currents are needed, with the additional advantage that it is only a two-terminal device. Magnetoresistors are inexpensive and have wide industrial applications. They have proven to be useful for scientific measurements; with temperature stabilization an accuracy of 0.1% can be achieved (Weld and Mace, 1970; Brouwer et al., 1992).

Search Coils. Search coils are flux meters that consist of a pick-up coil and an integrator. The pick-up or search coil is exposed to a changing magnetic field. The induced voltage V is proportional to the change of the flux Φ threading the coil, i.e., its number of turns, N, and the coil area, A.

$$V = -NA\frac{dB}{dt} \tag{22}$$

With an NA of 1 m^2, a field change of 1 T will generate 1 V. The coil can be inserted in a field, i.e., moved from a place with $B = 0$ to the place where the field is to be measured, or vice versa. The difference between the two measurements will give a good indication of the accuracy of the chosen setup. In the case of a solenoid, a search coil is moved from a shielded space ($B = 0$) outside the magnet to the center, and the voltage is measured via an integrator. The coil is then moved back into the shielded space, and the voltage is measured again. In a perfect setup, both measurements should give the same result. Obvious requirements are that the induced voltages in the leads be negligible, and that the mechanical positioning of the coil and its angle with the field at the magnet center be precise. There are alternate solutions to the described method, which are to place the coil in the magnet center and ramp the magnet, or to flip the coil around its axis perpendicular to the magnetic field vector.

The accuracy of the search coil method depends on the knowledge of NA and the characteristics of the integrator. Commercially available fluxmeters incorporate state-of-the-art electronic integrators with drifts as low as ±1 μV/min. In addition to dc field measurement, they can measure ac fields at frequencies up to 100 kHz; the accuracy is several percent. For high-accuracy measurements, special attention must be paid to temperature control of the integrator, shielding, low-frequency noise, thermal voltages, and geometry changes of the search coil due to insertion into cryogenic fluids. Very high accuracies may be obtained with differential measurement techniques wherein a pair of search coils is connected in opposition with one coil moving and the other one fixed. Magnet current fluctuations can be eliminated this way. A large variety of search coil geometries and complex harmonic coil systems have been developed. For instance, the flux ball (Brown and Sweer, 1995) makes it possible to perform point measurements in inhomogeneous fields.

A search coil may also be rotated at some frequency, ω, and an ac voltage will develop. The value of B can then be determined by measuring either the peak-to-peak voltage, rms or with a phase-sensitive detector. The complete system is called a rotating-coil gaussmeter and is commercially available.

Search coils or pick-up coils are the ideal way to measure pulsed fields. Because of the high fields and the short times, relatively high voltages are generated and attention has to be given to insulation. For integration, simple passive RC integrators matching the cable impedance can be used.

Nuclear Magnetic Resonance. Nuclear magnetic resonance is a technique that can measure the magnitude of magnetic fields to high precision, down to parts per billion (NUCLEAR MAGNETIC RESONANCE IMAGING). In commercial NMR systems an important subsystem is the field frequency NMR lock. This device uses NMR signals to detect and control the drift of the magnetic field, measure and eliminate variations in the magnetic field intensity across the sample, and measure the average field intensity at the sample position.

NMR measures the motion of nuclear magnetization in a magnetic field. The nuclear magnetization comes from the nuclei (i.e., hydrogen, phosphorus, sodium, and silicon, to name a few) that possess a magnetic moment μ ($\mu = \gamma I$ with γ gyromagnetic ratio, I nuclear angular momentum.) When nuclei having a magnetic moment are placed in a magnetic field, H, the nuclei align parallel to the field. The collection of aligned nuclear moments gives rise to a magnetic moment.

For a magnetic moment M in a uniform and static magnetic field H the equation of motion (Slichter, 1992) is

$$\frac{d\mathbf{M}}{dt} = \gamma[\mathbf{M} \times H] \tag{23}$$

The solutions to this equation are:

$$\begin{aligned} M_y(t) &= M_y(0)\cos(2\pi ft) + M_x(0)\sin(2\pi ft) \\ M_x(t) &= M_x(0)\cos(2\pi ft) - M_y(0)\sin(2\pi ft) \\ M_z(t) &= M_z(0) \end{aligned} \tag{24}$$

where x, y, and z refer to directions in the laboratory frame and the magnetic field is in the z direction. The frequency f is given by $\gamma|H|$. For typical magnetic fields of 1 to 30 T, the frequency is in the radiofrequency (rf) range, 1 MHz to 1 GHz.

The motion of the nuclear magnetization is precession about the magnetic field at the frequency γH. If the precessing magnetization occurs in a coil, a voltage is induced in the coil with the frequency f. The technically simple task of measuring the frequency of the induced voltage and a knowledge of the gyromagnetic ratio, γ, gives a measurement of $|H|$. The gyromagnetic ratio (γ) of a proton is 42.5759 [MHz/T]; it is known to an accuracy of better than 10^{-7}.

In the case of a slightly inhomogeneous field, the equations given above become

$$\begin{aligned} M_y(t) &= [M_y(0)\cos(2\pi ft) + M_x(0)\sin(2\pi ft)]G(t) \\ M_x(t) &= [M_x(0)\cos(2\pi ft) - M_y(0)\sin(2\pi ft)G(t)] \\ M_z(t) &= M_z(0) \end{aligned} \qquad (25)$$

where $G(t)$ is a function expressing the fanning out of the xy magnetization due to the distribution of magnetic fields in this inhomogeneous field. $G(t)$ is a direct measurement of the magnetic field homogeneity in the sample volume.

In the case of a time-varying magnetic field, the frequency given by $\gamma|H|$ will be time dependent. Slowly varying magnetic fields will be seen as a time dependence in the frequency f.

There are two commonly employed interferometric techniques to detect NMR signals: pulse and continuous wave (CW; Fukushima and Roeder, 1981). Both techniques make use of high-stability rf sources, with a stability of 10^{-8}/day.

In pulse NMR, the sample is irradiated with a short, intense rf pulse at the rf source frequency to nutate the magnetization away from the z axis and establish the precessing magnetization. Following the rf pulse, the voltage induced in the coil is detected, amplified, and mixed with the signal from the frequency source to yield a signal at the difference frequency, $f_{rf} - \gamma|H|$ (see NUCLEAR MAGNETIC RESONANCE IMAGING). This signal is digitized and analyzed, providing information about the magnitude, homogeneity, and time stability of the magnetic field.

In CW NMR, the sample is continuously irradiated with a low-level rf signal to maintain a precessing component of the magnetization and an audiofrequency magnetic field to modulate the magnetic field. Two coils aligned in quadrature, or balanced coils, are used to remove or reduce the irradiating rf signal. The resulting signal from the precessing magnetization is amplified and mixed with the rf source signal to yield a signal that is then further demodulated in a lock-in amplifier. The lock-in amplifier output signal can be analyzed for information about the magnitude, homogeneity, and time stability of the magnetic field.

The biggest limitation to using NMR in measuring magnetic fields is that the field must be reasonably homogeneous and stable. The spatial variations, field gradients, temporal variations, and field fluctuations must be less than 0.01%. NMR is not usually done in fields less than 1 T because the signal intensity is proportional to the square of the field. There is no limit to the largest fields that can be measured by NMR. The technical problems of the high frequencies due to large fields are circumvented by choosing nuclei with a lower γ.

LITERATURE CITED

Baca, A., Bacon, J., Boebinger, G., Boenig, H., Coe, H., Eyssa, Y., Kihara, K., Lesch, B., Li, L., Manzo, M., Mielke, C., Schillig, J., Schneider-Muntau, H., and Sims, J. 1999. First 100 T non-destructive magnet. 16th International Conference on Magnet Technology. *IEEE Transactions on Superconductivity*. In press.

Brouwer, G., Crijns, F. J. G. H., König, A. C., Lubbers, J. M., Pols, C. L. A., Schotanus, D. J., Freudenreich, K., Ovnlee, J., Luckey, D., and Wittgenstein, F. 1992. Large scale application of magneto resistors in the magnetic field measuring system of the L3 detector. *Nucl. Instrum. Methods* A313:50–62.

Brown, W. F. and Sweer, J. H. 1995. The flux ball. *Rev. Sci. Instrum.* 16:276–279.

Coey, J. M. D. and Cugat, O. 1994. Construction and evaluation of permanent magnet variable flux sources. *In* Proceedings of 13th International Workshop on Rare Earth Magnets and their Applications, pp. 41–54. University of Birmingham, Birmingham, U.K.

Crow, J. E., Parkin, D. M., Schneider-Muntau, H. J., and Sullivan, N. 1996. The United States National High Magnetic Field Laboratory: Facilities, science and technology. *Physica B* 216: 146–152.

Eyssa, Y., Bird, M. D., Gao, B. J., and Schneider-Muntau, H. J. 1998. Design and stress analysis of Florida bitter resistive magnets. *In* Proceedings of 15th International Conference on Magnet Technology, October 1997 (L. Lin, G. Shen, and L. Yan, eds.). pp. 660–663. Science Press, Beijing, China.

Eyssa, Y., Markiewicz, W. D., and Miller, J. 1997. Quench, thermal, and magnetic analysis computer code for superconducting solenoids. Proceedings of Applied Superconductivity Conference, 1996. Pittsburgh, Pa. *IEEE Trans. Magn.* vol. 7, no. 2, pp. 153–162.

Fraden, J. 1993. Handbook of Modern Sensors. American Institute of Physics, New York.

Fukushima, E. and Roeder, B. 1981. Experimental Pulse NMR. Addison-Wesley, Reading, Mass.

Gerthsen, C. 1966. Gerthsen Physik, 9th ed. Springer-Verlag, Berlin.

Henrichsen, K. N. 1992. Classification of Magnetic Measurement Methods. CERN RD/893-2500, pp. 70–83. European Laboratory for Particle Physics, Geneva, Switzerland.

Herlach, F. (ed.) 1985. Topics in Applied Physics, Vol. 57: Strong and Ultrastrong Magnetic Fields and Their Applications. Springer-Verlag, Berlin.

Herlach, F. and Perenboom, J. A. A. J. 1995. Magnet laboratory facilities worldwide—an update. *Physica B* 211:1–16.

Killoran, N. 1997. Recent progress on high homogeneity magnets at Oxford Instruments. *In* High Magnetic Fields: Applications, Generation, Materials (H.J. Schneider-Muntau, ed.). pp 269–280. World Scientific, Singapore.

Leupold, H. 1996. Static applications. *In* Rare Earth Iron Permanent Magnets (J. M. D. Coey, ed.). Clarendon Press, Oxford.

Li, L., Cochran, V. G., Eyssa, Y., Tozer, S., Mielke, C. H., Rickel, D., Van Sciver, S. W., and Schneider-Muntau, H. J. 1998. High-field pulse magnets with new materials. *In* Proceedings of the VIIIth International Conference on Generation of Megagauss Fields, Tallahassee, Fla. World Scientific, Singapore. In press.

Markiewicz, W. D., Vaghar, M. R., Dixon, I. R., and Garmestani, H. 1999. Generalized plane strain analysis of superconducting solenoids. *J. Appl. Phys.* 86(12):7039–7051.

Montgomery, D. B. 1969. Solenoid Magnet Design. John Wiley & Sons, New York.

Popovic, R. S. 1995. Hall Effect Devices. Adam Hilger (now IOP), New York.

Rubin, L. G., Nelson, D. R., and Sample, H. H. 1975. Characterization of two commercially available Hall effect sensors for high magnetic fields and low temperatures. *Rev. Sci. Instrum.* 46: 1624–1631.

Rubin, L. G., Brandt, B. L., Weggel, R. J., Foner, S., and McNiff, E. J. 1986. 33.6 T dc magnetic field produced in a hybrid magnet with Ho pole pieces. *Appl. Phys. Lett.* 49:49–51.

Schneider-Muntau, H. J. (ed.) 1997. High Magnetic Fields: Applications, Generation, Materials. World Scientific, Singapore.

Schott, C. and Popovic, R. S. 1997 Linearizing integrated hall devices. *In* Proceedings of Transducers, June 1997, Chicago, Ill.

Schott, Ch. and Popovic, R. S. 1999. Integrated 3 Hall magnetic field sensor. Proceeding of Transducers 99, Sendai, Japan. pp. 168–181.

Slichter, C. P. 1992. Principles of Magnetic Resonance. Springer-Verlag, New York.

Weinstein, R., Ren, Y., Liu, J., Sawh, R., and Parks, D. 1997. Permanent high field magnets of high temperature superconductor. *In* High Magnetic Fields: Applications, Generation, Materials (H. J. Schneider-Muntau, ed.). pp 99–115. World Scientific, Singapore.

Weld, E. and Mace, P. R. 1970. Temperature stabilized magnetoresistor for 0.1% magnetic field measurement. *In* Proceedings of 1970 Third International Conference on Magnet Technology. Hamburg, Germany, pp. 1377–1391.

HANS J. SCHNEIDER-MUNTAU,
Florida State University
Tallahassee, Florida

J. M. D. COEY
University Of Dublin,
Trinity College
Dublin, Ireland
(Permanent Magnets)

PHIL KUHNS,
National High Magnetic Field Laboratory
Tallahassre, Florida
(NMR Sensors)

LARRY RUBIN
Francis Bitter Magnet Laboratory
Cambridge, Massachusetts
(Search Coils)

CHRISTIAN SCHOTT
Swiss Federal Institute of Technology
Lausanne, Switzerland
(Hall Sensors)

MAGNETIC MOMENT AND MAGNETIZATION

INTRODUCTION

The intrinsic magnetic properties of materials refer to

1. the origin, magnitude, and directions of atomic magnetic dipole moments;
2. the nature of the interaction between atomic magnetic dipole moments on neighboring atoms in the solid;
3. whether these result in collective magnetic phenomena; and
4. the resulting temperature dependence of the magnetization.

Materials exhibiting collective magnetic response possess additional intrinsic magnetic properties, including the strength of the coupling of magnetic moments to one another and to crystallographic directions, magnetoelastic coupling coefficients, and the temperature(s) at which magnetic phase transformations occur. The intrinsic magnetic properties of species at surfaces and interfaces are known to be distinctly different from those of the bulk in many cases. In the following discussion, we briefly review the theory of intrinsic magnetic properties, specifically of dipole moment and magnetization, as well as theory and examples of collective magnetic response.

MAGNETIC FIELD QUANTITIES

A discussion of the magnetic properties of materials begins with the definition of macroscopic field quantities. The **magnetic induction** (flux density), **B**, is related to the magnetic field, **H**, through the relationship

$$\mathbf{B} = \mu_0\mathbf{H}(\text{in SI units}) \quad \text{or} \quad \mathbf{B} = \mathbf{H}(\text{in cgs units}) \qquad (1)$$

where μ_0 is the **permeability of the vacuum**, which is taken as 1 in cgs units. (Henceforth, we will drop the SI and cgs annotations, which should be understood where two alternative equivalent forms of the same relation are given). In a magnetic material, the magnetic induction can be enhanced or reduced by the material's **magnetization** (defined as dipole moment per unit volume) so that

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \quad \text{or} \quad \mathbf{B} = \mathbf{H} + 4\pi\mathbf{M} \qquad (2)$$

where the magnetization, **M**, is expressed in linear response theory as

$$\mathbf{M} = \chi_m\mathbf{H} \qquad (3)$$

where χ_m is called the "magnetic susceptibility." We can further express the induction $\mathbf{B} = \mu_r\mathbf{H}$ as

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H} \quad \text{or} \quad \mathbf{B} = (1 + 4\pi\chi_m)\mathbf{H} \qquad (4)$$

and we see that the relative permeability μ_r can be expressed as

$$\mu_r = \mu_0(1 + \chi_m) \quad \text{or} \quad \mu_r = 1 + 4\pi\chi_m \qquad (5)$$

Where μ_r represents an enhancement factor of the flux density in a magnetic material. If we have $\chi_m < 0$ we speak of **diamagnetism** and $\chi_m > 0$ (and no collective magnetism) **paramagnetism**.

The definition of magnetization as magnetic dipole moment per unit volume suggests that discussion of a material's magnetization involves identification of the microscopic origin of the dipole moments as well as the

number of dipole moments per unit volume. In the following, we discuss five common origins of magnetic dipole moments. These include

1. dipole moments due to diamagnetic (orbital) currents in closed shells;
2. local magnetic dipole moments due to the Hund's rule ground state of atoms or ions with unfilled outer electronic shells;
3. itinerant electron magnetic dipole moments due to the splitting of spin-up and spin-down energy bands;
4. dipole moments due to atomic clusters (in spin glasses and superparamagnets); and
5. dipole moments due to macroscopic shielding currents in superconductors.

ATOMIC ORIGIN OF MAGNETIC DIPOLE MOMENTS

Closed and Core Shell Diamagnetism: Langevin Theory of Diamagnetism

Diamagnetism is an atomic magnetic response due to closed outer-shell orbits and core electrons. In our discussion of diamagnetism, we consider atomic moments. If we have a collection of like atoms, all of which have the same atomic moments, then the magnetization can be described as:

$$\mathbf{M} = N_a \mu_{atom} \tag{6}$$

where $\mathbf{M}$ is the magnetization vector, N_a is the number of dipole moments per unit volume, and μ_{atom} is the atomic dipole moment vector.

The classical picture of core and closed outer-shell diamagnetism begins with the physical interpretation of a dipole moment. The dipole moment of an orbiting electron has a magnitude IA where I is the "current" in the closed loop electronic orbit and $A = \pi r_2$ is the area swept out by the orbiting electron with orbital radius r. The magnetic dipole moment μ is directed normal to this area. Recognizing that $\mathbf{J}dV = Idr$, and $\mathbf{J}$ is the current density, the dipole moment is equivalently expressed as

$$\mu_{orbit} = \int \mathbf{r} \times \mathbf{J}\, dV \tag{7}$$

Now, if we wish to construct an atomic magnetic dipole moment we must consider the moment due to Z electrons that orbit the nucleus. Assuming that all Z electrons orbit the nucleus with the same angular frequency, ω_0, then we express the current, I, associated with charge $q = Ze$ as

$$I = \frac{dq}{dt} = \frac{Ze\omega_0}{2\pi} \tag{8}$$

The orbital atomic magnetic dipole moment is

$$\mu_{orbit} = \frac{Ze\omega_0}{2} \langle r^2 \rangle \tag{9}$$

where $\langle r^2 \rangle$ is the expectation value of r^2 for the orbit. Finally,recognizing that the angular momentum vector is expressed as $\mathbf{L} = \mathbf{r} \times m\mathbf{v}$, (where m is the electron mass and $\mathbf{v}$ its velocity vector) and has magnitude $mvr = m\omega_0 r^2$ again directed orthogonal to the current loop, then we express a fundamental relationship between magnetic dipole moment and the angular momentum vector

$$\boldsymbol{\mu}_{orbit} = \frac{e}{2m}\mathbf{L} \quad \text{or} \quad \boldsymbol{\mu}_{orbit} = \frac{e}{2mc}\mathbf{L} \tag{10}$$

We can also express a spin dipole moment associated with spin angular momentum.

The previous arguments will give us the ammunition we need to discuss permanent atomic dipole moments that are central to the phenomenon of paramagnetism. In general, however, for closed shells, although electrons do orbit the nuclei, the net current associated with their motion is 0 because of cancellation of their summed orbital angular momentum (i.e., $\mathbf{L} = 0$). On the other hand, even for a closed shell, in the presence of an applied field we do induce a net current. By **Lenz's law** this current results in a moment that opposes the applied field. The characteristic frequency of this circulating induced current is called the "Larmor frequency" and has the value

$$\omega_L = \frac{eB}{m} \text{ or } \frac{eB}{mc} \tag{11}$$

We can calculate the induced moment as the current multiplied by the area, recognizing that different electrons can be in different orbits, and therefore we calculate the expectation value of the square of the orbital radius as

$$\langle R^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle \tag{12}$$

Now for an isotropic environment

$$\langle x^2 \rangle = \langle y^2 \rangle = \langle z^2 \rangle = \frac{\langle R^2 \rangle}{3} \tag{13}$$

where the induced moment is thus given by

$$\mu_{induced} = -\frac{Ze}{2}\left(\frac{eB}{m}\right)\frac{\langle R^2 \rangle}{3} = \frac{Ze^2B}{6m}\langle R^2 \rangle \tag{14}$$

$\langle R^2 \rangle$ is the expectation value of r^2 averaged over all orbits; the negative sign reflects Lenz's law. This may be associated with the negative diamagnetic susceptibility (for N atoms/volume)

$$\chi + \frac{N\mu_{induced}}{B} + \frac{-NZe^2}{6m}\langle R^2 \rangle \tag{15}$$

which well describes the diamagnetism of core electrons and of closed-shell systems. Typically, molar diamagnetic susceptibilities are on the order of $\chi_m = 10^{-6}$ to 10^{-5} cm^3/mol.

Atomic and Ionic Magnetism: The Origin of Local Atomic Moments

In discussion to this point, we have described angular momentum due to orbital motion and spin. Both of these angular momentum terms give rise to magnetic dipole moments and in both cases the magnetic dipole moments are quantized. These magnetic dipole moments will be crucial to the future discussion of magnetism. The physical description of a magnetic dipole moment is in terms of an electron current (circulating) multiplied by the area of the current loop that, in turn, can be related to the quantity $(e/2m)$ multiplied by the angular momentum vector

$$\mu = \frac{e}{2m}\Pi \quad \text{or} \quad \mu = \frac{e}{2mc}\Pi \tag{16}$$

where Π is a generic angular momentum vector. In the classes both of orbital and spin angular momentum the angular momentum is quantized, i.e., its projections along the field axis, z, have an eigenvalue that takes on integral or half-integral values. For orbital angular momentum

$$L_z = m_l\hbar \quad \text{and} \quad \mu = l\left(\frac{e\hbar}{2m}\right)$$
$$l = 0, 1, 2, 3 \ldots \tag{17}$$

The quantity $(e\hbar/2m)$ takes on special significance as the fundamental unit of magnetic moment that is called the "Bohr magneton"

$$\mu_B = \left(\frac{e\hbar}{2m}\right) = 9.27 \times 10^{-24} \text{ amp-m}^2 \tag{18}$$

In addition to the orbital moment of an electron, it is also possible to have an additional contribution to the magnetic moment of an electron, i.e., that due to spin. Spin is a quantum mechanical property, although we can view it semiclassically if we consider the electron as a particle with a rotating surface current. If we considered this problem classically, we would arrive at a spin moment identical to that which is derived in the quantum mechanical description, exactly 1 μ_B. The spin angular momentum is quantized as $S = 1/2$ and its possible projections on the z-axis of the chosen coordinate system (i.e., the field axis) are $m_s = \pm 1/2$. The two spin eigenvalues are denoted spin up and spin down, respectively.

Insight into the filling of atomic orbitals, and especially spectral assignments, can be gained by exploring Hund's rules. This allows us to fully describe the ground state multiplet including the m_l and m_s eigenstates, and allows us to calculate the total orbital, **L**, total spin, **S**, and overall total angular momentum, **J**. The magnitudes of the total orbital and spin angular momenta are constructed as quantum mechanical sums of the angular momentums over a multielectron shell containing n electrons

$$L = \sum_{i=1}^{n} (m_l)_i S = \sum_{i=1}^{n} (m_S)_i \tag{19}$$

The magnitude of the total angular momentum vector $(\mathbf{J} = \mathbf{L} + \mathbf{S})$, J, is given by: $J = |L - S|$ for less than half-filled shells and $J = |L + S|$ for greater than half-filled shells.

To determine the occupation of eigenstates of **S**, **L**, and **J**, we use Hund's rules. For a closed electronic shell $\mathbf{J} = \mathbf{L} = \mathbf{S} = 0$. For an open-shell atom: (1) Fill m_s states $(2l + 1 - \text{fold degenerate})$ in such a way as to maximize total spin, and (2) fill m_l states so as to maximize $|\mathbf{L}|$. To illustrate these rules, consider the ions of the transition metal series, TM^{2+}, i.e., ions that have given up $2s$ electrons to yield the $3d^n$ outer-shell configuration shown in Figure 1. The ground state J, L, and S quantum numbers for rare earth, RE^{3+}, ions are shown in Figure 1.

In conventional paramagnetic response we consider atomic dipole moments that are permanent and arise from unfilled electronic shells. Another type of paramagnetic response, *Pauli paramagnetism*, is associated with

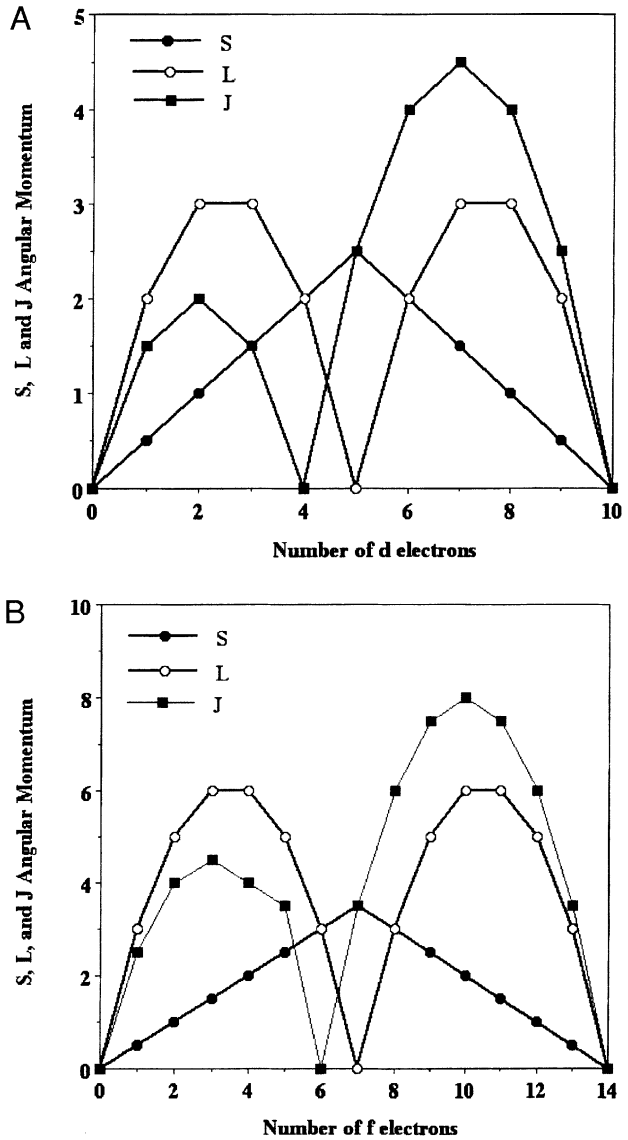


Figure 1. Ground state J, L, and S quantum numbers for the (**A**) transition metal ions, TM^{2+}, and (**B**) rare earth ions, RE^{3+}.

itinerant electrons in energy bands. Paramagnetic response is distinguished from ferromagnetic response in that there is no internal field present to align the permanent local atomic moments in the absence of an applied field. We can understand the origin of the atomic magnetic dipole moment present for paramagnetic atoms by considering again Hund's rules for unfilled shells. The (orbital) angular momentum operator can be expressed as

$$\mathbf{L} = \frac{\hbar}{i}(\mathbf{r} \times \nabla) \tag{20}$$

For a single electron, the magnitude of $\mathbf{L}$ is quantized as in Equation 17, $\mathbf{L} = l[*\hbar(l = 0, 1, 2\ldots)$ and the orbital magnetic dipole moment is also quantized in integer multiples of μ_B. To determine the occupation of states, and therefore $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{J}$, for a given open-shell multielectron atom, we use Hund's rules. Examples of Hund's rules ground states for transition metals (TM) and rare earth (RE) ions are shown in Table 1, including $L = |\mathbf{L}|, S = |\mathbf{S}|$, and $J = |\mathbf{J}|$ angular momenta for each element.

Describing L, S, and J for a given element is termed determination of the ground-state multiplet. This multiplet is written more compactly using the spectroscopic term symbol

$$^{2S+1}L_J \tag{21}$$

where L is the alphabetic symbol for orbital angular momentum ($L = 0 \equiv S$, $L = 1 \equiv P$, $L = 2 \equiv D$, $L = 3 \equiv F$, etc.) and $2S + 1$ and J are the numerical values of the same. For example, Cr^{3+} with $L = 3, S = 3/2$ and $J = 3/2$ would be assigned the term symbol: $^4F_{3/2}$. We can further relate the permanent local atomic moment with the total angular momentum, $\mathbf{J}$, quantum number

$$\mu = \gamma\hbar\,\mathbf{J} = -g\mu_B\mathbf{J} \tag{22}$$

where γ is called the "gyromagnetic factor" and $g = g(J, L, S)$ is called the "Lande g-factor" and is given by

$$g = g(J, L, S) = \frac{3}{2} + \frac{1}{2}\left[\frac{S(S+1) - L(L+1)}{J(J+1)}\right] \tag{23}$$

The Lande g factor accounts for the precession of the angular momentum components (Fig. 2). For identical ions with angular momentum J we can define an effective magnetic moment in units of μ_B

$$p_{\text{eff}} = g(J, L, S)[J(J+1)]^{1/2} \tag{24}$$

For certain solids (e.g., transition metal solids) the orbital angular momentum can be quenched as a result of crystal

Table 1. Ground-State multiplets of Free TM and RE Ions

		Normal State				n_{eff}			
Shell	No. e^-	Ion	S	L	J	Term	Calc. from $g[J(J+1)]^{1/2}$	Obs.	Calc. from $2[S(S+1)]^{1/2}$
	1	Ti^{3+},V^{4+}	1/2	2	3/2	$^2D_{3/2}$	1.55	1.70	1.73
	2	V^{3+}	1	3	2	$^3F_2$	1.63	2.61	2.83
	3	V^{2+},Cr^{3+}	3/2	3	3/2	$^4F_{3/2}$	0.77	3.85	3.87
	4	Cr^{2+},Mn^{3+}	2	2	0	$^5D_0$	0	4.82	4.90
d	5	Mn^{2+},Fe^{3+}	5/2	0	5/2	$^6S_{5/2}$	5.92	5.82	5.92
	6	Fe^{2+}	2	2	4	$^5D_4$	6.70	5.36	4.90
	7	Co^{2+}	3/2	3	9/2	$^4F_{9/2}$	6.63	4.90	3.87
	8	Ni^{2+}	1	3	4	$^3F_4$	5.59	3.12	2.83
	9	Cu^{2+}	1/2	2	5/2	$^2D_{5/2}$	3.55	1.83	1.73
	10	Cu^{+},Zn^{2+}	0	0	0	$^1S_0$	0	0	0
	1	Ce^{3+}	1/2	3	5/2	$^2f_{5/2}$	2.54	2.51	
	2	Pr^{3+}	1	5	4	$^3H_4$	3.58	3.56	
	3	Nd^{3+}	3/2	6	9/2	$^4I_{9/2}$	3.62	3.3- 3.71	
	4	Pm^{3+}	2	6	4	$^5I_4$	2.68	—	
	5	Sm^{3+}	5/2	5	5/2	$^6H_{5/2}$	0.85(1.6)*	1.74	
	6	Eu^{3+}	3	3	0	$^7F_0$	0(3.4)*	3.4	
f	7	Gd^{3+},Eu^{3+}	7/2	0	7/2	$^8S_{7/2}$	7.94	7.98	
	8	Tb^{3+}	3	3	6	$^7F_6$	9.72	9.77	
	9	Dy^{3+}	5/2	5	15/2	$^6H_{15/2}$	10.63	10.63	
	10	Ho^{3+}	2	6	8	$^5I_8$	10.60	10.4	
	11	Er^{3+}	3/2	6	15/2	$^4I_{15/2}$	9.59	9.5	
	12	Tm^{3+}	1	5	6	$^3H_6$	7.57	7.61	
	13	Yb^{3+}	1/2	3	7/2	$^2F_{7/2}$	4.53	4.5	
	14	Lu^{3+},Yb^{3+}	0	0	0	$^1S_0$	0	—	

*Values from van Vleck (1932).

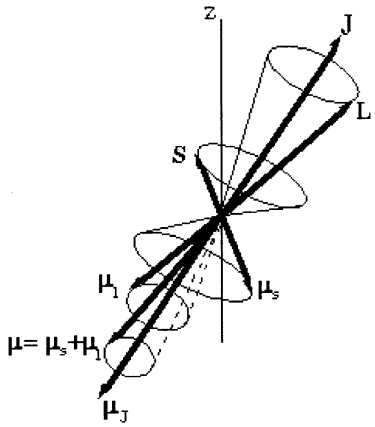


Figure 2. Vector (analogous to planetary orbit) model for the addition of angular momentum with spin angular momentum precessing around the orbital moment that precesses about the total angular momentum vector. The parameter z is the field axis (axis of quantization).

field effects. This implies $L = 0$ and that only spin angular momentum needs to be considered, i.e., $J = S$. In this case $g = 2$ and $p_{\text{eff}} = 2[S(S+1)]^{1/2}$. The addition of angular momentum is conveniently explained by the vector addition diagram of Figure 2.

Collective Magnetism

In magnetic materials, the field and temperature dependence of the magnetization is determined by the nature of the coupling of the atomic magnetic dipole moments. In the case of paramagnetism, the local dipole moments are uncoupled, resulting in a Curie law dependence of the magnetic susceptibility (see Introduction). Ferromagnetic response is an example of correlated or collective magnetism. To describe ferromagnetism we can start with the premise (as was the case for paramagnetism) of **permanent local atomic dipole moments** or consider an **energy band picture of the magnetic state**. Unlike the case for paramagnetic response, however, ferromagnetic response is distinct in that oriented magnetic dipole moments remain coupled even in the absence of an applied field. As a result, a ferromagnetic material possesses a nonzero or **spontaneous magnetization**, $\mathbf{M}_s$, over a macroscopic volume, called a "domain," containing many atomic sites, even for $\mathbf{H} = 0$. Ferromagnetism is said to be a collective phenomenon, since individual atomic moments interact so as to promote alignment with one another. The interaction between individual atomic moments, which gives rise to the collective phenomenon of ferromagnetism, has been explained alternately by two models as follows.

1. **Mean-Field Theory** considers the existence of a nonlocal internal magnetic field, called the Weiss field, which acts to align magnetic dipole moments even in the absence of an applied field, $\mathbf{H}_a$.
2. **Heisenberg Exchange Theory** considers (usually) a local (often nearest neighbor) interaction between atomic moments (spins), which acts to align adjacent moments even in the absence of a field.

Both of these theories will be considered below, since both capture the essence of the phenomenon of ferromagnetism. The two theories do lead to somewhat different pictures of certain aspects of the collective phenomena and the ferromagnetic phase transformation. Both theories lend themselves to quite convenient representations of other collective magnetic phenomena such as antiferromagnetism, ferrimagnetism, helimagnetism, etc. An understanding of the nature of the coupling of magnetic dipole moments allows for the description of the field and temperature dependence of the magnetization as discussed below.

The Band Theory of Magnetism

The **band theory of solids** considers the broadening of localized atomic states with discrete eigenvalues into a continuum of states for more itinerant electrons over a range of energies. The theory allows for calculation of energy dispersion (i.e., energy as a function of wave vector) and orbital-angular-momentum-specific and spin-resolved densities of states. The calculation of spin-resolved energy bands and densities of states allows for the description of atom-resolved magnetic dipole moments and therefore spontaneous magnetization of elemental and alloy magnetic solids. Among the successes of the band theory description of magnetic properties are the following:

1. The prediction of nonintegral or half-integral atomic dipole moments and resulting ground state magnetizations in metals and alloys.
2. The prediction that band widths and **exchange splittings** (energy differences between spin-up and spin-down bands) are intimately dependent on magnetic coordination number and atomic volume.

Table 2 summarizes 0 K and room temperature (where applicable) magnetizations and atomic dipole moments for some important transition metal and rare earth elemental magnets. Also shown are **Curie temperatures** (ferromagnetic ordering temperatures, discussed below) which are not ground-state properties directly calculable from band theory.

A simple schematic view of the band theory of solids and especially the existence of spin-polarized bands exists in the **Stoner theory** of magnetism. The Stoner model of ferromagnetism is equivalent to the free-electron model, with the added feature of spin polarized densities of states.

Table 2. Spontaneous and Room Temperature Magnetizations, Magnetic Dipole Moments, and Curie Temperatures for Elemental Ferromagnets

Element	M_s (Room Tem.)	M_s (0 K)	μ (0 K)/μ_B	T_c (K)
Fe	1707G	1740G	2.22	1043
Co	1400G	1446G	1.72	1388
Ni	485G	510G	0.606	627
Gd	—	2060G	7.63	292
Dy	—	2920G	10.2	88

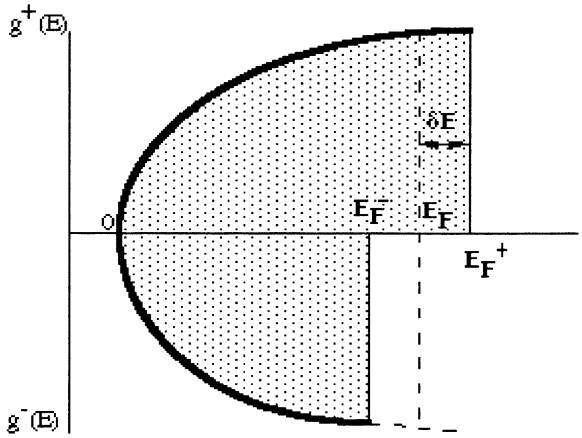


Figure 3. Spin-up and spin-down densities of states in the Stoner theory of ferromagnetism, taken to have a common zero of energy and different Fermi energies in the two spin bands.

With reference to their respective zeroes of energy, the spin-up and spin-down bands, respectively, are depicted in Figure 3, where E_F^+ and E_F^- are the spin-up and spin-down Fermi levels, respectively, E_F is the Fermi level in the non-spin-polarized state and $\delta E = E_F^+ - E_F = E_F - E_F^-$. Letting an exchange interaction energy between spin-up and spin-down electrons be

$$H_{ex} = Un_+n_- \tag{25}$$

where n_+ and n_- are as defined as spin-up and spin-down electron densities as in Equation 31, and U is the exchange energy. The magnetization is $M = (n_+ - n_-)\mu_B$. We define a reduced magnetization, $m = M/n\mu_B = (n_+ - n_-)/n$, and $n = n_+ + n_-$.

ΔE_M, the difference in exchange energy between the magnetic and the nonmagnetic states, can be expressed as

$$\Delta E_M = Un_+n_- - \frac{1}{4}Un^2m^2 \tag{26}$$

and is opposed by an increased kinetic energy:

$$\Delta E_K = \frac{1}{2}(n_+ - n_-)\delta E = \frac{1}{2}nm\delta E \tag{27}$$

where δE reflect the splitting (exhange) between the two spin bands. The total change in energy is given by

$$\Delta E_T = \frac{n^2m^2}{4g(E_F)}[1 - g(E_F)U] \tag{28}$$

so that the criterion for the stability of the magnetic state (determined by minimizing ΔE_T with respect to m) can be expressed as

$$g(E_F)U \geq 1 \tag{29}$$

Finally, the Pauli paramagnetic susceptibility in the Stoner theory is given by

$$\chi = R\chi_0 \tag{30}$$

where $\chi_0 = 2N\mu_0\mu B^2 g\ (E_F)$ is the 0 K susceptibility, N is the number of magnetic atoms per unit volume and $R = [1 - Ug(E_F)]^{-1}$ is the so-called **Stoner Enhancement factor**.

More sophsiticated band structure techniques use the so-called one-electron and local density approximations—e.g., the Layer Korring Kohn Rostoker (LKKR) technique (MacLaren et al., 1989), whose results we consider below. These treat the electronic potential more accurately, but still not at the level of consideration of all individual electron-electron interactions. These result, for example, in densities of states with more structure than in the simple free-electron model and in a more accurate description of the magnetic states in solids. A detailed discussion of band-structure techniques is beyond the scope of this unit. However, Figure 4 illustrates some examples of the results of band theory in describing magnetic dipole moments in solids.

Figure 4A shows the famous **Slater-Pauling curve** that illustrates the variation of the mean atomic magnetic dipole moment as a function of composition in transition metal alloy systems. Figure 4C and D show spin-resolved densities of states $g_+(E)$ and $g_-(E)$ for Co and Fe atoms, in an equiatomic FeCo alloy, as a function of energy (where the Fermi energy, E_F, is taken as the zero of energy). The number of spin-up, n_+, and spin-down, n_-, electrons in each band can again be calculated by integrating these densities of state

$$n_+ = \int_0^{E_F^+} g_+(E)dE \quad \text{and} \quad n_- = \int_0^{E_F^-} g_-(E)dE \tag{31}$$

The Fermi energies, E_F, are the same and the zeroes of energy different for the two spin bands. The atom-resolved (i.e., Fe or Co) magnetic dipole moments can be calculated as (where a unit volume density of states is implied)

$$\mu = (n_+ - n_-)\mu_B \tag{32}$$

Knowledge of atomic volumes or alloy density, then, allows for the direct calculation of the alloy magnetization. Figure 4B shows the band theory prediction of average dipole moment in $Fe_{1-x}Co_x$ alloys as a function of x to be in good quantitative agreement with the experimentally derived Slater-Pauling curve. It should be noted that the moments determined are spin-only moments. Thus, for good quantitative agreement, the orbital moments must also be included. These are expected to be small, however, since the orbital angular momentum is nearly quenched in $3d$ transition metal alloys. Thus, we would expect that the inclusion of the orbital moments could increase the total moment by approximately $0.1 - 0.2\,\mu_B$.

Spin Glass and Cluster Magnetism

Another distinct and interesting magnetic response occurs in **spin glass materials**. Spin glasses can show interesting manifestations of cluster magnetism as well as distinct collective magnetic response. If we consider dilute magnetic impurities in a non-magnetic host, the previously

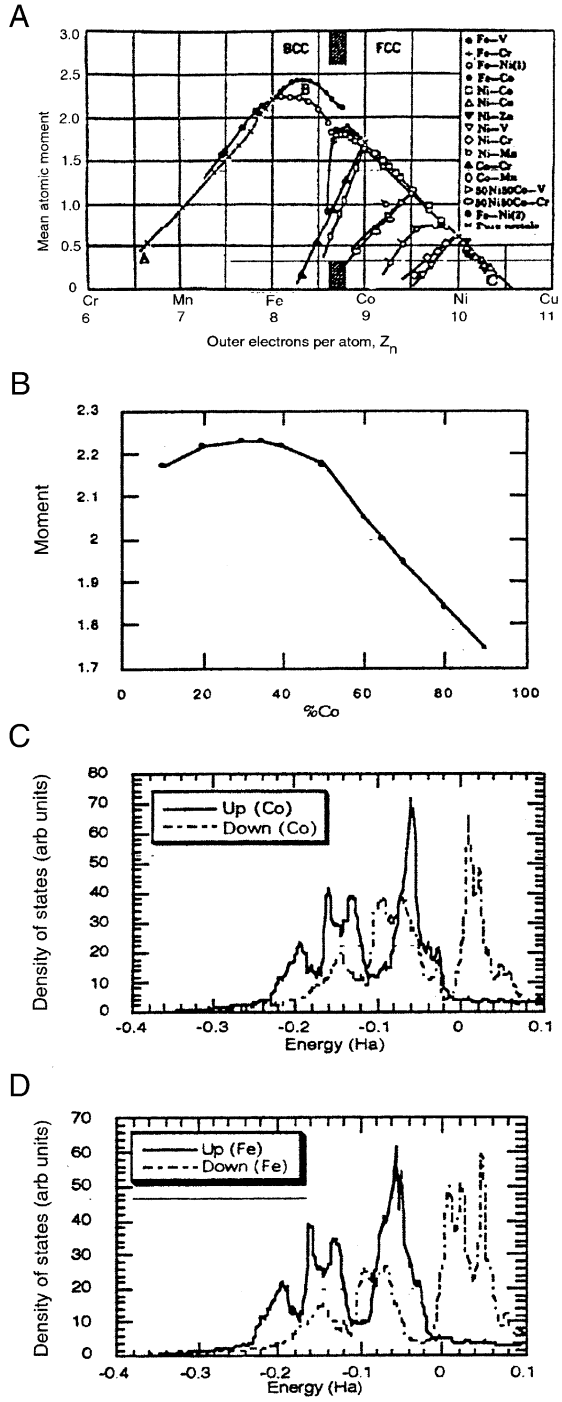


Figure 4. (**A**) Slater-Pauling curve for TM alloys. (After Bozorth.) (**B**) Spin-only Slater-Pauling curve for the ordered Fe-Co alloy as determined from LKKR band structure calculations. Density of states for Co (**C**) and Fe (**D**) in the ordered equi-atomic alloy (MacLaren et al., 1999). The Fermi level is taken as the zero of energy.

discussed theory of paramagnetism can be used to describe local moments and the temperature dependence of the magnetization for noninteracting moments. If, however, the local dipole moments couple though exchange or dipolar interactions, then a distinct "frozen" magnetic state can arise with cooperative or collective magnetism, which has features similar to ferromagnetic or antiferromagnetic states, but with other manifestations that are decidedly distinct. The current view in the evolving description of spin glasses is that the glassy or frozen state constitutes a new state of magnetism.

A spin glass is a magnetic system in which the geometric arrangement of magnetic ions, frozen in structural disorder, and/or random exchange interactions result in "frustration". Frustration refers to the inability of the magnetic moment configuration on a given magnetic site to simultaneously satisfy the constraints of exchange interactions with neighboring atoms. As a result, traditional long-range ferromagnetic or antiferromagnetic long range order is not possible in spin glasses. Below a freezing (spin glass) transition temperature, T_g, an ordered state is said to exist. The precise description of this state is still controversial, although a variety of universal experimental observations identify the spin glass state. These include the following.

1. A large difference between the field cooled and zero-field cooled magnetizations (in low fields) for temperatures below the spin glass freezing temperature, T_g.
2. Strong magnetic hysteresis below the spin glass transition temperature.
3. A peak (cusp) in the alternating current (ac) magnetic susceptibility.
4. A change in the magnetic heat capacity at T_g.

Spin glass behavior can also involve interactions between clusters of magnetic atoms with a distinct "cluster moment" as opposed to an atomic dipole moment. In such a case the previous arguments made based on atomic dipole moments need only be modified to consider cluster dipole moments. Figure 5 illustrates several experimental manifestations of spin glass response in an Al-Mn-Ge quasicrystalline alloy. These include different field cooled (FC) and zero-field cooled (ZFC) susceptibilities, magnetic relaxation, a cusp in the ac susceptibility, and magnetic contributions to the specific heat below the spin glass freezing temperature.

Magnetization of Superconductors

Superconductivity is the unique and complicated phenomenon in which the "free" electrons in a conductor pair up and move cooperatively below a critical temperature, T_c, called the "superconducting transition temperature." This pairing constitutes a magnetic phase transition involving only the electrons in the superconducting material (i.e., atom positions do not change below T_c). Paired electrons are able to correlate their motion with the vibrating crystalline lattice (phonons) so as to avoid electronic scattering by the lattice. This is accompanied by remarkable changes in both the electrical transport and the magnetic properties of the superconducting material. Notable among these changes are the loss of electrical resistivity below T_c, and exclusion of magnetic flux (the **Meissner effect**).

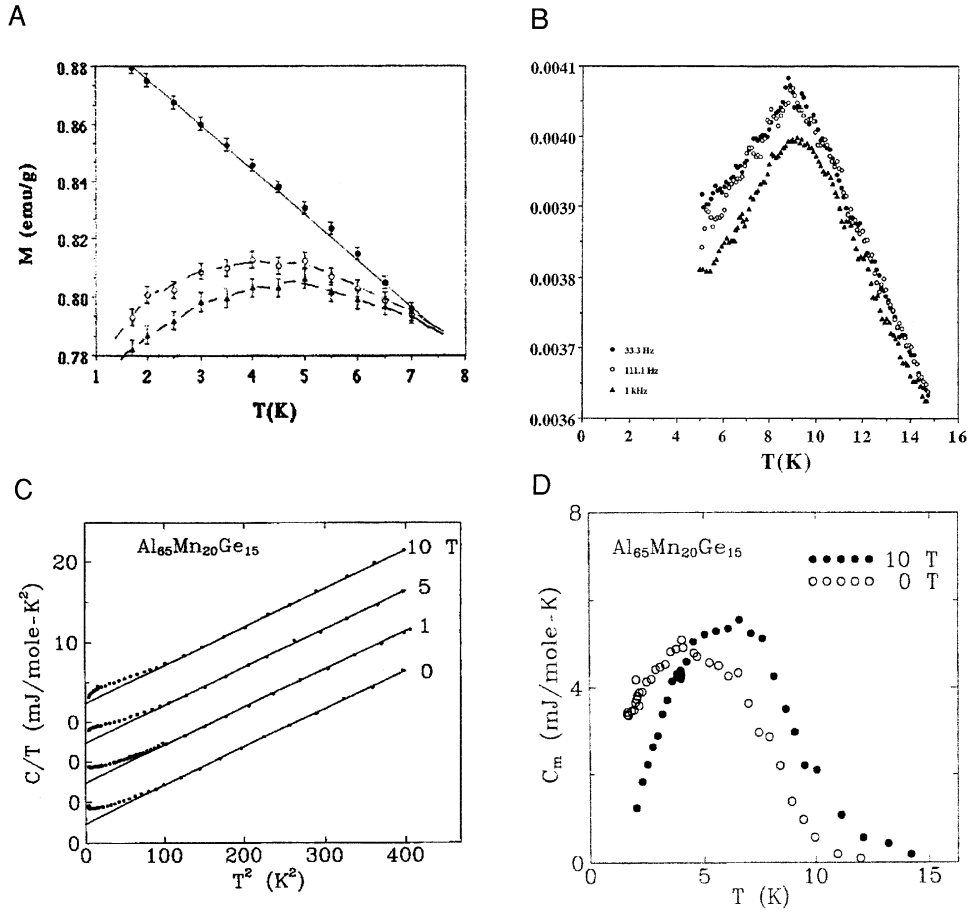


Figure 5. (**A**) Low T field cooled (filled circles) and zero field cooled [after 1 s (filled triangles) and 180 s (open circles)] susceptibilities for $Al_{65}Mn_{20}Ge_{15}$ showing spin glass freezing at a temperature $T_g \approx 8$ K. (**B**) The real component of the ac susceptibility, χ', (arbitrary units) as a function of T, $H = 10$ Oe rms and $\omega = 33.3$, 111.1 and 1000 Hz, respectively, showing a cusp (or maximum) in $\chi'(T)$. (**C**) Specific heat versus T plotted as C/T versus T^2 at 0, 1, 5, and 10 T. The data have been vertically offset for clarity. Solid lines indicate the nonmagnetic background specific heat; (**D**) The magnetic specific heat C_M plotted versus T at $H = 0$ (open circles) and 10 T (solid circles). Data from Hundley et al. (1992).

A fundamental experimental manifestation of superconductivity is the Meissner effect (exclusion of flux from a superconducting material). This observation preceded and motivated the development of London electrodynamics (London and London, 1935; London, 1950). The Meissner effect identifies the superconducting state as a true thermodynamic state and distinguishes it from perfect conductivity. This is illustrated in Figure 6, which compares the magnetic flux distribution near a perfect conductor and a superconductor, respectively, for conditions where the materials are cooled in a field (FC) or for cooling in zero-field (ZFC) with subsequent application of a field. Note the difference between flux density distributions for a perfect conductor and superconductor when field cooled. After zero-field cooling, however, the superconductor and perfect conductor have the same response. Both the perfect conductor and superconductor exclude magnetic flux lines in the ZFC case. This can be understood as resulting from diamagnetic screening currents which oppose dB/dt in accordance with Lenz's law. It is the first case, that of field cooling, that distinguishes a superconductor from a perfect conductor. As illustrated, the flux profile in a perfect conductor does not change on cooling below a hypothetical temperature where perfect conductivity occurs (since there is no dB/dt). On the other hand, a superconductor expels magnetic flux lines upon

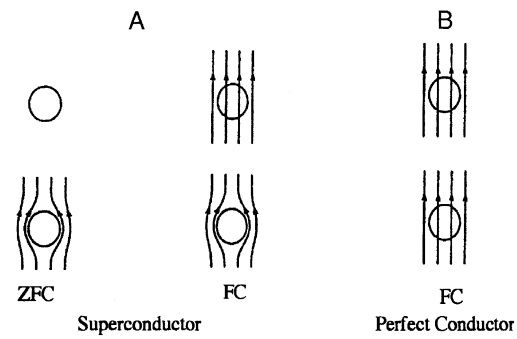


Figure 6. Flux densities for a superconductor, (**A**), zero-field cooled (ZFC) or field-cooled (FC) to $T < T_c$ and for a perfect conductor, (**B**), which is field cooled (FC).

field cooling, distinguishing the superconducting state from perfect conductivity. The extent to which field is excluded in a real type II superconductor (see below) is determined by magnetic flux pinning. In a clean superconductor (i.e., without pinning), flux is excluded to maintain $B = 0$ in the superconductor. This Meissner effect implies that:

$$\mathbf{B} = 0 = \mathbf{H} + 4\pi M \tag{33}$$

or that the magnetic susceptibility, $\chi = M/H = -1/4\pi$ for a superconductor.

Equilibrium superconductors are also distinguished by their critical magnetic field, H_c. At any temperature below the critical temperature, there is a critical magnetic field above which the superconductor becomes a normal conductor with finite resistance. Thus, if a magnetic field is applied to the superconducting sample, the shielding currents increase to maintain perfect diamagnetism and, at a critical value of the magnetic field, the sample becomes a normal conductor. The critical magnetic field of a superconductor varies with temperature, reaching a maximum at 0 K.

It should be noted that two distinct types of superconductors have been found. Until this point we have focused on type I superconductors, which lose superconductivity at $H_c(T)$. In a type I superconductor, perfect diamagnetism is observed until $H = H_c(T)$ where superconductivity (and diamagnetism) is abruptly lost. There is a second type of superconductor that goes through a mixed state with the application of a field, where some areas are superconducting while regions at the core of magnetic vortices (flux lines) are normal. These are called type II superconductors and the higher-temperature superconductors belong to this group. Type II superconductors were not recognized as being another class of superconductors until the work of Abrikosov (1957). A type II superconductor has perfect diamagnetism until a critical field called the lower critical field, H_{c1}. However, unlike a type I material, superconductivity does not disappear at H_{c1}. Instead, the superconducting and normal states coexist (by virtue of entrance of quantized magnetic vortices—flux lines—into the material) in equilibrium until another critical field, the upper critical field, H_{c2}, is reached where superconductivity is lost entirely.

The coexistence of the superconducting and normal states in the so-called mixed state is accomplished by letting single (quantized) magnetic flux lines into the sample for $H > H_{c1}$. Figure 7 shows the magnetic phase diagram for a type I and type II superconductor. In practical superconductors, it is important to pin these magnetic flux lines so that they do not dissipate energy by motion resulting from a Lorentz force interaction with the supercurrent density. In the newly discovered high-temperature superconductors, anisotropic crystal structures and short superconducting coherence lengths make a **vortex melting transition**, $H^*(T)$, possible that adds further complexity to the H-T phase diagram.

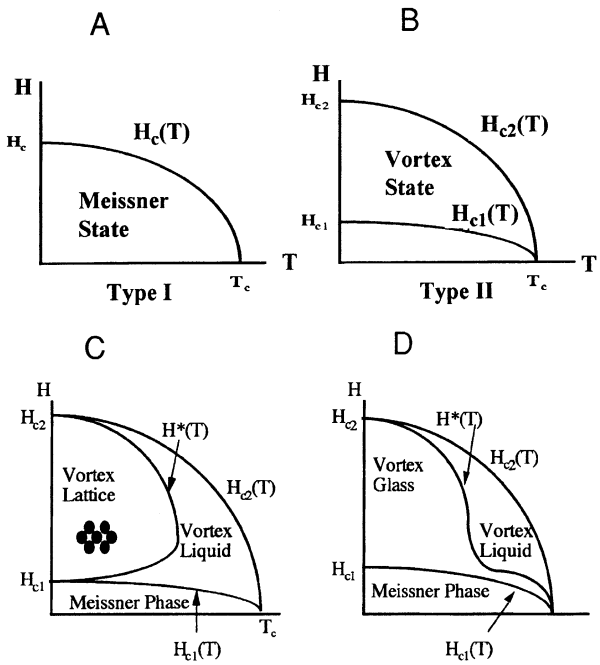


Figure 7. (**A**) Magnetic phase diagram for a type I superconductor; (**B**) magnetic phase diagram for a type II superconductor; (**C**) magnetic phase diagram for a type II superconductor with vortex melting to a vortex liquid state; and (**D**) magnetic phase diagram for a type II superconductor with a vortex glass state (McHenry and Sutton, 1994).

COUPLING OF MAGNETIC DIPOLE MOMENTS

Classical and Quantum Theories of Paramagnetism

Paramagnetism results from the existence of permanent magnetic moments on atoms as discussed above. In the absence of a field, a permanent atomic dipole moment arises from the incomplete cancellation of the electron's angular momentum vector. In a paramagnetic material in the absence of a field, the local atomic moments are uncoupled. For a collection of atoms in the absence of a field, these atomic moments will be aligned in random directions so that $\langle \mu \rangle = 0$ and therefore $\mathbf{M} = 0$. We now wish to consider the influence of an applied magnetic field on these moments. Consider the induction $\mathbf{B}$ in our paramagnetic material arising from an applied field $\mathbf{H}$. Each individual atomic dipole moment has a potential energy in the field $\mathbf{B}$ given by

$$U_p = -\mu \cdot \mathbf{B} = -\mu B \cos\theta \tag{34}$$

The distinction between the classical and quantum theories of paramagnetism lies in the fact that continuous values of θ, and therefore continuous projections of $\mathbf{M}$, on the field axis are allowed in the classical theory. In the quantum theory only discrete values of θ and projected moments, μ, are allowed, consistent with the quantization of angular momentum. The quantum theory of paramagnetism is discussed in the next section. Notice that in either case the potential energy predicted by Equation 34 is minimized when local atomic moments and induction, $\mathbf{B}$, are parallel.

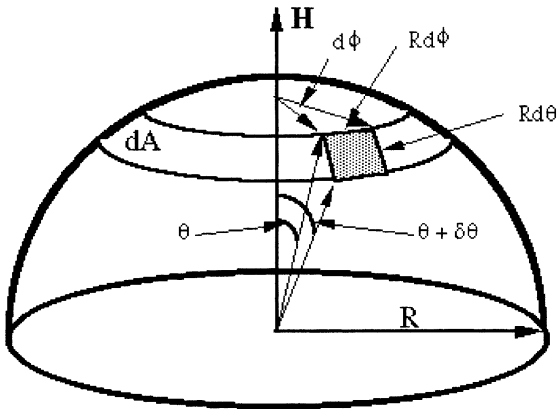


Figure 8. Distribution of magnetic dipole moment vector angles with respect to the field axis.

We are now interested in determining the temperature-dependent magnetic susceptibility for a collection of local atomic moments in a paramagnetic material. In the presence of an applied field, at 0 K, all atomic moments in a paramagnetic material will align themselves with the induction vector, **B**, so as to lower their potential energy, U_{p}. At finite temperatures, however, thermal fluctuations will cause misalignment, i.e., thermal activation over the potential energy barrier leading to a temperature dependence of the susceptibility, $\chi(T)$. To determine $\chi(T)$ for a classical paramagnet, we express the total energy of a collection of atomic magnetic dipole moments as

$$U_{\mathrm{p}}^{\mathrm{total}} = \sum_{atoms,i} -\mu_{\mathrm{m}}^{\mathrm{atom}}(\mu_0 \mathrm{H})\cos\theta_{\mathrm{i}} \tag{35}$$

where θ_{i} is the angle between the ith atomic dipole moment and B. If we consider (1) a field along the z axis, and (2) the set of vectors $\mathrm{n}_i = \mu/|\mu_i|$, which are unit vectors in the direction of the ith atomic moment, then to determine $\chi(T)$ we wish to discover the temperature distribution function of the angle θ_i. The probability of a potential energy state being occupied is given by Boltzman statistics

$$p = C\exp\left[\frac{-E_p}{kT}\right] = C\exp\left[\frac{-\mu_{\mathrm{m}}^{\mathrm{atom}}(\mu_0 \mathrm{H})\cos\theta}{k\mathrm{T}}\right] \tag{36}$$

where $p = p(E_{\mathrm{p}}) = p(\theta)$. As shown in Figure 8, the number of spins for a given θ at $T = 0\,\mathbf{K}, H = 0$ is given by $dn = 2\pi sin\theta\, d\theta$, since all angles are equally probable. At finite T, H

$$dn = C\exp\left[\frac{-\mu_{\mathrm{m}}^{\mathrm{atom}}(\mu_0 \mathrm{H})\cos\theta}{kT}\right]2\pi\sin\theta\, d\theta \tag{37}$$

and

$$N = \int_0^{\pi} dn \tag{38}$$

gives the number of spins. We calculate the average projected moment (along the axis of **B**) as

$$\langle\mu_{\mathrm{m}}^{\mathrm{atom}}\rangle = \frac{\int_0^{2\pi}\mu\cos\theta\, dn}{\int_0^{2\pi} dn} = \frac{\int_0^{2\pi} C\exp\left[\frac{-\mu_{\mathrm{m}}^{\mathrm{atom}}(\mu_0 \mathrm{H})\cos\theta}{kT}\right]\mu_{\mathrm{m}}^{\mathrm{atom}}\cos\theta\cdot 2\pi\sin\theta d\theta}{\int_0^{2\pi} C\exp\left[\frac{-\mu_{\mathrm{m}}^{\mathrm{atom}}(\mu_0 \mathrm{H})\cos\theta}{kT}\right]2\pi\sin\theta d\theta} \tag{39}$$

and using the substitution $a = -\mu_{\mathrm{m}}^{\mathrm{atom}}\,\mu_0 \mathrm{H}/kT$, then

$$\frac{\langle\mu_{\mathrm{m}}^{\mathrm{atom}}\rangle}{\mu_{\mathrm{m}}^{\mathrm{atom}}} = \frac{\int_0^{2\pi}\exp[a\cos\theta]\cos\theta\sin\theta\, d\theta}{\int_0^{2\pi}\exp[a\cos\theta]\sin\theta\, d\theta} \tag{40}$$

and evaluation of the integrals reveals

$$\frac{\langle\mu_{\mathrm{m}}^{\mathrm{atom}}\rangle}{\mu_{\mathrm{m}}^{\mathrm{atom}}} = \coth(a) - \frac{1}{a} = L(a) \tag{41}$$

where $L(a)$ is called the Langevin function. The Langevin function has two interesting attributes as illustrated in Figure 9

$$(1)a \overset{\lim}{\to} \infty L(a) = 1 (2) a \overset{\lim}{\to} 0 \frac{dL(a)}{da} = \frac{1}{3} \tag{42}$$

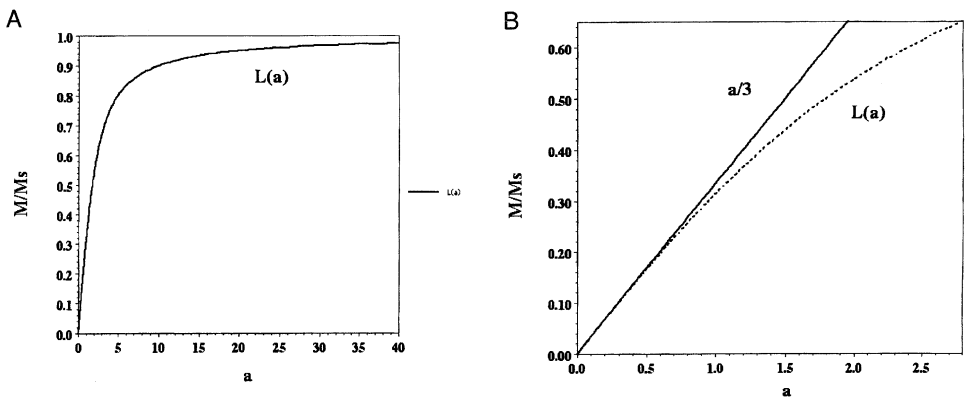


Figure 9. (**A**) Langevin function and (**B**) its low-temperature limiting form.

To calculate the magnetization we remember that M is defined as the total dipole moment per unit volume (we are interested in the component of **M** parallel to **B**), thus

$$\mathbf{M} = N_{\mathrm{m}}\langle\mu_{\mathrm{m}}^{\mathrm{atom}}\rangle = N_{\mathrm{m}}\mu_{\mathrm{m}}^{\mathrm{atom}}L(a) \tag{43}$$

where N_{m} is the number of magnetic dipoles per unit volume. In the large a limit $L(a) = 1$, and we infer that the saturation magnetization, M_{s}, is given by

$$M_{\mathrm{s}} = N_{\mathrm{m}}\mu_{\mathrm{m}}^{\mathrm{atom}} \tag{44}$$

and

$$M = M_{\mathrm{s}}L(\mathrm{a}) \tag{45}$$

in the low a limit (low field, high temperature), $L(a) \sim a/3$ and

$$M = M_{\mathrm{s}}\left(\frac{1}{3}\frac{\mu_0\mu_{\mathrm{m}}^{\mathrm{atom}}}{k}\right)\cdot\frac{H}{T} \tag{46}$$

and

$$\chi = \frac{M}{H} = \frac{N_{\mathrm{m}}\mu_0(\mu_{\mathrm{m}}^{\mathrm{atom}})^2}{3kT} = \frac{C}{T} \tag{47}$$

which is the Curie law of paramagnetism. Notice that if we know N_{m} (the concentration of magnetic atoms) from an independent experimental measurement, then an experimental determination of $\chi(T)$ allows us to solve for C as the slope of χ versus $1/T$ and therefore $\mu_{\mathrm{m}}^{\mathrm{atom}}\cdot\mu_{\mathrm{m}}^{\mathrm{atom}}$ is associated with the local effective moment as given by Hund's rules.

To further describe the response of the **quantum paramagnet**, we again consider an atom left with a permanent magnetic dipole moment, μ, due to its unfilled shells. We can now consider its magnetic behavior in an applied field. The magnetic induction, **B**, will serve to align the atomic dipole moments. The potential energy of a dipole oriented at an angle θ with respect to the magnetic induction is given by Equation 34 to yield

$$U_{\mathrm{p}} = -\boldsymbol{\mu}\cdot\mathbf{B} = -\mu\mathrm{B}\cos\theta = g\mu_{\mathrm{B}}\mathbf{J}\cdot\mathbf{B} \tag{48}$$

Now, our quantum mechanical description of angular momentum tells us that $|J_z|$, the projection of the total angular momentum on the field axis must be quantized, i.e.:

$$U_{\mathrm{p}} = (g\mu_{\mathrm{B}})m_{\mathrm{J}}\mathrm{B} \tag{49}$$

where $m_{\mathrm{J}} = J, J-1\ldots -J$ and $J = |\mathbf{J}| = |\mathbf{L}+\mathbf{S}|$ may take on integral or half-integral values. In the case of spin only, $m_{\mathrm{J}} = m_{\mathrm{s}} = \pm 1/2$. The quantization of J_z requires that only certain angles θ are possible for the orientation of μ with respect to **B**. The ground state corresponds to $\boldsymbol{\mu}\|\,\mathbf{B}$. However, with increasing thermal energy it is possible to misalign μ so as to occupy excited angular momentum states. If we consider the simple system with spin only we can consider the Zeeman splitting for spin only. Where eigenstates are split by $\mu_{\mathrm{B}}B$, the lower-lying state corresponds to $m_{\mathrm{J}} = m_{\mathrm{s}} = -1/2$ with the spin moment parallel to the field. The higher energy state corresponds to $m_{\mathrm{J}} = m_{\mathrm{s}} = 1/2$ and an antiparallel spin moment. For this simple two level system we can use Boltzmann statistics to describe the population of these two states. Now, if we have N such isolated atoms per unit volume in a field then we can define

$$N_1 = N_{\uparrow} = D\exp\left[\frac{\mu_{\mathrm{B}}B}{kT}\right] N_2 = N_{\downarrow} = D\exp\left[\frac{-\mu_{\mathrm{B}}B}{kT}\right] \tag{50}$$

where D is an unspecified constant. Recognizing that $N = N_1 + N_2$. Finally the net magnetization (dipole moment/volume) is then $M = (N_1 - N_2)\,\mu$ or

$$M = \frac{N}{D(e^x + e^{-x})}\cdot D(e^x - e^{-x})\mu = N\mu\frac{e^x - e^{-x}}{e^x + e^{-x}} = N\mu\ \tanh(x) \tag{51}$$

where $\mathrm{x} = \mu_{\mathrm{B}}B/kt$ notice that for small values of x, tanh(x) $\sim x$ can approximate μ

$$M = N\frac{\mu^2 B}{kT} \quad \text{and} \quad \chi = \frac{M}{B} = \frac{N\mu^2}{kT} \tag{52}$$

This expression relating χ to $1/T$ is called the "paramagnetic Curie law."

For the case where we have both spin and orbital angular momentum, we are interested in the J quantum number and the $2J+1$ possible values of M_J, each giving a different projected value (J_z) of **J** along the field (z) axis. In this case we no longer have a two level system but instead a $(2J+1)$-level system. The $2J+1$ projections are equally spaced in energy. Considering a Boltzmann distribution to describe thermal occupation of the excited states, we find that $\mu\cdot\mathbf{B} = (J_z/J)\mu_{\mathrm{B}}B$, and

$$\mu = \sum_j \frac{J_z}{J}\mu_{\mathrm{B}}\exp\left[\frac{J_{\mathrm{Z}}}{J}\frac{\mu_{\mathrm{B}}B}{kT}\right]$$

and

$$N = \sum_J \exp\left[\frac{J_{\mathrm{Z}}}{J}\frac{\mu_{\mathrm{B}}B}{kT}\right] \tag{53}$$

so that finally

$$M = N\mu = \frac{\sum_j \frac{J_{\mathrm{Z}}}{J}\exp\left[\frac{J_{\mathrm{Z}}}{J}\frac{\mu_{\mathrm{B}}B}{kT}\right]}{\sum_j \exp\left[\frac{J_{\mathrm{Z}}}{\mathrm{J}}\frac{\mu_{\mathrm{B}}B}{kT}\right]} = N_g J\mu_{\mathrm{B}}B_{\mathrm{j}}(x)\left(x = \frac{gJ\mu_{\mathrm{B}}B}{kT}\right) \tag{54}$$

where $B_J(x)$ is called the Brillouin function and is expressed

$$B_{\mathrm{J}}(x) = \frac{2J+1}{J}\coth\left[\frac{(2j+1)x}{2J}\right] - \frac{1}{2J}\coth\left[\frac{x}{2J}\right] \tag{55}$$

Note that for $J = 1/2, B_J(x) = tanh(x)$ as before. Now for $x \ll 1$

$$B_J(x) \sim \frac{(2J+1)^2 - 1}{3(2J^2)} - \frac{J(J+1)}{3J^2}x \tag{56}$$

For small x, we then see that

$$\chi = \frac{M}{B} = \frac{Ng^2\mu_B^2 J(J+1)}{3kT} = \frac{np_{\text{eff}}^2}{3kT} \tag{57}$$

where $p_{\text{eff}} = g[J(J+1)]^{1/2}\,\mu_B$ is called the "effective local moment." Now this expression is again called a "Curie law" with

$$\chi = \frac{C}{T} \quad \text{and} \quad C = \frac{Np_{\text{eff}}^2}{3k} \tag{58}$$

Experimentally derived magnetic susceptibility versus T data can be plotted as $1/\chi$ versus T to determine C (from the slope) and therefore p_{eff} (if the concentration of paramagnetic ions is known). Figure 10 shows the behavior of M versus H/T for a paramagnetic material (GdC_x nanoparticles). At low temperature $M(x) \sim x/3$. $M(H)$ is well described by a Brillouin function with a characteristic H/T scaling bringing curves from different temperatures into coincidence.

Superparamagnetism is a form of cluster or fine particle magnetism in which the magnetic entities consist of large clusters or fine particles of magnetic atoms. Conventional internal magnetic order exists within the clusters (i.e., ferromagnetism, antiferromagnetism, etc.), but the clusters themselves need not be coupled for ferromagnetic techniques to measure magnetic ordering transitions. From the standpoint of magnetic dipoles and magnetization, superparamagnetic response differs in that the dipole moments are associated with the atoms in aggregate in the entire fine particle, and coupling with a field is determined using the conventional theory of paramagnetism, but with particle moments. Superparamagnetism is distinguished from cluster spin glass behavior in that no interaction between the fine particles is assumed in the former.

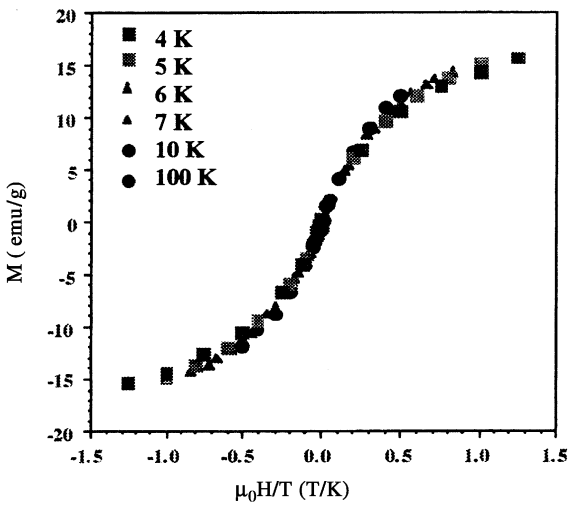


Figure 10. Magnetic response of Gd^{3+}ions in Gd_2C_3 nanocrystals exhibiting H/T scaling (Diggs et al., 1994).

Superparamagnetic response also concerns the probability of thermally activated switching of magnetic fine particle moments. This thermally activated switching can be described by an Arrhenius law for which the zero field activation energy barrier is $K_u\langle V\rangle$, where $\langle V\rangle$ is the particle volume and K_u is a (uniaxial) magnetic anisotropy energy density which ties the magnetization vector to a particular direction (crystallographic or otherwise). The switching frequency becomes larger for smaller particle size, smaller anisotropy energy density, and at higher T. Above a blocking temperature, T_B, the switching time is less than the experimental observation time, and net magnetization is lost, i.e. the coercive force becomes zero. Above T_B the magnetization scales with field and temperature in the same way as does a classical paramagnetic material, with the exception that the inferred dipole moment is a particle moment and not an atomic moment. $M(H,T)$ can be fit to a Langevin function, L, using the relation: $M/M_0 = L(x) = \coth(x) - 1/x$ where M_0 is the 0 K saturation magnetization and $x = \mu H/kT$. The particle moment, μ, is given by the product $M_s\langle V\rangle$ where M_s is the saturation magnetization and $\langle V\rangle$ is the average particle volume.

Below the blocking temperature, hysteretic response is observed for which the coercivity (field width of the magnetization curve) has a temperature dependence: $H_c = H_{c0}[1 - (T/T_B)^{1/2}]$. In the theory of superparamagnetism, the blocking temperature represents the temperature at which the metastable hysteretic response is lost for a particular experimental time frame. In other words, below the blocking temperature, hysteretic response is observed since thermal activation is not sufficient to allow the immediate alignment of particle moments with the applied field. For spherical particles with a uniaxial anisotropy axis, the rotational energy barrier to alignment is $K_u\langle V\rangle$. For hysteresis loops taken over ~ 1 h, the blocking temperature should roughly satisfy the relationship: $T_B = K_u\langle V\rangle/30k$. The factor of 30 represents $\ln(\omega_0/\omega)$, where ω is the inverse of the experimental time constant ($\sim 10^{-4}$ Hz) and ω_0 a switching attempt frequency (~ 1 GHz).

Ferromagnetism

Mean-Field Theory for Ferromagnets. We begin our discussion of ferromagnetism by introducing mean-field theory as introduced by Pierre Weiss in 1907. Weiss postulated the existence of an **internal magnetic field** (the Weiss field), $\mathbf{H}_{\text{INT}}$, which acts to align the atomic moments even in the absence of an external applied field, $\mathbf{H}_a$. The basic assumption of the mean field theory is that this internal field is nonlocal and is directly proportional to the magnetization of the sample

$$\mathbf{H}_{\text{INT}} = \lambda\mathbf{M} \tag{59}$$

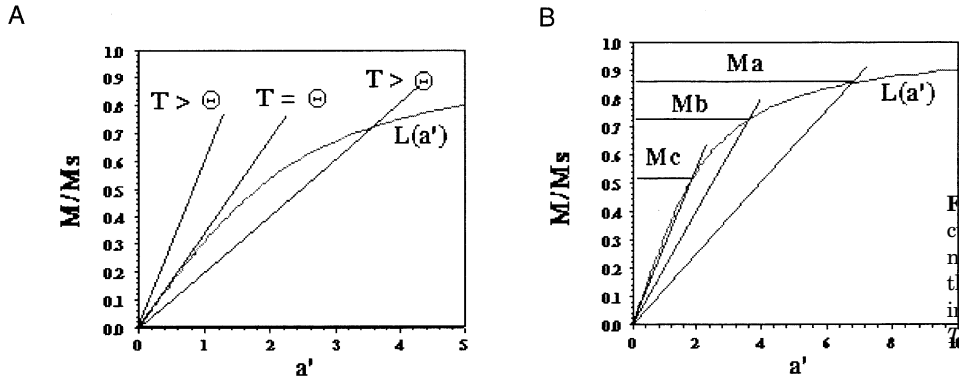


Figure 11. (**A**) Intersection between the curves 7.49a and 7.49b for $T < T_c$ gives a nonzero stable ferromagnetic state (**B**) the locus of $M(T)$ can be determined by intersections at various temperatures $T < T_c$.

where the constant of proportionality, λ, is called the "Weiss molecular field constant" (we can think of it as a crystal field constant). We now wish to consider the effects on ferromagnetic response of application of an applied field, $\mathbf{H}_a$, and the randomizing effects of temperature. We can treat this problem identically to that of a paramagnet, except that in this case we must now consider the superposition of the applied and internal magnetic fields. By analogy we conclude that

$$\langle \mu_m \rangle = \mu_m^{atom} L(\alpha') \tag{60}$$

where $\alpha' = \mu_0 \mu_m^{atom}[H + \lambda M]$, for a collection of classical dipoles. Also, $M_s = N_m \langle \mu_m^{atom} \rangle$ and

$$\frac{M}{N_m \mu_m^{atom}} = \frac{M}{M_S} = L\left\{\frac{\mu m \mu_0}{kT}[H + \lambda M]\right\} \tag{61}$$

where this rather simple expression represents a formidable transcendental equation to solve. Under appropriate conditions, this leads to solutions for which there is a nonzero magnetization (**spontaneous magnetization**) even in the absence of an applied field. We can show this graphically considering $M(H = 0)$ and defining the variables

$$\alpha' = \frac{\mu_m^{atom} \mu_0}{kT}(\lambda M) \tag{62}$$

which is dimensionless [and $M(0) = M_s L(\alpha')$] and

$$\Theta = \frac{N_m (\mu_m^{atom})^2 \mu_0 \lambda}{3k} \tag{63}$$

Notice that θ has units of temperature. Notice also that

$$\frac{\alpha'}{\Theta} = \frac{3}{T}\left(\frac{M(0)}{M_S}\right) \text{ so that : } \frac{M(0)}{M_S} = \frac{\alpha' T}{3\Theta} = L(\alpha') \tag{64}$$

The two equations for reduced magnetization represented in Equation 64 [i.e., $M(0)/M_s = \alpha' T/3\Theta$ and $M(0)/M_s = L(\alpha')$] can be solved graphically, for any choice of T by considering the intersection of the two functions $\alpha'/3(T/\Theta)$ and $L(\alpha')$. As is shown in Figure 11, for $T \geq \Theta$, the only solutions for which Equation 64 is simultaneously satisfied are when $M = 0$, i.e., no spontaneous magnetization and paramagnetic response. For $T < \Theta$ we obtain solutions with a nonzero, spontaneous, magnetization, the defining feature of a ferromagnet. For $T = 0$ to $T = \Theta$ we can determine the spontaneous magnetization graphically as the intersection of our two functions $\alpha'/3(T/\Theta)$ and $L(\alpha')$. This allows us to determine the zero-field magnetization $M(0)$ as a fraction of the spontaneous magnetization as a function of temperature. As shown in Figure 12, $m(t) = M(0, T)/M_s$, where $t = T/\Theta$, decreases monotonically from 1 at 0 K to 0 at $T = \Theta$, where Θ is called the **ferromagnetic Curie temperature**. At $T = \Theta$, we have a phase transformation from ferromagnetic to paramagnetic response, which can be shown to be second order in the absence of a field. In summary, mean-field theory for ferromagnets predicts the following:

1. For $T < \Theta$, collective magnetic response gives rise to a spontaneous magnetization even in the absence of a magnetic field. Spontaneous magnetization defines a ferromagnet.
2. For $T > \Theta$, the misaligning effects of temperature serve to completely randomize the direction of the atomic moments in the absence of a magnetic field. The loss of any spontaneous magnetization defines the return to paramagnetic response.

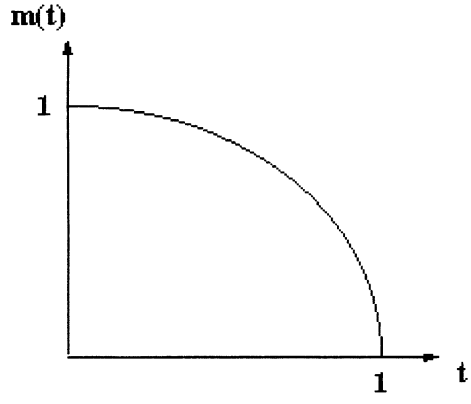


Figure 12. Reduced magnetization versus reduced temperature as derived from Figure 11.

3. In the absence of a magnetic field the ferromagnetic to paramagnetic phase transition is second order (it is first order in the presence of a magnetic field).

As a last ramification of the mean-field theory, we consider the behavior of magnetic dipoles in the paramagnetic state $T > \Theta$ with the application of a small field, H. We wish to examine the expression for the paramagnetic susceptibility $\chi(T) = M(H,T)/H$. Here again we can assume that we are in the small argument regime for describing the Langevin function so that

$$\frac{M(H,T)}{M_S} = L[\alpha'] = \frac{\alpha'}{3} = \frac{\mu_m^{atom}\mu_0}{3kT}(H + \lambda M) \tag{65}$$

and

$$M = \frac{\left[\frac{N_m(\mu_m^{atom})\mu_0}{3K}\right]H}{T - \frac{N_m(\mu_m^{atom})^2\mu_0\lambda}{3k}} = \frac{CH}{T-\Theta} \tag{66}$$

Thus the susceptibility, $\chi(T)$, is described by the so-called Curie-Weiss law

$$\chi = \frac{M}{H} = \frac{C}{T-\Theta} \tag{67}$$

where, as before, C is the Curie constant and Θ is the Curie temperature. Notice also that

$$\frac{\Theta}{C} = \lambda \tag{68}$$

It is possible to determine the size of the atomic moment and molecular field constant from high-temperature $\chi(T)$ data. Plotting $1/\chi$ versus T allows for determination of a slope ($= 1/C$), from which μ_m^{atom} can be determined, and an intercept, $-\Theta/C = -\lambda$. A further influence of a magnetic field on the magnetic phase transition is that the ferromagnetic to paramagnetic phase transition is not as sharp at $T = \Theta$ in a field, exhibiting a Curie tail that reflects the ordering influence of field in the high-temperature paramagnetic phase.

Antiferromagnetism

For a simple **antiferromagnet**, like body-centered cubic (bcc) Cr, e.g., spins on adjacent nearest neighbor atomic sites are arranged in an antiparallel fashion below an ordering temperature, T_N, called the "Neel temperature." The susceptibility of an antiferromagnet does not diverge at the ordering temperature but instead has a weak cusp (Fig. 13). The mean field theory for antiferromagnets considers two sublattices, an A sublattice for which the spin moment is down, and a B sublattice for which the spin moment is up. We can express, in mean-field theory, the internal fields on the A and B sites, respectively

$$\begin{aligned} \mathbf{H}_A^{INT} &= -\lambda_{BA}\mathbf{M}_B \\ \mathbf{H}_N^{INT} &= -\lambda_{AB}\mathbf{M}_A \end{aligned} \tag{69}$$

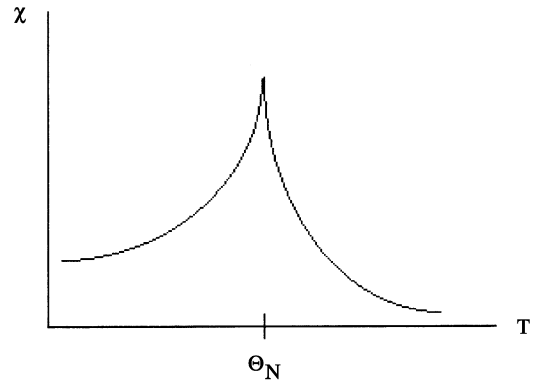


Figure 13. T-dependence of the magnetic susceptibility for an antiferromagnetic material.

where, by symmetry, $\lambda_{AB} = \lambda_{BA}$, $\mathbf{M}_B$ and $\mathbf{M}_A$ are the magnetizations of the $A$and B sublattices. Using the paramagnetic susceptibilities χ_p^A and χ_p^B

$$\mathbf{M}_A = \chi_p(\mathbf{H}_a + \mathbf{H}_A^{INT}) = \frac{C_A}{T}(\mathbf{H}_a + \lambda_{AB}\mathbf{M}_B) \tag{70}$$

$$\mathbf{M}_B = \chi_p(\mathbf{H}_a + \mathbf{H}_B^{INT}) = \frac{C_B}{T}(\mathbf{H}_a + \lambda_{AB}\mathbf{M}_A) \tag{71}$$

Now, for an antiferromagnet the moments on the A and B sublattices are equal and opposite so that $C_A = C_B = C$. Rearranging Equations 70 and 71 we get

$$T\mathbf{M}_A + C\lambda_{AB}\mathbf{M}_B = C\mathbf{H}_a \tag{72}$$

$$C\lambda_{AB}\mathbf{M}_A + T\mathbf{M}_B = C\mathbf{H}_a \tag{73}$$

In the limit as $\mathbf{H}_a \to 0$, these two equations have nonzero solutions (spontaneous magnetizations) for $\mathbf{M}_A$ and $\mathbf{M}_B$ only if the determinant of the coefficients vanishes; i.e.,

$$\begin{vmatrix} T & \lambda C \\ \lambda C & T \end{vmatrix} = 0 \tag{74}$$

and we can solve for the ordering temperature

$$\Theta_N = \lambda C \tag{75}$$

For $T > \Theta_N$ the susceptibility is given by

$$\chi = \frac{2CT - 2\lambda C^2}{T^2 - (\lambda C)^2} = \frac{2C}{T + \lambda C} = \frac{2C}{T + \Theta_N} \tag{76}$$

Ferrimagnetism

The mean-field theory for antiferromagnets is easily generalized to simple AB **ferrimagnetic** alloys. Here the magnitude of the moment on the A and B sublattices need not be the same and therefore $C_A \neq C_B$ in the limit as $\mathbf{H}_a \to 0$ and the determinant of the coefficients is

$$\begin{vmatrix} T & \lambda C_B \\ \lambda C_A & T \end{vmatrix} = 0 \tag{77}$$

The ferrimagnetic order temperature is determined to be

$$\Theta = \lambda\sqrt{C_A C_B} \tag{78}$$

and the magnetic susceptibility for $T > \Theta$ becomes

$$\chi = \frac{(C_A + C_B)T - 2\lambda C_A C_B}{T^2 - \Theta^2} \tag{79}$$

with curvature in $1/\chi$ versus T characteristic of a ferrimagnet.

The mean-field theory for spinel ferrimagnets is even richer. In the spinel structure, discussed in more detail below, magnetic cations may occupy octahedral A or tetrahedral B sites, respectively. In these systems however, the cations are close enough that A-A, A-B and B-B exchange interactions are possible. What is more, the sign of all of these exchange interactions is thought to be negative. By symmetry, however, if the A and B sites couple antiferromagnetically then the A-A and B-B pairs are aligned in a parallel fashion, i.e., even though their exchange interaction is negative, a stronger antiferromagnetic interaction between A and B sites might win out. This is commonly the case for ferrites, but may be complicated if one or the other of the sites is occupied by a nonmagnetic ion or if there is disparate temperature dependence between the A-A, B-B and A-B exchange interactions, respectively.

Heisenberg Model and Exchange Interactions

The **Heisenberg model** considers ferromagnetism and the defining spontaneous magnetization to result from nearest neighbor exchange interactions that act to align spins in a parallel configuration. The Heisenberg model can be further generalized to account for atomic moments of different magnitude, i.e., in alloys, and for exchange interactions that act to align nearest-neighbor moments in an antiparallel fashion or in a noncollinear relationship. Let us consider first the Heisenberg ferromagnet. Here we assume that the atomic moments on nearest neighbor sites are coupled by a nearest-neighbor exchange interaction giving rise to a potential energy

$$U_p = J_{ex}\mathbf{S}_i \cdot \mathbf{S}_{i,i+1} \tag{80}$$

between identical spins at sites i and $I + 1$ in a one-dimensional (1D) lattice. For identical spins

$$U_p = 2J_{ex}S^2\cos(\theta_{i,i+1}) \tag{81}$$

which for $J_{ex} > 0$ favors parallel alignment of the spins. For a linear chain of N spins (where N is large or exploits the periodic Born–Von Karmon boundary condition) and the total internal energy is

$$U_{TOT} = -2JS^2\sum_{i=1}^{N}\cos(\theta_{i,i+1}) \tag{82}$$

which for $J_{ex} > 0$ is minimized for a configuration in which all the spins are aligned in a parallel ferromagnetic configuration.

Exchange interactions result from the spatially dependent energy functional for electrons with parallel or antiparallel spins (or canted spins in more complicated models). For the hydrogen molecule, e.g., as shown in Figure 14, the configuration with spins aligned parallel is stable at larger interatomic separations and that with spins aligned antiparallel is stable at smaller interatomic separations. The choice of spin configuration depends on the relationship between the cross-over radius and the equilibrium separation of the atoms. In 3d transition metal solids, the famous **Bethe-Slater curve**, as shown in Figure 14B predicts the sign of the exchange interaction in 3d transition metal solids. In both cases, the interplay between electron-electron Coulomb interactions and the constraints of the Pauli exclusion principle determine the sign of the exchange interaction. In transition metal solids, a measure of the overlap between nearest-neighbor d orbitals is given by the ratio of the atomic to the 3d ionic (or nearest-neighbor) radius.

We will have further occasion to investigate the Heisenberg ferromagnet in future discussions. At this point, however, it is useful to consider the Heisenberg model prediction for other magnetic ground states. For example, notice that if $J_{ex} < 0$ an antiparallel configuration for adjacent spins in a 1D chain is predicted, consistent with an anti-ferromagnetic ground state as shown in Figure 15B. We can deduce a ferrimagnetic ground state as illustrated in Figure 15C and observed, e.g., for ferrites when we have $J_{ex} < 0$ and two magnetic sublattices, e.g., a and b, for which $\mu_a \neq \mu_b$. In three-dimensional (3D) systems it is possible to relax the restrictions of nearest-neighbor exchange and it is also possible to have noncollinear exchange interactions as shown in Figure 15D for a helimagnet. In certain ferrite systems (e.g., Mn_2O_3), it has been observed that Mn atoms in the octahedral and tetrahedral sites in face-centered cubic (fcc) interstices of the oxygen anion sublattice couple in a Yafet-Kittel triangular configuration.

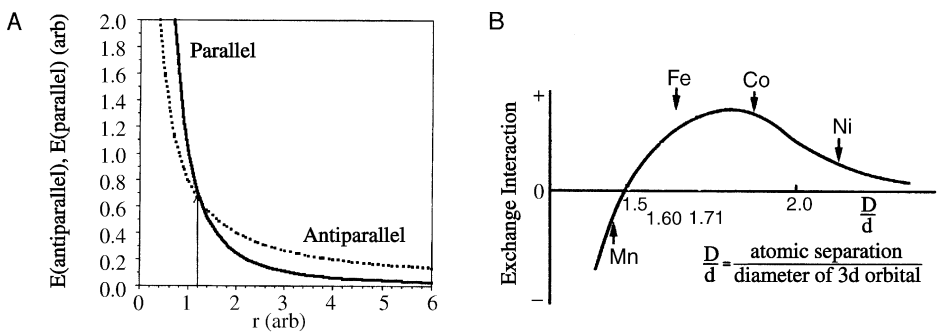


Figure 14. (**A**) Energies for parallel and antiparallel spins in the hydrogen molecule as a function of interatomic separation and (**B**) the Bethe-Slater curve predicting the sign of the exchange interaction in 3d transition metal solids.

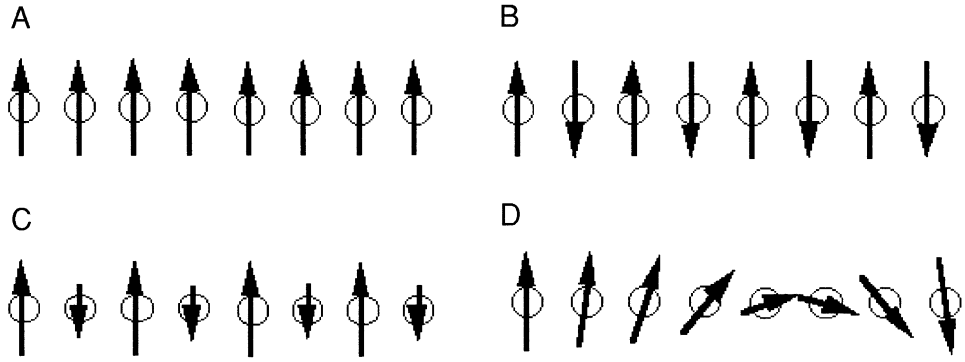


Figure 15. Spin configurations in a variety of magnetic ground states: (**A**) ferromagnet, (**B**) antiferromagnet, (**C**) ferrimagnet, (**D**) noncollinear spins in a helimagnet.

In the context of the ferrites and other oxides we may distinguish between direct and indirect exchange. We have described ferrimagnetic and Yafet-Kittel triangular spin configurations between neighboring magnetic cation sites. This is an example of an indirect exchange mechanism, since it must be transmitted through intervening nearest-neighbor oxygen sites. In fact, the exchange interaction is transmitted through overlap between magnetic d orbitals on the cation sites and the p orbitals of oxygen. This particular p orbital transmitted indirect exchange interaction is given the name "superexchange"and illustrated in Figure 16.

A form of indirect exchange that has recently been shown to be important in, e.g., magnetic/nonmagnetic multilayers, is the oscillatory **RKKY exchange** (Fig. 17). This is mediated through the conduction electron gas usually associated with nonmagnetic atoms but sometimes associated, e.g., with *sp* conduction electrons and magnetic electrons in rare earths. This indirect exchange is transmitted by polarization of the free electron gas. We can consider a polarization plane wave as emanating from a magnetic ion located at a position $r = 0$. The first waves influenced by a polarizing field are those with wave vector $\mathbf{k} = \mathbf{k}_F$ (at the Fermi surface) and therefore the sign of the exchange interaction should oscillate spatially like $\cos(\mathbf{k}_F \cdot \mathbf{r})$ as well as decaying exponentially as a function of r. To determine the sign of the indirect exchange interaction between the magnetic ion at $r = 0$ and at $r = r$ we need only calculate $\cos(\mathbf{k}_F \cdot \mathbf{r})$.

Other Forms of Collective Magnetism

Helimagnetism is a form of collective magnetism which refers to the spiral or helical arrangement of atomic magnetic dipole moments in certain metals, alloys, and salts. Ferromagnets and antiferromagnets can be considered as helimagnets with helical angles of $\Theta = 0$ and $\pi/2$, respectively. Examples of helimagnets include MnO_2 with a nonconical helix for $T < 86$ K; $MnAu_2$ which also has a non-conical helix for 85 K $<$ T $<$ 179 K and ferromagnetism for $T < 85$ K. Er has a conical helix for $T < 20$ K and has a complex oscillation for $20 < T < 53$ K and is a sinusoidal antiferromagnet for 53 K $< T <$ 85 K. $MnCr_2O_4$ has a complex conical helix, for $T < 18$ K and is a simple ferrimagnet for 18 K $< T <$ 43 K.

Helimagnetism is found in many rare earth metals and certain transition metal salts. Considering for example a hexagonal magnet, if we couple succesive basal planes we can identify J_1 as the exchange interaction between nearest-neighbor planes, J_2 between next nearest planes, etc. If the helical angle between successive basal planes is Θ_h then:

$$j\Theta_h = \Theta_{i+j} - \Theta_I \tag{83}$$

i.e., the angular difference between a plane i and $i + j$ is $j\Theta$h the exchange energy is proportional to

$$J_0 + 2J_1 \cos\Theta_h + 2J_2 \cos(2\Theta_h) + \cdots \tag{84}$$

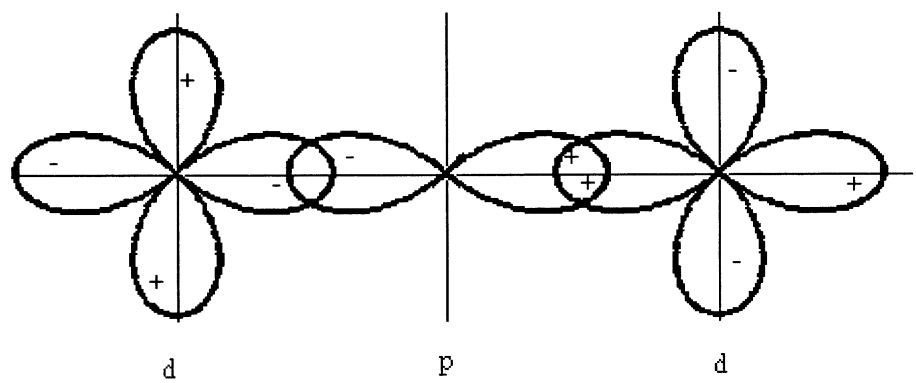


Figure 16. Superexchange interaction of magnetic ion d orbitals mediated through an oxygen p orbital.

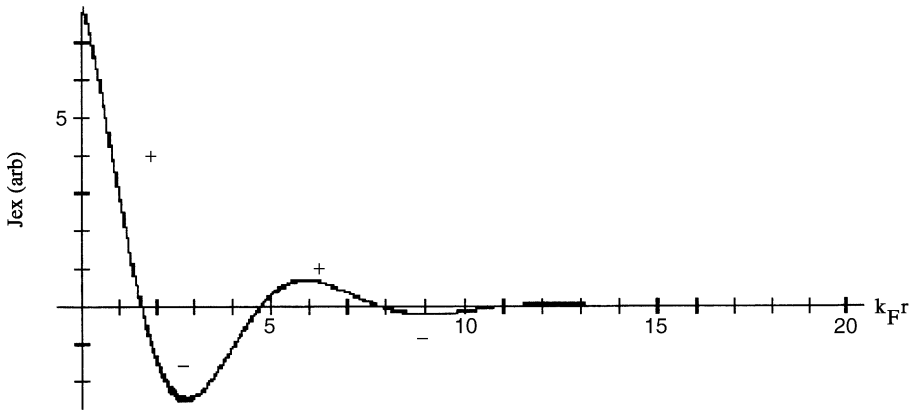


Figure 17. Spatial dependence of the RKKY exchange interaction.

keeping only two terms in the expansion and mimizing the exchange energy reveals

$$\Theta_h = \cos^{-1}(-J_1/2J_2) \quad (85)$$

for the helical angle.

CONCLUSIONS

In this unit, the basic notions of magnetic dipole moments and magnetization in solids have been summarized. In particular, the origin of magnetic dipoles from an atomic standpoint, a band structure standpoint, and a shielding current (superconductors) picture has been described. The coupling of magnetic dipoles, the field and temperature dependence of the magnetization, and simple magnetic phase transitions have been illustrated. This unit serves in defining certain basic magnetic phenomena which will be elaborated on, illustrated and extended in subsequent units. Magnetic phenomena are rich and varied; the fundamental principles of the origin, coupling and vector summation of magnetic dipole moments are at the heart of a comprehensive understanding of magnetic phenomena.

LITERATURE CITED

Abrikosov, A. A. 1957. *Soviet Phys.* JETP 5:1174.

Diggs, B., Zhou, A., Silva, C., Kirkpatrick, S., Nuhfer, N. T., McHenry, M. E., Petasis, D., Majetich, S. A., Brunett, B., Artman, J. O., and Staley, S. W. 1994. Magnetic properties of carbon-coated rare earth carbide nanocrystals produced by a carbon arc method. *J. Appl. Phys.* 75:5879.

Hundley, M. F., McHenry, M. E., Dunlap, R. A., Srinivas, V., and Bahadur, D. 1992. Magnetic moment and spin glass behavior in an $Al_{65}Mn_{20}Ge_{15}$ quasicrystal. *Philos. Mag. B* 66–239.

MacLaren, J. M., Crampin, S., Vvedensky, D. D., and Eberhart, M. E. 1989. *Phys. Rev. Lett.* 63:2586–2589.

MacLaren, J. M., Schulthess, T. C., Butler, W. H., Sutton, R. A., and McHenry, M. E. 1999. Calculated exchange interactions and Curie temperature of equiatomic B2 FeCo. *J. Appl. Phys.* 85: 4833–4835.

McHenry, M. E. and Sutton, R. A. 1994. Flux pinning and dissipation in high temperature oxide superconductors. *Prog. Mater. Sci.* 38:159.

van Vleck, J. 1932. The Theory of Electric and Magnetic Susceptibilities. p. 243. Oxford University Press, New York.

KEY REFERENCES

Solid-State Physics Texts—Introductory (Undergraduate) Level

Eisberg, R. and Resnick, R. 1947. Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles, John Wiley & Sons, New York.

Rosenberg, H. M. 1990. The Solid State. Oxford Science Publications, Oxford.

Solymar, L. and Walsh, D. 1998. Lectures on the Electrical Properties of Materials. Oxford University Press, New York.

Solid-State Physics Texts—Intermediate to Graduate Level

Ashcroft, N. W. and Mermin, N. D. 1976. Solid State Physics. Holt, Rinehart and Winston, New York.

Blakemore, J. S. 1981. Solid State Physics. Cambridge University Press, Cambridge.

Burns, G. 1985. Solid State Physics. Academic Press, San Diego.

Kittel, C. 1976. Introduction to Solid State Physics. John Wiley & Sons, New York.

Physics of Magnetism

Chikazumi, S. 1978. Physics of Magnetism. Kreiger, Malabar, Fla.

Keefer, F. 1982. Helimagnetism. *In* McGraw-Hill Encyclopedia of Physics (S.P. Parker, ed.). McGraw-Hill, New York.

Morrish, A. H. 1965. The Physical Principles of Magnetism. John Wiley & Sons, New York.

Mydosh, J. A. 1993. Spin Glasses, An Experimental Introduction. Taylor and Francis, London.

White, R. M. 1983. Quantum Theory of Magnetism, Springer Series in Solid State Physics 32. Springer-Verlag, Berlin.

Magnetic Materials

Berkowitz, A. E. and Kneller, E. 1969. Magnetism and Metallurgy, Vol. 1. Academic Press, San Diego.

Bozorth, R. M. 1951. Ferromagnetism. Van Nostrand, New York.

Chen, C.-W. 1986. Metallurgy of Soft Magnetic Materials. Dover, New York.

Cullity, B. D. 1972. Introduction to Magnetic Materials. Addison-Wesley, Reading, Mass.

Kouvel, J. S. 1969. *In* Magnetism and Metallurgy 2 (A. E. Berkowitz and E. Kneller, eds.) p. 523. Academic Press, New York.

McCurrie, R. A. 1994. Ferromagnetic Materials: Structure and Properties. Academic Press, London.

Fine Particle Magnetism

Bean, C. P. and Livingston, J. D. 1959. *J. Appl. Phys.* 30:120S-129S.

Bitter, F. 1931. On inhomogeneities in the magnetization of ferromagnetic materials. *Phys. Rev.* 38:1903–1905.

Bitter, F. 1932. Experiments on the nature of ferromagnetism. *Phys. Rev*. 41:507–515.

Luborsky, F. E. 1961. *J. Appl. Phys.* 32:171S–183S.

Energy Band Theory of Magnetism

Kubler, J. and Eyert, V. 1995. Electronic structure calculations. *In* Electronic and Magnetic Properties of Metals and Ceramics. Materials Science and Technology, A Comprehensive Treatment, Vol. 3 (K. H. J. Buschow, ed.) pp. 1–146. VCH Publishers, New York.

MacLaren et al., 1989. See above.

MacLaren, J. M., McHenry, M. E., Eberhart, M. E., and Crampin, S. 1990. Magnetic and electronic properties of Fe/Au multilayers and interfaces. *J. Appl. Phys.* 67:5406.

McHenry, M. E., MacLaren, J. M., Eberhart, M. E., and Crampin, S. 1990. Electronic and magnetic properties of Fe/Au multlayers and interfaces. *J. Mag. Mater.* 88:134.

Alloy Magnetism

Pauling, L. 1938. The nature of interatomic forces in metals. *Phys. Rev.* 54:899–904.

Slater, J. C. 1936. The ferromagnetism of nickel. *Phys. Rev.* 49:537–545.

Slater, J. C. 1936. The ferromagnetism of nickel. II. *Phys. Rev.* 49:931–937.

Slater, J. C. 1937. Electronic structure in alloys. *J. Appl. Phys.* 8:385–390.

MICHAEL E. MCHENRY
Carnegie Mellon University
Pittsburgh, Pennsylvania

THEORY OF MAGNETIC PHASE TRANSITIONS

THERMODYNAMICS

The existence of magnetic order (collective magnetism) appearing in materials below a particular ordering temperature (e.g., the Curie temperature, T_C, or the Neel temperature, T_N) points to a class of physical phenomena which are described as magnetic phase transitions. The thermodynamics of these phase transitions can be described by energy functions (expressed in terms of intensive or extensive variables), in terms of magnetization–temperature ($M-T$) phase diagrams, or in terms of critical exponents that describe the variation of thermodynamic properties (as a function of the order parameter) as the ordering temperature is approached. In this section, we begin by describing the thermodynamics of magnetic ordering transitions, the order of the transitions, critical exponents, and thermodynamic variables. We further describe magnetic phase transitions within the framework of the mean field or Landau theory of phase transitions with discussion of several magnetic equations of state.

The discussion of thermodynamic properties begins with the definition of thermodynamic potential functions and their derivatives. For simplicity, we can take pressure and volume as being held constant and consider only magnetic work terms. In this case the entropy, S, and magnetization, M, are extensive variables, the temperature, T, and field, H, are intensive variables, and the different energy functions are the internal energy, $U(S,M)$; the enthalpy, $E(S,H)$; the Gibbs free energy, $G(T,H)$; and the Helmholtz free energy, $F(T,M)$. The differentials of each are:

$$\begin{aligned} dU &= TdS + HdM \qquad & dE &= TdS - MdH \\ dG &= -SdT - MdH \qquad & dF &= -SdT + HdM \end{aligned} \tag{1}$$

Two specific heats C_M and C_H are defined as:

$$C_M = -T\left(\frac{\partial^2 F}{\partial T^2}\right)_M, \qquad C_H = -T\left(\frac{\partial^2 G}{\partial T^2}\right)_H \tag{2}$$

The adiabatic and isothermal susceptibilities are given, respectively, by:

$$\chi_S = \left(\frac{\partial M}{\partial H}\right)_S = \left(\frac{\partial^2 E}{\partial H^2}\right)_S, \qquad \chi_T = \left(\frac{\partial M}{\partial H}\right)_T = -\left(\frac{\partial^2 G}{\partial H}\right)_T \tag{3}$$

Phase transitions reflect discontinuities in the free energy function. The order of a phase transition is defined in terms of the smallest derivative of the free energy function for which a discontinuity occurs at the ordering temperature. The magnetization at constant temperature:

$$M = \left(\frac{\partial G}{\partial H}\right)_T \tag{4}$$

in an equilibrium ferromagnet is discontinuous at $H = 0$ for $T < T_C$ (Figure 1). At T_C, M is continuous, but the susceptibility is discontinuous between the ferromagnetic and paramagnetic states. The Helmholtz free energy (Figure 1A) has a minimum only at $M = 0$ for $T > T_C$. For $T < T_C$, two minima occur at $\pm M_s$, the value of the spontaneous magnetization. The $H = 0$ ferromagnetic transition is second order, at T_C, while for $H \neq 0$ the transition is first order.

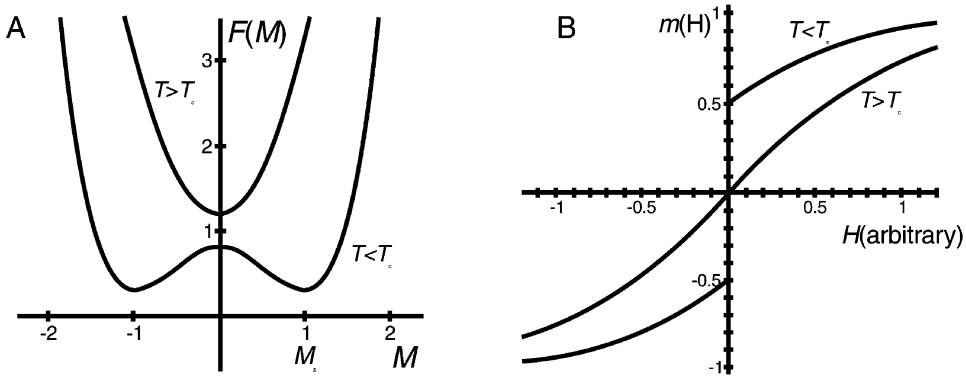


Figure 1. Schematic depiction of (**A**) magnetic Helmholtz free energy isotherms and (**B**) reduced magnetization, $m(H)$ for $T < T_C$ and $T > T_C$ (Ausleos and Elliot, 1983).

From the above discussion it is clear that the spontaneous magnetization, M, or the reduced magnetization, $m = M/M_s$, will serve as the order parameter in the ferromagnetic phase transition. In antiferromagnetic, ferrimagnetic, or helimagnetic systems the local vector reduced magnetization, $m(r)$, must be taken as the order parameter, since the magnetization is not spatially uniform on an atomic scale.

LANDAU THEORY OF MAGNETIC PHASE TRANSITIONS

In future discussions we will rely on the so-called Landau theory of the magnetic phase transitions. In the Landau theory, the Helmholtz free energy is expanded in a Taylor series where, for symmetry reasons, only even powers of the order parameter, M, are kept in the expansion. The temperature dependence is described in terms of T-dependent expansion coefficients. Near T_C, where the order parameter is small, only a few terms need to be kept in the expansion. Considering such an expansion, truncated at two terms and adding in a term to reflect the magnetostatic contributions (i.e., potential energy of the magnetization in the internal field), the magnetic Helmholtz free energy (at constant temperature) can be expressed as:

$$F_M = \frac{1}{2}A(T)M^2 + \frac{1}{4}B(T)M^4 - \mu_0 MH \tag{5}$$

where $A(T)$ and $B(T)$ are the temperature dependent expansion coefficients in the Landau theory. An equation of state can be derived by minimizing the free energy with respect to magnetization:

$$\frac{\partial F_M}{\partial M} = 0 = A(T)M + B(T)M^3 - \mu_0 H \tag{6}$$

that results in the expression:

$$M^2 = \frac{-A(T)}{B(T)} + \frac{\mu_0 H}{B(T)M} \tag{7}$$

By invoking a microscopic model of the magnetization (of which there are many), one can determine the coefficients $A(T)$ and $B(T)$, specifically. For example, in the Stoner theory of itinerant ferrromagnets, the magnetic equation of state and coefficients are expresssed by the following equations (Gignoux, 1995):

$$M(H,T)^2 = M(0,0)^2\left[1 - \frac{T^2}{T_C^2} + \frac{2\chi_0 H}{M(H,T)}\right] \tag{8}$$

$$A(T) = \frac{-\mu_0\left[1 - (T^2/T_C^2)\right]}{2M(0,0)\chi_0} \qquad B(T) = \frac{-\mu_0}{2M(0,0)^2\chi_0} \tag{9}$$

where $M(0,0)$ is the $H = 0$, $T = 0$ magnetization and χ_0 is the magnetic susceptibility at 0 K. Notice that the Landau theory predicts a (Pauli) paramagnetic state, $M = 0$ (no spontaneous magnetization), when the coefficient $A(T) > 0$. On the other hand, a stable spontaneous magnetization is predicted for $H = 0$ when $A(T) < 0$. This spontaneous magnetization is:

$$M^2 = \frac{-A(T)}{2B(T)} \tag{10}$$

In the presence of a field, and for the case $A(T) < 0$, this gives rise to a stable spontaneous magnetization, which is described by the equation of state (see Equation 7). This suggests that plotting different isotherms of M^2 versus H/M allows for the determination of $A(T)$ and $B(T)$. Moreover, since $A(T_C) = 0$, i.e., A vanishes at the Curie temperature, T_C, then the Curie temperature can be determined as the isotherm with 0 as its intercept. Plots of M^2 versus H/M isotherms are called Arrott plots and are discussed below as a method for determining the magnetic ordering temperature from magnetization measurements.

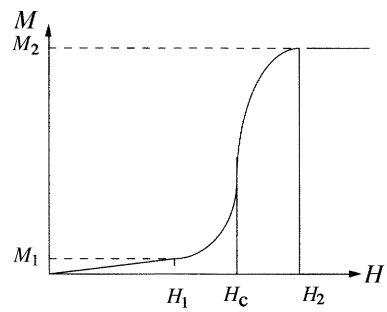


Figure 2. Magnetization curve metamagnetic response. M_1 is the external magnetization in the paramagnetic state; M_2 is the saturation magnetization in the metamagnetic state.

Table 1. Ordering Temperatures for Some Selected Magnetic Materials

Ferrimagnets	Curie Temperature (T_C), K	Ferrimagnets	Curie Temperature (T_C), K	Antiferromagnets	Neel Temperature (T_N), K	Helimagnets	(T_H), K
CO	1388	Fe_3O_4	858	NiO	600	MnO_2	84
Fe	1043	$MnFe_2O_4$	573	Cr	311	$MnAu_2$	363
Ni	627	Y_3FeO_{12}	560	Mn	95	Er	20
Gd	289	γ-Fe_2O_3	948	FeO	198		

Another free energy minumum exists, which gives rise to a so-called metamagnetic state. If $A(T) > 0$ but small, a Pauli paramagnetic state is predicted in zero field. However if $B(T) < 0$ it is possible to have a minimum in the Helmholtz free energy at $H \neq 0$. This minimum may in fact be deeper than that at $M = 0$. In such a case, application of a field causes the system to choose the second minimum (at M_2) giving rise to an $M(H)$ curve as depicted in Figure 2 for this so called metamagnetic response.

CRITICAL EXPONENTS IN MAGNETIC PHASE TRANSITIONS

One of the goals of thermodynamic treatments of magnetic phase transitions is to determine critical exponents associated with the phase transition. This involves describing power law exponents for the temperature dependence of thermodynamic quantities as the ordering transition is approached. Determination of critical exponents and scaling laws allows for closed-form representations of thermodynamic quantities with significant usefulness in prediction and/or extrapolation of thermodynamic quantities (Callen, 1985). To describe such scaling-law behavior, the reduced temperature variance, ε, is defined as:

$$\varepsilon = \frac{(T - T_C)}{T} \tag{11}$$

which approaches 0 as $T \rightarrow T_C$ from below or above. For $T > T_C$ and $H = 0$, the specific heat (C, which can be C_M or C_H) and isothermal susceptibility obey the scaling laws:

$$C \sim \varepsilon^{-\alpha} \quad \text{and} \quad \chi_T \sim \varepsilon^{-\gamma} \tag{12}$$

For $T < T_C$, the specific heat, the magnetization and the isothermal susceptibility obey the scaling laws:

$$C \sim \varepsilon^{-\alpha'} \quad \chi_T \sim (-\varepsilon)^{-\gamma'} \quad \text{and} \quad M \sim (-\varepsilon)^{\beta} \tag{13}$$

At T_C the critical isotherm is described by

$$H \sim |M|^{\delta} \tag{14}$$

Thermodynamic arguments, such as those of Rushbrooke (see Callen, 1985), allow us to place restrictions on the critical exponents. For example, defining

$$\alpha_H = \left(\frac{\partial M}{\partial T}\right)_H \tag{15}$$

and using thermodynamic Maxwell relations, it is possible to show that:

$$\chi_T(C_H - C_M) = T\alpha_H^2 \tag{16}$$

which implies

$$C_H > \frac{T\alpha_H^2}{\chi_T} \tag{17}$$

This then requires that, as $T \rightarrow T_C$ from below:

$$\alpha' + 2\beta + \gamma' \geq 2 \tag{18}$$

Furthermore, if C_M/C_H approaches a constant ($\neq 1$) then two more inequalities may be expressed:

$$\alpha' + \beta(1 + \delta) \geq 2 \quad \text{and} \quad \gamma' \geq \beta(\delta + 1) \tag{19}$$

In the Landau theory, it can be determined that:

$$\alpha' = \alpha = 0; \quad \beta = 1/2; \quad \gamma' = \gamma = 1; \quad \text{and} \quad \delta = 3 \tag{20}$$

If we consider a vector magnetization, and allow for a spatially varying local magnetization, then the Landau theory must be extended to a more complicated form. Further, it is often necessary to add terms $\sim(\nabla \mathrm{M})^2$ to the energy functional. In such cases, $T_C = T_C(\mathbf{k})$ and $\chi = \chi(\mathbf{k})$ can be taken as reciprocal space expansions and the spatial dependence of the susceptibility, $\chi(\mathbf{r})$, can be determined as the Fourier transform of $\chi(\mathbf{k})$. In these cases a correlation length, ζ, for the magnetic order parameter is defined, which diverges at T_C. Further discussion of the spatial dependence of the order parameter is beyond the scope of this unit.

Table 1 summarizes ordering temperatures for selected materials with a variety of types of magnetic order. These various local magnetic orders and the exchange interactions that give rise to them are discussed in TECHNIQUES TO MEASURE MAGNETIC DOMAIN STRUCTURES. For most of that discussion, we will consider ferromagnetic ordering only.

LITERATURE CITED

Ausleos, M. and Elliot, R. I. 1983. Magnetic Phase Transitions. *In* Springer Series in Solid State Sciences, vol 48 (M. Cardona, P. Fulde, K. von Klitzing, and H.-J. Queisser, eds.). Springer-Verlag, New York.

Callen, H. B. 1985. Thermodynamics and an Introduction to Thermostatistics. John Wiley & Sons, New York.

Gignoux, D. 1995. Magnetic Properties of Metallic Systems. Electronic and Magnetic Properties of Metals and Ceramics. *In* Materials Science and Technology: A Comprehensive Treatment, Vol. III (K. H. J. Buschow, ed.). VCH, Weinheim.

KEY REFERENCES

Ausleos and Elliot, 1983. See above.

Collection of articles dealing with statistical mechanics and theory of magnetic phase transitions.

Callen, 1985. See above.

Advanced undergraduate or graduate text on thermodynamics. This is notable for its discussion of magnetic work terms, phase transitions, critical phenomena, etc., as well as its excellent exposition on classical thermodynamics and statistical mechanics.

Mattis, D. C. 1981. The Theory of Magnetism. *In* Springer Series in Solid-State Sciences, vol. 17 (M. Cardona, P. Fulde, K. von Klitzing, and H.-J. Queisser, eds.). Springer-Verlag, New York.

Advanced text on the theory of magnetism with strong basis in the quantum mechanical framework. Much interesting historical information. Callen, 1985. See above.

MICHAEL E. MCHENRY
DAVID E. LAUGHLIN
Carnegie Mellon University
Pittsburgh, Pennsylvania

MAGNETOMETRY

INTRODUCTION

All materials possess a magnetic moment; techniques for measuring that bulk macroscopic property are defined here as magnetometry. This unit will review the most common techniques for measuring the total magnetic moments of small samples (volume ≤ 1 cm^3 and/or mass ≤ 1 g). Several factors contribute to the bulk magnetic moment of a sample. Essentially all materials show weak diamagnetism from filled electronic core states of the atoms. Metallic materials show an additional contribution to their diamagnetism from the orbital motion of the otherwise degenerate spin-up and spin-down conduction electrons as well as Pauli paramagnetism from splitting of the conduction bands. The largest contribution to the magnetic moment of materials comes from unpaired localized spins of elemental constituents—unpaired $3d$ and sometimes $4d$ and $5d$ electrons in the case of the transition metal series, $4f$ electrons for the rare earths, and $5f$ electrons for the actinide constituents.

It is important for many reasons to know a material's magnetic moment, a thermodynamic quantity. For strongly magnetic materials, it is probably the most useful and important physical property, determining their utility in applications. For magnetically ordered systems, magnetic moment measurements provide information about spin structure, anisotropy (in the case of nonpolycrystalline samples), and phase transitions. For paramagnetic systems, the total moment at high fields or the temperature dependence of the magnetic susceptibility can yield a measure of the moment per chemical constituent or magnetic impurity concentration, whether introduced purposefully or as impurities. The temperature dependence of the moment yields information on interactions between the paramagnetic ions or with the lattice. Certain features of the electronic structure of metallic materials can be determined from magnetic susceptibility, such as the density of states at the Fermi surface. For superconductors, high field magnetization can provide information about the critical field and critical current density.

This unit is restricted to the centimeter-gram-second (cgs) system of units for the magnetic moment, which is given in electromagnetic units (emu), where 1 emu $= 10^{-3}$ A-m^2 [International (SI) System of Units]. To determine the magnetization $\mathbf{M}$ (G/cm^3), where $\mathbf{B} = \mathbf{H} + 4\pi\mathbf{M}$, multiply moment per gram (emu/g) by the density (g/cm^3). The magnetic moment, magnetic field ($\mathbf{H}$), magnetic induction ($\mathbf{B}$), and associated conversion factors may be found in MAGNETIC MOMENT AND MAGNETIZATION of this part.

Recent reviews of magnetometry are given by Foner (1994) and by Flanders and Graham (1993) and recent developments are often published in the *Review of Scientific Instruments*. Proceedings of the annual Intermag Conferences appear in a special fall issue of the *IEEE Transactions on Magnetics*, and those of the annual Magnetism and Magnetic Materials Conference generally appear in a special spring issue of the *Journal of Applied Physics*.

PRINCIPLES OF THE METHOD

There are two direct techniques for measuring the magnetic moment of materials: one utilizes the detection of a change in magnetic flux produced by a sample, and the other utilizes the detection of a change in the force experienced by a sample. An analog device, a transducer, couples the change in flux or force to electronic circuitry for signal processing and readout, usually in the form of a direct-current (dc) voltage proportional to the sample moment. The transducer is usually a configuration of pickup coils for flux-based measurements. For force measurements, the transducer will most likely be an electromechanical balance or piezoelectric device. Often a magnetometer system operates in some null detection mode using feedback,- which is proportional to the magnetic moment of the sample being measured.

Flux measurements are usually performed with a configuration of pairs of series-opposing pickup coils. Most of these magnetometers utilize the strategy of moving the sample with respect to the pickup coils in order to measure the moment of the sample while minimizing contributions from the background magnetic field. The induced electromotive force (emf) in a pickup coil is generated from a temporal change in magnetic flux, obeying Faraday's law: $E(\mathrm{V}) = -N \times 10^{-8}(d\Phi/dt)$, where N is the number of turns in the coil and the magnetic flux, $\Phi(\mathrm{G\text{-}cm}^2) = \mathbf{B}a = (\mathbf{H} + 4\pi\mathbf{M})a$, a being the effective area of the coil

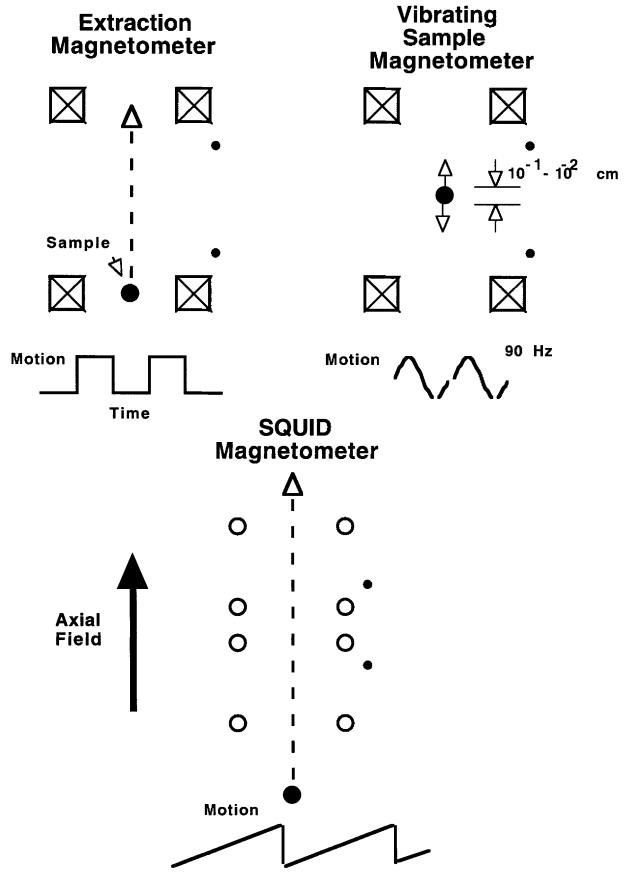


Figure 1. Schematic representation of three flux-based magnetometers and the time dependence of the sample motion for each. The planes of the coil windings are perpendicular to the field for all these examples (axial configuration). The figures show the cross-section of the coils (not to scale) and the dots indicate coil polarity. The pickup coils of the extraction and vibrating sample magnetometers are multiturn, but those for the SQUID magnetometer are constructed with very few turns to minimize inductance and maximize coupling to the low-impedance SQUID transducer.

normal to **B**. A change in the magnetic flux can also be induced in a sample by changing an externally controlled variable such as magnetic field strength or temperature.

Figure 1 is a schematic representation of three flux-based magnetometers: the extraction, vibrating sample, and superconducting quantum interference device (SQUID) magnetometers. In each case the sample is moved relative to series-opposing pickup coils, which detect the flux changes. The oscillatory sample motion for the vibrating-sample magnetometer is very small and rapid, the motion for the extraction magnetometer is stepwise and slower and confined to the coil dimensions, and in general the SQUID motion extends beyond the coil dimensions to measure peak-to-peak flux change. The pickup coils of the first two magnetometers are multiturn in order to maximize the induced voltage, but for the SQUID, the pickup coils have only a few turns in order to minimize the inductance and maximize the coupling to the extremely low impedance SQUID transducer.

All force magnetometers depend on the difference in energy experienced by the sample in a magnetic field gradient. Figure 2 is a schematic representation of four force-related magnetometers: the traditional Faraday, the cantilever, the piezoresistive, and the alternating-gradient magnetometers. In general, assuming the (spherical) sample composition is homogeneous, isotropic, and in a vacuum, the force for a gradient along the z axis is given by

$$\mathbf{F}\,(\text{dynes}) = -\nu\chi\mathbf{H}(dH/dz) \tag{1}$$

if χ, the magnetic susceptibility, is field independent, i.e., $\mathrm{M} = \chi\mathbf{H}$. Here $\mathbf{M}$ is the magnetization per cubic centimeter and ν is the volume (in cubic centimeters). If the sample is ferromagnetic and above saturation, $\mathbf{M} = \mathbf{M}_s$, the force is given by

$$\mathbf{F} = -(\nu/2)\mathbf{M}_s(d\mathbf{H}/dz) \tag{2}$$

If the sample is inherently anisotropic or has shape anisotropy, effects of the x and y gradients can contribute to

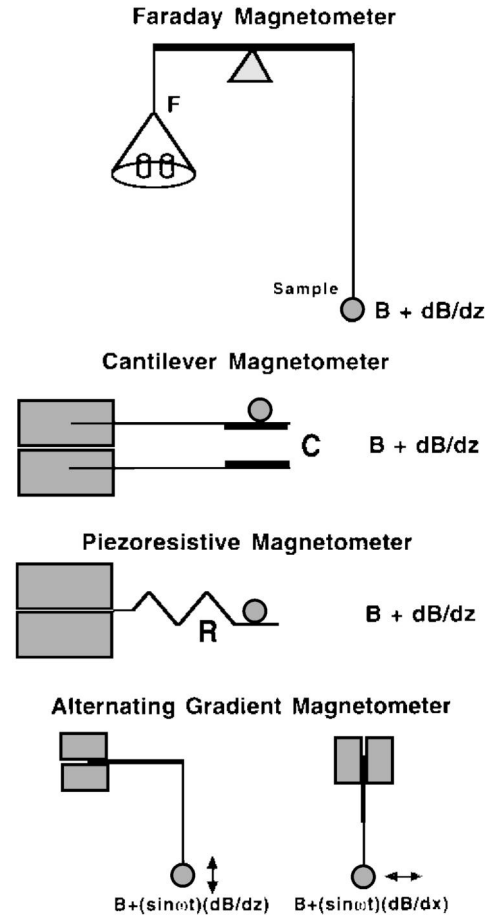


Figure 2. Schematic representation of four force-based magnetometers and the time and spatial field variation for each. In all cases, a feedback mechanism is often used to allow null detection of the sample moment. The Faraday magnetometer generally accommodates somewhat larger samples than the other three systems. The piezoresistive element is indicated by the resistor and the piezoelectric sensors for the AGM are shown as thick lines.

the measured force. Those components of the gradient are always present (divergence $\mathbf{B} = 0$). For force measurements, one needs to know the field and the gradient and also be able to reproduce accurately the position of the sample. The sensitivity is a strong function of field and gradient. Samples with high anisotropy or highly nonspherical shape can be problematic due to the transverse field gradients. Wolf (1957) has shown that, under certain conditions, even the sign of the measured force can be incorrect.

External magnetic fields are usually required for magnetometers (see GENERATION AND MEASUREMENT OF MAGNETIC FIELDS). The two common sources of external fields for magnetometry are iron core electromagnets, which provide fields up to ~20,000 G (2 T), and superconducting solenoids, which provide much higher fields, up to 150,000 G (15 T) or more. For many purposes the field from the electromagnets is ample. These can be turned on in a few minutes and furnish easy access to the field volume for manipulation and modifications. Superconducting solenoids generally require liquid helium for operation, take several hours to cool to operating temperatures, and offer a more restricted access to the experimental volume. Bitter solenoids, resistive solenoids that can provide fields to 30 T, and hybrid magnets, a combination of Bitter solenoids surrounded by superconducting solenoids providing even larger fields, are accessible at national laboratory facilities. Such user-oriented facilities include those in the United States, France, Japan, The Netherlands, and Poland.

It is important for the experimenter to estimate the sample moment prior to measurement because this may dictate which technique is suitable. A 10-mg polycrystalline Ni spherical calibration sample will display a moment of 0.55 emu at saturation (above 6000 G) at 300 K and 0.586 emu at $T = 4.2$ K. A 10-mg pure Pd sample will have a moment of ~0.007 emu at the same field and temperature. The moment of a superconducting sample or ferromagnetic material depends not only on the applied field and temperature but also on the history of the sample in field. A 10-mg bulk Pb superconducting sample cooled to below its transition temperature in a zero field will show a moment of ~ -0.007 emu (diamagnetic) in 100 G, but it will have a much smaller moment above the critical field. A given magnetometer can cover the range from a few emu to 10^{-6} emu or less. It should be stressed that ultimate sensitivity may not always be the most important criterion, given other constraints. Clearly, a sample moment of 1 emu is fairly large and easily measured.

PRACTICAL ASPECTS OF THE METHOD

Flux Magnetometers

Flux-Integrating Magnetometer. The basic instrument for detecting a change in sample moment or magnetic flux change, $\Delta\phi$, in a pickup coil is the fluxmeter, essentially an operational amplifier with capacitive feedback C (Flanders and Graham, 1993). A pickup coil having N turns and a resistance R_c in which the sample is located is connected in series with the fluxmeter input resistance R_i. The fluxmeter integrates the voltage change given by Faraday's law, V (in volts)$dt = -N \times 10^{-8}\, d\phi$ (G-cm^2). The sensitivity of a direct-reading fluxmeter depends on N, the time constant (which involves C, R_i, and R_c), but is also limited by drift in the input, often generated by thermal emfs. The sensitivity and utility depend heavily on coil design and the values chosen for R_i and C. Commercially available high-speed digital voltmeters effectively replace the operational amplifier-based fluxmeter and deal more effectively with the drift problem. Variation of the integration time generally involves a trade-off between lower noise and higher drift (longer integration times) or lower drift and higher noise level (shorter integration times). Krause (1992) described a magnetometer that used a stepping motor to move the sample in a few seconds between opposing pickup coils in series [the same coils used for alternating-current (ac) susceptometry] so that the instrument carried out both dc and ac measurements. An extraction magnetometer works on the principle of relatively rapid sample removal from or insertion into the interior of a pickup coil.

Vibrating-Sample Magnetometer. Commercially available since the mid-1960s, the vibrating sample magnetometer (VSM) was first described by Foner (1956). Most commonly, the sample is mechanically vibrated in zero or uniform applied field at <100 Hz, far from mechanical resonance (amplitude of a few tenths of a millimeter to a few millimeters) at a point midway between pairs of series opposing pickup coils, where the signal is a maximum and is insensitive to position to first order. The coils detect the ac component of the field produced by the moving sample, and their ac output is detected by a lock-in amplifier employing a phase-sensitive detector. The dc output of the VSM is proportional to $d\Phi/dt$ and is given by

$$V = G\sigma\omega A \cos(\omega t) \tag{3}$$

for small sinusoidal drive, where G is a geometric factor for the coils, σ the sample moment, and A and ω are the amplitude and angular frequency of the sample motion, respectively. The mechanical drive itself is normally an electromechanical device like a loudspeaker, less commonly a motor drive.

The most common VSMs use a transverse-coil configuration where the vibration axis (z axis) is perpendicular to the field supplied by an electromagnet and the pickup coils are arranged to sense the moment along the field direction. For higher fields, transverse-field superconducting magnets have been used. By rotating the drive unit about the z axis, the magnetic moment versus the angle is obtained in the x-y plane. Various more complex pickup coil arrangements have been used to sense the components of the magnetization vector at any angle to the applied field for anisotropic materials or magnetic thin films (Foner, 1959; Bowden, 1972; Zieba and Foner, 1982; Pacyna and Ruebennauer, 1984; Mallinson, 1992; Bernards, 1993; Ausserlechner et al., 1994). Recently, these arrangements have been called vector or torque VSMs. A second popular VSM arrangement is the axial configuration, where the vibration axis is parallel to an axial field, most often furnished by a superconducting magnet. These coil

configurations are shown in Figure 1. The VSM allows sample magnetic moments from 10^{-6} to many emu to be measured over field ranges up to 30 T in resistive Bitter solenoids and higher fields in hybrid magnets (Foner, 1996) over temperatures from 0.5 to 1000 K or more. Sensitivities as high as 10^{-9} emu have been quoted for closely coupled pickup coils (Foner, 1975). The VSM also has been adapted for high-pressure measurements (Guertin and Foner, 1974) and for very low frequency ac susceptibility measurements (down to 10^{-3} Hz) to observe the frequency dependence of the spin-freezing temperatures of spin-glasses (Guyot et al., 1980). The VSM takes data continuously, so that rapid sweeps of field and temperature can be made.

It is possible to use slow drive magnetometers using flux integration techniques in special circumstances. One technique, developed by Foner and McNiff (1968) for observation of small changes in a large signal, moves the sample beyond the maximum signal in the pickup coils, thus avoiding centering problems. This overextended sample drive motion is sometimes utilized in the commercial SQUID magnetometer.

SQUID Magnetometer. The SQUID is an extremely low impedance, high-sensitivity flux detector capable of operating over a wide range of frequencies (from dc to gigahertz). The magnetometer detects the flux produced by a sample passing slowly through a pickup coil. The SQUID does not measure the field from the sample directly but is coupled inductively to superconducting pickup coils located in the high-field region. Changes in flux set up a persistent current in the pickup coil–SQUID transformer circuit. A feedback, which is generated to either null the persistent current in the superconducting circuit or change the flux in the SQUID, is proportional to the sample moment. The sample is moved slowly through the pickup coils. The SQUID must be kept in a low-field region and is thus shielded by a superconducting shield. The shield not only protects the SQUID from high stray magnetic fields from the superconducting magnet but also reduces ambient laboratory stray field interference. A modification for low-frequency ac susceptibility measurements down to 10^{-3} Hz is discussed by Sager et al. (1993). Superconducting solenoids almost always furnish the field because the pickup coils must be in an extremely stable field while data are taken, so the magnet operates in a persistent mode while the sample is translated between the pickup coils. The peak-to-peak flux changes are measured or the flux changes over the traversal are fitted to a calculated profile. To change the field, the superconducting pickup coil circuit is made nonsuperconducting so that the induced currents are small during the time the field is changed. First, the superconducting magnet is made nonpersistent, and the field is changed, then returned to a persistent mode when the field drift becomes small, after which the pickup coil circuit is made superconducting again. Large field changes produce longer magnet drift times. This sequential procedure is automated in commercial instruments, and although data taking is slow, the systems can be programmed to run for several days before the liquid helium needs replenishment. The programs run sequences of temperature and field without intervention once the sample is mounted. Data are taken point by point at each temperature and/or field.

The SQUID itself is capable of detecting flux changes on the order of $10^{-3}\phi_0$, where ϕ_0, the flux quantum, is given by 1.07×10^{-7} G-cm^2 ($= hc/2e$). The sensitivity of the magnetometer is typically $\sim 10^{-7}$ emu, with higher sensitivities possible using special closely coupled coil configurations. Commercial SQUID magnetometers are currently available with fields to 9 T over a range of 2 to 400 K (1200 K with furnace insert). The system can also include a low-frequency VSM mode, and the SQUID can also be adapted for ac susceptibility measurements.

Force Magnetometers

DC Faraday Magnetometer. Historically, the Faraday magnetometer used an iron core magnet with tapered pole pieces to provide the gradient, from 100 to 1000 Oe/cm. Assuming $dH/dz = 500$ Oe/cm, the balance must be able to detect changes in mass weight of -5 μg (5×10^{-3} dynes) to detect changes on the order of 10^{-5} emu in a 1-cm^3 sample (a very large sample by laboratory standards). To avoid sample displacement when the external field or temperature is changed, the force is generally compensated by a small electromagnet exerting an equal force on a magnetic material on the opposite end of the balance. The feedback to maintain sample position is the direct readout proportional to the sample magnetic moment. Thermal isolation of the transducer from the sample environment is essential for all force transducers.

Accurate and reproducible sample positioning in the field is essential for proper utilization of the Faraday magnetometer. The sample, which is suspended from the balance by a fiber, may have a hang-down weight attached by another fiber to furnish a force to minimize transverse force displacements. Care must be taken to avoid convection currents. Faraday magnetometers have been adapted to solenoidal fields, i.e., superconducting magnets or resistive Bitter solenoids, with field gradient coils controlled separately from the main solenoid. The sample may be placed away from the center of the magnet to capture a large field gradient, but there is a penalty in reduction of the applied field. Commercial Faraday magnetometers are currently available covering a temperature range of 1.5 to 300 K.

Alternating-Gradient Magnetometer. Zijlstra (1970) first described a force magnetometer that attached a small sample to a thin reed in a uniform field and applied an independently controlled ac gradient field tuned to the mechanical resonance of the sample plus the reed. This device is called a vibrating-reed or alternating-gradient magnetometer (AGM). The system is operated typically at a few hundred hertz, above most environmental noise frequencies, and Q values may be over 100. Detection of the vibration motion (normally a few micrometers) is done optically or by means of a piezoelectric bimorph

transducer (Flanders, 1988). The AC gradient is restricted by power dissipation in the gradient coils to ~100 Oe/cm and can be applied transverse or longitudinal to the vertical support. Data can be accumulated continuously with changing field or temperature. Lock-in detection is used to measure the voltage from the transducer locked to the gradient coil frequency.

The AGM has received considerable use in the magnetic tape recording industry because of its ease of operation and its sensitivity, which is quoted at $\sim 10^{-8}$ emu. A commercial instrument using a bimorph transducer has recently been modified by O'Grady et al. (1993) and evaluated for operation between 5 and 300 K using He gas cooling, with sensitivity affected by flow rate from 2×10^{-5} emu at 5 K to 10^{-7} emu at 300 K. Because the resonant system is sensitive to damping, it is not operated directly in a cryogen. The bimorph is sensitive to temperature and the resonant supports are short (a few centimeters), so low-temperature arrangements are not conventional. O'Grady et al. (1993) also examined spurious effects of the gradients on hysteresis loop parameters of thin magnetic films, including squareness, irreversibility, and magnetic viscosity.

Cantilever Magnetometer. Originally developed to operate at the very low temperature of a dilution refrigerator, the cantilever magnetometer employs a thin, flexible platform or beam as a sensing element of magnetic force or torque, with the sample rigidly attached to the free end of the cantilevered beam (Chaparala et al., 1993). In a uniform field, an anisotropic sample will experience a torque and a field gradient is required to provide force for an isotropic sample, so both effects may occur simultaneously. The beam deflection may be detected via capacitance, strain, or optics. The former is the most common. A fixed electrode is placed under the metallized cantilever, leaving a small ($\ll$1-mm) gap. With a cantilever area of 1 cm^2, deflection of <0.1 nm can be measured. Typical sensitivity ranges are 10^{-1} to 10^{-5} N-m for torque and 10^{-7} to 10^{-2} N (10 μg to 1 g weight equivalent) for force, though higher sensitivities have been reported. An alternative is to detect displacement with a piezoresistive sensor (Rossel et al., 1996). Another variation uses a diaphragm in a cantilever arrangement in a uniform magnetic field combined with a modulated gradient field (Bindilatti and Oliveira, 1994). It is also a type of AGM working nonresonantly.

A silicon cantilever magnetometer is a micromachined device that incorporates integrated calibration and null detection, as well as simultaneous electrical resistivity capability. An advantage of the cantilever over other methods is its use from room temperature down to millikelvin temperatures, and the mechanical characteristics change by <10% over this range. Cantilevers have been employed in fields from 10^{-3} to 30 T and in pulsed fields.

Other Alternative Magnetometry Techniques

Many alternative magnetometry methods have been developed for specific applications. These include flux and force measurements as well as indirect techniques that depend on a known functional relationship between the magnetic properties and some other parameter, e.g., resistance, optical transmission, or strain. Many such methods have been reviewed by Foner (1967, 1981) and Zjilstra (1967).

High-Pressure Magnetometry. Most commercially available magnetometer systems have been adapted for use with devices for exerting high pressure on the samples. The VSM and the Faraday balance were adapted by Guertin and Foner (1974) and Wohlleben and Maple (1971), respectively, to incorporate a high-hydrostatic-pressure cell. Both systems operate with hydrostatic pressures to ~1 GPa. SQUID magnetometers have also been adapted for high-pressure use, as has the vibrating-coil magnetometer (Smith, 1956; Ishizuka et al., 1995).

Vibrating-Coil Magnetometer. In a vibrating-coil magnetometer (VCM), the sample is held motionless but the series-opposing pickup coils are vibrated. While this technique may be useful for samples that cannot conveniently be moved, it is extremely difficult to eliminate unwanted pickup from the background field even when it is not varying. The earliest application by Smith (1956) was for measuring a magnetic material in a high-pressure cell. A more recent application of the VCM measured the magnetic moment of a sample sandwiched between diamond anvils under extremely high pressure (Ishizuka et al., 1995). The pickup coils are just outside the pressure cell near the high-pressure region. The coils, which are driven by a piezoelectric bimorph, are coupled to a SQUID.

Rotating-Sample Magnetometer. In a rotating-sample magnetometer (RSM), the sample moves in a circular orbit with the field perpendicular to the plane of the circle (Flanders and Graham, 1993). Pickup coils detect the periodic change in flux, and the Gaussian-like pulse produced is composed of a fundamental frequency (oscillation frequency) and several harmonics, all of which can be detected separately with a lock-in amplifier. The output of the RSM is proportional to the sample moment and frequency of rotation as well as the coupling to the pickup coils. While the sensitivity may be comparable to that of the VSM, the RSM requires more space for operation and the calibration will depend on sample size and shape in a complicated way. The RSM may be used to directly detect magnetic anisotropy by configuring the pickup coils to detect the moment that is perpendicular to the axis of rotation. For this measurement, the sample is a circular cylinder with its axis coincident with the rotation axis.

Surface Magnetometry. Though this unit is restricted to the review of measurements of the bulk sample magnetic moment, very useful surface measurements can be used to analyze properties of materials in situ and nondestructively. Selected examples are mentioned below.

Kerr Effect Magnetometry. The plane of polarization is rotated when a beam of polarized light is transmitted along the magnetization direction in a magnetized medium. When transmitted through the sample, it is called

the Faraday effect; when reflected from a surface, it is called the Kerr effect. The rotation of the polarization depends on frequency and is proportional to the effective path length of the light, the Kerr constant of the material, and the magnetization. Assuming all else is fixed, the Kerr effect measures the relative magnetization on the surface. By varying the field, hysteresis loops of the relative magnetization are generated. Automated Kerr magnetometers are used to scan the surface of large recording discs. It should be noted that the measurement probes the optical penetration depth, which may not reflect the bulk properties, e.g., if there are surface strains.

Hall Effect Sensor. A Hall effect sensor is a convenient and often quite compact element that can be used for measuring magnetic field. The Hall voltage V_H across a flat slab with thickness d is given by $V_H = RIH_0 \sin \Phi d$, where I is the currents, H_0 is the applied field, Φ is the angle between them, and R is the Hall constant given by $1/nec$, where n is the number of carriers per cubic centimeter, e is the elctronic charge, and c is the velocity of light. Note that V_H is maximum when it is measured along a direction that is perpendicular to the current and both are perpendicular to H_0. The sign of V_H depends on the direction of H_0 and is proportional to H_0, not dH_0/dt. Semiconductors are often used for sensors because n is small and V_H is large. By applying an AC current, the thermoelectric voltages, which are large for small n, are minimized. Typical sensors can be quite small; commercial elements have active areas of 1 mm^2 or less and can be compensated so that R is field and temperature independent to better than 1% for H_0 up to 23 T between 1.5 and 300 K (Sample and Rubin, 1976). Because these sensors are small, they fit into small regions as field sensors, but care must be used to assure that H_0 is perpendicular to the flat area. The Hall device can be used as a field sensor in magnetometers (Flanders and Graham, 1993; see DEEP-LEVEL TRANSIENT SPECTROSCOPY). If the Hall sensor can be placed in very close proximity to the surface of a magnetic material, it detects the normal component of **B** at the surface averaged over the sensor surface. Several researchers have used Hall probes to observe the magnetic properties of ferromagnetic and superconducting materials by scanning the surface.

Other Surface Probes. There are many surface probes used for special applications. Very useful are means to nondestructively evaluate ferromagnetic materials. When a magnetic field is applied, imperfections such as cracks produce local gradients that can be detected. One example, called a "magnetoscope," analyzes the hysteresis loop (Eichmann et al., 1992) to determine parameters such as applied stress, residual stress, and fatigue damage. Very powerful techniques involve derivatives of the scanning probe microscopes such as the magnetic force microscope for detecting surface magnetism at the atomic level.

Procedures for Magnetometry Measurements

Calibration. In principle, the response of both flux and force techniques can be calculated from first principles—flux from the law of induction and knowledge of the coil's area-turns, force from energy considerations and knowledge of the applied field profile. However, small variations in geometry, stray fields, or electronic or transducer sensitivity drift make this approach difficult, so that routine calibration by a specific sample is highly recommended. Good operating procedures dictate measurement with a calibration sample both before and after a series of measurements. A standard sample should be set aside specifically for calibration and testing of the magnetometer starting when the system is initially put into operation. (As an extra precaution, it is a good idea from time to time to measure the moment of just the sample support structure.)

A convenient and reliable calibration material is a spherical sample of pure, polycrystalline Ni with a mass ~10 mg measured in $H_0 > 10{,}000$ G (1 T) at saturation where the moment is independent of field (Eckert and Sievert, 1993). The calibration values are 58.6 emu/g at 4.2 K and 55.1 emu/g at 300 K. Nickel can be obtained with high purity at low cost, and it does not oxidize. Any appreciable increase in the moment above saturation could signal sample chamber contamination, e.g., by solid O_2 (see below). A paramagnetic material can be used as an alternative, but the calibration depends on knowing both the temperature and field.

Sample Size. When a sample is sufficiently small, a magnetic dipole approximation may be assumed. If the sample size approaches the size of the pickup coils in a flux measurement or subtends a very large portion of the sample chamber in a force measurement, corrections must be made. Large samples give results that are more sensitive to position than small samples. Two approaches for correcting for sample size can be used: (1) measure the large sample moment and use data from a small sample of the same material at a convenient point for normalization and (2) calculate the errors for a specific sample size and coil dimensions. Such corrections have been tabulated for VSMs (Zieba and Foner, 1982; Ausserlechner et al., 1994). For example, for a thin rod with normalized length $c/a_o = 0.5$, where a_o is the radius of a thin optimized coil pair and c is the length parallel to the coil axis, the error is ≈3% (Zieba and Foner, 1982). Corrections for force techniques with large or nonspherical or anisotropic samples require inclusion of x and y gradients as well as the effect of large z gradients.

Sample Position. It is essential to know sample position precisely for the full temperature and field range of measurement, because to some degree all the techniques discussed are sensitive to sample position. For flux measurements, uncertainty in sample position can be minimized but not eliminated by pickup coil design. It is important to know these characteristics for a given system so that estimates of reliability and errors of the data can be made. Force methods require knowledge of the various position dependencies of the field and field gradients. Tests should be conducted to assure that the sample is positioned where it is thought to be! Some of these effects can be large. Calculations for an axial field configuration SQUID system show that errors of up to 30% can occur if

the sample is sufficiently displaced radially (Miller, 1996). The dimensions of sample support structures including long support rods will in general vary with temperature, so it is good practice to compensate for this when necessary by readjusting the sample position at various temperatures.

Sample Temperature. Most magnetometers operating at temperatures below room temperature use He gas cooling to establish the temperature, for $5 < T < 300$ K. The temperature sensor should be as close to the sample as possible. It is also essential that the magnetic field dependence of the sensor calibration be known to high accuracy; this is particularly important at low temperatures. Sample temperature stability is achieved more slowly at high temperatures because the heat capacities of sample, support structure, and sensor increase rapidly for $T > 20$ K. Operation for $1.5 < T < 4.2$ K requires liquid ^{4}He and reduced pressures; extension to liquid ^{3}He temperatures ($0.3 < T < 1.5$ K) and lower using dilution refrigerators requires special adaptations. Measurements at high temperatures ($T > 300$ K) require a thermally shielded oven in the field volume. The furnace should be noninductively wound to minimize background effects. In addition, sample mounts and support elements must be chosen to withstand these temperatures and add minimum magnetic contributions.

METHOD AUTOMATION

Several turnkey automated DC magnetometer systems are commercially available, and more are under development. These systems use desktop computers for convenient data acquisition, control of independent variables such as temperature and magnetic field, data analysis, and real-time display of data for on-line assessment. Most commercial SQUID systems are fully automated with respect to magnetic field and temperature programming (using gas flow control), and typical experimental runs may extend for days without intervention. More recently, some vendors have provided similar automation for VSM systems, which generally have somewhat higher external magnetic fields. The inherent instability of the sample position in force detection magnetometers implies that temperature control may be more difficult; thus gas flow temperature control is generally not employed in these systems.

DATA ANALYSIS AND INITIAL INTERPRETATION

The following simulation using isothermal magnetization data for a reference (calibration) sample and a test material serves to illustrate the method for analyzing magnetic moment data. In this case, a small, pure, polycrystalline Ni sphere is used as the reference sample with data taken at 4.2 K. Refer to Figure 3.

At $T = 4.2$ K, the saturation moment of Ni is 58.6 emu/g. (It is ~5% smaller at room temperature.) Therefore, the measured moment at a field sufficient to saturate the

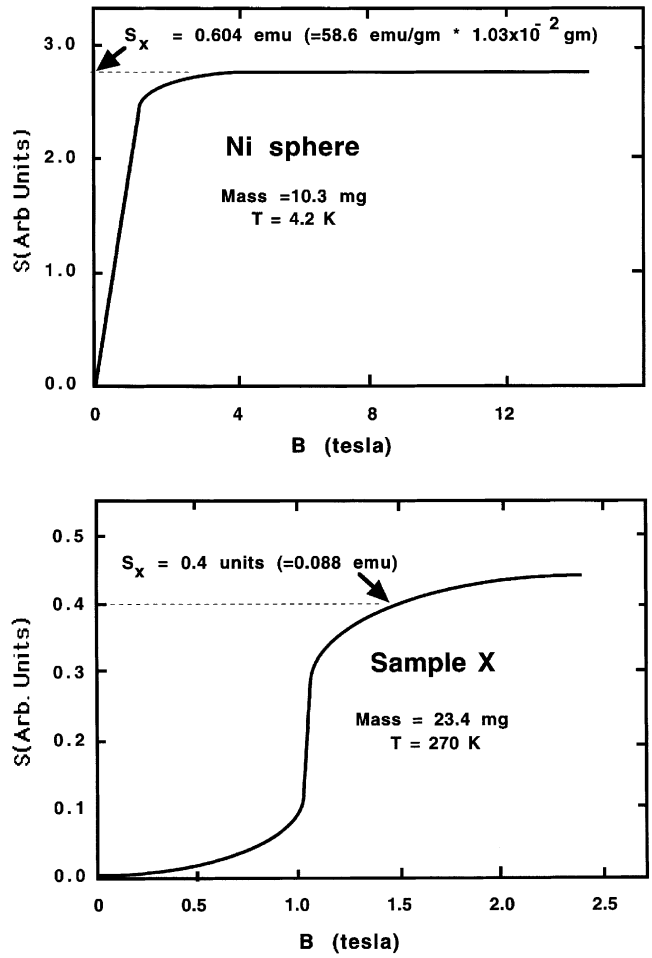


Figure 3. Simulation showing how the saturation moment of a polycrystalline Ni sphere may be used to calibrate the magnetization of an unknown, sample X.

magnetization, S_{Ni}, will be $S_{Ni} = Km_{Ni}\sigma_{Ni}$, where K is a constant of proportionality dependent on the pickup coil geometry, m_{Ni} the mass, and σ_{Ni} (= 58.6 emu/g) the saturation moment. This provides the calibration of the signal, S, with which all other samples can be measured. From Figure 3, the test point ($S_x = 0.4$ units) yields a total moment of ~0.088 emu [$= (0.4/2.75) \times 0.604$] or $\sigma_x = 3.75$ emu/g using $m_x = 2.34 \times 10^{-2}$ g.

Sample X, the test sample, represents a material that appears to have a metamagnetic transition near room temperature at about $\mathbf{H} = 1$ T.

Both the reference sample and the test sample are presumed to be small relative to the pickup coils in the case of a flux detection magnetometer (sample dimensions less than about half the inner dimension of the pickup coil). For force detection magnetometers, (1) the sample dimensions should be small relative to the change in the field and field gradient, and (2) this procedure assumes a constant field gradient, otherwise a field dependent correction for gradient is required.

Although this illustration involves isothermal magnetization, isofield magnetization as a function of temperature

(which in the limit of zero field is the magnetic susceptibility) can also be used to calibrate an unknown sample moment provided both temperature and moment are known. Corrections for larger samples (causing K to be slightly different in the two cases) may be made, and the reader is referred to Zieba and Foner (1982).

PROBLEMS

System Automation

Automation and computer control of turnkey instruments tend to produce the trusting "black box" syndrome. The appropriate response is to trust and then verify by regularly making measurements with the standard calibration sample, especially before and after a given set of measurements. A record of the readouts for the standard is useful to observe whether the system is degrading with time. Temperature sensors and amplifiers can drift with time: the calibrations of temperature and field should also be tested periodically.

Several issues arise from system automation that deserve attention, including the possibility of overshooting the target temperature and/or field. The system control may cause the sample temperature to overshoot (or undershoot if decreasing) relative to the target temperature. In systems that are magnetically reversible (nonhysteretic), paramagnetic, or weakly diamagnetic, this is not a problem, but for ferromagnetic or superconducting systems with magnetic hysteresis, this can yield spurious results. Overshooting or undershooting the external magnetic field relative to the target can also be problematic.

Contamination of Sample Chamber

It is important to assure that the sample chamber in the magnetometer is contaminant free, and to this end, the volume should be cleaned at convenient intervals. A few of the most common contaminants are discussed below.

Generally, the most ubiquitous source of contamination is Fe in the form of small particles or oxides (dirt) that may stick to the sample or container. An Fe film [1 nm (10 Å) over 1 cm^2] left by a machine tool on a sample mount will show a moment of $\sim 2 \times 10^{-4}$ emu at saturation. Iron contamination is easily observed as a magnetization nonlinearity as it is magnetized. Most surface Fe can be removed easily through etching in a dilute HCl solution.

A common source of contamination is superconducting solder at low temperatures and fields ($T < 6\,\mathrm{K}$ and $\mathbf{H} < 1000\,\mathrm{G}$). Superconducting contaminants such as solder can produce spurious diamagnetic or paramagnetic signals depending on their location relative to the transducer and on their history of exposure to external fields. The use of "nonmagnetic" stainless steels in sample mounts or cryogenic vessels is common. However, most stainless steels are strongly paramagnetic, especially at low temperatures. If mechanically worked or brazed, some stainless steels can become ferromagnetic.

Finally, for low-temperature measurements, it is important to assure that no air (oxygen) is trapped in the sample tube. Below ~80 K, oxygen (O_2) may condense on the sample or sample mount. Oxygen is paramagnetic, spin $S = 1$, and becomes solid below ~50 K, depending on partial pressure. Oxygen can accumulate slowly with time through leaking sliding seals and during sample changes. The only solution to O_2 contamination is to warm up the contaminated region, pump out the gas, and backfill with high-purity He gas. Nonsaturation of a Ni standard magnetization and/or erratic values of the Ni saturation moment may indicate solid O_2 contamination. Unfortunately the magnetism literature is replete with evidence for solid O_2 contamination of data. For example, the amount of O_2 gas in 1 mm^3 of air at STP will yield a moment of $\sim 2 \times 10^{-4}$ emu at $T = 4.2$ K and $\mathbf{H} = 10{,}000$ G (1 T).

Sample Support Structure

All support structures for sample mounts will show magnetic moments. Their contributions can be minimized in several ways for flux techniques, including minimizing the mass of the sample container and mount, minimizing a uniform cross-section support rod, and extending this support rod symmetrically well beyond the pickup coils so that no flux change is detected as the rod itself is translated. As stated above, it is a good practice to measure the moment of the empty sample mount from time to time. If it is generally <10% of the sample moment, it can be subtracted accurately to give reliable sample moment results.

Phase Considerations

In most ac signal detection systems, a lock-in amplifier provides the dc output signal proportional to the sample moment. Generally, the in-phase lock-in signal is proportional to the sample moment, but the out-of-phase component should be monitored, as it may change due to extraneous signals, and in some instances can overwhelm the in-phase signal. Eddy currents induced in the sample or sample mount (see below) may be detected as out-of-phase components relative to the in-phase sample moment.

Magnetic Field Inhomogeneities

In flux-detection methods where the sample is generally moved, inhomogeneities in the external magnetic field may lead to several problems: if the sample is highly conductive (a good metal), motion of the sample through a field inhomogeneity will induce eddy currents that may be detected at the transducer as a false moment. This problem is particularly severe if the sample is encased in a metallic high-pressure vessel and the eddy currents are induced in the pressure vessel. In general, eddy currents induced in the sample or sample mount and hysteresis losses will produce out-of-phase components that can be erroneously detected as real sample moments. Although the sample motion excursion is small for the VSM, motion through field inhomogeneities will produce an out-of-phase signal. For the extraction magnetometer, even though the excursion is large, waiting for the eddy

currents to decay is an option. Another option is to have a uniform magnetic field over the region of the sample motion. Another problem arises if the sample is in powder form: moving an unconstrained and strongly magnetic powdered sample through a field inhomogeneity can cause the powder to move relative to the container, thereby giving a false magnetization value. In general, any powdered or liquid sample should be fully constrained in a container for magnetometry measurements.

ACKNOWLEDGMENTS

The authors are grateful for the contribution on cantilever magnetometry by M. J. Naughton, formerly of the Physics Department, State University of New York at Buffalo, and currently in the Physics Department, Boston College, Chestnut Hill, Mass. R. P. Guertin wishes to thank his hosts at the National High Magnetic Field Laboratory, Tallahassee, Fla., where he was on leave for the 1996–1997 academic year, during which much of this unit was written.

LITERATURE CITED

Ausserlechner, U., Steiner, W., and Kasperkovitz, P. 1994. Vector measurement of the magnetic dipole moment by means of a vibrating sample magnetometer. *IEEE Trans. Magn.* 30: 1061–1063.

Bernards, J. P. C. 1993. Design of a detection coil system for a biaxial vibrating sample magnetometer and some applications. *Rev. Sci. Instrum.* 64:1918–1930.

Bindilatti, V. and Oliveira, N., Jr. 1994. Operation of a diaphragm magnetometer in a plastic dilution refrigerator with an oscillating field gradient. *Physica B* 194–196:63–64.

Bowden, G. J. 1972. Detection coil systems for vibrating sample magnetometers. *J. Phys. E* 5:1115–1119.

Chaparala, M., Chung, O. H., and Naughton, M. J. 1993. Capacitance platform magnetometer for thin film and small crystal superconductor studies. *A.I.P. Conf. Proc.* 273:407–413.

Eckert, D. and Sievert, J. 1993. On the calibration of vibrating sample magnetometers with the help of nickel reference samples. *IEEE Trans. Magn.* 29:3001–3003.

Eichmann, A. R., Devine, M. K., Jiles, D. C., and Kaminski, D. A. 1992. New procedures for the in situ measurement of the magnetic properties of materials: Applications of the magnescope. *IEEE Trans. Magn.* 28:2462–2464.

Flanders, P. J. 1988. An alternating-gradient magnetometer. *J. Appl. Phys.* 63:3940–3945.

Flanders, P. J. and Graham, C. D., Jr. 1993. DC and low frequency magnetic measurement techniques. *Rep. Prog. Phys.* 56:431–492.

Foner, S. 1956. Vibrating sample magnetometer. *Rev. Sci. Instrum.* 27:548.

Foner, S. 1959. Versatile and sensitive vibrating-sample magnetometer. *Rev. Sci. Instrum.* 30:548–557.

Foner, S. 1967. Special measurement techniques. *J. Appl. Phys.* 38:1510–1519.

Foner, S. 1975. Further improvements in vibrating sample magnetometer sensitivity. *Rev. Sci. Instrum.* 46:1425–1426.

Foner, S. 1981. Review of magnetometry. *IEEE Trans. Magn.* MAG-17:3358–3363.

Foner, S. 1994. Measurement of magnetic properties and quantities. *In* Encyclopedia of Applied Physics, Vol. 9 (G.L. Trigg, ed.). pp. 463–490. VCH Publishers, New York.

Foner, S. 1996. The vibrating sample magnetometer: Experiences of a volunteer. *J. Appl. Phys.* 79:4740–4745.

Foner, S. and McNiff, E. J., Jr. 1968. Very low frequency integrating vibrating sample magnetometer (VLFVSM) with high differential sensitivity in high DC fields. *Rev. Sci. Instrum.* 39: 171–179.

Guertin, R. P. and Foner, S. 1974. Application of a vibrating sample magnetometer to magnetic measurements under hydrostatic pressure. *Rev. Sci. Instrum.* 45:863–864.

Guyot, M., Foner, S., Hasanain, S. K., Guertin, R. P., and Westerholt, K. 1980. Very low frequency AC susceptibility of the spin-glass PrP_{1-y} ($y \leq 1$). *Phys. Lett. A* 79:339–341.

Ishizuka, M., Amaya K., and Endo, S. 1995. Precise magnetization measurements under high pressure in the diamond-anvil cell. *Rev. Sci. Instrum.* 66:3307–3310.

Krause, J. K. 1992. Extraction magnetometry in an AC susceptometer. *IEEE Trans. Magn.* 28: 3063–3071.

Mallinson, C. 1992. Gradient coils and reciprocity. *IEEE Trans. Magn.* 27:4398–4399.

Miller, L. L. 1996. The response of longitudinal and transverse pickup coils to a misaligned magnetic dipole. *Rev. Sci. Instrum.* 67:3201–3207.

O'Grady, K., Lewis, V. G., and Dickson, D. P. E. 1993. Alternating gradient force magnetometry: Applications and extensions to low temperatures. *J. Appl. Phys.* 73:5608–5613.

Pacyna, A. W. and Ruebennauer, K. 1984. General theory of vibrating magnetometer with extended coils. *J. Phys. E* 17: 141–154.

Rossel, C., Bauer, P., Zech, D., Hofer, J., Willemin, M., and Keller, H. 1996. Active microlevers as miniature torque magnetometers. *J. Appl. Phys.* 79:8166–8173.

Sager, R. E., Hibbs, A. D., and Kumar, S. 1993. Using SQUIDS for AC measurements. *IEEE Trans. Magn.* 28:3071–3077.

Sample, H. H. and Rubin, L. G. 1976. Characterization of three commercially available Hall effect sensors for low temperatures and fields to 23 T. *IEEE Trans. Magn.* MAG-12:810–812.

Smith, D. O. 1956. Development of a vibrating coil magnetometer. *Rev. Sci. Instrum.* 27:261–268.

Wohlleben D. and Maple, M. B. 1971. Application of the Faraday method to magnetic measurements under pressure. *Rev. Sci. Instrum.* 42:1573.

Wolf, W. P. 1957. Force on an anisotropic paramagnetic crystal in an inhomogeneous field. *J. Appl. Phys.* 28:780–781.

Zieba, A. and Foner S. 1982. Detection coil, sensitivity function and magnetic geometry effects for vibrating sample magnetometer. *Rev. Sci. Instrum.* 53:1344–1354.

Zilstra, H. 1967. Experimental Methods of Magnetism, Vols. 1 and 2 (E.P. Wohlfarth, ed.). John Wiley and Sons, New York.

Zilstra, H. 1970. A vibrating Reed magnetometer for microscopic samples. *Rev. Sci. Instrum.* 41:1241–1243.

KEY REFERENCES

Flanders and Graham, 1993. See above.

An extensive review of dc and low frequency magnetism measuring techniques.

Foner, 1994. See above.

A thorough review of magnetometry emphasizing low frequency techniques.

Zilstra, 1967. See above.

A review of magnetism measuring techniques introducing the alternating gradient magnetometer.

ROBERT P. GUERTIN
Tufts University
Medford, Massachusetts

SIMON FONER
Massachusetts Institute of Technology
Cambridge, Massachusetts

THERMOMAGNETIC ANALYSIS

INTRODUCTION

The most common method of measuring magnetic ordering temperatures is through the use of thermomagnetic analysis techniques. These techniques rely on methods previously described (GENERATION AND MEASUREMENT OF MAGNETIC FIELDS, MAGNETOMETRY, and THEORY OF MAGNETIC PHASE TRANSITIONS). In essence, thermomagnetic analysis involves sensitive measurement of a magnetic dipole moment in a zero or an applied field as a function of temperature. The components of the system therefore include an appropriate magnetometer. Superconducting quantum interference device (SQUID), vibrating-sample, and Faraday-balance magnetometers are commonly used in the field (see MAGNETOMETRY). Large axial fields can be applied using a superconducting solenoid; smaller axial fields can be applied using wound-metal solenoids. Transverse fields can be applied using split-coil superconducting solenoids or electromagnets. The details of these are discussed in previous units. Measurement of the temperature dependence of magnetic properties requires careful design of the sample chamber and methods for heating and cooling, as well as accurate thermometry for monitoring temperature. Measurement of magnetic transitions of the types described in the abovementioned units may entail the measurement of magnetic properties at temperatures ranging from cryogenic to $\sim$1100°C (the T_c of elemental Co).

PRACTICAL ASPECTS OF THE METHOD

The data illustrated in the figures in this unit, probing magnetic phase transitions, were taken using either a SQUID magnetometer or a vibrating sample magnetometer. We give an abbreviated discussion of these systems here; generation of fields and measurement of dipole moments are treated in more detail in GENERATION AND MEASUREMENT OF MAGNETIC FIELDS and MAGNETOMETRY, respectively. A typical SQUID magnetometer is equipped with a 5.5-T superconducting solenoid, temperature control between 1.7 and 400 K (using a nitrogen-jacketed liquid He Dewar), and could measure dipole moments from 10^{-8} to 2 emu. High-moment options can allow measurement to several hundred electromagnetic units. Options for higher fields, transverse fields, and high temperatures ($\sim$800 K) all exist. Vibrating-sample magnetometers are typically equipped with a 1- to 2-T electromagnet, though superconducting solenoids can be used for larger fields. Options for inert gas ovens up to 1273 K exist. Modern sensitive pick-up coils allow for measurement of moments between 5×10^{-6} and 1×10^{4} emu. Options for fields up to 9 T (using a superconducting solenoid) and temperatures as low as 1.5 K using a liquid He Dewar are possible.

The principle of operation of a vibrating sample magnetometer (VSM) is based on placing a magnetic sample in a uniform magnetic field. The sample dipole moment is made to undergo a periodic sinusoidal motion at a fixed frequency using a transducer drive head to vibrate a sample rod. The vibrating magnetic dipole moment (through Faraday's law of induction) induces a voltage in a sensitive set of pick-up coils placed between the pole pieces of the electromagnet. This signal, proportional to the magnetization, is amplified and monitored. The VSM is calibrated using a standard sample of known moment such as a Ni (ferromagnetic) or Pt (diamagnetic) sphere.

The term SQUID is an acronym for superconducting quantum interference device. Superconducting and magnetic transition temperature (T_c), magnetic susceptibility, magnetic hysteresis, and magnetic relaxation can be measured using a SQUID magnetometer. During a measurement of magnetic moment, the previously centered sample is moved through a set of signal coils. The differential sensitivity of a typical SQUID is $\sim 10^{-8}$ emu and T stability below 100 K is $\pm 0.5\%$. Field inhomogeneity in the solenoidal magnet of the magnetometer can cause superconducting samples to undergo minor hysteresis loops during travel through the SQUID coils. Therefore, a scan length of 3 cm is used in order to minimize the inhomogeneity to $<0.05\%$.

The principles of operation of a typical SQUID magnetometer system for measuring a sample are briefly summarized as follows (Quantum Design, 1990).

1. The sample is transported through highly balanced second-derivative detection coils inducing a voltage proportional to dM/dt and therefore to M at constant frequency.
2. The signal detected in the second-derivative coils is coupled to the SQUID sensor through a superconducting isolation transformer. Voltage readings sensed by the detection coils are converted into a corresponding magnetic flux signal via the mutual inductance between the SQUID loop and coils.
3. After the raw data have been collected, the magnetic moment of the sample is computed by numerical methods using the previously determined system calibration factors. The output data thus consist of T, H, measured magnetic moment, time, and percentage error, etc. Since multiple measurements

are made in one scan, and since it is possible to perform several scans, a mean and standard deviation of the moment can be ascertained under a given set of measurement conditions.

One practical aspect of taking magnetization data using a superconducting solenoid is to recognize that superconducting solenoids can trap appreciable fields, i.e., on the order of 100 Oe for a 5- to 10-T magnet. Therefore, if a low-field measurement is desired, it is necessary to oscillate remnant fields out of the solenoid. Even with this procedure it is not possible to guarantee much less than ~1 Oe of trapped fields. For more stringent field requirements, normal conductor solenoids can be used; however, the ~0.3-Oe field of the earth must also be considered.

In the case of a SQUID magnetometer, it is possible to have additional transverse SQUIDS to measure orthogonal components of the magnetization. For a VSM, additional transverse pick-up coils can be used for the same purpose. Rotation options also exist for varying the angle of the applied field with respect to the sample reference frame for both devices. These require a greater degree of sophistication for axial (solenoidal) fields than for transverse fields.

There are several important issues in sample preparation for thermomagnetic measurements. These include issues of size, shape, weight, density, and crystallographic orientation. The sample size governs whether the magnetic sample can be considered as a point dipole, which is important in the measurement and modeling of the magnetic moment from the specimen. Sample size and weight, of course, also determine the size of a dipole moment, and must be considered along with the sensitivity of the magnetometer in order to measure statistically significant data. Sample shape is extremely important in understanding magnetic response. Demagnetization effects due to surface dipoles are purely a geometrical effect. As a result of demagnetization, the sample experiences local internal fields that may differ markedly from the externally applied field. Since most magnetic measurement techniques measure dipole moments, in order to calculate the sample magnetization the sample volume or weight and density must be known precisely (see MASS AND DENSITY MEASUREMENTS). The unsaturated magnetization can be strongly dependent on crystallographic orientation in anisotropic materials; it is therefore desirable to know the field orientation with respect to the crystallographic axes in single-crystal specimens and to employ methods to average over orientations in the case of powder samples. Because the application of a field in directions orthogonal to the magnetization results in a torque on a magnetic sample, samples need to be appropriately immobilized so as to prevent rotation in the field. Common morphologies for magnetic specimens include (1) powders, (2) plates, films or foils, and (3) crystalline parallelopipeds. To correct for demagnetization effects, ellipsoidal specimens are useful because the demagnetization factor is known analytically.

For cylindrical samples in a transverse field or for noncylindrical geometries, demagnetization effects need to be considered. The internal field can be expressed as

$$H_i = H_a - DM \tag{1}$$

where D is the demagnetization factor (e.g., $\frac{1}{3}$ for a sphere or $\frac{1}{2}$ for a cylinder or infinite slab in a transverse field). It follows that

$$H_i = \frac{H_a}{(1-D)} \tag{2}$$

and

$$M = \frac{-1}{4\pi(1-D)}H \tag{3}$$

For noncylindrical geometries, demagnetization effects imply that the internal field can be concentrated to values that exceed the applied field H_a.

METHOD AUTOMATION

All commercially available SQUID magnetometers have computer-control modules for sweeping field and temperature and measuring magnetic moment using general-purpose interface bus (GPIB) interfaces and/or proportional/integral differential (PID) temperature controllers. For controlling the field with an electromagnet, the current to the electromagnet coils is controlled and the field monitored with a Hall probe. In an electromagnet, control of a bipolar power supply is crucial to ensure the field changes sign smoothly through zero. For field control in a superconducting solenoid, a solenoidal field current calibration is used. Superconducting solenoids can be swept or latched in a persistent current mode.

Proportional/integral differential temperature controllers coupled with thermocouple measurement and gas-flow controllers allow for the monitoring and control of temperature in both the SQUID and VSM systems. It is therefore possible with both devices to measure the magnetization as a function of field or temperature. It is common to measure isothermal magnetization curves or magnetization at fixed field as a function of temperature. Computerized GPIB control allows for monitoring of field temperature and magnetic dipole moment during the course of an experiment.

For monitoring magnetic phase transitions, efficient heating and cooling are required as well as accurate thermometry for measuring temperature (see THERMOMETRY). Typical high-temperature oven assemblies on VSMs employ an electrically heated tube assembly with vacuum and reflective thermal insulation. With ~60 W of power, a maximum temperature of 1273 K can be reached. Components such as sample holders and oven parts are made from nonmagnetic components. The sample chamber can be evacuated and back-filled with inert gas to prevent sample oxidation. Chromel-alumel thermocouples, mounted in the heater, are used to monitor and facilitate temperature control. For more sensitive and accurate measurements, a thermocouple can be mounted directly in the sample zone.

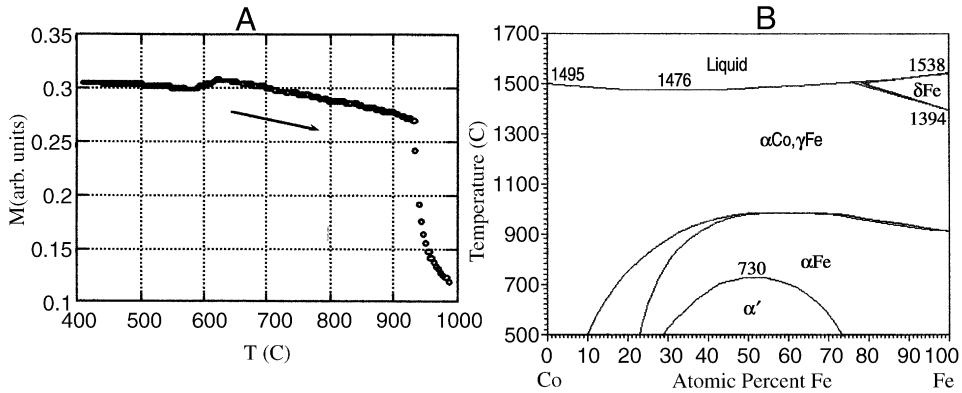


Figure 1. (**A**) $M(T)$ data for C-arc synthesized equiatomic Fe-Co nanoparticles measured using a Lakeshore Model 7300 VSM (Lakeshore Cryotronics, 1995) and oven assembly at 500 Oe, on a second heating cycle to 995°C after initial heating to 920°C (Turgut et al., 1997). (**B**) Fe-Co phase diagram (produced using TAPP software from ES Microware).

EFFECTS OF STRUCTURAL PHASE TRANSFORMATIONS

A complicating aspect of the measurement of magnetic phase transitions is the existence of other structural phase transformations that influence the magnetic properties. These include such transformations as crystallization processes in amorphous magnets, order-disorder, and martensitic or other structural phase transitions. As an example of such complications, consider thermomagnetic data for equiatomic Fe-Co nanoparticles as shown in Figure 1. Fe-Co alloys undergo an order-disorder transformation at a temperature of 730°C at the composition $Fe_{50}Co_{50}$, with a change in structure from the disordered α-BCC(A1) to the ordered α′-CsCl(B2)-type structure. At T > 900°C, $Fe_{50}Co_{50}$ transforms from the magnetic BCC phase to a nonmagnetic FCC λ phase.

Figure 1 shows $M(T)$, at $H = 500$ Oe, for equiatomic FeCo nanocrystalline powders measured on a second heating cycle after initial heating to 920°C. The influence of ordering is observed prominently in features that are quite similar to thermomagnetic observations for bulk $Fe_{49}Co_{49}$ V_2 alloys. These features include an increase in a discontinuity in $M(T)$ at ~600°C (due to chemical ordering) and a return to the extrapolated low temperature branch of the curve at the disordering temperature of ~730°C. Also of note in Figure 1 is that, at ~950°C, we observe an abrupt drop in the magnetization, which corresponds to the $\alpha \Rightarrow \gamma$ structural phase transformation. At $H = 500$ Oe, we observe a prominent Curie tail above this transformation temperature, indicating paramagnetic response. It can be concluded from the abruptness in the drop of $M(T)$ that the $Fe_{0.5}Co_{0.5}$ has a Curie temperature exceeding the $\alpha \Rightarrow \gamma$ phase transformation temperature. As such, a magnetic phase transition is not observed; instead it can be concluded that (ferro)magnetic α-$Fe_{0.5}Co_{0.5}$ transforms to non(para)-magnetic γ-$Fe_{0.5}Co_{0.5}$. This is corroborated in the differential thermal analysis data (see DIFFERENTIAL THERMAL ANALYSIS AND DIFFERENTIAL SCANNING CALORIMETRY), where the order-disorder and $\alpha \Rightarrow \gamma$ phase transformations have also been clearly observed.

Another interesting feature of magnetic phase transitions occurs in materials with large amounts of structural disorder, in which deviations in the exchange interactions can cause dramatic differences in the shape of the thermomagnetic response. The extreme case of this is in amorphous materials that do not have any structural long-range order. In amorphous or disordered ferromagnets, we use the premise that the mean field theory results discussed in MAGNETIC MOMENT AND MAGNETIZATION must be modified to account for deviations in the exchange interactions (due to deviations in atomic positions) in the amorphous alloy. In this limit

$$m = \frac{1}{2}\{B_J[x(1+\delta)] + B_J[x(1-\delta)]\} \qquad (4)$$

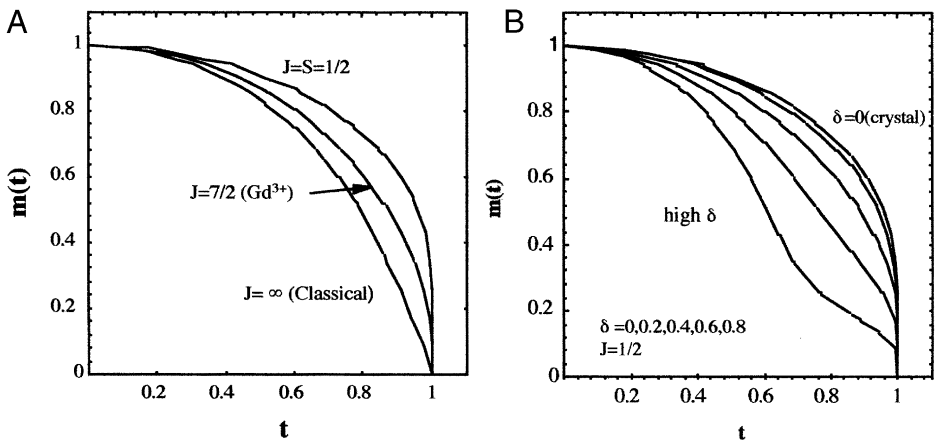


Figure 2. Comparison of magnetization as a function of reduced temperature, t, (**A**) in a spin-only ferromagnet (using a $J = 1/2$ Brillouin function), for $J = 7/2$, and for the classical limit $J = 8$; and (**B**) in an amorphous magnet with $J = 7/2$ and different values of the exchange fluctuation parameter δ. (Courtesy of Hiro Iwanabe, Carnegie Mellon University.)

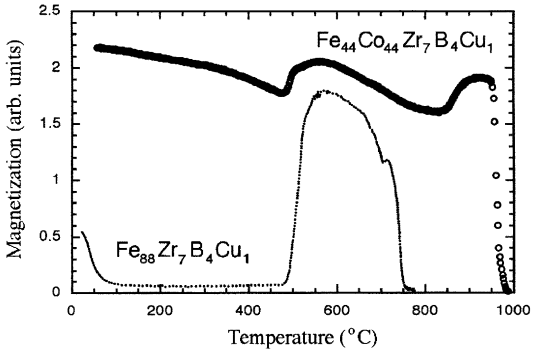


Figure 3. $M(T)$ (Willard et al., 1998, 1999) for an alloy with a NANOPERM composition $Fe_{88}Zr_7B_4Cu$ and an alloy with a HITPERM composition, $Fe_{44}Co_{44}Zr_7B_4Cu$.

where δ is the root mean square deviation in the exchange interaction

$$\delta\sqrt{\frac{\Delta Jex^2}{\langle Jex^2\rangle}} \tag{5}$$

Figure 2A compares mean field results for $M(T)$ in the classical, spin only, and total angular momentum representations. Figure 2B shows mean field results for $M(T)$, taking into account the disorder-induced fluctuations in the exchange parameter in hypothetical amorphous alloys. This construction then predicts a quite remarkable change in the mean field theory thermomagnetic response for amorphous magnets (Gallagher et al., 1999).

Viewed in another light, magnetic measurements can also offer sensitive probes of structural phase transitions. Examples of the use of such probes are too numerous to detail. However, by way of example, we consider the use of thermomagnetic data to probe the Curie temperature of an amorphous ferromagnet and reentrant ferromagnetism associated with crystallization of the same. For a $Fe_{88}Zr_7B_4Cu$ (NANOPERM) and $Fe_{44}Co_{44}Zr_7B_4Cu$ (HITPERM) alloy, $M(T)$, measured with a Lakeshore Cryotronics VSM with oven assembly, is illustrated in Figure 3. For the $Fe_{88}Zr_7B_4Cu$ alloy, the disappearance of M at $\sim$100°C reflects the Curie temperature of this alloy. The reentrant magnetization at $\sim$500°C reflects the crystallization of the initially amorphous material with the crystallization product being α-Fe, with an $\sim$770°C Curie temperature. For the $Fe_{44}Co_{44}Zr_7B_4Cu$ (HITPERM) alloy, $M(T)$ does not disappear prior to crystallization, so that reentrant ferromagnetism is not observed. However, the crystallization event at 500°C results in an $\alpha(\alpha')$-FeCo phase. This phase has a structural phase transition ($\alpha \Rightarrow \gamma$), as discussed above, at 985°C. The $M(T)$ response therefore probes both the amorphous $\Rightarrow$ crystalline and $\alpha \Rightarrow \gamma$ phase transition in this material.

DATA ANALYSIS AND INITIAL INTERPRETATION

Figure 4 shows magnetization versus temperature data for a 0.25-cm^3 cubic Gd single crystal taken in a field of 10 Oe. Several points are worth noting in interpreting such data.

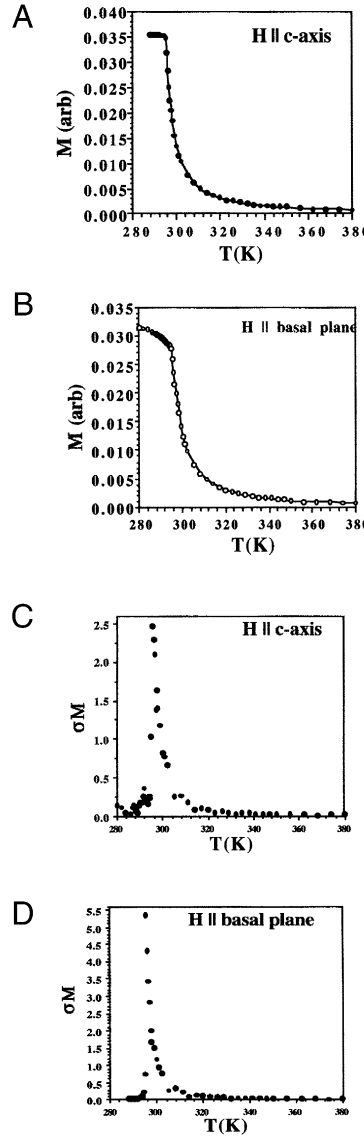


Figure 4. Magnetization-versus-temperature data for a cube-shaped (0.25-cm^3) Gd single crystal taken in a field of 10 Oe for (**A**) $H \parallel$ c and (**B**) $H \parallel$ basal plane, respectively (data taken using a Quantum Design MPMS SQUID magnetometer; units scaled by 10^4). The graphs in (**C**) and (**D**) show the variation of the temperature dependence of standard deviation of successive measurements using the SQUID measurement system. (All data from R. A. Dunlap and M. E. McHenry, unpub. observ.)

First, with appropriate signal averaging and averaging of several measurements, it is possible to obtain standard deviations of the dipole moment two orders of magnitude smaller than the mean. The standard deviations are in fact larger at intermediate values of M and display a λ-type anomaly reminiscent of the temperature dependence of the specific heat, which will be discussed below. This could be attributed to dynamical (heating) effects as the sample passes through the ferromagnetic to paramagnetic transition. Second, even in a small field (i.e., 10 Oe), a significant broadening of the transition (i.e., a Curie tail) is observed for this large sample. This broadening prevents an accurate determination of T_c by extrapolation to the

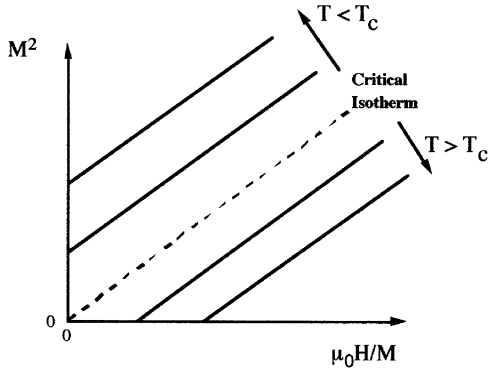


Figure 5. Schematic representation of the method of determining ordering temperatures using the Arrot plot construction.

$M = 0$ axis. On the other hand, fitting the data to a Curie-Weiss law does in fact allow for accurate determination of T_c. Gadolinium is an example of an anisotropic ferromagnet, in which its *c* axis is the easy direction of the magnetization. As a result, $M(T)$, below T_c, is markedly different for fields applied parallel to or perpendicular (i.e., in the basal plane) to the *c* axis. This anisotropy has been illustrated in the fact that a T_c for H parallel to the *c* axis differs by ~1 K from that determined with H in the basal plane (Geldart et al., 1989).

Data on the inverse susceptibility versus T for the same crystal used for Figure 4 has been used to yield a Curie temperature (paramagnetic) of 293.6 and 292 K, respectively for H parallel to *c* and to the *ab* basal plane (Geldart et al., 1989). The inverse susceptibility data was fit using a power law of the form $\chi_c^{-1}(T) = At^{\gamma}$ and $\chi_b^{-1}(T) = B + Ct^{\gamma}$ where A, B, C, γ, y, T_c^c and T_c^{basal} were taken as free parameters. The critical exponents γ and y were determined to be 1.23 and 1.01, respectively. These fits revealed a difference in T_c of ~1.5 K that was attributed to a difference in in-plane and out-of-plane dipolar interactions, which will not be discussed in detail here.

Another analytic method for determining ordering temperatures from thermomagnetic data is by assuming a Landau theory equation of state and fitting data with so called Arrott plots. In Arrott plots, shown schematically in Figure 5 isotherms of M^2 are plotted as a function of H/M in accordance with the Landau theory equation of state. These isotherms consist of a set of parallel lines, with the isotherm corresponding to the Curie temperature passing through the origin. This is a particularly sensitive method for determining the Curie temperature.

LITERATURE CITED

Gallagher, K. A., Willard, M. A., Laughlin, D. E., and M. E. McHenry, 1999. Distributed exchange interactions and temperature-dependent magnetization in amorphous $Fe_{88-x}Co_xZr_7B_4Cu_1$ alloys. *J. Appl. Phys.* 85.

Geldart, D. J. W., Hargraves, P., Fujiki, N., and Dunlap, R. A. 1989. Anisotropy of the critical magnetic susceptibility of gadolinium. *Phys. Rev. Lett.* 62:2728.

Lakeshore Cryotronics. 1995. Model 7200 Vibrating Sample Magnetometer User's Manual. Lakeshore Cryotronics, Westerville, Ohio.

Quantum Design. 1990. Model MPMS/MPMS Application Notes/Technical Advisories. Quantum Design, San Diego, Calif.

Turgut, Z., Huang, M.-Q., Gallagher, K., Majetich, S. A., and McHenry, M. E. 1997. Magnetic evidence for structural-phase transformations in Fe-Co alloy nanocrystals produced by a carbon arc. *J. Appl. Phys.* 81:4039.

Willard, M. A., Huang, M.-Q., Laughlin, D. E., McHenry, Cross, J. O., Harris, V. G., and Franchetti, C. 1999. Magnetic properties of HITPERM $(Fe,Co)_{88}Zr_7BaCu_1$ magnets. *J. Appl. Phys.* 85:4421–4423.

Willard, M. A., Laughlin, D. E., McHenry, M. E., Thoma, D., Sickafus, K., Cross, J. O., and Harris, VG. 1998. Structure and magnetic properties of $(Fe_{0.5}Co_{0.5})_{88}Zr_7B_4Cu_1$ nanocrystalline alloys. *J. Appl. Phys.* 84:6773

KEY REFERENCES

Ausleos, M. and Elliot, R. I. (eds.) 1983. Magnetic Phase Transitions, Springer Series in Solid State Sciences 48. Springer-Verlag, New York.

Collection of articles dealing with the statistical mechanics and theory of magnetic phase transitions.

Birgeneau, R. J., Tarvin, J. A., Shirane, G., Gyorgy, E. M., Sherwood, R. C., Chen, H. S., and Chien, C. L. 1978. Spin-wave excitations and low-temperature magnetization in the amorphous metallic ferromagnetic $(Fe_xNi_{1-x})_{75}P_{16}B_6Al_3$. *Phys. Rev. B* 18: 2192–2195.

Article discussing the temperature dependence of the magnetization in amorphous alloys.

Boll, R. 1994. Soft Magnetic Metals and Alloys. *In* Materials Science and Technology, A Comprehensive Treatment, Vol. III (K. H. J. Buschow, ed.) pp. 399–450.

Recent review article on soft magnetic materials covering principles, structure, processing, materials and alloy spectrum, and properties. Much tabulated data included.

Chappert, J. 1982. Magnetism of Amorphous Metallic Alloys. *In* Magnetism of Metals and Alloys (M. Cyrot, ed.) North-Holland Publishing, New York.

Review chapter covering the theory of amorphous magnetism with some summary of experimental observations.

Chen, C. W. 1986. Magnetism and Metallurgy of Soft Magnetic Materials. Dover, New York.

Classic text discussing magnetic materials with a strong emphasis on process, structure, and properties relationships. Much data on alloy spectrum and properties.

Mattis, D. C. 1981. The Theory of Magnetism, Springer Series in Solid-State Sciences 17. Springer-Verlag, New York.

Advanced text on the theory of magnetism with strong basis in the quantum mechanical framework. Much interesting historical information.

Pfeifer, F., Radeloff, C. 1980. Soft magnetic Ni-Fe and Co-Fe alloys: Some physical and metallurgical aspects. *J. Magn. Magn. Mater.* 19:190–207.

Properties of FeCo and NiFe magnetic materials.

Rajkovic, M. and Buckley, R. A. 1981. Ordering transformations in Fe-50Co based alloys. *Metal Science*. Vol. 21.

Properties of FeCo magnetic materials.

MACHAEL E. MCHENRY
Carnegie Mellon University
Pittsburgh, Pennsylvania

TECHNIQUES TO MEASURE MAGNETIC DOMAIN STRUCTURES

INTRODUCTION

Critical to understanding the magnetic properties and the technological application of magnetic materials is the ability to observe and measure magnetic domain structure. The measurement challenge is quite varied. Magnetic fluids can be used to decorate and make simple domain structures visible while highly sophisticated electron microscopy-based methods may be needed to explore and image nanometer-scale magnetic phenomena. No one method will solve all of the important domain imaging problems. For that reason, we will discuss the bases of all the important methods and the practical implications of each.

For purposes of comparison, several important characteristics of each domain imaging technique we discuss are briefly summarized in Table 1. However, the table should only be used as a broad outline of the different methods and any choice of method should include consultation with the relevant sections below and the references therein.

There are several general comments that apply to every domain imaging method discussed in this unit. First, each method is related either to a sample's magnetization or to the magnetic field generated by that magnetization. The former observation is of great usefulness in understanding or developing new materials, and the latter is of practical importance in devices. Second, all of the methods are limited to the topmost few micrometers of the sample, and are therefore dominated by magnetic structure related to the surface. None provide the domain structure of the interior of a thick sample where bulk magnetic microstructure is fully developed. Third, all of the methods require smooth, clean, damage-free surfaces. In general, the surfaces need to have at least the equivalent of a mirror polish, otherwise, topography and surface stress will affect the measurement to varying degrees. Fourth, correlations exist between several of the parameters in each technique we describe. For example, it may not be possible to achieve the ultimate sensitivity specified at the highest resolution possible. Finally, almost all of the techniques use digital signal acquisition and can benefit from the use of modern image processing tools.

What follows are expanded discussions of the basis of each technique, its practical aspects, sample preparation required, possible sample modifications, and problems that are specific to the technique described. Each section contains references to more complete treatments of the technique and recent examples of its use.

BITTER PATTERN IMAGING

Principles of the Method

Bitter or powder pattern imaging of domain structures is arguably the oldest domain imaging technique, and still probably the simplest method to apply. In Bitter pattern imaging, a thin colloidal suspension of magnetic particles is painted on a magnetic surface. The particles collect and agglomerate in regions where large stray fields from the sample are present, typically over domain walls. The decorated domain walls can then be imaged using an optical microscope, or, if higher resolution is required, an electron microscope. Figure 1 shows an example of a Bitter pattern image of bits written on a magnetic hard disc.

The Bitter technique is strictly a stray-field decoration technique. The patterns provide no information about the magnitude or direction of the magnetization, but in materials with sufficiently large external fields, Bitter patterns can quickly provide information about the size and shape of any domains that may be present. Bitter pattern imaging has been reviewed by Kittel and Galt (1956) and Craik (1974).

Practical Aspects of the Method

The resolution of the Bitter method depends primarily on both the size of the individual or agglomerated particles in the colloid and the resolution of the microscope used to image the patterns. Historically a researcher had to be skilled in the preparation of Bitter solutions, but today there are several commercial suppliers of Bitter solution colloids, sometimes referred to as ferrofluids (Ferrofluids from Ferrofluidics and Lignosite FML from Georgic Pacific). By selecting the appropriate Bitter solution, these commercial ferrofluids can reveal magnetic structures down to the resolution limit of the optical microscope. Higher resolutions can be attained by imaging the decorated sample with a scanning electron microscope (Goto and Sakurai, 1977) or by using fine-grained sputter-deposited films for decoration (Kitakami et al., 1996).

Contrast in the Bitter method depends on sufficiently large stray magnetic field gradients to collect the magnetic particles. Although certain ferrofluids can be sensitive to stray fields as small as a few hundred A/m, the stray fields outside of some high-permeability or low-anisotropy materials may still be too small to image. For this reason, the technique generally works better with higher coercivity magnets or perpendicularly magnetized samples. In fact, a small perpendicular applied magnetic field is frequently used to improve the contrast. As with any magnetic field–sensitive domain imaging technique, deriving the magnetic structure from the observed Bitter pattern image can be difficult, since the external magnetic fields may be the result of nonlocal variations in the sample's magnetization.

Sample Preparation

The Bitter method requires some sample preparation, because the technique does not separate the magnetic from the topographic contrast. Bulk samples are usually prepared by mechanical of polishing the surface, followed by chemical or electropolishing to remove residual surface strains. Thin films deposited on polished substrates can also be used. Aside from the smoothness, there are few constraints on the types of magnets that can be studied. The samples can be conductors or insulators, and thin nonmagnetic coatings are allowed. On the downside, the ferrofluid can contaminate the surface, leaving behind residues that may not be easily removed.

Table 1. Magnetic Domain Imaging Techniques

	Bitter	Magnetooptic	MFM[a]	Lorentz	DPC[b]	SEMPA[c]	Holography	XMCD[d]	TXMCD[e]	SPLEEM[f]
Principles of method										
Contrast Origin[g]	Grad B_{ext}	M	Grad B_{ext}	B	B	M	B,Φ_B	M	M	M
Quantitative	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Practical aspects										
Best resolution (nm)	100	200	40	~10	~2	20	~5	300	30	20
Typical resolution (nm)	500	1000	100	50	20	200	20	500	60	40
Information depth (nm)	500	20	20–500	Sample thickness	Sample thickness	2	Sample thickness	2–20	Sample thickness	1 s
Acquisition time	0.03 s	10^{-8} 1 s	5 to 30 min	0.04 to 30 s	5 to 50 s	1 to 100 min	0.03 to 10 s	0.03 s to 10 s	3 s	1 s
Insulators	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No
Vacuum requirement[h]	None	None	None	HV	HV	UHV	HV	UHV	None	UHV
Complexity	Low	Moderate	Moderate	Moderate	Mod./high	High	High	High	High	High
Commercially available	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No
Cost[i]	1	50–500	150	200–1000	1000	800	1300	300+	300+	1000
Sample preparation										
Sample thickness (nm)	No limit	No limit	No limit	<150	<150	No limit	<150	No limit	<100	No limit
Special smoothness	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Clean surface required	No	No	No	No	No	Yes	No	No	No	Yes
Specimen modification										
Maximum applied external field (kA/m)	No limit	No limit	800	500 (vert.) 100 (horiz.)	500 (vert.) 100 (horiz.)	None	100	None	No limit	None
Problems										
Topographic feedthrough	Yes	Yes	Yes	Some	some	No	Some	No	No	No
Crystallographic feedthrough	No	No	No	Yes	Yes	No	Yes	No	Not tested	Yes

[a] Magnetic force microscopy (MFM).
[b] Differential phase contrast (DPC).
[c] Scanning electron microscopy with polarization analysis (SEMPA).
[d] X-rays magnetic circular dichroism (XMCD).
[e] Transmission XMCD.
[f] Spin polarized low-energy electron microscopy (SPEEM).
[g] Abbreviations: Grad B_{ext}, gradient of the external magnetic flux density; M, sample magnetization; B, magnetic flux density; Φ_B, magnetic flux.
[h] Thousands of U.S. dollars.

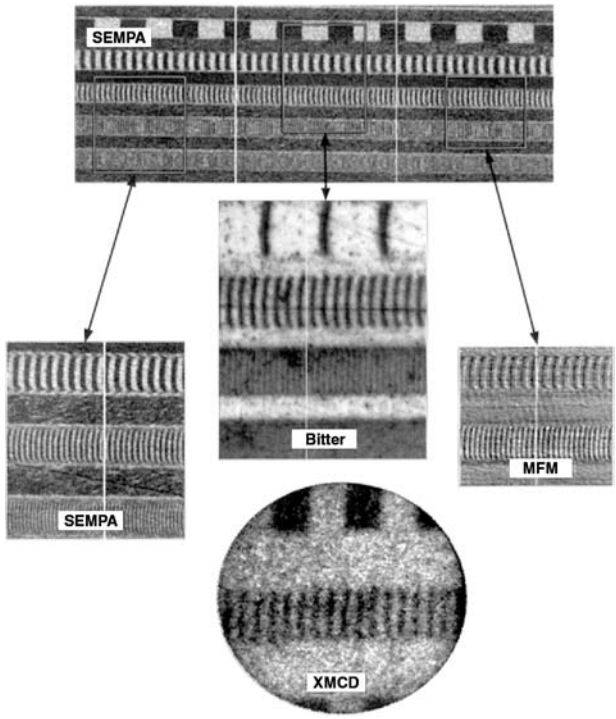


Figure 1. An illustration of several magnetic imaging techniques using a pattern written on magnetic storage media. The test pattern is composed of horizontal tracks, each ~10 μm wide and containing a series of magnetization reversals, or "bits." The bit length ranges from 10.0 to 0.2 μm. The bit length and spacing of the large bits in the XMCD electron yield image is 10 μm (Tonner et al., 1994). Note that the Bitter and XMCD images are from a different, but similar, test sample.

Specimen Modification

As long as the magnetic particles remain in solution, the Bitter pattern can be used to examine the motion of domains and domain walls while applying a magnetic field. The response time of the ferrofluid may be rather slow, however, due to the viscosity of the colloidal solution. In some cases it may take several minutes for the Bitter pattern to come to complete equilibrium. The ability to observe domain wall motion while applying a magnetic field is essential when trying to separate magnetic structure from sample topography.

MAGNETO-OPTIC IMAGING

Principles of the Method

The weak interaction between polarized light and a material's magnetization leads to several useful optical domain imaging methods (Craik, 1974). Although the physics of the magneto-optic interactions can be rather complicated (Hubert and Schäfer, 1998), at the most basic level, magneto-optic imaging is simply based on the rotation of the plane of polarization of linearly polarized light upon reflection from, or transmission through, a magnetic material. The magneto-optic effect in transmission is usually referred to as the Faraday effect and in reflection as the Kerr effect. In both cases, the domain contrast in the image is directly related to the magnitude and direction of the magnetization in the sample.

Practical Aspects of the Method

The relative geometry between the magnetization direction and the direction of the transmitted or reflected light determines which component of the magnetization vector will be visible in a particular magneto-optic image. When light is incident normal to the sample surface in either the "polar" Faraday or Kerr mode, domains that are magnetized perpendicular to the surface are imaged. In-plane magnetization can be detected in the Faraday or Kerr modes using oblique illumination. In the longitudinal Kerr effect, the in-plane magnetization component lies in the scattering plane of the light, while in the transverse Kerr effect the magnetization is perpendicular to the scattering plane. Contrast in the polar mode is greatest at an angle of incidence of 0°, while the longitudinal and transverse Kerr effects are greatest at ~60° angle of incidence.

Typically, magneto-optic images can be generated either by using conventional imaging optics or by rastering a finely focused laser spot across the sample surface. In either case, high- quality strain-free optics and polarizers with high extinction ratios are preferred, because the magneto-optic effects are quite small. The Faraday and polar Kerr modes provide the most contrast, while the contrast in the longitudinal Kerr mode is so small that additional electronic signal processing is usually required to separate the magnetic contrast from the nonmagnetic background (Schmidt et al., 1985; Argyle, 1990; Trouilloud et al., 1994).

The resolution of the magneto-optic image is determined by the resolution of the optical imaging system, or, in the case of rastered laser imaging, by the size of the focused laser spot. Typical resolution is therefore ~1 μm, but can be improved to 0.3 μm in a high-quality optical microscope using an oil-immersion objective and blue light illumination. Further improvements in resolution can be achieved by using various forms of scanned near-field optical imaging, but the contrast mechanisms in these methods are not yet well understood (Betzig et al., 1992; Silva and Schultz, 1996). The sampling depth in the Kerr imaging mode is determined by the penetration depth of the light and is ~20 nm in a metal. Therefore, the technique is moderately surface sensitive and can be used to image magnetic domains in thin films that are only a few monolayers thick, as well as to image domains that are covered by sufficiently thin films of materials that are normally opaque.

Sample Preparation

The small sampling depth and topographic sensitivity of the Kerr imaging mode requires preparing samples that

have optically flat, damage-free surfaces. Bulk samples can be prepared by mechanical polishing, followed by chemical polishing or annealing to remove the remaining damage. High-quality surfaces for imaging can also be generated by evaporation or electrodeposition of thin films on flat, polished substrates. Samples may be coated by thin, nonmagnetic films without significantly affecting the magneto-optic images. In fact, appropriate antireflective coatings can be applied to samples in order to increase the magneto-optical contrast.

Specimen Modification

Perhaps the greatest advantage of magneto-optic imaging is the speed with which magnetic images can be acquired. Video-rate imaging of domain dynamics is routinely achieved using standard arc lamp illumination. Stroboscopic imaging using pulsed laser illumination can reveal domain wall motion that occurs over time scales as short as a few nanoseconds (Petek et al., 1990; Du et al., 1995). In fact magneto-optic imaging can yield a great deal of information about magnetization dynamics in a magnetic material or device, since arbitrarily large magnetic fields may be applied to the sample while imaging. Figure 2 shows a magneto-optic image of a thin-film recording head. This image highlights the domain wall motion within the head by taking the difference between images with the current in the head coils reversed.

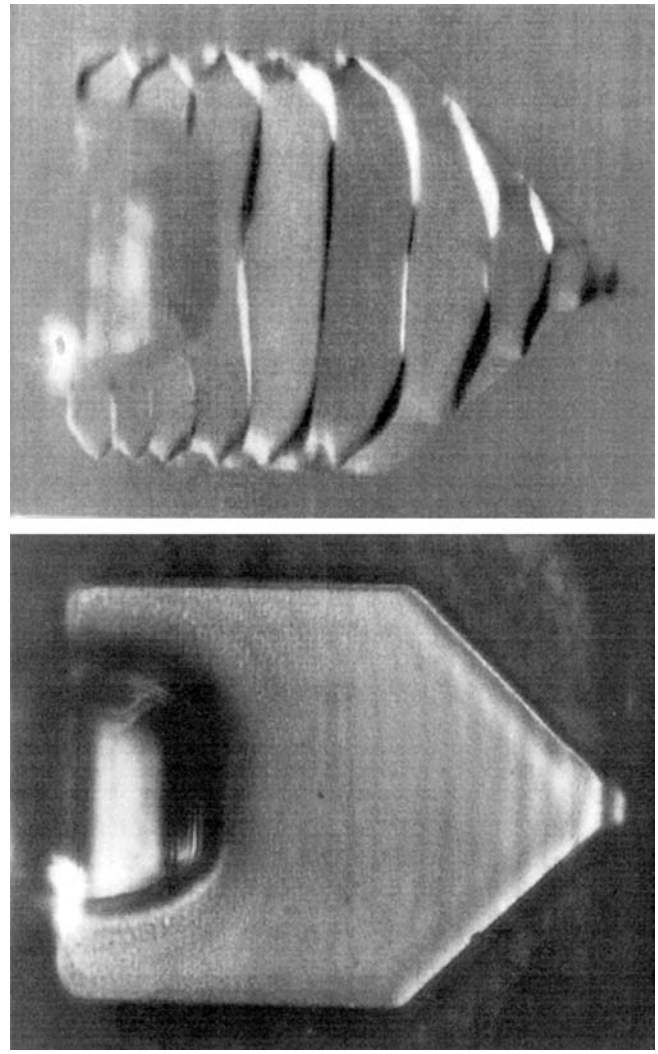


Figure 2. Magneto-optic image of a thin-film recording head. The lower panel shows the raw optical image showing the sample topography. The top panel shows the difference between two magneto-optical images taken with opposite currents driving the head magnetization. The white and black regions reveal how the domain walls have moved. The head is ~100 μm across. Photo courtesy of B. Argyle.

Problems

The major difficulty with the Kerr imaging mode is having to separate the magnetic image from a potentially larger nonmagnetic background and from the intensity variations due to the sample's topography. Various approaches have been used to improve the contrast and eliminate topographic feedthrough. Antireflective coatings may be applied to the sample (Hubert and Schäfer, 1998). The fact that the sense of the Kerr rotation is independent of the direction of incidence may also be exploited by using wide-angle illumination and a segmented optical detector (Silva and Kos, 1997). The most common method is to measure, using sufficiently precise instrumentation, the difference between the images taken before and after a magnetic field is applied to reverse the magnetization. In practical terms, this means that magneto-optic imaging is best applied to samples where the magnetization may be changed by applying a field, such as in the thin-film magnetic recording head shown in Figure 2. On the other hand, high-quality images of static magnetic domain patterns that one cannot or does not wish to alter, such as written bits in recording media, are difficult to acquire.

Finally, there are several new developments in magneto-optical imaging that are worth noting because of their potential future impact. First of all, near-field optical techniques are being used to overcome the resolution limits of conventional diffraction-limited optics. By using scanned apertures or tips in close proximity to a surface, resolutions on the order of $\lambda/10$ have been achieved (Betzig et al., 1992; Silva and Schultz, 1996). Second, magneto-optic indicator films have been developed as an alternative to Bitter pattern imaging (Nikitenko et al., 1996). In this method a thin, free-standing garnet film is placed against a magnetic sample and the resulting domain pattern, induced by the sample's stray field, is imaged using a conventional polarized-light microscope. Compared with Bitter imaging, this relatively simple and inexpensive technique has the advantage of faster response to applied magnetic fields and no sample contamination. Finally, intense laser illumination has made possible imaging using second-harmonic Kerr effects (Kirilyuk et al., 1997). The second-harmonic effects result from the nonlinear optical response of certain materials, enabling this mode

to image structures, such as domains in antiferromagnets, that are not visible with conventional Kerr imaging.

MAGNETIC FORCE MICROSCOPY

Principles of the Method

Magnetic force microscopy has become one of the most widespread tools for studying the magnetic structure of ferromagnetic samples and superconductors (Rugar and Hansma, 1990; Sarid, 1991). The technique is based on the forces between a very small ferromagnetic tip attached to a flexible cantilever and the inhomogeneous stray magnetic field immediately outside a sample of interest. As the magnetic tip is scanned over a magnetic sample, these minute forces are sensed in any of a variety of ways to provide maps related to the magnetic field above the sample.

In the implementation of an MFM, there are a very large number of choices to be made that influence, in a fundamental way, how the observed image should be related to the magnetic field. For example, an MFM in which the tip is magnetically relatively soft (i.e., the tip magnetization is modified by its interaction with the sample) provides a very different image of the magnetic field than one which utilizes a tip that is magnetically relatively hard (i.e., the tip magnetization does not change in response to the magnetic field of the sample). Similarly, the distribution of magnetic moments in the tip can have a pronounced qualitative effect on the imaging. Whereas a point-like magnetic particle on the tip may best be modeled as a simple magnetic dipole, a tip that is longer or more like a needle may be best modeled as a magnetic monopole. Furthermore, the MFM may sense either the net deflection of the tip in the field of the sample or may sense changes in either the amplitude, frequency, or phase of the vibrational motion of a tip oscillating resonantly in the field above the sample. One direct result of the complex and multifaceted interaction between the tip and sample is that it is very difficult to determine the field distribution from an MFM image.

Because of the huge number of possible implementations implied by these (and many other) available choices, and further because the state-of-the-art for magnetic force microscopy is currently being developed very rapidly in numerous laboratories, it is not possible in the current context to provide a thorough or complete report on the status of magnetic force microscopy. Instead, we will present a general description of an MFM in one form readily available commercially.

Rather than directly sensing the deflection of the flexible cantilever due to magnetic forces acting on the magnetic cantilever tip, MFMs typically detect changes in the resonant vibrational frequency of the cantilever. In this mode of operation, the cantilever is electrically driven to oscillate, with the driving frequency controlled to track very precisely the resonant vibrational frequency of the cantilever. When they are in close proximity, the scanning tip and sample surface generally experience an attractive net force (e.g., from van der Waals interactions). In addition to this attractive force, there will be a force of magnetic origin, which will be either attractive or repulsive depending on the relative orientation of the tip magnetization and the magnetic field gradients above the sample. Where the magnetic force is also attractive, the cantilever is deflected further towards the surface. This effectively stiffens the cantilever, raising its natural resonant frequency. If the magnetic force is repulsive, the cantilever is deflected less strongly towards the sample, effectively softening the cantilever and lowering its resonant frequency. Because the resonant frequency of the cantilever can be determined with very high precision, that frequency provides a convenient means by which to monitor local variations in the magnetic field gradients.

However, because the magnetic forces between tip and sample are very small, one general problem that all MFMs must address is the separation of the image contrast that arises from magnetic forces from the image contrast that is due to other (stronger) short-range physical forces (Schoenenberger et al., 1990).

Figure 1 shows, as an inset, a magnified region of a test pattern written on magnetic storage media as imaged with a commercial magnetic force microscope. The written bits are clearly visible and highlight the fact that the MFM is primarily sensitive to gradients in the magnetic field. Regions where the magnetization changes from left to right (or right to left) are visible as white (or dark) lines. This contrast would reverse with a magnetization reversal of the MFM tip.

Practical Aspects of the Method

There are several practical aspects to MFM imaging that should be noted. First, it is generally not known with confidence what underlying contrast mechanism gives rise to an MFM image. The signal is generally proportional to spatial derivatives of the stray field above the sample, but which spatial derivative in which direction depends on the such factors as the details of the magnetic moment distribution within the tip, tip/sample interactions, operating mode of the MFM signal detection and control electronics, vibrational amplitude of the cantilever, and lift height of the MFM scan relative to the AFM topographic scan. As a consequence, quantitative interpretation of MFM images is generally very difficult (Hug et al., 1998). Even qualitative interpretation can sometimes be very uncertain. Because the MFM can often be configured to give contrast that is proportional to the magnetic field as well as to various spatial derivatives, one is often uncertain whether observed magnetic contrast is due to a magnetic domain or to the domain wall between two domains. As with many scanned-tip microscopes, navigation on a sample can be difficult. The MFM always operates at very high magnification, so it can be difficult to find specific isolated features for study.

As mentioned above, the effects of surface topography on magnetic force microscopy can be pronounced. One common solution to this problem is to acquire the magnetic image in two steps. In the first step, the microscope is operated as a conventional atomic force microscope (AFM; not discussed here) so as to determine in detail the topographic

profile along one scan line. The microscope then retracts the tip and, using the previously determined line profile, rescans the same line with the magnetic tip at a small but constant height (typically 20 to 200 nm) above the sample. In this second scan, the longer-range magnetic forces still affect the cantilevered tip, but the effect of the short-range forces is minimized. This effectively provides a map only of variations in the local magnetic field, free of topographic contrast.

Even with such methods to compensate for the effects of surface topography, samples must be very smooth for these methods to be effective. Surfaces that have significant roughness or surface relief can be very difficult to image with the MFM. Sometimes the AFM prescan is not adequate because the surface relief is too great for the AFM tip positioning to be reliable. Even if the AFM topography scan is successful, other problems arise if the combination of tip vibrational amplitude and surface texturing is so large that the tip contacts the sample during the MFM scan.

One very useful feature of the MFM is that, because the MFM senses the stray magnetic field rather than the sample magnetization directly, it is easy to see the magnetic structure even through relatively thick nonmagnetic and even insulating overlayers.

Data Analysis and Initial Interpretation

Significant image processing is generally required to aid in the interpretation of MFM images. The relevant image-processing steps are typically integrated into the commercial MFM instrument controllers. One example is the abovementioned subtraction of contrast due to surface topography. This subtraction, however, has the side effect that domain walls that run parallel to the scan direction can be much more difficult to image than walls running at a significant angle to the scan direction.

Sample Preparation

One of the very attractive features of the MFM is that minimal surface preparation is required. Samples that are smooth and flat enough to image with an atomic force microscope can generally be studied with MFM as well.

Problems

As mentioned above, consideration must always be given to the extent to which the magnetic structure of the tip or sample is modified in response to the magnetic field of the other.

TYPE I AND TYPE II SCANNING ELECTRON MICROSCOPY

Principles of the Method

It is often possible to image the magnetic domain structure of a sample with a conventional scanning electron microscope (SEM) (SCANNING ELECTRON MICROSCOPY). Two distinct contrast mechanisms, referred to as Type I and Type II magnetic contrast, have been described (Reimer, 1985; Newbury et al., 1986). Both are based on the deflection (due to the Lorentz force) of electrons moving in a magnetic field. Two important features distinguish between Type I and Type II contrast. First, whereas Type I contrast involves imaging low-energy secondary electrons ejected from the sample, Type II contrast arises from high energy, backscattered electrons. Second, whereas Type I contrast arises from the deflection of the secondary electrons by stray fields outside of the sample, Type II contrast relies on deflection of the incident and backscattered electrons by magnetic fields within the samples.

Practical Aspects of the Method

There are relative advantages and disadvantages to both types of magnetic contrast. The primary advantages are that no special modifications are required either for the scanning electron microscope or the sample. Essentially any conventional scanning electron microscope can be used for Type I contrast. The only special requirement, which may involve modification to the secondary electron detector, is the positioning of the detector to one side of the sample so as to preferentially detect secondary electrons that are ejected toward rather than away from the detector. Some detectors have a bias voltage for secondary-electron collection that is so high that its directional sensitivity is inadequate to observe Type I contrast. If the microscope includes (as is very common) a detector for backscattered electrons, then Type II contrast is also feasible. The primary disadvantages are that spatial resolution is limited, and that it is sometimes difficult to distinguish between magnetic contrast of either type and other types of signal contrast.

For Type I contrast to be optimized, the primary beam energy typically should be <10 keV in order to produce the most secondary electrons. Because contrast depends on the geometry of the microscope's electron detector relative to the sample magnetic structure, tilt and rotation control of the sample are necessary to optimize contrast. Spatial resolution is limited typically to ~ 1 μm due to the spatial extent of the external magnetic fields. In some circumstances, the sample tilt and rotation also allow Type I contrast to be used for quantitative imaging of the magnetic fields outside of a sample (Wells, 1983).

Type II contrast is maximized by using the highest available beam energy and by tilting the sample surface to $\sim 50^\circ$ relative to the incident beam. The contrast varies from $\sim 0.1\%$ at 20 keV to $\sim 1\%$ at 200 keV. Because the contrast is relatively low, high primary beam currents are required. The spatial resolution is then determined both by the size of the primary beam and by the escape volume of the back-scattered electrons. For the direction parallel to the tilt axis, this resolution varies from ~ 1 μm at lower energies to ~ 2 μm at 200 keV. For the direction perpendicular to the tilt axis, the resolution is significantly degraded and is roughly determined by the penetration depth for the incident electron. For 200-kV electrons, that resolution is ~ 10 μm.

Type II magnetic contrast, of all the imaging techniques described, can have the greatest information depth. At

200 keV, the electrons penetrate ~15 μm of material, with the maximum back-scattering intensity coming from a depth of ~9 μm. Consequently, Type II is less surface sensitive than the other imaging techniques. Further, exploitation of the dependence of the penetration depth on primary beam energy provides a coarse method for magnetization depth profiling.

Data Analysis and Initial Interpretation

No significant data analysis is required, but interpretation is generally restricted to qualitative studies of domain sizes and shapes.

Sample Preparation

No special preparation is required, other than polishing for the removal of topographic features that would otherwise compete with the relatively weak magnetic contrast.

LORENTZ TRANSMISSION ELECTRON MICROSCOPY

Principles of the Method

This form of very high resolution magnetic microscopy gets its name from the Lorentz force, which causes a deflection of the electron trajectory for a beam traveling perpendicular to a magnetic field. The Type I and Type II methods, described previously, also are based on the Lorentz force. However, while Type I and Type II imaging are performed on the surface of a sample in an SEM, Lorentz microscopy makes use of either a conventional transmission electron microscope (CTEM) or a scanning transmission electron microscope (STEM), and involves transmission through thinned samples (TRANSMISSION ELECTRON MICROSCOPY and SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). Within the general category of Lorentz microscopy, three distinct imaging techniques (Chapman, 1984; McFayden et al., 1992) can be identified, i.e., Fresnel, Foucault, and DPC microscopy.

By using a CTEM in the Fresnel mode (Heyderman et al., 1995), narrow regions of relatively high and low intensity are formed at positions that correspond to domain walls; this results from small deflections caused by the Lorentz force acting on a defocused beam of electrons transmitted through the sample. Consider a thinned magnetic sample with a domain geometry of parallel strips, with the magnetizations in each strip lying in-plane and in alternating directions parallel to the domain walls separating them. The Lorentz force will, depending on the magnetization direction of each domain, deflect the beam slightly toward one wall or the other. As a consequence, the domain will be seen to have a bright wall on one side and a dark wall on the other. Figure 3A shows an example of Fresnel imaging of a metallic glass.

In the Foucault mode (Heyderman et al., 1995) using a CTEM, the image remains in focus and a diffraction pattern is formed at one of the aperture planes of the microscope. By displacing the aperture, it is possible to block

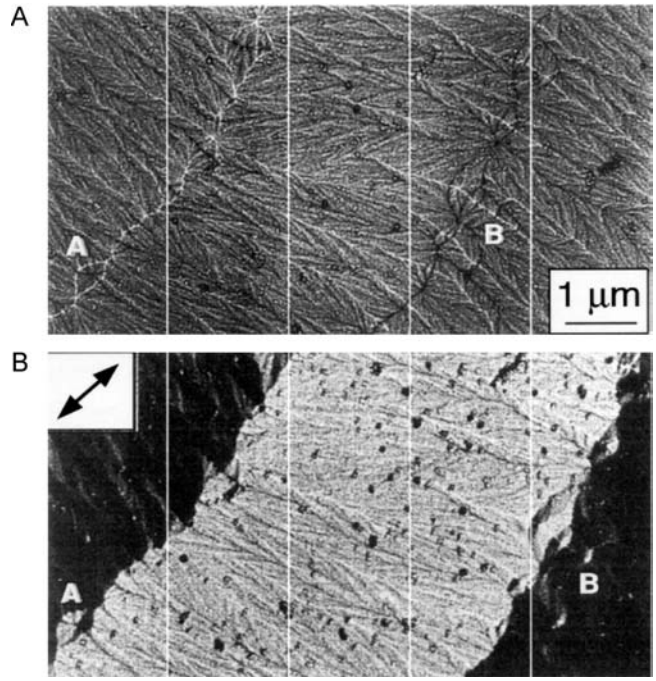


Figure 3. (**A**) A Fresnel image of a metallic glass magnetic material showing a bright and a dark domain wall. (**B**) A Foucault image of the same area, where the contrast depends on the magnetization parallel to the domain walls, as indicated by the arrows. In addition to the main 180° domains, these techniques clearly reveal small variations in the magnetization, i.e., magnetization ripple (Heyderman et al., 1995).

electrons that have been deflected in one direction by the magnetic field. In the image subsequently formed, the brightness within each domain will depend on the direction of its magnetization. Figure 3B shows an example of Foucault imaging.

In the DPC mode (McVitie et al., 1997), a STEM is modified by the addition of a quadrant electron detector and matching electronics. A focused scanning probe beam is deflected on passing through the thin sample, and the extent of the Lorentz deflection is measured from the differences and ratios of currents incident on the quadrants of the electron detector. A quantitative measure of the magnetic field lying perpendicular to the electron's path through the sample is possible.

Practical Aspects of the Method

These methods, particularly DPC, are capable of excellent resolution, approaching 2 nm for magnetic structures in the best of circumstances and 10 to 20 nm for more typical applications. Physical structure can be seen with even higher resolution. Image acquisition times are typically tens of seconds for computer-controlled DPC image acquisition.

The methods are quite sensitive, being able to detect just a few layers of Fe in cases where the crystallographic contrast does not interfere. Since deflections are caused by

the magnetic field within the sample, the information depth is the full sample thickness with equal weighting for all depths. Note, however, that magnetic fields outside the sample surface can also cause deflections that could modify the image.

The complexity of these methods is moderately high. They all use a TEM and may involve preparing a sample by thinning, which, it should be remembered, can have an effect on the domain structure. TEMs suitable for use in these modes are readily available commercially. The highest-resolution systems, with field emission sources and a variety of attachments, can cost one million dollars, although the cost to add a magnetic imaging capability to an existing TEM would be much less. In the DPC method, the electronics necessary to derive, display, and store the difference signals are not commercially available. Also, some technique must be implemented to "descan" the probe beam on the detector plane, i.e., remove the contribution to the differential detector signal that comes from a scanning probe beam rather than magnetic deflection.

Method of Automation

In the DPC method, the four quadrant detectors (or eight if a scheme to separate magnetic and nonmagnetic contrast is used) are connected to preamplifiers and an amplifier/mixer to derive multiple signals, including the sum of all channels, and the sums and differences between various quadrants. These signals are displayed in real time and filtered, digitized, and stored in a computer for later analysis.

Data Analysis and Initial Interpretation

Images produced in the Fresnel and Foucault modes immediately display the domain walls or domains, respectively. The dimensions of the domains can be determined quantitatively from these images. The DPC method uses the digitized signals stored for each x,y point in the image; the magnitude, direction, and curl of the magnetic induction can be calculated and displayed. Additional displays, including color wheel representations of direction and histograms showing the distribution of fluctuations in direction, are also readily available.

Sample Preparation

Only very thin samples can be analyzed. This method is generally applicable for samples $<$150 nm thick. As with other electron-based methods, only conductive samples can be imaged.

Common to all the Lorentz methods described above is the requirement that the microscope not generate sufficient magnetic field at the sample to modify the domain pattern. This problem can be addressed by either turning off the objective lens, moving the sample away from the strongest field position, or using a low-field lens specifically designed for the purpose. Unfortunately, all of these solutions will degrade the spatial resolution somewhat.

Specimen Modification

Magnetic fields can be purposefully applied to the sample, so long as they do not significantly affect the operation of the microscope. In a typical case, magnetic fields can be applied to the sample if held below 500 kA/m in the direction along the microscope column and 80 kA/m perpendicular to it. Modification of the sample due to beam heating effects is usually considered negligible.

SCANNING ELECTRON MICROSCOPY WITH POLARIZATION ANALYSIS

Principles of the Method

Scanning electron microscopy with polarization analysis is a magnetic-imaging technique based on the measurement of the spin polarization of secondary electrons ejected from a ferromagnetic sample by an incident beam of high-energy electrons (Scheinfein, et al., 1990). These secondary electrons have a spin polarization that is determined by their original spin polarization, i.e., the magnetization in the bulk sample. An image of the magnetic domain microstructure at the surface can thus be generated by a measurement of the secondary electron spin polarization at each point as a tightly focused electron beam is scanned in a raster fashion over a region of interest.

For a $3d$ transition metal such as iron, both the bulk magnetic properties and the emitted low-energy secondary electrons are dominated by the valence electrons, with remaining electrons behaving essentially like an inert core. One can, then, roughly predict the degree of secondary-electron spin polarization in iron from the observed magnetic moment of atoms in the bulk (2.22 Bohr magnetons) and the number (8) of valence electrons. One expects a spin polarization of $\sim 2.22/8 = \sim 0.28$, in good agreement with experimental observations. Similarly, good predictions can be made for cobalt (0.19) and nickel (0.05) as well.

The polarization detectors typically used for SEMPA (Scheinfein et al., 1990) provide simultaneous determination of the electron polarization projected onto two orthogonal axes. Frequently, those axes are parallel to the surface of the sample, so that SEMPA measures the in-plane surface magnetization. SEMPA detectors can also be constructed to measure the magnetization perpendicular to the surface, along with one in-plane component. However, magnetic anisotropy at the surface generally forces the magnetic moments for most samples to lie in the plane of the surface, so that out-of-plane magnetization is rarely observed.

There are several general features of SEMPA that deserve particular attention in comparison with other magnetic-imaging techniques. First, whereas many magnetic-imaging techniques are sensitive to the magnetic fields either inside or outside of the sample, SEMPA determines the sample magnetization directly. Second, because the physical basis for magnetic contrast is well understood, the magnetic images can be quantitatively analyzed. Third, the magnetization signal is very large. A huge number of secondary electrons are generated in a typical SEM

scan, and even in a low-moment material like nickel, the magnetization signal is 5% of the total secondary electron current. Unfortunately, as discussed below, the inherently low efficiency of spin polarization detectors makes measurement of this large signal quite difficult. Fourth, because the magnetic resolution is essentially given by the focused diameter of the primary electron beam, very high spatial resolutions (~20 nm) can be achieved. Fifth, the electron polarization detector simultaneously provides images of both the magnetization and the secondary electron intensity. The intensity images provide information about the physical structure of the sample under study. Hence, a comparison of the magnetic and topographic images can provide insight into the influence of physical structure on the magnetic domain structure. Finally, because the escape depth of low-energy secondary electrons is short (a few nanometers), SEMPA measures the magnetic properties of the near-surface region and is thus ideally suited to studies of magnetic thin films or of magnetic properties peculiar to the surface region.

An insert in Figure 1 shows a magnetic image of a test pattern in hard disk media as imaged by SEMPA. The details of the magnetic structure are clearly visible, both in the recorded tracks and in the randomly magnetized regions between the tracks.

Practical Aspects of the Method

There are several practical aspects of SEMPA which determine its applicability to specific magnetic-imaging problems. Most notably, because the attenuation length of the low-energy secondary electrons is typically sub-nanometer, SEMPA is sensitive to only the outermost few atomic layers. Hence, magnetic domain contrast is rapidly attenuated by nonmagnetic overlayers. Such overlayers could be the surface lubricant of magnetic hard disk media, protective non-magnetic capping layer of a magnetic multilayer, or merely the naturally occurring adsorbed gases and contaminants that are always present on surfaces exposed to the atmosphere. Prior to magnetic imaging, any nonmagnetic overlayers should be removed, generally by ion sputtering. In order to minimize the effects of surface contamination after cleaning, SEMPA studies must be performed under ultrahigh vacuum conditions.

One very useful side effect of SEMPAs high surface sensitivity is the ability to measure the magnetic properties of minute amounts of magnetic material. One can readily measure the magnetic properties of iron at sub-monolayer coverages. At ~1 monolayer coverage, with the primary electron beam focused into a 10-nm spot, SEMPA is measuring the magnetic properties of only ~1000 iron atoms.

As alluded to above, one practical consideration is the inherently low efficiency of electron polarization detectors. At present, most polarization detectors rely on differences, which are due to electron-spin orientation, in the cross-section for scattering of the electrons under investigation from a heavy nucleus. Because the overall scattering probability is low, and also because the difference due to electron spin is small, the overall efficiency of polarization detectors is $\sim 10^{-4}$. Consequently, data-acquisition times can be quite long, depending on the intensity of the primary electron beam and the inherent secondary electron polarization of the sample under study.

Data Analysis and Initial Interpretation

Data analysis in SEMPA is rather straightforward. The data are collected as the magnetization vector component along two orthogonal directions, typically in the sample plane. These components, however, can be readily combined to generate the net magnetization

$$|M| = \sqrt{M_x + M_y} \tag{1}$$

and the magnetization direction

$$\theta = \tan^{-1}\left(\frac{M_y}{M_x}\right) \tag{2}$$

While the magnetization direction is determined uniquely, the magnitude of the magnetization is determined only to within a constant. The absolute efficiency of most electron-polarization analyzers is not known with accuracy, so individual detectors require calibration. Furthermore, as mentioned above, the dependence of the secondary electron spin polarization on sample magnetization is known only approximately. However, because both of these factors are essentially constant, it is possible to make quantitative interpretations concerning changes in the degree of magnetization over the sample under study.

Because the magnetization signal from electron spin polarization detectors can be very small, instrumental artifacts can sometimes influence the magnetization images in significant ways. However, such instrumental artifacts are generally either reduced to negligible levels or measured and accounted for in the image analysis (Kelley et al., 1989).

Sample Preparation

Because SEMPA measures the magnetic properties of only the outermost atomic layers, samples should be atomically clean. The preferred method is ion sputtering. However, although the magnetic contrast may be significantly reduced, it is sometimes possible to observe domain structure through very thin layers. For example, one can observe domain structures essentially unchanged through a few monolayers of gold.

The collection of low-energy secondaries and the preservation of their spin orientation severely limit the acceptable size of any magnetic fields at the sample. Typically, stray fields must be nonvarying and of the order of the earth's magnetic field or smaller. Apart from special sample geometries that allow nearly complete elimination of stray magnetic fields, one is essentially restricted to zero-field measurements.

HOLOGRAPHY

Principles of the Method

Magnetic domain imaging is an important application of electron holography (Tonomura, 1994), in part because this technique makes possible both direct visualization and quantitative measurement of magnetic flux, Φ_B. In electron holography, a high-energy, e.g., 200-keV, field-emission transmission electron microscope (TEM) or scanning transmission electron microscope (STEM) is used to form an interference pattern from electrons that can reach the detector via two alternative paths, one of which contains the sample specimen. The electrons can be thought of as waves emitted from a very small source with a very narrow energy distribution, i.e., the beam has high degrees of both spatial and temporal coherence. The beam is divided into two parts using an electron biprism, with one half of the beam passing through the thin specimen to be studied while the "reference" half of the beam continues unimpeded. When the two halves of the beam are recombined on a film detector, following electron optical magnification, an interference pattern is observed. If there were no sample, the interference pattern would consist of equally spaced parallel lines. These lines are located at positions where the lengths of the two possible paths differ by an integral number of electron wavelengths. If there is a sample in one path, then the sample may introduce phase shifts in the electron wave passing through it, and those phase shifts will result in a local modification of the interference pattern.

When electron holography is used for magnetic imaging (Osakabe et al., 1983; Tonomura, 1994; Mankos et al., 1996a,b), interference is observed between electron waves in the reference path and those passing through a magnetic sample. However, an additional phase shift is introduced between the two waves. This phase shift is proportional to the total magnetic flux enclosed by the two alternative electron paths. For this case, known as the absolute method, the lines in the interference pattern can be directly interpreted as flux lines. In fact, the normalization is such that a flux of 4.1×10^{-15} Wb, i.e., h/e, flows in the space between two adjacent contour lines. An alternative, differential mode can also be used. In this mode, both beams go through the magnetic sample with slightly different paths. The phase difference, being proportional to the flux within the path enclosed by both beams, will be constant for a uniformly magnetized material and sensitive to a region of rapid change in enclosed flux, e.g., a domain wall. Figure 4 shows the flux distribution both inside and adjacent to a strip of magnetic tape measured using the absolute method in a TEM.

Practical Aspects of the Method

Using electron holography, high spatial resolution is possible; a resolution of 10 nm has been demonstrated. Since this is a transmission method, the magnetic sample must be uniformly thinned and the information obtained will represent values averaged over a typical sample thickness of 50 nm. Samples of up to 150 nm in thickness are generally possible. One must consider the possibility that the necessary thinning of the sample may affect the magnetization distribution in the sample being imaged. Only conductive samples can be measured. The flux enclosed in the electron path is quantitatively measured with a minimum sensitivity of $\sim 10^{-16}$ Wb. The instrument used is an extension of a high-quality transmission electron microscope. As such, it is a rather complex device. A commercial version designed for holography is available at a cost of $\sim$1.3 million dollars.

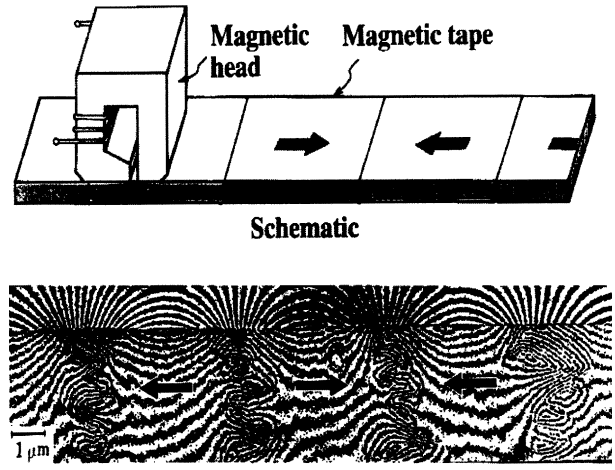


Figure 4. Top: A schematic diagram showing an inductive recording head writing domains of in-plane magnetization on a magnetic recording tape. Bottom: An interference micrograph, obtained using electron holography, illustrating the flux distribution both outside (top 25% of image) and inside (bottom 75%) this image of a 45-nm thick Co film (Osakabe et al., 1983).

Method Automation

A variety of image capture methods are available. Photographic film can be used or a CCD TV camera can provide a real-time display or permit transfer of the image to videotape to record dynamic information—e.g., the variation in response with a changing magnetic field.

Data Analysis and Initial Interpretation

The hologram that results from the two-beam interference in a microscope equipped for electron holography is an interferogram that directly displays the changes in phase that result from the magnetic flux of the sample. A major strength of this method is that these interference fringes can be directly interpreted as representing magnetic lines of force. Occasionally, additional sensitivity is required to see small variations in the enclosed magnetic flux. In this case, the phase-amplification technique (Tonomura, 1994) is used to increase, within limits set by the overall signal-to-noise ratio, the number of interference lines for a set amount of flux, i.e., to increase the sensitivity.

Sample Preparation

The sample must be uniformly thinned to $<$150 nm. If the absolute method is going to be used, the sample should be

prepared so that the area to be measured is either near the sample edge or near a hole in the sample through which the reference beam can pass.

Specimen Modification

It is possible to have a local magnetic field of <100 kA/m at the sample.

Problems

This is a complex method requiring a very significant investment in equipment, sample preparation, and image reconstruction and analysis. It has been applied to a wide range of problems, including small particle and multilayer magnetism, observation of domain walls, and imaging of fluxons in a superconductor.

X-RAY MAGNETIC CIRCULAR DICHROISM

Principles of the Method

X-ray magnetic circular dichroism is a recently developed technique made possible by the availability of intense, tunable synchrotron radiation from a new generation of storage rings. The effect depends on the relative orientation of the x-ray photon angular momentum and the sample magnetization. The difference in the absorption of x-rays, a phenomenon know as x-ray dichroism, is maximum when magnetization of the material and the photon angular momentum, or helicity, are parallel and antiparallel. The x-ray photon transfers angular momentum to the photoelectron excited from a spin-orbit split core level. For transition metals, XMCD measurements typically involve transitions from spin-orbit split p states (L_2 and L_3 core levels) to valence d states which in a ferromagnet are spin polarized. The transition probability depends on the spin orientation of the d final states, and hence on the magnetization.

A key advantage of XMCD for imaging magnetic domains is the elemental specificity that derives from the process being tied to an absorption event at a particular core level. In principle, it is possible to correlate the magnetic measurement with other core-level measurements, which give information on the local site, symmetry, and chemical state. Because the spin and orbital moments can be determined from XMCD measurements, the magnitude of the magnetization can also be quantitatively determined.

For domain imaging, the magnetic x-ray dichroism is monitored either by measuring the total electron yield (Stöhr et al., 1993; Tonner et al., 1994; Schneider, 1996) or, in the case of thin samples, by directly measuring the transmitted x-ray flux (Fischer et al., 1996). In the first method, the secondary electrons emitted from the magnetic material are electron-optically imaged onto a channel plate. The intensity in this magnified image is proportional to the x rays absorbed at the point where the electrons originate, and therefore on the relative orientation of the magnetization and the photon helicity at that point. Thus, XMCD gives a spatially resolved image of the magnetization direction. In the second method, the differential x-ray absorption is imaged in a transmission x-ray microscope to give the magnetization image. In either imaging mode, the image shows the projection of the magnetization along the photon propagation direction. In the electron-imaging mode, the surface is illuminated at oblique incidence, so the in-plane magnetization is predominantly imaged. In the transmission mode, the magnetization perpendicular to the sample surface is measured.

Recently, x-ray magnetic linear dichroism (XMLD) has also been observed (Hillebrecht et al., 1995). The difference in the absorption of linearly polarized x rays is the x-ray analog of the transverse magneto-optic Kerr effect (magnetization perpendicular to the plane of incidence). Although the image contrast obtained with linearly polarized light is about ten times less than with circularly polarized light, an image of the magnetization orthogonal to the light direction can be measured.

Practical Aspects of the Method

The lateral best spatial resolution achieved to date in the electron-imaging mode is 200 nm. Typical resolution is $\sim$500 nm. Magnetization images of a monolayer-thick magnetic film can be obtained. Changes in images can be observed in real time on a video monitor; acquisition time for a high-quality image is a few minutes. An image of recorded bits on a Co-Pt magnetic hard disk derived from the magnetic dichroism at the Co L-edge is shown in Figure 1. If the incident x-ray energy is tuned away from the absorption edges, the dichroism is absent and a secondary electron image of the topography is obtained. The nonuniform response of the optical system can be removed by normalizing to this topography image. The acquisition time and lateral resolution can be improved with increased synchrotron radiation flux, which is becoming available from insertion devices on new, higher-power storage rings. The ultimate resolution would then be limited by the resolution of the imaging electron optics, which has been demonstrated at 10 nm in low-energy electron microscopy (LEEM), as discussed in the next section.

Transmission XMCD (TXMCD) is an even more recent development. The lateral spatial resolution achieved in the first magnetic imaging experiments was 60 nm. Spatial resolution of 20 nm can be achieved in a transmission x-ray microscope. Information is provided about the bulk properties integrated over the path of the transmitted x rays through the sample. In contrast to the electron-imaging mode, there are no constraints on the application of magnetic fields. A 1024×1024 pixel image can be acquired in 3 s.

The equipment for magnetic imaging in either mode involves sophisticated electron or x-ray optics, and would be considered complex. The cost is $\sim$\$300,000, not including the required synchrotron radiation beam line.

Method Automation

Magnetic domain imaging using XMCD in either the electron imaging or transmission mode can be automated in

the sense that the incident wavelength variation and image acquisition are under computer control; in principle both methods could be set up for automatic data acquisition over several hours. In reality, however, both methods are sufficiently new that their application should not be considered routine or automatic.

Data Analysis and Initial Interpretation

There is sufficient contrast in XMCD that domains can be seen in the raw images as they are acquired. The contrast can be improved by subtracting properly scaled images taken at the L_2 and L_3 edges. This also removes possible artifacts due to topography, which can be present in the electron images. Using the elemental specificity of XMCD, it is possible to measure separately the contributions of different elements in a complicated system to the magnetization image.

Sample Preparation

The sample-preparation requirements are somewhat different in the two imaging modes. For electron imaging, the sample should be conductive; for an insulator, it should be not more than ~100 nm thick. The information sampling depth is material dependent; it is ~2 nm for transition metals and on the order of 10 nm in insulators. Even though the sampling depth for a transition metal ferromagnet is fairly short, there are no strict requirements on sample cleanliness, because the relative intensity of the secondary electrons is unchanged as they exit through a surface contamination layer such as carbon. Magnetic images can be obtained from surfaces with an rms roughness on the order of 100 nm.

A major consideration in the transmission imaging mode is the sample thickness, which should be selected for 30% to 40% x-ray transmission, which implies thicknesses on the order of 100 nm for transition metals. Because magnetic microstructure is thickness dependent, this imaging technique is particularly well suited when the sample to be measured is of the appropriate thickness. On the positive side, the sample need not be conducting, and the measurement is insensitive to surface contamination or moderate roughness.

SPIN POLARIZED LOW-ENERGY ELECTRON MICROSCOPY

Principles of the Method

Low-energy electron microscopy (LEEM) is a relatively new technique for surface imaging that uses electron lenses as in conventional electron microscopes to image elastically backscattered low-energy (1 to 100 eV) electrons (Bauer, 1994, 1996). In the case of crystalline samples, contrast is most often produced by diffraction. Other contrast mechanisms include interference in overlayers or between terraces of different height on the surface. Topographic features can be resolved with submonolayer vertical resolution and lateral resolution of 5 to 10 nm. In addition to measurements of surface topography, LEEM has been used to study phase transitions, sublimation and growth, and the interaction of the surface with foreign atoms such as adsorbates or segregated species. A particular strength of LEEM is its rapid image acquisition over a large field of view (5 to 10 μm diameter), which allows surface processes to be observed in real time.

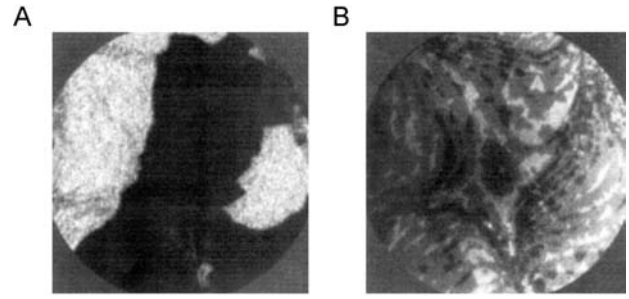


Figure 5. (**A**) SPLEEM image of 5-monolayer thick Co film on a W(110) single-crystal surface. The polarization is in-plane and collinear with the uniaxial magnetization. Taken at an energy of 1.5 eV, to optimize the magnetic contrast, with a field of view of 8 μm. (**B**) LEEM image of the same region at an energy of 3.6 eV showing the atomic scale roughness of the Co film. Courtesy of E. Bauer.

Magnetic contrast can be obtained by replacing the conventional electron gun with a spin polarized electron gun that produces a beam of electrons with a preferential orientation of the electron spins. In such a SPLEEM measurement, there is an additional interaction, the exchange interaction, between the incident electron spin $\mathbf{s}$ and the net spin density of a ferromagnetic or ferrimagnetic material. The contribution to the scattering resulting from the exchange interaction is proportional to $\mathbf{s} \bullet \mathbf{M}$, where $\mathbf{M}$ is the sample magnetization. The spin-dependent scattering is largest at an energy near a band gap in the spin-split band structure such that electrons of one spin but not the other are reflected. The greatest magnetic sensitivity is achieved by reversing the electron spin polarization direction in order to measure the normalized difference between the intensity $\mathbf{I}^{\uparrow\uparrow}$ with $\mathbf{s}$ and $\mathbf{M}$ parallel and $\mathbf{I}^{\uparrow\downarrow}$ with $\mathbf{s}$ and $\mathbf{M}$ antiparallel. This normalized spin-dependent asymmetry $\mathrm{A} = (\mathbf{I}^{\uparrow\uparrow} - \mathbf{I}^{\uparrow\downarrow})/(\mathbf{I}^{\uparrow\uparrow} + \mathbf{I}^{\uparrow\downarrow})$ has the advantage that it is independent of, but correlated with, the topographic contrast, $\mathbf{I}^{\uparrow\uparrow} + \mathbf{I}^{\uparrow\downarrow}$, which is measured independently at the same time. Examples of a magnetic and topographic image are shown in Figure 5. In a typical magnetic material, the magnitude of the magnetization is constant, but the direction varies from one domain to another or within a domain wall. Using a spin-rotation device in the incident beam, it is possible to get a quantitative measurement of the direction of the magnetization in a SPLEEM image.

Practical Aspects of the Method

SPLEEM is applicable to ferromagnetic or ferrimagnetic materials, i.e., materials with a net spin density. With electrons as the probe, samples should have sufficient conductivity to avoid charging. The sensitivity is such that domains in as little as two monolayers (ML) of Co can be imaged at a lateral resolution of 40 nm. The best spatial resolution demonstrated for SPLEEM at the present time is ~20 nm. Changes in the magnetization can be observed in real time as other parameters such as

temperature or film thickness are changed. Application of a magnetic field is usually not practical because a stray field would disturb the low-energy electrons involved in the imaging. The dynamical imaging frequency is ~1 Hz. The acquisition time for an 8 bit 512 × 512–pixel image of, for example, 5 ML Co, is 1 to 3 sec. The technique is developing and the resolution improving with the further development of a rather special objective lens, a cathode lens in which the specimen is one of the electrodes.

Method Automation

Data are usually acquired by measuring the difference between the images taken for opposite directions of the incident electron-beam polarization. Imaging can be automated in the sense that the digital image can be obtained as conditions, e.g., temperature or spin polarization direction of the incident beam, are varied under computer control.

Data Analysis and Initial Interpretation

From any particular SPLEEM image, one gets a picture of the magnetization along the chosen direction of spin polarization. By rotating the incident polarization, one can obtain a measure of the three components of the magnetization, M_x, M_y, and M_z, which are proportional to the asymmetries A_x, A_y, and A_z measured along each spin polarization direction. The distribution of magnetization directions can be plotted and analyzed in light of, e.g., aniso-tropies expected to determine the magnetization direction. A consistency check on the data is obtained by computing the relative magnitude of the magnetization, $M = (M_x^2 + M_y^2 + M_z^2)^{1/2}$, which is expected to be constant over the image.

Sample Preparation

Because low-energy electrons are central to this measurement, it is surface sensitive, with an energy- and material-dependent probing depth that ranges from a few monolayers to ~10 monolayers. The experiment therefore takes place in ultrahigh vacuum, and the usual surface-science techniques are used to prepare a clean surface. Because of the longer probing depths at very low energies, it may be possible to obtain an image from a surface covered by an adsorbate monolayer if the adsorbate does not affect the magnetization. Most magnetic materials are not adversely affected by low-energy electron beams. This technique is well suited for *in situ* ultrahigh vacuum measurements such as the observation of the development of magnetic domains as a function of film thickness.

SCANNING HALL PROBE AND SCANNING SQUID MICROSCOPES

Principles of the Method

Scanning Hall probe (Chang et al., 1992; Oral et al., 1996) and scanning superconducting quantum interference device (SQUID) techniques (Kirtley et al., 1995) give quantitative measurements of the stray magnetic field perpendicular to and just outside a surface (see HALL EFFECT IN SEMICONDUCTORS). Each has been demonstrated for domain imaging, although originally designed for measurement of magnetic flux in superconductors. Neither is commercially available, and with the limited spatial resolution demonstrated to date, the choice of either of these techniques would likely be driven by the need for noninvasive quantitative measurements with high sensitivity. Both the Hall and SQUID probes use scanning tunneling microscopy positioning techniques to achieve close proximity to the surface and for scanning.

Practical Aspects of the Method

The scanning Hall probe is more versatile than the SQUID. It consists of a submicron Hall sensor, manufactured in a two-dimensional electron gas, and gives a voltage output proportional to the magnetic field perpendicular to the sensor. It has been operated at temperatures from that of liquid He to room temperature (with a reduction in sensitivity) and does not have a limitation on the ambient magnetic field. Typical spatial resolution is 1000 nm, and the best resolution achieved is 350 nm. The magnetic field sensitivity depends on the speed or bandwidth of the measurement. At a temperature of 77 K, the sensitivity is 3×10^{-8} T/(Hz)$^{1/2}$ times the square root of the bandwidth given in hertz. Hence, a measurement with a sensitivity of 3×10^{-6} T is possible with a 10-kHz bandwidth. High spatial and magnetic-field resolution images require a few minutes. Lower-resolution images can be acquired in a few seconds. The scanning SQUID is typically operated at liquid He temperature in low (<8 kA/m) ambient magnetic fields. A spatial resolution of 10 μm has been demonstrated with a magnetic field sensitivity of 10^{-10} T/Hz$^{1/2}$.

LITERATURE CITED

Argyle, B. E. 1990. A magneto-optic microscope system for magnetic domain studies. *In* Proceedings of the 4th International Symposium on Magnetic Materials and Devices, Vol. 908 (L. T. Romankiw and D. A. Herman, eds.). pp. 85–95. Electrochemical Society, Pennington, N.J.

Bauer, E. 1994. Low energy electron microscopy. *Rep. Prog. Phys.* 57:895–938.

Bauer, E. 1996. Low energy electron microscopy. *In* Handbook of Microscopy (S. Amelinckx, D. van Dyck, J. F. van Landuyt, and G. van Tendeloo, eds.). VCH, Weinheim, Germany.

Betzig, E., Trautman, J. K., Wolfe, R., Gyorgy, E. M., Finn, P. L., Kryder, M. H., and Chang, C. H. 1992. Near-field magneto-optics and high density data storage. *Appl. Phys. Lett.* 61:142–144.

Chang, A. M., Hallen, H. D., Harriott, L., Hess, H. F., Kao, H. L., Kwo, J., Miller, R. E., Wolfe, R., vander Ziel, J., and Chang, T. Y. 1992. Scanning Hall probe microscopy. *Appl. Phys. Lett.* 61:1974–1976.

Chapman, J. N. 1984. The investigation of magnetic domain structures in thin film foils by electron microscopy. *J. Phys. D* 17: 623–47.

Craik, D. J. 1974. The observation of magnetic domains. *In* Methods of Experimental Physics, Vol. 11, (R. V. Coleman, ed.). pp. 675–743. Academic Press, New York.

Du, M., Xue, S. S., Epler, W., and Kryder, M. H. 1995. Dynamics of the domain collapse in the true direct overwrite magneto-optic disk, *IEEE Trans. Magn.* 31:3250–3252.

Fischer, P., Schütz, G., Schmahl, G., Guttmann, P., and Raasch, D. 1996. Imaging of magnetic domains with the X-ray microscope at BESSY using X-ray magnetic circular dichroism. *Z. Phys. B* 101:313–316.

Goto, K. and Sakurai, T. 1977. A colloid-SEM method for the study of fine magnetic domain structures. *Appl. Phys. Lett.* 30:355–356.

Heyderman, L. J., Chapman, J. N., Gibbs, M. R. J., and Shearwood, C. 1995. Amorphous melt spun ribbons and sputtered thin films—investigation of the magnetic domain structures by TEM. *J. Mag. Mag. Mater.* 148:433–445.

Hillebrecht, F. U., Kinoshita, T., Spanke, D., Dresselhaus, J., Roth, Ch., Rose, H. B., and Kisker, E. 1995. New magnetic linear dichroism in total photoelectron yield for magnetic domain imaging. *Phys. Rev. Lett.* 75:2224–2227.

Hubert, A. and Schäfer, R. 1998. Magnetic Domains: the Analysis of Magnetic Microstructures. Springer-Verlag, Berlin.

Hug, H. J., Stiefel, B., van Schendel, P. J. A., Moser, A., Hofer, R., Güntherodt, H.-J., Porthun, S., Abelmann, L., Lodder, J. C., Bochi, G., and O'Handley, R. C. 1998. Quantitative magnetic force microscopy on perpendicularly magnetized samples. *J. Appl. Phys.* 83:5609–5620.

Kelley, M. H., Unguris, J., Scheinfein, M. R., Pierce, D. T., and Celotta, R. J. 1989. Vector imaging of magnetic microstructure. *In* Proceedings of the Microbeam Analysis Society Meeting (P. E. Russell, ed.). pp. 391–395. San Francisco Press, San Francisco, Calif.

Kirilyuk, V., Kirilyuk, A., and Rasing, Th. 1997. New mode of domain imaging: Second harmonic generation microscopy. *J. Appl. Phys.* 81:5014–5015.

Kirtley, J. R., Ketchen, M. B., Tsuei, C. C., Sun, J. Z., Gallagher, W. J., Yu-Jahnes, L. S., Gupta, A., Stawiasz, K. G., and Wind, S. J. 1995. Design and applications of a scanning SQUID microscope. *IBM J. Res. Dev.* 39:655–668.

Kitakami, O., Sakurai T., and Shimada, Y. 1996. High density recorded patterns observed by high-resolution Bitter scanning electron microscope method. *J. Appl. Phys.* 79:6074–6076.

Kittel, C. and Galt, J. K. 1956. Ferromagnetic domain theory. in Solid State Physics, Vol. 3 (F. Seitz and D. Turnbull, eds.). pp. 437–564. Academic Press, New York.

Mankos, M., Cowley, J. M., and Scheinfein, M. 1996b. Quantitative micromagnetics at high spatial resolution using far-out-of-focus STEM electron holography. *Phys. Status Solidi A* 154:469–504.

Mankos, M., Scheinfein, M. R., and Cowley, J. M. 1996a. Electron holography and Lorentz microscopy of magnetic materials. *Adv. Imaging Electron Phys.* 98:323–426.

McFadyen, I. R. and Chapman, J. N. 1992. Electron microscopy of magnetic materials. *Electron Microsc. Soc. Am. Bull.* 22: 64–75.

McVitie, S. and Chapman, J. N. 1997. Reversal mechanisms in lithographically defined magnetic thin film elements imaged by scanning transmission electron. *Microsc. Microanal.* 3: 146–153.

Newbury, D. E., Joy, D. C., Echlin, P., Fiori, C. E., and Goldstein, J. I. 1986. Advanced Scanning Electron Microscopy and X-Ray Microanalysis. Plenum, New York.

Nikitenko, V. I., Gornakov, V. S., Dedukh, L. M., Kabanov, Yu, P., Khapikov, A. F., Bennett, L. H., Chen, P. J., McMichael, R. D., Donahue, M. J., Swartzendruber, L. J., Shapiro, A. J., Brown, H. J., and Egelhoff, W. F. 1996. Magneto-optical indicator film study of the magnetization of a symmetric spin valve. *IEEE Trans. Magn.* 32:4639–4641.

Oral, A., Bending, S. J., and Henini, M. 1996. Real-time scanning Hall probe microscopy. *Appl. Phys. Lett.* 69:1324–1327.

Osakabe, N., Yoshida, K., Horiuchi, Y., Matsuda, T., Tanabe, H., Okuwaki, T., Endo, J., Fujiwara, H., and Tonomura, A. 1983. Observation of recorded magnetization pattern by electron holography. *Appl. Phys. Lett.* 42:746–748.

Petek, B., Trouilloud, P. L., and Argyle, B. 1990. Time-resolved domain dynamics in thin-film heads. *IEEE Trans. Magn.* 26: 1328–1330.

Reimer, L. 1985. Scanning Electron Microscopy. Springer-Verlag, Berlin.

Rugar, D. and Hansma, P. 1990. Atomic force microscopy. *Phys. Today* 43:23–30.

Sarid, D. 1991. Scanning Force Microscopy. Oxford University Press, New York.

Scheinfein, M. R., Unguris, J., Kelley, M. H., Pierce, D. T., and Celotta, R. J. 1990. Scanning electron microscopy with polarization analysis (SEMPA). *Rev. Sci. Instrum.* 61:2501–2506.

Schmidt, F., Rave, W., and Hubert, A. 1985. Enhancement of magneto-optical domain observation by digital image processing. *IEEE Trans. Magn.* 21:1596–1598.

Schneider, C. M. 1996. Imaging magnetic microstructures with elemental selectivity: Application of magnetic dichroisms. *In* Spin-Orbit Influenced Spectroscopies of Magnetic Solids (H. Ebert and G. Schütz, eds.). pp. 179–196, Springer-Verlag, Berlin.

Schoenenberger, C., Alvarado, S. F., Lambert, S. E., and Sanders, I. L. 1990. Separation of magnetic and topographic effects in force microscopy. *J. Appl. Phys.* 67:7278–7280.

Silva, T. J. and Kos, A. B. 1997. Nonreciprocal differential detection method for scanning Kerr-effect microscopy. *J. Appl. Phys.* 81:5015–5017.

Silva, T. J. and Schultz, S. 1996. A scanning near-field optical microscope for the imaging of magnetic domains in reflection. *Rev. Sci. Instrum.* 67:715–725.

Stöhr, J., Wu, Y., Hermsmeier, B. D., Samant, M. G., Harp, G. R., Koranda, S., Dunham, D., and Tonner, B. P. 1993. Element-specific magnetic microscopy with circularly polarized X-rays. *Science* 259:658–661.

Tonner, B. P., Dunham, D., Zhang, J., O'Brien, W. L., Samant, M., Weller, D., Hermsmeier, B. D., and Stöhr, J. 1994. Imaging magnetic domains with the X-ray dichroism photoemission microscope. *Nucl. Instrum. Methods A* 347:142–147.

Tonomura, A. 1994. Electron Holography. Springer-Verlag, Berlin.

Trouilloud, P. L., Petek, B., and Argyle, B. E. 1994. Methods for wide-field Kerr imaging of small magnetic devices. *IEEE Trans. Magn.* 30:4494–4496.

Wells, O. C. 1983. Fundamental theorem for type-1 magnetic contrast in the scanning electron-microscope (SEM). *J. Microscopy-Oxford* 131:RP5–RP6.

R. J. Celotta
J. Unguris
M. H. Kelley
D. T. Pierce
National Institute of Standards and Technology
Gaithersburg, Maryland

MAGNETOTRANSPORT IN METALS AND ALLOYS

INTRODUCTION

Electronic transport properties are fundamental to the classification of materials. The behaviors of the electrical resistivity ρ, the thermal conductivity κ, and the thermopower S are used to define whether a material is a metal, a semiconductor, or an insulator. Studies of how ρ, κ, and S vary with impurity content (alloying), magnetic field $\mathbf{B}$, sample size, deformation, etc., provide insight into the nature of current carriers and how they are scattered. Studies of magnetoresistance, the variation of ρ with $\mathbf{B}$, can yield additional information about electronic structure, the current carriers, and their scattering. In systems involving magnetic metals, $\rho(\mathbf{B})$ has technical applications to magnetic sensing and memory (Heremens, 1993a). The Hall effect and the thermopower S can often be used to infer the sign of the charge of the majority current carriers, and the thermal conductivity κ, while closely related to ρ, can manifest differences that contain significant information. The Hall effect in semiconductors also finds use in sensors (Heremens, 1993a,b), but in metals it is smaller and usually of more interest for the physical insight it can provide into both nonmagnetic and magnetic metals. Changes with magnetic field in the thermal conductivity and thermoelectric coefficients of metals are usually small and difficult to measure. They have provided useful information about physical phenomena such as many-body contributions to thermoelectricity (see, e.g., Averback et al., 1973; Thaler et al., 1978) and giant magnetoresistance in granular alloys (Piraux et al., 1993) and magnetic multilayers (Tsui et al., 1994) but are much less studied than the resistivity and Hall effect. Transport and magnetotransport measurements of metals and alloys are made over a wide range of temperatures extending from the lowest achievable temperature to the liquid state.

This unit focuses on four measured quantities in solid metals and alloys in the presence of a temporally constant and spatially uniform magnetic field $\mathbf{B}$—the electrical resistance R, the Hall resistance R_H, the thermal conductance K, and the thermopower S. The basic theory underlying these quantities applies also to liquid metals and alloys. The discussion begins with definitions and general information, including how to relate the measured quantities to the fundamental properties ρ, the Hall coefficient R_0, κ, and S. These are followed by a brief description of the behaviors of ρ, κ, and S in zero magnetic field and, in some more detail, those of $\rho(\mathbf{B})$, $R_H(\mathbf{B})$ and, to a lesser extent, $\kappa(\mathbf{B})$ and $S(\mathbf{B})$. Many more quantities than these four can be defined and measured, some of which have provided important information about metals (see, e.g., Fletcher, 1977; Thaler et al., 1978). A few will be briefly discussed at the end of the following section; for more information see Jan (1957), Scanlon (1959), and Blatt et al. (1976).

PRINCIPLES OF THE METHOD

Measured Quantities and Their Relation to Intrinsic Ones

To define the measured quantities and relate them to intrinsic ones, consider a sample in the form of a thin foil of width W and thickness t (Fig. 1), with the x axis along the foil length, the y axis along the foil width, and transverse $\mathbf{B}$ in the direction B_z. The cross-sectional area of the foil is $A = Wt$. An electric current I_x or a heat current $\dot{Q}_x$ is sent through the sample, and a voltage V or a temperature difference ΔT is measured across a length L (or the width W) of the sample. Experimental conditions are assumed such that the electrical or heat current densities $j = I/A$ or $\dot{q} = \dot{Q}/A$ are uniform across the sample area A.

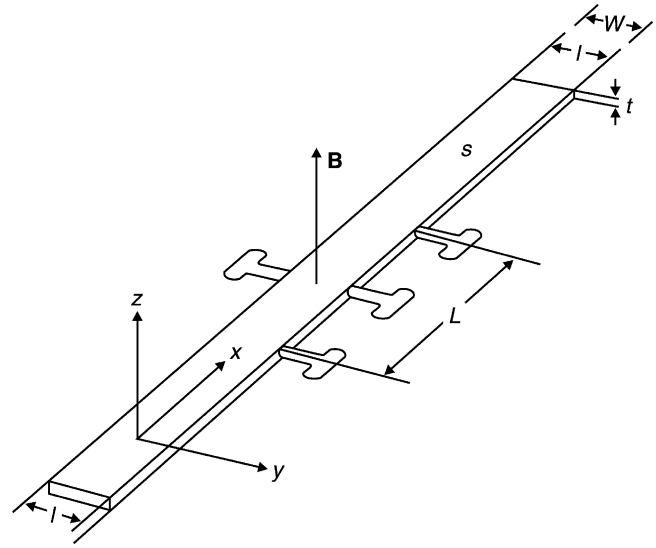


Figure 1. Sample foil (s) of width W and thickness t; l indicates the current leads.

Electrical Resistance. Electrical resistance R for a length L of the foil is measured by passing a current I_x through the foil and measuring the ratio of the voltage difference V_x across L to I_x:

$$R(\text{ohms}) = V_x(\text{volts})/I_x(\text{amperes}) \tag{1}$$

Whether or not a field $\mathbf{B}$ is present, the intrinsic quantity is the resistivity ρ, related to R by

$$R = \rho L/A \tag{2}$$

For measurements in a magnetic field, it is useful to define the magnetoresistance MR as the ratio of the change in resistance due to application of $\mathbf{B}$ over the initial resistance:

$$\text{MR}(\mathbf{B}) = \Delta R(\mathbf{B})/R(0) = [R(\mathbf{B}) - R(0)]/R(0) \tag{3}$$

For Equation 3, the current distribution need not be uniform. So long as the distribution does not change with the magnitude of $\mathbf{B}$, MR($\mathbf{B}$) will be an intrinsic quantity, i.e., geometric effects will cancel out.

See Crawford Dunlap (1959), Pearson (1959), Meaden (1965), Rossiter (1987), and Rossiter and Bass (1994) for more detailed discussion.

Hall Resistance. Hall resistance R_H is defined as the ratio of the voltage V_y in the y direction produced by a current I_x in the x direction when the field is oriented along the z direction:

$$R_H = V_y/I_x \quad \text{for} \quad B = B_z \tag{4}$$

The intrinsic quantity in the Hall effect is the Hall coefficient, R_0, defined by

$$R_H = \rho_H/t = R_0 B/t \tag{5}$$

By symmetry, R_H should be zero at $B = 0$.

See Fritzsche (1959), Hurd (1972), and Chien and Westgate (1980) for more detailed discussion.

Thermal Conductance. Thermal conductance K is defined as the ratio of the rate of heat current, $\dot{Q}_x$, through the sample to the difference between the hot (H) and cold (C) temperatures, $\Delta T_x = T_H - T_C$, across the length L,

$$K = \dot{Q}_x(\text{watts})/\Delta T_x(\text{kelvin}) \quad \text{for } I_x = 0 \tag{6}$$

The intrinsic quantity, the thermal conductivity κ, is determined by

$$K = \kappa A/L \tag{7}$$

See Pearlman (1959) and Tye (1969) for more detailed discussion.

Thermopower. Three thermoelectric coefficients, the thermopower S, the Peltier heat Π, and the Thomson coefficient μ, are coupled by the Kelvin relations:

$$S = -\Pi/T \tag{8a}$$

and

$$\mu = T(dS/dT) \tag{8b}$$

As Π and μ (the only one of the three that can be directly measured on a single material) are generally more difficult to measure reliably, experimentalists usually study S. To measure S_A or Π_A of a metal A requires use of a thermocouple (Fig. 2) composed of two different metals, A and B, electrically connected at both ends of A. For the temperature difference between the thermocouple ends, $\Delta T_x = T_H - T_C, S_{AB} = S_B - S_A$ is defined as positive when $\Delta V_x = V_H - V_C$ is positive in

$$S_{AB}\,(\text{volts/kelvin}) = \Delta V_x(\text{volts})/\Delta T_x(\text{kelvin}) \quad \text{for} \quad I_x = 0 \tag{9}$$

The thermopower $S_A(T)$ can be found if B is a reference material for which $S_B(T)$ is known. At low enough temperatures, the use of a material B that is superconducting is especially convenient, since its thermopower is zero and remains so up to whatever magnetic field turns it fully normal. The standard reference materials for zero field

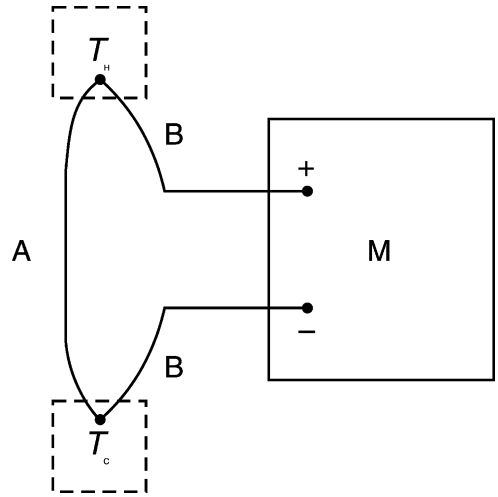


Figure 2. Apparatus used to measure the thermopower S_{AB}. A thermocouple of metals A and B with a hot temperature T_H at one junction and a cold temperature T_C at the other is connected to a device M that measures the output voltage V_{AB}.

measurements are Pb up to room temperature (Roberts, 1977) and Pt or W up to 1600 to 1800 K (Roberts et al., 1985), their absolute values of S having been determined by measuring their Thomson coefficients. To directly determine Π_A requires measuring the rate of reversible heat emission or absorption at the AB junction when a current I_x is passed through the junction.

Other transport coefficients, such as the ratio of V_x to the heat input $\dot{Q}_x$, can also be defined and measured (see, e.g., Stone et al., 1976).

See Frederikse et al. (1959), Barnard (1972), and Blatt et al. (1976) for more detailed discussion of thermoelectric coefficients.

Transport Equations

It is assumed that the sample shape and the measuring geometry provide uniform current densities of both charge $\mathbf{j}$ and heat $\dot{\mathbf{q}}$ flowing through the sample. Because a uniform $\mathbf{B}$ does no work on moving electrons, the form of the transport equations does not change when $\mathbf{B}$ is applied, but the transport coefficients become functions of $\mathbf{B}$ and must be treated as tensors even for a completely homogeneous material. If either $\mathbf{j}$ or $\dot{\mathbf{q}}$ is not uniform, one must start instead from Laplace's equation (Jackson, 1962); two examples are given below (see Problems section).

There are two important alternative forms of the transport equations in rectangular coordinates. One involves measured quantities and relates the electric field $\mathbf{E}$ and heat flow rate density $\dot{\mathbf{q}}$ to the applied current density $\mathbf{j}$ and temperature gradient ∇T.

$$\mathbf{E} = \rho(\mathbf{B}) \cdot \mathbf{j} + \mathbf{S}(\mathbf{B}) \cdot \nabla T \tag{10a}$$

$$\dot{\mathbf{q}} = \Pi(\mathbf{B}) \cdot \mathbf{j} - \kappa(\mathbf{B}) \cdot \nabla T \tag{10b}$$

The electrical resistivity tensor elements are $\rho_{ik}(\mathbf{B}) = E_i/j_k$ with $\nabla T = 0$, the thermopower elements are $S_{ik}(\mathbf{B}) = -E_i/\nabla_k T$ for $\mathbf{j} = 0$, and the thermal conductivity

elements are $\kappa_{ik}(\mathbf{B}) = \dot{q}_i/\nabla_k T$ for $\mathbf{j} = 0$. Tensor elements of each quantity usually depend upon both the magnitude and direction of $\mathbf{B}$, as well as upon the directions of $\mathbf{j}$ and ∇T relative to the crystallographic axes of the metal. If the metal is magnetic, the tensor elements depend also upon its magnetization $\mathbf{M}$ and usually upon its history (hysteresis). Magnetotransport analysis can be quite complex.

The Kelvin-Onsager relations (Jan, 1957) for the elements of the Peltier heat tensor Π and the thermopower tensor $\mathbf{S}$ reduce the number of independent tensors in Equations 10 from four to three via $\Pi_{ki}(-\mathbf{B}) = TS_{ik}(\mathbf{B})$, and Onsager's relations give $\rho_{ik}(\mathbf{B}) = \rho_{ki}(-\mathbf{B})$ and $\kappa_{ik}(\mathbf{B}) = \kappa_{ki}(-\mathbf{B})$. Thus, when transport equations are applicable, the diagonal components of the $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ tensors should remain unchanged under reversal of $\mathbf{B}$, but the off-diagonal components should reverse sign. Consider, for example, the Hall bar geometry of Figure 1. Then $R_x = V_x/I_x$ is the resistance, $\mathrm{MR} = [R_x(\mathbf{B}) - R_x(0)]/R_x(0)$ is the magnetoresistance, and $R_\mathrm{H} = V_y/I_x$ is the Hall resistance. Note that these are all defined under isothermal conditions: $\nabla T = 0$. As in Equation 2, R and R_H are related to the resistivities in Equation 10 by

$$R = \rho_{xx} Ll A \tag{11a}$$

and

$$R_\mathrm{H} = \rho_{xy}/t \tag{11b}$$

The difference in these relations occurs because of the difference between $V_x = E_x L$ and $V_y = E_y W$. To determine ρ_{xx}, one must measure L, W, and t of the foil, but for ρ_{xy}, only t.

In the Hall bar geometry, the relative orientations of I_x and B_z are fixed. But if the sample is a single crystal, the magnitudes of R_x and R_H, and even the forms of their variations with the magnitude of $\mathbf{B}$, will generally depend upon which crystallographic axis lies along x and the angle between the other crystallographic axes and z (see Practical Aspects of the Method for a discussion of expected behaviors in a magnetic field).

The alternative form of Equations 10 more applicable to calculations relates $\mathbf{j}$ and $\dot{\mathbf{q}}$ to $\mathbf{E}$ and ∇T:

$$\mathbf{j} = \boldsymbol{\sigma}(\mathbf{B}) \cdot \mathbf{E} - \varepsilon(\mathbf{B}) \cdot \nabla T \tag{12a}$$

$$\dot{\mathbf{q}} = T\varepsilon^\mathrm{t}(-\mathbf{B}) \cdot \mathbf{E} - \mathbf{K}(\mathbf{B}) \cdot \nabla \mathbf{T} \tag{12b}$$

where ε^t indicates the transpose—$\varepsilon^\mathrm{t}_{ij} = \varepsilon_{ji}$. Each of the tensor elements in Equations 12 is related to one or more of the tensor elements in Equations 10 by tensor (matrix) inversion. The simplest case is

$$\rho = \sigma^{-1} \tag{13}$$

because R is measured with $\nabla T = 0$. The relation between κ and $\mathbf{K}$ is more complex,

$$\kappa = \mathbf{K} - T\varepsilon^\mathrm{t} \cdot \sigma^{-1}\varepsilon \tag{14}$$

because $\boldsymbol{K}$ is measured with $\mathbf{j} = 0$. Fortunately, in metals, the second term on the right in Equation 14 is usually small. Lastly, as $\mathbf{S}$ is measured with $\mathbf{j} = 0$, the thermoelectric coefficients are related by

$$\mathbf{S} = \boldsymbol{\sigma}^{-1} \cdot \varepsilon \tag{15}$$

It is important to remember that Equations 10 to 15 are tensor relations. That is, to calculate the tensor on the left, all of the elements of the tensors on the right must be known. In even the simplest cases, such as Equation 13, very different dependences of the elements $\sigma_{ij}(\mathbf{B})$ for different materials can lead to very different behaviors for ρ_{xx} and ρ_{xy}, which generally depend upon the size of $\mathbf{B}$—more specifically, upon whether the product of the cyclotron frequency $\omega_\mathrm{c} = eB/m$ (m is the electron mass) and the transport scattering time τ are such that $\omega_\mathrm{c}\tau \ll 1$ (low-field limit) or $\omega_\mathrm{c}\tau \gg 1$ (high-field limit) (Abrikosov, 1988). Some examples of different behaviors are given below (see Practical Aspects of the Method).

Since each transport tensor has nine elements, and these elements depend upon the direction of $\mathbf{B}$, many different magnetotransport coefficients can be defined (Jan. 1957). A few that are relevant to the precautions necessary for reliable measurements of $\boldsymbol{\rho}(\mathbf{B})$, $\mathbf{R_H}$, $\boldsymbol{\kappa}(\mathbf{B})$, and $\mathbf{S}(\mathbf{B})$ (see Problems section, below) will be described briefly here.

Consider the foil sample of Figure 1 with current leads (l) composed of a different metal from the sample (s), and make the current density small enough so that Joule heating (proportional to j^2) is negligible. For $B = B_z$, passing a current I_x through the sample will generate not only the desired voltages $E_x L = V_x = R_{xx} I_x$ and $E_y W = V_y = R_{yx} I_x$ (Equations 10a and 11) but also the Thomson heat $\dot{Q}_x = (\Pi^l_{xx} - \Pi^s_{xx})l_x$ at one s-l contact and absorb the same heat at the other (Equation 10b). If the sample is in a vacuum, the off-diagonal elements of the thermal conductivity and thermoelectric tensors will lead to both temperature gradients and voltages being established not only along the length of the sample ($\nabla_x T$ and voltage change ΔV_x) but also across its width ($\nabla_x T$ and V_y). The production by a longitudinal current of a transverse temperature gradient is called the Ettingshausen effect and of a transverse voltage the Nernst-Ettingshausen effect. The comparable production of a transverse temperature gradient by a longitudinal heat current is called the Righi-Leduc effect and of a transverse voltage a second Nernst-Ettingshausen effect. Because of these effects, the protocols for measuring R and R_H will specify that the sample be kept as nearly isothermal as feasible. An alternating current should automatically give isothermal conditions, since the reversible heats will average to zero.

PRACTICAL ASPECTS OF THE METHOD

Expected Behaviors of ρ, κ and *S* in Zero Field

For completeness, the expected forms of ρ, κ, and S in a zero magnetic field will be briefly summarized. Most metals have cubic crystal structures—simple cubic (sc), face-centered cubic (fcc), or body-centered cubic (bcc)—the high symmetry of which reduces the tensors $\boldsymbol{\rho}$, $\boldsymbol{\kappa}$ and $\mathbf{S}$ to scalars ρ, κ, and S, independent of the directions of the charge

or heat current flows relative to the crystallographic axes. Metals with the simplest noncubic structures—e.g., hexagonal close packed—require two independent tensor elements, one for current flow in the basal plane and one for flow perpendicular to that plane (Bass, 1982; Foiles, 1985). Alloys with more complex structures may require more elements. Here we limit ourselves to cubic metals, or to polycrystalline samples, so that ρ, κ, and S are scalars. To determine the forms of ρ, κ, and S, a model is needed for the scattering of electrons in a metal. The simplest is the Drude-Sommerfeld "free electron" model (Ashcroft and Mermin, 1976), where the electron Fermi surface is taken to be spherical and electrons are assumed to be scattered with a single relaxation time τ that will generally be different if the scattering is dominated by impurities or by quantized lattice vibrations (phonons). This model is useful for estimating the magnitudes of ρ, κ, and S and for qualitatively describing their behaviors. It rarely works quantitatively, however, and can even fail qualitatively. The characteristic behaviors of ρ, κ, $L = \rho\kappa/T$, and S are illustrated in Figure 3 for a high-purity ("pure") metal (solid curves) and for a dilute alloy (dashed curves).

For free electrons scattered only by phonons that remain in thermal equilibrium, solution of the transport equations (Equations 12, above) predicts a pure metal resistivity, $\rho_p(T)$, that varies approximately linearly with T, except at low temperatures, where it should vary approximately as T^5 (Ashcroft and Mermin, 1976; Kittel, 1986). To a first approximation, adding a small impurity concentration c should just add a constant term proportional to c, giving $\rho_t(T,c) = \rho_p(T) + \rho_0(c)$—Matthiessen's rule. Deviations from Matthiessen's rule are examined by Bass (1972) and values of $\rho_0(c)$ for a variety of hosts and impurities are given by Bass (1982). Of course, even the highest-purity metal has residual impurities, so its resistivity at very low temperature will not be zero, unless the metal becomes superconducting. Figure 3A illustrates the expected behaviors for such a pure metal and for a dilute alloy.

Analysis of κ and S is complicated by the fact that they are measured in the presence of a temperature gradient and that both electrons and phonons can carry heat through a metal.

If the phonons remain at thermal equilibrium (i.e., carry no heat), then from the free electron solution of Equations 12 above, the Wiedemann-Franz ratio $L = \kappa\rho/T$ should be constant at the Lorenz number $L = L_0 = 2.45$ W Ω/K^2 (Ashcroft and Mermin, 1976), both at high temperatures where the phonon Debye energy is $\leq k_B T$ and at low temperatures where impurity scattering dominates, and should fall below L_0 in between. Kittel (1986) and Ashcroft and Mermin (1976) list values of L at 273 and 373 K for a wide variety of metals, and most lie within several percent of L_0. Combining these predictions for L with expectations of $\rho \propto T$ at high T and $\rho = \rho_0$ at low T requires $\kappa = \kappa_0$ at high T and $\kappa \propto T$ at low T. The expected behaviors of κ for a pure metal and dilute alloy are illustrated in Figure 3B, with those for L in the inset. For a concentrated alloy, the electronic contributon to κ can become so small that the heat carried by the flow of nonequilibrium phonons becomes significant. A high magnetic field also tends to reduce the electronic κ, thereby increasing the relative importance of phonons. In fact, application of a high field has been used to determine the phonon contribution to κ (see, e.g., Pernicone and Schroeder, 1975), which should be nearly independent of **B**.

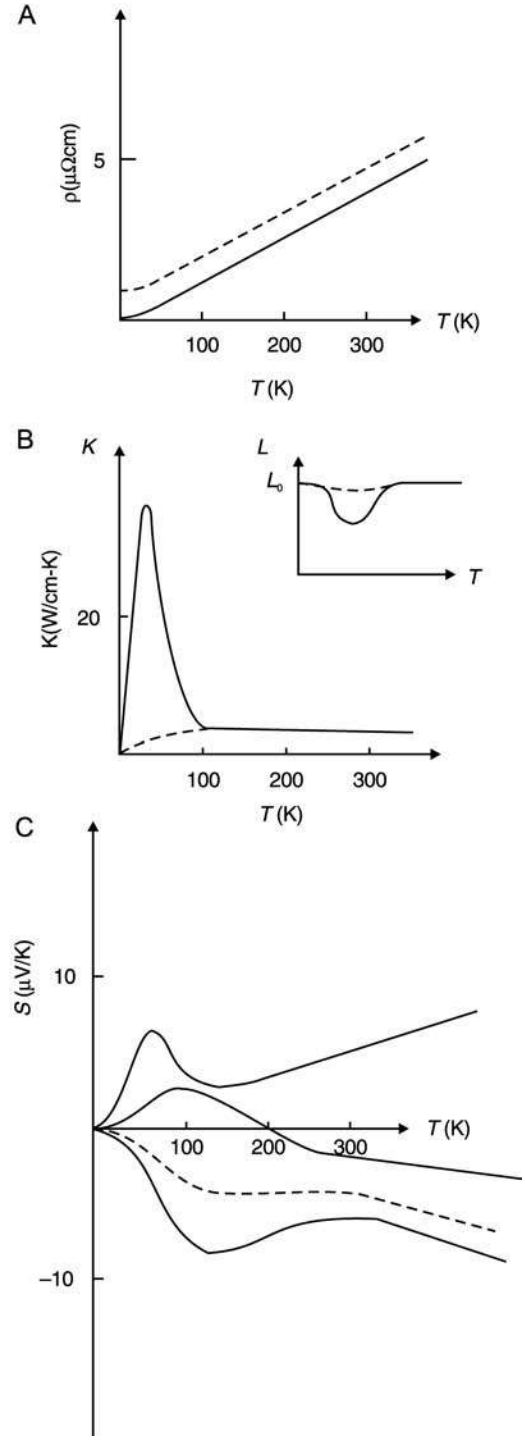


Figure 3. Schematics of the temperature dependences of ρ, κ, L, and S for high-purity metals (solid curves) and dilute alloys (dashed curves).

For phonons in equilibrium, the free electron solution of Equations 12 predicts that S should be negative and linear

in T when the primary source of scattering is either impurities (low temperature) or phonons (high temperature), with different coefficients of the linear term for the two cases (Blatt et al., 1976). In fact, S usually has a more complex form that can be approximated (Blatt et al., 1976) as the sum of an electronic thermopower term S_e that is often nearly linear in T and a phonon-drag thermopower term S_g (due to the nonequilibrium flow of phonons resulting from the applied temperature gradient) that goes through a maximum as T increases. Unlike κ, for which the effect of phonon flow is usually small in metals, S_g can predominate over a substantial temperature range, as illustrated in Figure 3C. While the predicted negative free electron S_e is seen in some metals, such as Al, others, such as the noble metals Ag, Au, and Cu, have positive S_e. Such an "inverse" sign is attributed to deviations of the Fermi surface from a free electron sphere. As shown in Figure 3C, both S_e and S_g can be negative or positive. The dashed curve shows that adding impurities can change both S_e and S_g; while such addition usually reduces S_g, in some cases it can increase it.

Expected Behaviors of ρ, R_H, κ, and S in a Magnetic Field

The vast majority of magnetotransport studies have involved $\rho(\mathbf{B})$ and $R_H(\mathbf{B})$, both because they are easiest to measure and because their interpretation is usually most straightforward. The difficulty of measuring $\kappa(\mathbf{B})$, combined with the expectation that a generalization of the Wiedemann-Franz law will apply under those circumstances where it can be most simply interpreted, has limited studies of $\kappa(\mathbf{B})$. The combination of experimental difficulties, lack of simple interpretation, and expectation that changes will usually be relatively small, have limited studies of $S(\mathbf{B})$. The main focus here will thus be upon $\rho(\mathbf{B})$ and $R_H(\mathbf{B})$.

The Sommerfeld free electron model predicts $\rho(\mathbf{B}) = \rho(\mathbf{B} = \mathbf{0})$, i.e., MR = 0, and $R_H(\mathbf{B}) = -1/n_e ec$, where n_e is the electron density in the metal, e is the magnitude of the electron charge, and c is the speed of light in a vacuum. This prediction results from the assumption of a spherical Fermi surface with only a single relaxation time. Due to the Lorentz force $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ on a charge q moving with velocity $\mathbf{v}$, application of $\mathbf{B}$ causes electrons to be deflected from moving straight down the sample foil, but the boundary condition $j_y = 0$ leads to a pileup of electrons on one side of the foil. This produces an electric field E_y (Hall effect) that exactly counteracts the effect of the Lorentz force.

In most real metals, the presence of Brillouin zone boundaries leads to deviations of the Fermi surfaces from spherical, and electrons on different parts of the Fermi surface have different relaxation times, destroying the exact balance and giving nonzero $\rho(\mathbf{B})$ and deviations of $R_H(\mathbf{B})$ from $-1/n_e ec$. At low fields ($\omega_c \tau \ll 1$), the transport coefficients are determined by details of the ways in which the electrons are scattered. Any simple model with two or more different relaxation times predicts that $\rho_{xx}(\mathbf{B})$ should initially increase from $\rho_{xx}(0)$ as B^2. Examples of $\rho(\mathbf{B})$ for a wide variety of polycrystalline samples are shown in Figure 4 for a transverse field ($B = -B_z$) and in Figure 5

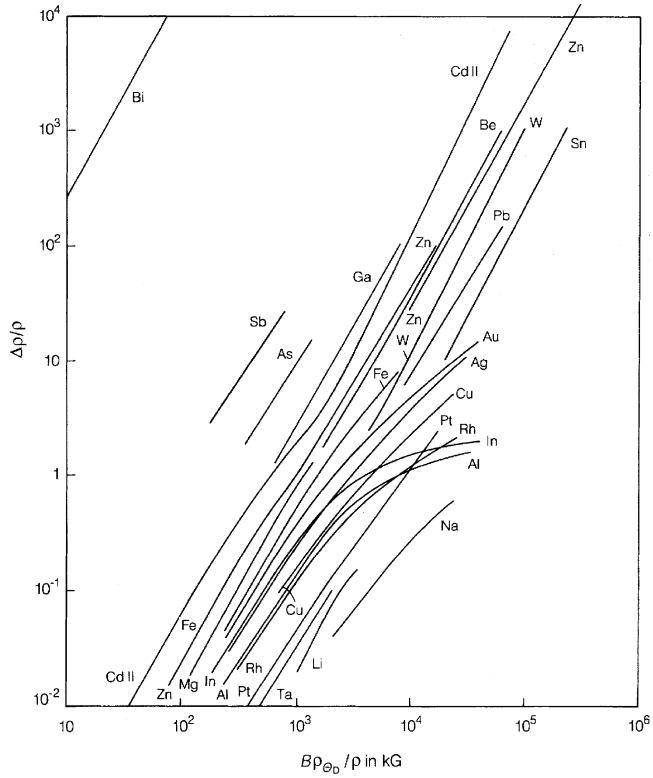


Figure 4. MR versus reduced transverse field $B\rho_\Theta/\rho_0$ for a wide variety of polycrystalline metals. ρ_0 is the residual resistivity at 4.2 K and ρ_Θ is the resistivity at the Debye temperature in the field. (From Olsen, 1962.)

for a longitudinal field ($B = B_x$). Usually $\rho(\mathbf{B})$ initially grows approximately as $\mathbf{B^2}$, but as $\mathbf{B}$ increases further, the behaviors differ for different metals. At high fields ($\omega_c \tau \gg 1$) the transport coefficients become dominated by Fermi surface topography, as explained below.

For several liquid metals and for the alkali metals, such as K, which are expected to have nearly spherical Fermi surfaces, R_0 lies close to the expected free electron values, as illustrated in Table 1 (for a more complete listing see Hurd, 1972). But in other metals, R_H often deviates from the free electron value—it can be positive and even vary with B. Figure 6 shows an interesting case: R_H for Al at 4.2 K (Forsvoll and Holwech, 1964). At low field, R_0 is negative and has the magnitude expected for a Fermi surface corresponding to a three-electron sphere. But as $\mathbf{B}$ increases, R_0 changes sign, finally "saturating" at high $\mathbf{B}$ at the value for the net single "hole" (positive charge) per atom expected for Al with a three-electron sphere embedded in a face-centered cubic Brillouin zone structure. The Hall coefficients of metals and alloys can also display significant temperature dependences (Hurd, 1972). An important additional complication occurs in magnetic metals, where R_H depends upon the sample magnetization M, leading to an "anomalous" Hall effect, where $\rho_H = R_0 B + \alpha R_s M$, where R_s is called the spontaneous Hall coefficient and the constant α depends upon the units used (O'Handley, 1972). Figure 7 schematically illustrates such behavior.

Table 1. Hall Coefficients of Some Solid and Liquid Metals

Metal	Solid (S) or Liquid (L)	Temperature (K)	Number of Valence e^-	R/R_{fe}	Reference
Na	L	371	1	0.98	Busch and Tieche, 1963
Cd	L	594–923	2	1.04	Enderby, 1963
Pb	L	603–823	4	0.88	Enderby, 1963
Sb	L	903–1253	5	1.14	Busch and Tieche, 1963
K	S	Low T, high $\mathbf{B}^a$	1	1.1	Ashcroft and Mermin, 1976
Ag	S	Low T, high $\mathbf{B}^a$	1	1.3	Ashcroft and Mermin, 1976
Al	S	Low T, high $\mathbf{B}^a$	3	−0.3	Ashcroft and Mermin, 1976

[a] Low temperature, high magnetic field.

At high fields, the dependences of $\rho(\mathbf{B})$ and $R_H(\mathbf{B})$ upon $\mathbf{B}$ should be qualitatively different depending on whether the metal is "compensated" ($n_e = n_h$) or "uncompensated" ($n_e \neq n_h$), where n_e and n_h are the electron and hole densities, and whether it contains open electron orbits in the plane perpendicular to $\mathbf{B}$ (Abrikosov, 1988). The expected behaviors are summarized in Table 2 and Figure 8. For an uncompensated metal with no open orbits in the plane perpendicular to $\mathbf{B}$, at high fields $\rho(\mathbf{B})$ should saturate in value and $R_H(\mathbf{B}) = R_0 B$ should become linear in B, with R_0 taking the simple value $R_0 = -1/(n_e - n_h)e$. In this case, the sign of R_0 specifies whether the dominant carriers are electrons (e) or holes (h). For a compensated metal, or one with open orbits in the plane perpendicular to $\mathbf{B}$, $\rho(\mathbf{B})$ can increase at high fields as B^2 and R_0 is not a constant. The metals for which $\rho(\mathbf{B})$ continues to increase as B^2 at high fields in Figure 4 are compensated. Figure 8 shows that at large enough fields high-purity metals display quantum oscillations. Thin wires or foils also display size effects (including oscillations) when B becomes so large that the cyclotron radius $r_c = v_F \omega_c$ (v_F is the Fermi velocity) becomes smaller than the wire or foil thickness (Olsen, 1962).

In magnetic materials, the quantities of interest can also vary either directly or indirectly with the magnetization $\mathbf{M}$. The anomalous Hall effect noted above is an example of a direct dependence on $\mathbf{M}$. Another is anisotropic magnetoresistance in a ferromagnetic film—where the resistance increases with B for $\mathbf{B}$ (and thus $\mathbf{M}$) parallel to $\mathbf{I}$ but decreases for $\mathbf{B}$ perpendicular to $\mathbf{I}$ (Bozorth, 1951). Giant (G) magnetoresistance (MR) in magnetic multilayers or granular magnetic materials is an example of an indirect dependence, where the MR varies with the relative orientations of the M_I for localized magnetic subcomponents of the sample (Levy, 1994). Magnetic materials also usually display hysteresis—the value of

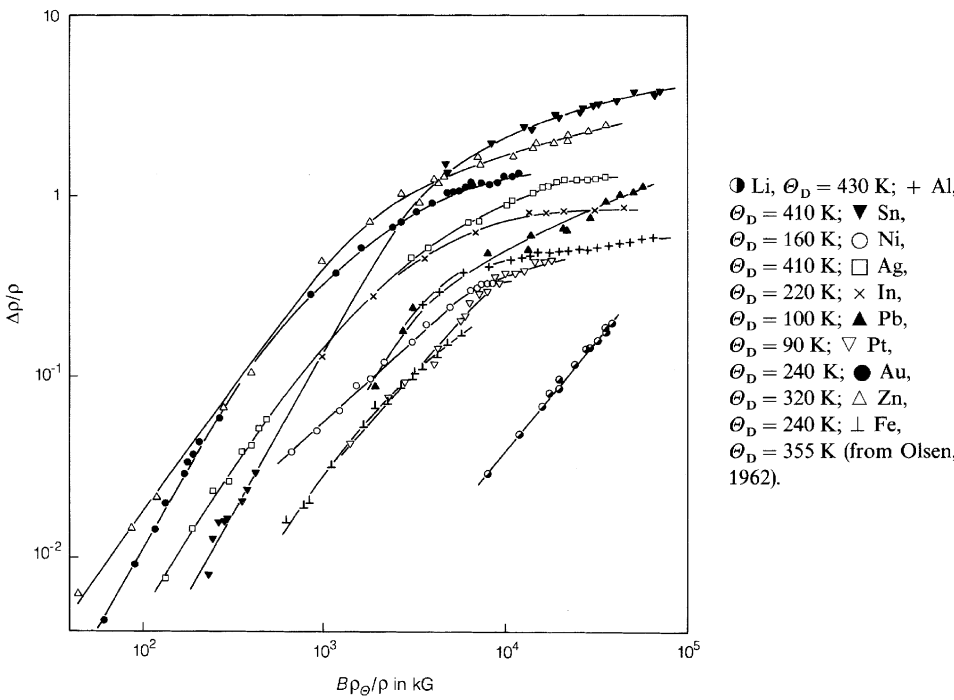


Figure 5. MR versus reduced longitudinal field $B\rho_\Theta/\rho_0$ for a wide variety of polycrystalline metals. ρ_0 is the resistivity at 4.2 K and ρ_Θ is the resistivity at the Debye temperature. The measurements were made in fields up to 200,000 G. (From Olsen, 1962.)

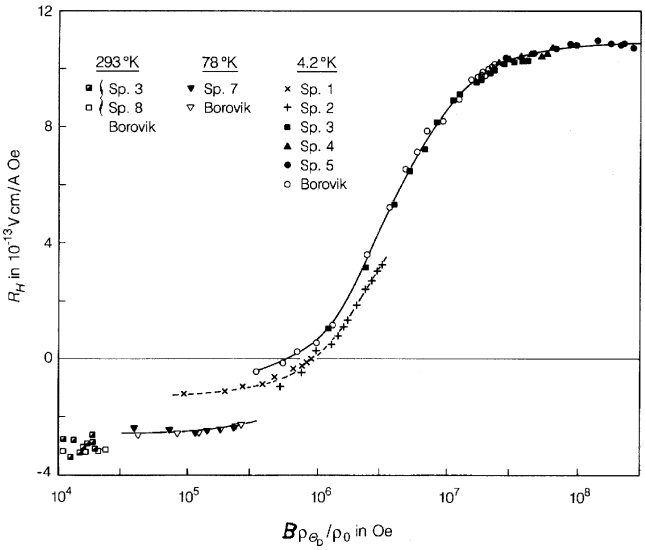


Figure 6. R_H versus reduced field $B\rho_\Theta/\rho_0$ at 4.2 K for Al and dilute Al alloys, showing an unusual variation due to real Fermi surface effects. ρ_0 is the residual resistivity at 4.2 K and ρ_Θ is the resistivity at the Debye temperature. Sp. 1 and Sp. 2 refer to dilute Al(Zn) alloys and the other symbols refer to pure Al samples of different thicknesses. (After Hurd, 1972, using the data of Forsvall and Holwech, 1964.)

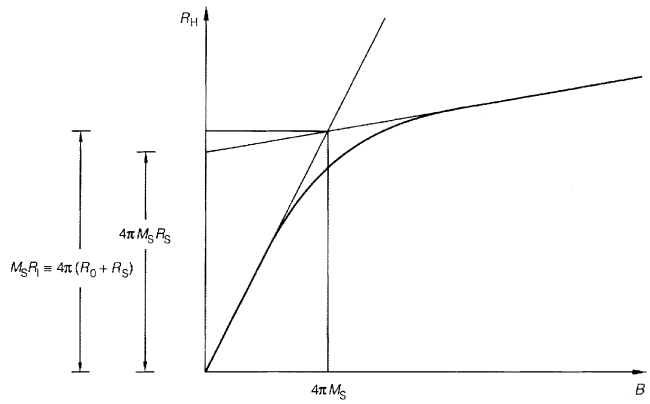


Figure 7. Schematic of R_H versus B for a magnetic metal illustrating the normal and anomalous Hall effects. (From Hurd, 1972.)

the transport coefficient depends upon the magnetic history of the sample (Bozorth, 1951). The Weidemann-Franz law is found to hold rather well for a magnetic granular AgCo alloy in a magnetic field (Piraux et al., 1993), and both granular AgCo and multilayer Co/Cu display significant changes in S with **B** (Piraux etal., 1993; Tsui et al., 1994). Figure 9 shows that the magnetothermopowers of Cu, Ag, and Au are small above 60 K, making them appropriate reference materials for measurements of $S(\mathbf{B})$ at such temperatures (Chiang, 1974).

METHOD AUTOMATION

The measured quantities in magnetotransport are just currents and voltages, which can be straightforwardly collected in a computer and converted to R, R_H, S, and K. In

Table 2. Types of Behavior of the Hall and Transverse Magnetoresistance Effects in Metals (from Hurd, 1972)

Type	Magnetoresistance[a]	Hall Coefficient	Nature of Orbits in Planes Normal to the Applied Field Eirection: State of Compensation[b]
Behavior in the high-field condition			
Single-crystal sample			
1	$\Delta\rho/\rho(0) \rightarrow$ saturation	$R \propto (n_e - n_h)^{-1}$	All closed; $N_e \neq N_h$
2	$\Delta\rho/\rho(0) \rightarrow B^2$ (transverse) $\Delta\rho/\rho(0) \rightarrow$ saturation (longitudinal)	$R \rightarrow$ const[c]	All closed; $N_e = N_h$
3	$\Delta\rho/\rho(0) \rightarrow$ saturation	$R \propto (n_e - n_h)^{-1}$	Negligible number of open orbits; $N_e \neq N_h$
4	$\Delta(0)\rho/\rho \rightarrow B^2 \sin^2\theta$[d]	$R \rightarrow$ const[c]	Open orbits in one direction only
5	$\Delta\rho/\rho(0) \rightarrow$ saturation	$R \propto B^{-2}$	Open orbits in more than one direction
Polycrystalline sample			
6	$\Delta\rho/\rho(0) \propto B$	$R \rightarrow$ const[c]	All types if the crystallites have random orientations
Behavior in the low-field conditions			
Single-crystal or polycrystalline sample			
7	$\Delta\rho/\rho(0) \propto B^2$	$R \rightarrow$ const[e]	It is irrelevant what type of orbit predominates

[a] Transverse magnetoresistance except as indicated for type 2.
[b] N_e and N_h are, respectively, the number of electrons and holes per unit cell of the Bravais lattice; n_e and n_h are, respectively, the density of electrons and holes in real space.
[c] The Hall coefficient is not related to the effective number of carriers in any simple manner.
[d] Note that $\Delta\rho/\rho(0) \rightarrow$ saturation when θ, the angle between the applied electric field and the axis of the open orbit in real space, is zero.
[e] The Hall coefficient depends upon the anisotrophy of the dominant electron-scattering process and the electron's velocity and effective mass at each point on the Fermi surface.

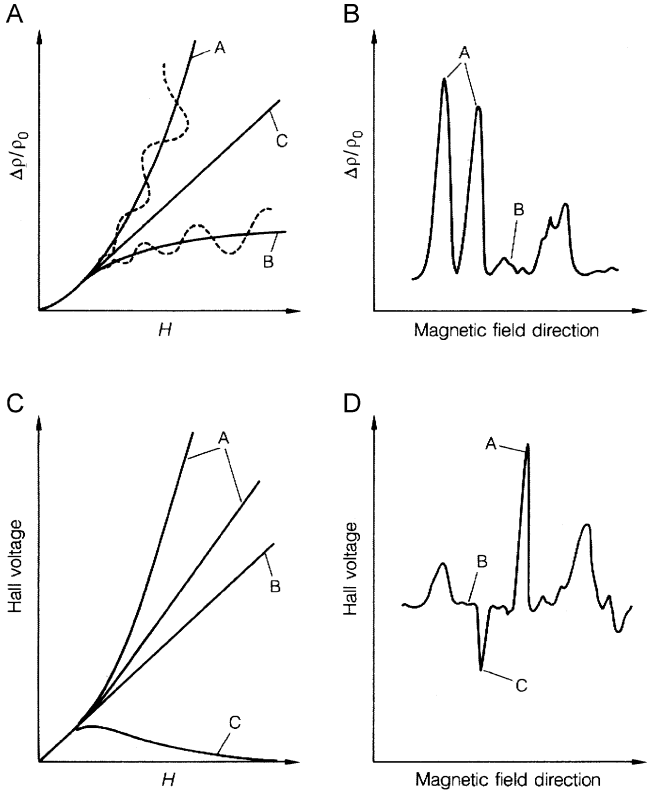


Figure 8. Types of field dependences observed experimentally for the transverse MR and the Hall resistance R_H (after Hurd, 1972). In parts (**A**) and (**B**), line A illustrates a quadratic (i.e., B^2) high-field variation, line B illustrates saturation, and line C illustrates a linear (i.e., B) variation, all of which are seen in polycrystalline samples (see Fig. 4). Quantum oscillations (dashed curves) may be superimposed upon behaviors A and C for single crystal samples. In parts **C** and **D**, line A illustrates high-field linear variations of different magnitudes, line B illustrates a linear variation extending all the way down to $B = 0$, and line C illustrates a B^{-1} high-field variation.

all cases, a separate current through the magnet that sets **B** determines the magnetic field. The ambient temperature is usually determined by a resistance thermometer involving one fixed current and a measured voltage. Determining $R(\mathbf{B})$ and $R_H(\mathbf{B})$ requires measuring the current sent through the sample and the voltage that it produces. Measuring $S(\mathbf{B})$ requires (1) the current sent through a heating resistor to produce the desired temperature difference, (2) the resulting thermoelectric voltage, and (3) either the voltage produced by a differential thermocouple or the voltages produced by two resistance thermometers through which fixed currents are passed. For $K(\mathbf{B})$ it is necessary to measure the current into the heating resistor and either the output voltage of a differential thermocouple or the voltages produced by two resistance thermometers.

DATA ANALYSIS AND INITIAL INTERPRETATION

For dc measurements, determining $R(\mathbf{B})$ and $R_H(\mathbf{B})$ requires averaging the voltages measured in four different configurations: forward and reversed current to eliminate spurious thermoelectric voltages and forward and reversed magnetic field to eliminate unwanted intermixing of $R(\mathbf{B})$ and $R_H(\mathbf{B})$. For ac measurements, only the field is reversed. For $S(\mathbf{B})$ and $K(\mathbf{B})$ the results for forward and reversed fields should be averaged. To ensure that no unwanted heating effects are present, one should check that $R(\mathbf{B})$, $R_H(\mathbf{B})$, $K(\mathbf{B})$, and $S(\mathbf{B})$ are the same for two different measures of electrical or heat currents.

Unless signals are so small that averaging techniques are necessary, once the steps just listed have been taken, analysis of $R(\mathbf{B})$, $R_H(\mathbf{B})$, $S(\mathbf{B})$, and $K(\mathbf{B})$ is straightforward, simply involving plotting the calculated quantity versus B at each temperature of interest.

SAMPLE PREPARATION

Figure 1 shows the preferred form of a foil or film for quantitatively determining $\rho(\mathbf{B})$ and $R_0(\mathbf{B})$. In practice: (1) a sample can also be in the form of a wire; (2) if only MR(**B**) is desired, the contacts can be made differently; and (3) alternative techniques for measuring R_H using the Corbino disc geometry, or Helicon waves, are also possible (Hurd, 1972). The sample form of Figure 1 can also be used in very high magnetic fields (>40 T), which can be reached only by pulsing the fields (see, e.g., Sdaq et al., 1993).

The conditions for measuring $R(\mathbf{B})$ and MR specify no temperature gradients along the length, width, or thickness of the sample. To facilitate this and to minimize sample distortion by the Lorentz force induced by the magnetic field, the sample is usually physically anchored to a thermally conducting but electrically insulating substrate and submerged in a constant-temperature fluid (at cryogenic temperatures, liquid helium; at higher temperatures, an inert gas such as helium or argon, or air if the sample and holder are chemically inert). To minimize heat flow in or out along the potential leads of Figure 1, these leads are thermally "heat sunk" to (but electrically insulated from) a metal binding post held at nearly the same temperature as the sample.

The sample form of Figure 1 can also be used to measure $K(\mathbf{B})$ and $S(\mathbf{B})$ by placing temperature sensors on the potential lead pads. For $S(\mathbf{B})$, the potential leads will then become the reference material B specified above. To minimize unwanted heat flow in or out of the sample, the sample wire or foil is suspended in vacuum (usually on an electrically insulating substrate) and, if the temperatures are high enough that radiation is significant, surrounded by a thermal shield. To measure either $K(\mathbf{B})$ or $S(\mathbf{B})$, a heat current $\dot{Q}$ is injected at one end of the sample to produce a temperature gradient. For $K(\mathbf{B}) = \dot{Q}/\Delta T$, all of the injected heat $\dot{Q}$ must flow through the sample and there must be no additional heat flow in or out. For $S(\mathbf{B}) = \Delta V_{AB}/\Delta T$, extraneous heat flows are less crucial, but it is important to keep the regions of the two thermocouple contacts at uniform temperature to avoid spurious "internal voltages" from these contact regions. The only major constraint occurs when measuring $K(\mathbf{B})$ for thin samples on a substrate; either the sample must be thick enough

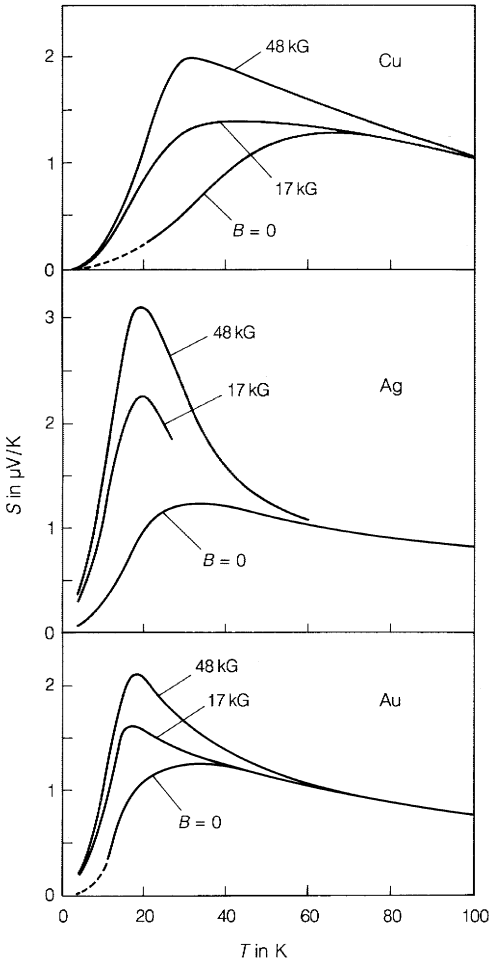


Figure 9. $-S(\mathbf{B})$ vs. T for Co, Ag, and Au at $B = 0$, 1.7 T, and 4.8 T. (From Chiang, 1974.)

so that it carries essentially all of the heat or the heat conduction of the substrate must be measured and corrected for.

PROBLEMS

Vibration of the sample or its leads in the presence of a magnetic field can induce unwanted voltages. Where possible, the sample should be anchored to a substrate. To minimize wire loops in which spurious currents may be induced, current and voltage lead pairs should be separately twisted together and, where possible, physically anchored.

The potential leads, and leads to any thermometers attached to the sample, can bring heat into or out of the sample. To minimize unwanted heat flow, such leads should be thin and heat sunk to bars maintained at a temperature close to that of the sample.

For both dc and ac currents, measurements with forward and reverse fields should be "averaged," defining $R = 1/2[R_H(+\mathbf{B}) + R_x(-\mathbf{B})]$ and $R_H = 1/2[R_H(+\mathbf{B}) - R_H(-\mathbf{B})]$.

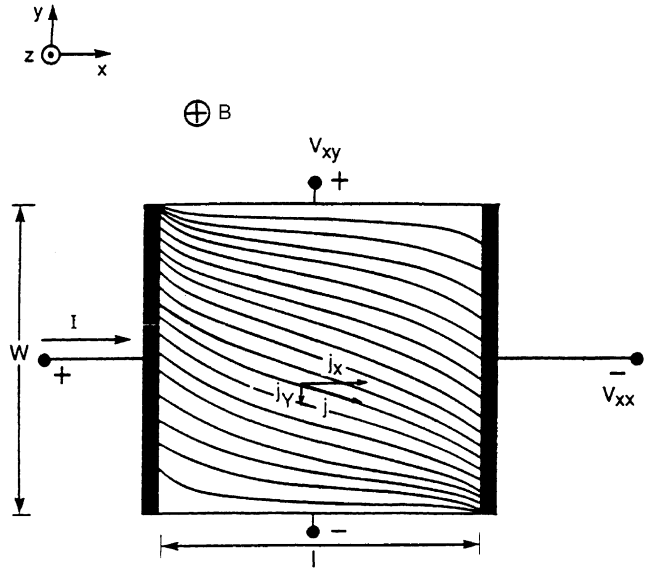


Figure 10. Cross-section of the distribution of current density j (curved lines) between two planar current electrodes (dark regions) for magnetic field $\mathbf{B}$ pointing into the plane of the page. V_{xx} is the longitudinal voltage and V_{xy} the transverse or Hall voltage. (After Heremens, 1993.)

Since R is expected to be unchanged by reversal of $\mathbf{B}$, whereas R_H is expected to reverse sign, reversal of $\mathbf{B}$ is used to define the experimental R and R_H, thereby correcting for any sample or lead misalignment that could intermix components of R_H and R_x. In practice, it is best to minimize such voltages by making the Hall voltage leads in Figure 1 as closely opposite to each other as possible.

Electrical and heat current uniformity through the sample is essential for high-accuracy determinations of $\rho(\mathbf{B})$, R_0, and $\kappa(\mathbf{B})$ (Hurd, 1972). To ensure uniformity to within 1%, the total sample length should be at least five times the sample width and at least three times the distance L between the voltage leads. For measuring $\rho(\mathbf{B})$ and R_0, the current contacts should extend over the entire sample width and should have low resistance compared to that of the sample. Care must be exercised in attaching voltage leads. Attaching leads of a different material directly to a wire or foil sample without the projections shown in Figure 1 can lead to disruption of uniform current flow in the vicinity of the contacts. Especially dangerous in such a case is use of superconducting contacts, which can short out the sample over a finite length. To minimize disrupting uniform current flow, it is usual for highest precision work to shape the sample so that its material projects beyond its main body, as shown in Figure 1. Contacts of a different metal, such as Cu, spot welded or soldered to the projections at least twice the probe width away from the sample body will then not signigicantly perturb the current distribution. For further details, see Hurd (1972).

Any temperature sensors that are located in the magnetic field should have either low or well-known B dependence.

A nonuniform magnetic field, or an inhomogeneous sample, can lead to nonuniform current densities that require solving Laplace's equation. Such complications are for the most part beyond the scope of the present unit, but two examples may be instructive. First, at low temperatures, a small groove on the surface of a high-purity film ($\omega_c\tau \gg 1$) will cause the current lines to deviate from straight paths, thereby inducing a Hall voltage along the direction of current flow and generating a linear magnetoresistance (Bruls et al., 1981). Second, a uniform **B** field can also produce a nonuniform current in a homogeneous sample that is not infinitely long and narrow. Figure 10 illustrates the current flow patterns that occur when a field **B** is applied perpendicular to a film of finite length L. When the film is short compared to its width W, the Hall effect is shorted out—the limiting version of such behavior is the Corbino disc geometry described by Hurd (1972). Only when the film is much longer than it is wide does the current flow become uniform over most of the sample. Conformal mapping techniques give the general relation for the resistance (Lippmann and Kuhrt, 1958):

$$R(B) = R(B=0)[1 + f(W/L)(\rho_{xy}/\rho_{xx})^2] \times \rho_{xx}(B)/\rho_{xx}(0) \tag{16}$$

where $f(W/L)$ is a known function of the ratio W/L. The first term ("1") in the brackets gives the usual field-dependent resistance and the second term gives the effect of the finite length L. In a free electron gas, ρ_{xx} should saturate and ρ_{xy} increase linearly with B. In Equation 16, $R(\mathbf{B})$ would then increase quadratically with B. A "short-wide" geometry is used to enhance the MR in semiconductor "Hall devices" (Heremens, 1993b).

ACKNOWLEDGMENT

This work was supported in part by U.S. National Science Foundation grant no. DMR-94-23795.

LITERATURE CITED

Abrikosov, A. A. 1988. Fundamentals of the Theory of Metals. North-Holland Publishing, Amsterdam.

Ashcroft, N. W. and Mermin, N. D. 1976. Solid State Physics. Saunders College, Pa.

Averback, R. S., Stephan, C. H., and Bass, J., 1973. Magnetic field dependence of the thermopower of dilute aluminum alloys. *J. Low Temp. Phys.* 12:319–346.

Barnard, R. D. 1972. Thermoelectricity in Metals and Alloys. Taylor and Francis, London.

Bass J. 1972. Deviations from Matthiessen's Rule. *Adv. Phys.* 21:431–604.

Bass, J. 1982. Electrical resistivity of pure metals and dilute alloys. *In* Metals: Electronic Transport Phenomena, Landolt-Bornstein New Series, Group III, Vol. 15a (K.-H. Hellwege and J. L. Olsen, eds.). pp. 1–288. Springer-Verlag, Berlin.

Blatt, F. J., Schroeder, P. A., Foiles, C. L., and Greig, D. 1976. Thermoelectric Power of Metals. Plenum, New York.

Bozorth, R. M. 1951. Ferromagnetism. Van Nostrand, New York.

Bruls, G. J. C. L., van Gelder A. P., van Kempen, H., Bass, J., and Wyder, P. 1981. Linear magnetoresistance caused by sample thickness variations. *Phys. Rev. Lett.* 46:553–556.

Busch, G. and Tieche, Y. 1963. *Phys. Kond. Mater.* 1:78–104.

Chaikin, P. M. and Kwak J. F. 1975. Apparatus for thermopower measurements on organic conductors. *Rev. Sci. Inst.* 46:218–220.

Chiang, C.-K. 1974. Thermopower of Metals in High Magnetic Fields. PhD thesis, Michigan State University.

Chien, C. L. and Westgate, C. R. 1980. The Hall Effect and Its Applications. Plenum, New York.

Crawford Dunlap, Jr., W. 1959. Conductivity measurements on solids. *In* Methods of Experimental Physics, Vol. 6B (K. Lark-Horovitz and V. A. Johnson, eds.). pp. 32–70. Academic Press, New York.

Enderby, J. E. 1963. *Proc. Phys. Soc. (London)* 81:772–779.

Fletcher, R., 1977. Righi-Leduc and Hall coefficients of the alkali metals. *Phys. Rev. B* 15:3602–3608.

Foiles, C. L. 1985. Thermopower of pure metals and dilute alloys. *In* Metals: Electronic Transport Phenomena, Landolt-Bornstein, New Series, Group III, Vol. 15b (K.-H. Hellwege and J. L. Olsen, eds.) pp. 48–206. Springer-Verlag, Berlin.

Forsvoll, K. and Holwech, I. 1964. Galvanomagnetic size effects in aluminum films. *Philos. Mag.* 9:435–450.

Frederikse, H. P. R., Johnson, V. A., and Scanlon, W. W. 1959. Thermoelectric effects. *In* Methods of Experimental Physics, Vol. 6B (K. Lark-Horovitz and V.A. Johnson, eds.). pp. 114–127. Academic Press, New York.

Fritzsche, H., 1959. The Hall Effect. *In* Methods of Experimental Physics, Vol. 6B (K. Lark-Horovitz and V. A. Johnson, eds.) pp. 145–159. Academic Press, New York.

Heremens, J. 1993a. Solid state magnetic field sensors and applications. *J. Phys. D: Appl. Phys.* 26:1149–1168.

Heremens, J. P. 1993b. Magnetic devices. *In* Encyclopedia of Applied Physics, Vol. 8 (G. Trigg, ed.). pp. 587–611. VCH Publishers, New York.

Hurd, C. M. 1972. The Hall Effect in Metals and Alloys. Plenum, New York.

Jackson, J. D., 1962. Classical Electrodynamics. John Wiley & Sons, New York.

Jan, J. P. 1957. Galvanomagnetic and thermomagnetic effects in metals. *In* Solid State Physics, Vol. 5 (F. Seitz and D. Turnbull, eds.). pp. 1–96. Academic Press, New York.

Kittel, C. 1986. Introduction to Solid State Physics, 6th ed. John Wiley & Sons, New York.

Levy, P. M., 1994. Giant magnetoresistance in magnetic layered and granular materials. *In* Solid State Physics Series, Vol. 47 (H. Ehrenreich and D. Turnbull, eds.). pp. 367–462. Academic Press, New York.

Lippmann, H. J. and Kuhrt, K. 1958. Geometrieeinfluss auf den Hall-effekt bei Halbleiterplatten. *Z. Naturforsch., A: Phys. Sci.* 13A:462–483.

Meaden, G. T. 1965. Electrical Resistance of Metals. Heywood Books, London.

Olsen, J. L. 1962. Electronic Transport in Metals. John Wiley & Sons, New York.

O'Handley, R. C. 1980. Hall effect formulae and units. *In* The Hall Effects and Its Applications (C. L. Chien and C. R. Westgate, eds.). pp. 417–419. Plenum Press, New York.

Pearlman, N. 1959. Thermal conductivity. *In* Methods of Experimental Physics, Vol. 6A (K. Lark-Horovitz and V. A. Johnson, eds.). pp. 385–406. Academic Press, New York.

Pearson, G. L. 1959. Magnetoresistance. *In* Methods of Experimental Physics, Vol. 6B (K. Lark-Horovitz and V. A. Johnson, eds.). pp. 160–165. Academic Press, New York.

Pernicone, J. and Schroeder, P. A. 1975. Temperature and magnetic field dependences of the electronic and lattice conductivities of tin from 1.3 to 6K. *Phys. Rev.* 11:588–599.

Piraux, L., Cassart, M., Jiang, J. S., Xiao, J. Q, and Chien, C. L. 1993. Magnetothermal transport properties of granular Co-Ag solids. *Phys. Rev. B* 48:638–641.

Roberts, R. B. 1977. The absolute scale of thermoelectricity. *Philos. Mag.* 36:91–107.

Roberts, R. B., Righini, F., and Compton, R.C., 1985. Absolute scale of thermoelectricity, III. *Philos. Mag.* B43, 1125.

Rossiter, P. L. 1987. The Electrical Resistivity of Metals and Alloys. Cambridge University Press, Cambridge.

Rossiter, P. L. and Bass, J. 1994. Conductivity in metals and alloys. *In* Encyclopedia of Applied Physics, Vol. 10 (G. Trigg, ed.). pp. 163–197. VCH Publishers, New York.

Scanlon, W. W. 1959. Other transverse galvanomagnetic and thermomagnetic effects. *In* Methods of Experimental Physics, Vol. 6B (K. Lark-Horovitz and V.A. Johnson, eds.). pp. 166–170. Academic Press, New York.

Sdaq, A., Broto, J. M., Rakoto, H., Ousset, J. C., and Raquet, B. 1993. Magnetic and transport properties of Ni/Ti, NiC/Ti, and Co/Cu multilayers at high fields. *J. Magn. Magn. Mater.* 121: 409–412.

Stone, E. L., III., Ewbank, M. D., Pratt, W. P., and Bass, J. 1976. Thermoelectric measurements on tungsten at ultra-low temperatures. *Phys. Lett.* 58A:239–241.

Thaler, B. J., Fletcher, R., and Bass, J., 1978. The low temperature, high field Nernst-Ettingshausen coefficient of Al. *J. Phys. F: Metal Phys.* 8:131–139.

Tsui, F., Chen, B., Barlett, D., Clarke, R., and Uher, C. 1994. Scaling behavior of giant magnetotransport in Co/Cu superlattices. *Phys. Rev. Lett.* 72:740–743.

Tye, R. P. 1969. Thermal Conductivity, Vol. 1. Academic Press, New York.

KEY REFERENCES

Abriksov, 1988. See above.

A good overview of theory.

Blatt et al., 1976. See above.

A comprehensive treatment of thermopower.

Hurd, 1972. See above.

A comprehensive treatment of the Hall effect.

Rossiter, 1987. See above.

A comprehensive treatment of resistivity.

JACK BASS
Michigan State University
East Lansing, Michigan

SURFACE MAGNETO-OPTIC KERR EFFECT

INTRODUCTION

In 1845, Michael Faraday found that the polarization plane of linearly polarized light rotates as the light is transmitted through a magnetized medium. Thirty-two years later the magneto-optic Kerr effect (MOKE) was discovered by John Kerr when he examined the polarization of light reflected from the polished surface of an electromagnet. These two monumental works heralded the beginnings of magneto-optics and form the foundation of its modern utilization (Dillon, 1971). As the field of surface and thin-film magnetism has emerged and blossomed in recent years (Falicov et al., 1990), so has the need for innovative approaches to study magnetic phenomena on the nanometer-thickness scale. The application of the Kerr effect to study the surface magnetism was introduced by Moog and Bader (1985) along with the acronym SMOKE to denote the surface magneto-optic Kerr effect. These researchers studied atomic-layer film thicknesses of iron grown epitaxially on Au(100). Magnetic hysteresis loops were obtained with monolayer sensitivity. Since then SMOKE has emerged as a premier surface magnetism technique of choice in many laboratories worldwide. The broad acceptance of the SMOKE technique stems from its simplicity and its ability to generate the "universal currency" in magnetism—the hysteresis loop. In addition, there are almost no materials limitations to this technique, as long as the sample surface is smooth enough to generate optical reflection. However, the magnitude of the SMOKE signal depends on the materials properties and on the optical wavelength. Although in this unit we describe the SMOKE technique for visible light with fixed wavelength, it is easy to extend the technique to wavelength-dependent measurements such as magneto-optic spectroscopy and spatially resolved measurements such as magneto-optic microscopy. Nevertheless, SMOKE has been applied successfully to address various contemporary topics in low-dimensional magnetism (Bader, 1991). The aim of this unit is to provide general background about the basic principles and experimental methods of the SMOKE technique. While much of the discussion is directed to nonspecialists, it is structured with the inclusion of mathematical exposition to describe magneto-optics of magnetic multilayers.

There are many challenges left to overcome in the quest to provide a complete magnetic characterization of ultrathin-film structures. Significant progress has been made recently to develop second-harmonic-generation (SHG) MOKE to distinguish the response of the surface and/or interfacial layer from the interior-layer or bulk response. A future direction could involve the combination of SMOKE with other techniques to enhance both spatial and time resolution so that small-scale processes, such as domain wall dynamics, can be investigated. The purpose of this unit is to provide background on the SMOKE technique with an emphasis on magnetic multilayers.

As described in the next section, the magneto-optic effect originates from the spin-orbit interaction. Therefore SMOKE should be regarded as one of many possible versions of this interaction. Others include the Faraday effect

technique, which is usually applied to optically transparent materials, and core-level magnetic dichroism (see X-RAY MAGNETIC CIRCULAR DICHROISM), which is an element-specific measurement.

PRINCIPLES OF THE METHOD

Phenomenological Explanation of the Origins of Magneto-Optic Effects

Maxwell (1873) expressed linear polarized light as a superposition of two circular polarized components. He realized that the Faraday effect is a consequence of different propagating velocities of the two circular modes. Thus, it is the difference in the dielectric constants for left- and right-circular polarized light that account for the Faraday rotation. This explanation remains the phenomenological description of the Faraday effect still presented in introductory physics textbooks. The dielectric property of a medium is characterized by a 3×3 tensor, ε_{ij}, with $i, j = 1, 2, 3$. In general, this dielectric tensor can be decomposed into symmetric and antisymmetric parts. The symmetric part can be diagonalized by an appropriate rotation of the coordinates. If the three eigenvalues are the same, the medium is isotropic, and the dielectric tensor is reduced to a dielectric constant. Otherwise, the medium is anisotropic. Nevertheless, the normal modes of the symmetric tensor are linearly polarized light along the three principal axes. Therefore, the symmetric part of the dielectric tensor does not give rise to the Faraday effect. Since the symmetric part of ε_{ij} is unimportant to the Faraday effect, we will always assume that it is isotropic.

To see the effect of the antisymmetric part of ε_{ij}, consider the dielectric tensor

$$\tilde{\varepsilon} = \varepsilon \begin{pmatrix} 1 & iQ_z & -iQ_y \\ -iQ_z & 1 & iQ_x \\ iQ_y & -iQ_x & 1 \end{pmatrix} \tag{1}$$

The two normal modes are the left- (L) circular polarized light ($E_y = iE_x$) with $\varepsilon_{\mathrm{L}} = \varepsilon(1 - \mathbf{Q} \cdot \hat{\mathbf{k}})$ and right- (R) circular polarized light ($E_y = -iE_x$) with $\varepsilon_{\mathrm{R}} = \varepsilon(1 + \mathbf{Q} \cdot \hat{\mathbf{k}})$ where $\mathbf{Q} = (Q_x, Q_y, Q_z)$ is known as the Voigt (1908) vector and $\hat{k}$ is the unit vector along the light propagation direction. Thus, it is the antisymmetric elements, $\mathbf{Q} = (Q_x, Q_y, Q_z)$, that produce the difference in the refraction indices between the two circularly polarized modes and the Faraday effect originates from the off-diagonal antisymmetric elements of the dielectric tensor. The complex Faraday rotation of the polarization plane after traveling a distance L within a magnetized medium is

$$\theta = \frac{\pi L}{\lambda}(n_{\mathrm{L}} - n_{\mathrm{R}}) = -\frac{\pi L n}{\lambda}\mathbf{Q} \cdot \hat{\mathbf{K}} \tag{2}$$

where n is the refractive index. The real part of the above formula gives Faraday rotation, and the imaginary part gives the Faraday ellipticity.

It is instructive to consider why an external magnetic field has a stronger effect on the polarization plane of light than an external electric field. This question can be addressed with a simple argument based on consideration of time-reversal symmetry. Under the time-reversal process, the electric field remains unchanged, while the magnetic field undergoes a change of sign. Thus, Onsager's relation (Landau and Lifshitz, 1960) gives $\varepsilon_{ij}(\mathbf{E}, \mathbf{H}) = \varepsilon_{ij}(\mathbf{E}, -\mathbf{H})$. Here $\mathbf{E}$ and $\mathbf{H}$ are the external electric and magnetic fields, respectively. By expanding ε_{ij} up to terms that are linear in the electric and magnetic fields, it is obvious that the antisymmetric part of ε_{ij} comes from the magnetic field. The magnetic field here provides an example of a special case of time-reversal symmetry breaking. In general, any quantity that breaks the time-reversal symmetry should generate antisymmetric elements of the dielectric tensor and, thus, a Faraday effect. This is why, for example, magneto-optic experiments were utilized in research on high-temperature superconductors to search for the hypothesized existence of anions, which are believed to break time-reversal symmetry (Kiefl et al., 1990; Wilczek, 1991).

Classical Explanation

The dielectric properties of a medium are determined by electron motion in response to an electromagnetic wave. Thus, an analysis of electron motion should offer a microscopic explanation of the magneto-optic effect. In the classical picture, each electron inside a medium is bonded to an atom (with nature frequency ω_0) and dissipates energy with a relaxation time τ. Consider an electromagnetic wave propagating in a medium with an external magnetic field H applied in the z direction. The equation of motion for an electron in the medium is then

$$\frac{d^2\mathbf{r}}{dt^2} + m\omega_0^2\mathbf{r} + \frac{m}{\tau}\frac{d\mathbf{r}}{dt} = -e\mathbf{E} - \frac{e}{c}\frac{d\mathbf{r}}{dt} \times H\hat{\mathbf{z}} \tag{3}$$

Here $\mathbf{E}$, H, c, t, $\mathbf{r}$, m, and e are the electrical field, magnetic field, speed of light, time, position of the electron, mass of the electron, and charge of the electron, respectively, and $\hat{\mathbf{z}}$ is a unit vector in the z-direction. Without the magnetic field present, a circular polarized electric field will drive the electron into circular motion with the same radius for both left- and right-circular polarized waves. Since the electric dipole moment is proportional to the radius of the circular orbit, there is no difference between the dielectric constants for left- and right-circular polarized waves, and, thus, there is no Faraday rotation. However, with a magnetic field applied in the propagation direction of the electromagnetic wave, there is an additional Lorentz force acting on each electron. This Lorentz force is directed toward or away from the center of the circle for left- or right-circular motions and thus shrinks or expands the radius of the electron orbit. The difference in the resultant radii of the left- and right-circular polarized modes yields the difference in the corresponding dielectric constants. Thus, it is the Lorentz force of the magnetic field that generates the Faraday effect.

From the above analysis, we see that the difference in the dielectric constants for the two circularly polarized modes is proportional to the strength of the Lorentz force, which itself is proportional to the magnetic field and the

frequency. Thus, it becomes apparent from Equation 2 that the Faraday rotation is directly proportional to the magnetic field and inversely proportional to the square of the wavelength λ of the light. Indeed, under the conditions $\omega_0 \gg \omega \ \omega_c$ ($\omega_c = eH/mc$ is the cyclotron frequency) and $\omega\tau \gg 1$, which hold for visible light, the Faraday rotation is given by

$$\theta = \frac{2\pi n e^3 \omega^2 LH}{m^2 c^2 \omega_0^4 \sqrt{1 + 4\pi n e^2/m\omega_0^2}} \qquad (4)$$

where n is the electron density. The rotation angle is proportional to the sample length L, the magnetic field H, and the square of the light frequency ω. This result conforms with experimental observations.

Quantum Description of the Magneto-Optic Effect in a Ferromagnet

The development of a quantum description of the magneto-optic effect was motivated by the particularly strong Faraday effect in ferromagnetic materials. To explain the magnitude of the Faraday rotation in a ferromagnet, which is of order $\sim 10^5$ deg/cm for magnetization parallel to the light propagation direction, an effective magnetic field of $\sim 10^7$ Oe has to be invoked. This value is of the order of the Weiss field that was postulated to account for the existence of ferromagnetism. The Weiss field was explained by Heisenberg to originate from the exchange interaction among the electrons. Heisenberg's exchange interaction correctly revealed the origin of magnetism as an effective magnetic field that aligns the individual spins. However, the problem remained that this field could not explain the Faraday effect, since it is not coupled to the electron motion, which determines the dielectric properties of a material. This difficulty was solved by Hulme (1932), who pointed out that the spin-orbit interaction couples the electron spin to its motion. Thus, the Faraday rotation in a ferromagnet could be understood.

The spin-orbit interaction $\sim(\nabla V \times \mathbf{p}) \cdot \mathbf{s}$ results from the interaction of the electron spin with the magnetic field it experiences as the electron moves through the electric field $-\nabla V$ inside a medium with momentum $\mathbf{p}$. This interaction couples the electron spin with its motion and thus connects the magnetic and optical properties of a ferromagnet. Note that the effect of a magnetic field on the electron motion manifests itself as $H_{\text{int}} \sim -(e/mc)\,\mathbf{p} \cdot \mathbf{A}$. Thus, to a certain extent, the spin-orbit interaction $\sim(\nabla V \times \mathbf{p}) \cdot \mathbf{s} = \mathbf{p} \cdot (\mathbf{s} \times \nabla V)$ can be thought of as an effective magnetic field vector potential $\mathbf{A} \sim \mathbf{s} \times \nabla V$ acting on the motion of the electron. For nonmagnetic materials, although a spin-orbit interaction is present, this effect is not strong. This is because the equal number of spin-up and spin-down electrons cancels the net effect. For ferromagnetic materials, however, the effect is pronounced because of the unbalanced electron spin populations.

Hulme calculated the two (right- and left-polarized) refraction indices. He used the Heisenberg model of a ferromagnet and the Kramers-Heisenberg dispersion formula, which gives the refraction index in terms of the eigenenergy and the matrix elements of the dipole moment operator with respect to the eigenfunctions of the system. He accounted for the difference of the two refraction indices by the energy splitting due to the spin-orbit interaction. He neglected, however, the change of the wavefunction due to the spin-orbit interaction. This theory is unsatisfying because the quenching of orbital angular momentum in transition metal ferromagnets gives no energy splitting. Kittel (1951) showed that it is the change of the wave functions due to the spin-orbit interaction that gives rise to a right order of magnitude of the difference of the two refraction indices. Argyres (1955) later gave a full derivation of the magneto-optic effect in a ferromagnet using perturbation theory. Subsequent works were performed to calculate the magneto-optic effect in different regimes (Shen, 1964; Bennet and Stern, 1965; Erskine and Stern, 1975).

Formalism of Magneto-Optic Kerr Effect in Multilayers

Since metallic magnetic materials absorb light strongly, it is more useful to measure the reflected than the transmitted signal to probe the magneto-optic effect. Thus, we concentrate on the Kerr effect in this section. The formalism, however, can readily be extended to include the Faraday effect. For a given magnetic multilayer system, the refraction tensor for each layer can be expressed by a 3×3 matrix. The goal is to calculate the final reflectivity along different polarization directions. The general method is to apply Maxwell's equations to the multilayer structure and to satisfy the boundary conditions at each interface. Zak et al. (1990, 1991) developed a general expression for the Kerr signal based on this method. The essential ingredients of this theory are two matrices that relate the electric fields at each boundary. These matrices are described in Appendix A, and the special case of the ultrathin limit is described in Appendix B.

PRACTICAL ASPECTS OF THE METHOD

There are three important Kerr configurations: polar, longitudinal, and transverse. The polar Kerr effect has a magnetic field applied in the normal direction to the film plane (Fig. 1A) and thus is sensitive to the perpendicular component of the magnetization. The longitudinal case has a magnetic field applied in the film plane and in the plane of the incident light (Fig. 1B) and thus should be sensitive to the in-plane component of the magnetization. The transverse Kerr effect also has a magnetic field applied in the film plane, but perpendicular to the incident plane of the light (Fig. 1C). The polar and longitudinal Kerr effects are linear in Q and yield complex rotation ϕ of the polarization of the light. The polar signal is typically an order of magnitude greater than the longitudinal signal due to the optical factors. The transverse Kerr effect requires a second-order expansion in Q and manifests itself by a small reflectivity change for p-polarized incident light. The criteria for choosing an experimental angle of incidence are governed by weighing practical geometric constraints of the experimental chamber against the optimal angles to maximize the signal (Zak et al., 1991).

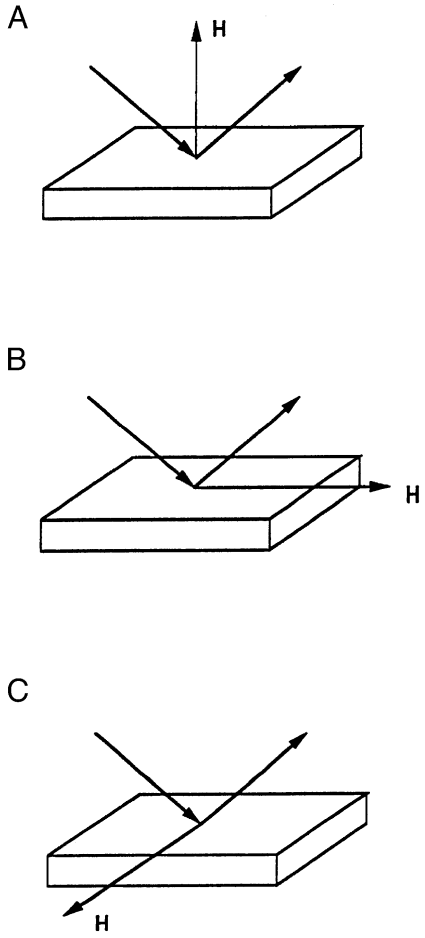


Figure 1. Three common configurations for the SMOKE measurements: (**A**) polar, (**B**) longitudinal, and (**C**) transverse Kerr effect.

An experimental SMOKE setup has the great advantage of simplicity (see Fig. 2). Typically a laser is used as a fixed-wavelength light source. The goal here would be to characterize the magnetic response of ultrathin layered systems by means of magneto-optics. However, if the goal were to extract magneto-optic properties per se, then the laser could be replaced by a lamp and monochromator in order to generate wavelength dependencies. (Synchrotron light sources can be used as well and are typical for magnetic dichroism studies in the x-ray regime; see MAGNETIC X-RAY SCATTERING.) Typically a low-power (few-milliwatt) laser suffices. It is highly desirable to use an intensity-stabilized laser, especially for monolayer studies. Crystal prism polarizers are useful both for defining the polarization and as an analyzer in front of the photodiode detector. Sheet polarizers can be used, but they have a lower extinction ratio when crossed than prism polarizers and so are not optimal for monolayer studies. Figure 2 outlines a straightforward DC detection scheme, although AC methods can be adopted. In that case typically the incident polarization is varied using a photoelastic modulator (Kliger et al., 1990), and the output of the photodiode detector is fed to a lock-in amplifier. Another common AC approach in optics involves chopping the incident beam in order to

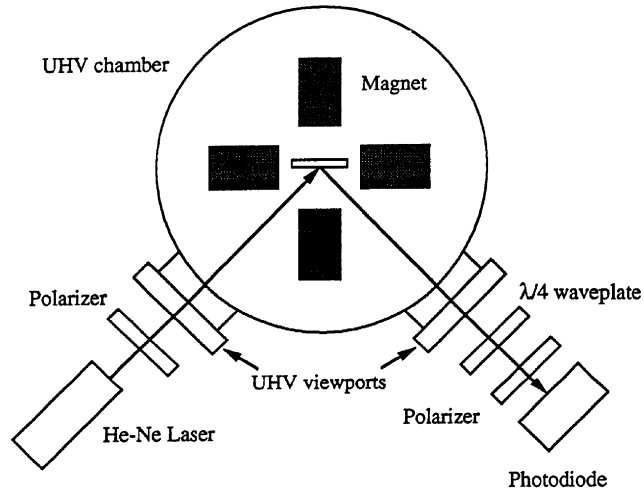


Figure 2. Schematic of the experimental setup of the SMOKE technique.

improve the signal-to-noise ratio. Finally, the magnet shown in Figure 2 consists of two split-coil solenoid pairs. Energizing either pair would generate a field in the film plane or perpendicular to it for longitudinal or polar measurements, respectively. Conventional or superconducting solenoids can be used depending on the field requirements of the experiment.

A linear p-polarized laser will serve as the light source for the purpose of our discussion (although s-polarized light could be used as well). After reflection from the sample surface, the light intensity is detected by a photodiode with a linear (analyzing) polarizer in front of it. The analyzing polarizer is set at a small angle δ ($\sim 1^\circ$ to 2°) from the extinction condition to provide a DC bias. Then the reflection intensity as a function of the externally applied magnetic field will generate the magnetic hysteresis loop (Fig. 2). Figure 3 shows an example of the hysteresis loop

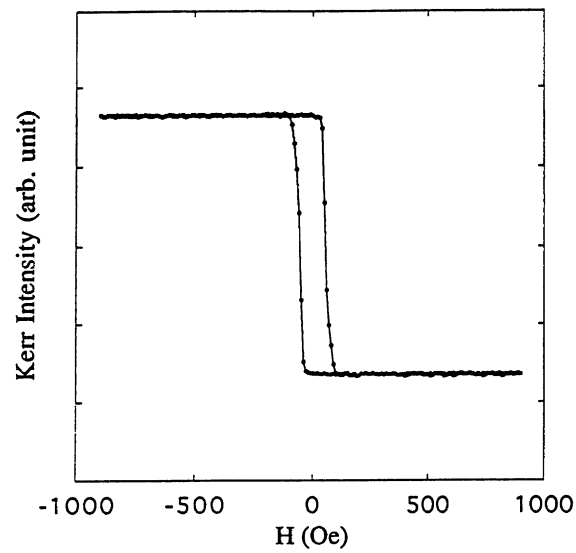


Figure 3. A SMOKE hysteresis loop taken from a 6-ML Fe film grown on Ag(100) substrate.

measured by SMOKE from a 6-ML Fe/Ag(100) film, where ML denotes monolayer. The figure illustrates that monolayer sensitivity is readily achieved in the SMOKE technique.

METHOD AUTOMATION

Standard spectroscopic automation can be used in the data collection by means of a laboratory computer. The computer would govern the sweep of the magnetic field by driving the bipolar power supply of the magnet via a voltage ramp. The computer would also simultaneously store the output of the photodiode detector by either digitizing an analog voltage output signal or by transferring the digital signal directly read by the detector electronics. Signal averaging can be performed by sweeping the field multiple times. A key feature of the SMOKE technique is that the optical elements are fixed and are not adjusted manually or under computer control.

DATA ANALYSIS AND INITIAL INTERPRETATION

To provide a quantitative analysis, the Kerr intensity for p-polarized incident light measured by the photodiode after the light has passed through the analyzing polarizer is

$$I = |E_p \sin\delta + E_s \cos\delta|^2 \approx |E_p \delta E_s|^2 \tag{5}$$

where $E_s/E_p = \phi' + i\phi''$ gives the Kerr rotation ϕ' and the Kerr ellipticity ϕ''. Then Equation 5 becomes

$$I = |E_p|^2|\delta + \phi' + i\phi''|^2 \approx |E_p|^2(\delta^2 + 2\delta\phi') = I_0\left(1 + \frac{2\phi'}{\delta}\right) \tag{6}$$

with

$$I_0 = |E_p|^2\delta^2 \tag{7}$$

as the intensity at zero Kerr rotation. Thus, the measured intensity as a function of the applied field yields a magnetic hysteresis loop. The maximum Kerr rotation ϕ'_m can be determined by the relative change of the Kerr intensity, $\Delta I/I_0$, upon reversing the saturation magnetization:

$$\phi'_m = \frac{\delta}{4}\frac{\Delta I}{I_0} \tag{8}$$

For *in situ* measurements an experimental difficulty that can mislead a nonexpert is that the ultrahigh-vacuum (UHV) viewport window (w) usually produces a birefringence, $\phi'_w + i\phi''_w$, that prevents the extinction from being realized. In this situation, a quarter-waveplate is usually placed before the analyzing polarizer to cancel the birefringence of the UHV viewport. The extinction condition can then be achieved if the principal axis of the quarter-waveplate makes an angle ϕ'_w to the p axis and if the polarization axis of the analyzing polarizer makes an angle $\pi/2 + \phi'_w - \phi''_w$ to the p axis. Then measured Kerr intensity is

$$I = |E_p|^2(\delta^2 + 2\delta\phi'') = I_0\left(1\frac{2\phi''}{\delta}\right) \tag{9}$$

Note that in this case the relative Kerr intensity determines the Kerr ellipticity rather than the rotation. The reason is that the quarter-waveplate produces a $\pi/2$ phase difference between the p and s components, so the analyzing polarizer "sees" $i(\phi' + i\phi'') = -\phi'' + i\phi'$ instead of $\phi' + i\phi''$. Thus, the quarter-wave plate interchanges the Kerr rotation with ellipticity. To measure the Kerr rotation, a half-waveplate can be used to replace the quarter-waveplate.

Experimental Result on Co/Cu Superlattices

It is instructive to provide an example of magneto-optic results in a magnetic multilayer system. Measurements on Co-Cu multilayers are presented for this purpose. Single crystals of both Co overlayers and Co/Cu superlattices were prepared by means of molecular beam epitaxy (MBE) using Cu(100) and Cu(111) single crystals as substrates and growth at room temperature. Detailed descriptions of the sample preparation appear elsewhere (Qiu et al., 1992). To provide a sense of the film quality in the present case, Figure 4 shows the intensity of the reflection high-energy electron diffraction (RHEED) during the growth of a Co/Cu(100) superlattice. More than 200 RHEED oscillations were observed during the growth of the entire superlattice, although the figure only shows results for initial and final bilayers. This is an indication of high-quality growth; each oscillation in intensity identifies the growth of an atomic monolayer.

The Kerr ellipticities of the multilayers and overlayers were measured *in situ* after each bilayer or Co dose was deposited. The results are plotted in Figure 5. Ellipticity data also are plotted, for comparison purposes, for Co deposited on a polycrystalline Cu substrate. Concentrating

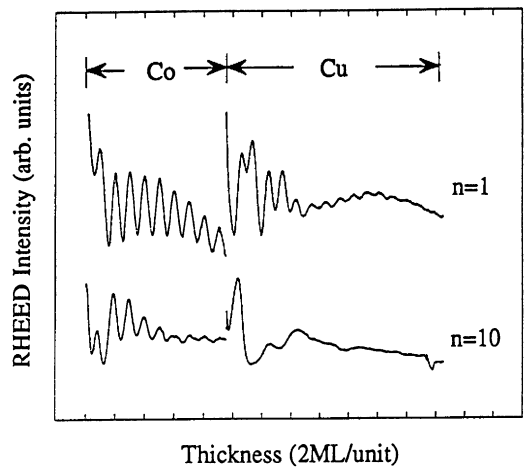


Figure 4. RHEED oscillations taken during the growth of a $[Co(9.5\ ML)/Cu(16\ ML)]_n$ superlattice on a Cu(100) substrate.

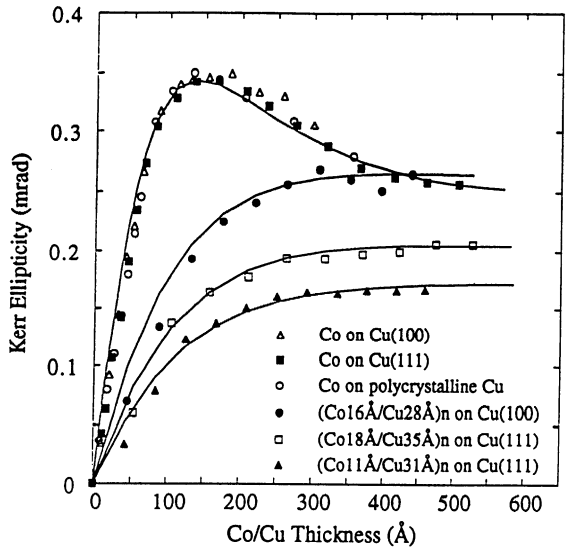


Figure 5. Kerr ellipticity measured for different samples. The solid lines are theoretical calculations.

first on the magneto-optic behavior of the Co overlayers on the Cu substrates, a linear increase in the ellipticity is observed in the ultrathin regime, reaching a maximum at ~120 Å of Co and finally dropping to a constant value for >400 Å thickness of Co. The initial rise is due to the increasing amount of Co present within the penetration depth of the light. In the thick regime (>400 Å of Co) the signal approaches a constant value since the absorption of light limits the depth sensitivity. In the intermediate regime the maximum in the ellipticity at ~120 Å of Co is due to the reflectivity changing from being dominated by Cu to being dominated by Co. Similar behavior is observed in the Fe/Au system (Moog et al., 1990). It is interesting to note that the ellipticity is independent of crystal orientation in the thickness range studied. In this example we can see the difference between MOKE and SMOKE. Traditionally, MOKE is defined as being independent of thickness, as opposed to the Faraday effect, which is proportional to thickness. For film thicknesses >400 Å we are in the traditional MOKE regime. But the initial linear rise of the magneto-optic signal is a characteristic of the SMOKE regime, which is encountered in the ultrathin limit.

For a quantitative analysis, we applied the formalism derived in the Appendix to simulate the results. The refractive indices we used are obtained from tabulations (Weaver, 1999) in the literature: $n_{Cu} = 0.249 + 3.41i$ and $n_{Co} = 2.25 + 4.07i$. We left the values of Q_1 and Q_2, where $Q = Q_1 + iQ_2$, for Co as free parameters to best fit the experimental curves and obtained $Q_1 = 0.043$ and $Q_2 = 0.007$. The calculated curves, depicted as solid lines in Figure 5, are in good overall agreement with the experimental data. In particular, the peaked behavior of the overlayers is faithfully reproduced.

The ellipticities of three epitaxial Co/Cu superlattices also appear in Figure 5. The superlattices are [Co(16 Å)/Cu(28 Å)]$_n$, grown on Cu(100), and [Co(11 Å)/Cu(31 Å)]$_n$ and [Co(18 Å)/Cu(35 Å)]$_n$, both grown on Cu(111). Their ellipticities in Figure 5 appear as a function of the total superlattice thickness. The superlattice ellipticities initially increase linearly in the ultrathin region and saturate in the thick regime, although there is no maximum at intermediate thicknesses. The lack of a maximum is readily understood because the reflectivity is not evolving from that of Cu to that of Co, as in the overlayer cases above. Instead, the reflectivity remains at an average value between the two limits, since both Co and Cu remain within the penetration depth of the light, no matter how thick the superlattice becomes. Using the Q value obtained from the Co overlayers, the Kerr ellipticities for the superlattices were calculated and plotted in Figure 5. The agreement with experimental data is obvious and is discussed further below.

To test the additivity law applicable to the ultrathin regime, the experimental data in Figure 5 were replotted in Figure 6 as a function of the magnetic Co layer thickness only (the Cu layers are ignored). In the ultrathin regime all the data lie on a single straight line. This result provides a demonstration and confirmation of the additivity law that states that the total Kerr signal in the ultrathin regime is a summation of the Kerr signal from each individual magnetic layer and is independent of the nonmagnetic spacer layers.

Despite the good overall semiquantitative agreement, the calculated ellipticity can be seen, upon close inspection, to exceed the experimental values in the ultrathin regime. For example, the calculated linear slope is 6.6 μrad/Å, while the experimental result yields only 4.3 μrad/Å. This systematic deviation can be due, for instance, either to the breakdown of the macroscopic description in the ultrathin region or to optical parameters (n and Q) that deviate from their bulk values.

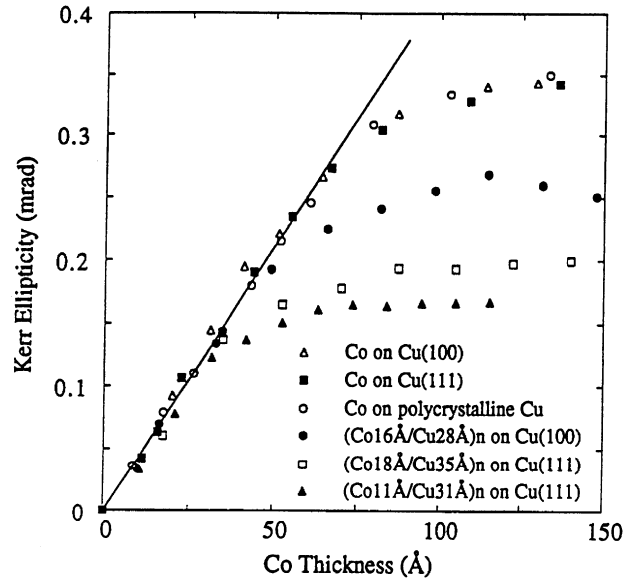


Figure 6. The additivity law shows that the Kerr signal in the ultrathin regime depends on the magnetic layer only.

SAMPLE PREPARATION

Sample preparation is governed by the same considerations as in other areas of surface science and thin-film growth: Vacuum conditions are maintained in the UHV realm, Auger spectroscopy (AUGER ELECTRON SPECTROSCOPY) is usually used to monitor surface cleanliness, and electron diffraction is used to monitor structural integrity. The types of evaporators might range from electron beam to thermal evaporators and are only limited by the creativity of the fabricator and the ability to control and monitor flux rates of different chemical species.

PROBLEMS

Several factors need to be considered in the evaluation of technical problems associated with the SMOKE technique. Experimentally, instability of the laser intensity is probably the main contributor to noise in the SMOKE signal. It is highly recommended to use intensity-stabilized lasers and/or lock-in amplifiers for the SMOKE measurements. A certain degree of vibration isolation is also important as in any optical measurement. As for the interpretation of SMOKE signals, one has to keep in mind that although the SMOKE rotation or ellipticity is proportional to the magnetization, the proportionality coefficient depends on the optical properties of the materials and is usually an unknown quantity for ultrathin films. Therefore, SMOKE is not a form of magnetometry that can determine the absolute value of the magnetic moment. This is an important fact, especially in studies of spin dynamics, because the SMOKE dynamics will not necessarily reflect the spin dynamics.

ACKNOWLEDGMENT

Work supported by the U.S. Department of Energy, Basic Energy Sciences-Materials Sciences, under contract DE-AC03-76SF00098 (at Berkeley) and W-31-109-ENG-38 (at Argonne).

LITERATURE CITED

Argyres, P. N. 1955. Theory of the Faraday and Kerr effects in ferromagnetics. *Phys. Rev.* 97:334–345.

Bader, S. D. 1991. SMOKE. *J. Magn. Magn. Mater.* 100:440–454.

Bennet, H. S. and Stern, E. A. 1965. Faraday effect in solids. *Phys. Rev.* 137:A448–A461.

Dillon, J. F., Jr. 1971. Magneto-optical properties of magnetic crystals. *In* Magnetic Properties of Materials (J. Smith, ed.). pp. 149–204. McGraw-Hill, New York.

Erskine, J. E. and Stern, E. A. 1975. Calculation of the M_{23} magneto-optical absorption spectrum of ferromagnetic nickel. *Phys. Rev. B* 12:5016–5024.

Falicov, L. M., Pierce, D. T., Bader, S. D., Gronsky, R., Hathaway, K. B., Hopster, H. J., Lambeth, D. N., Parkin, S. S. P., Prinz, G., Salamon, M., Schuller, I. K., and Victora, R. H., 1990. Surface, interface, and thin-film magnetism. *J. Mater. Res.* 5:1299–1340.

Hulme, H. R. 1932. The Faraday effect in ferromagnetics. *Proc. R. Soc.* A135:237–257.

Kiefl, R. F., Brewer, J. H., Affleck, I., Carolan, J. F., Dosanjh, P., Hardy, W. N., Hsu, T., Kadono, R., Kempton, J. R., Kreitzman, S. R., Li, Q., O'Reilly, A. H., Riseman, T. M., Schleger, P., Stamp, P. C. E., Zhou, H., Le, L. P., Luke, G. M., Sternleib, B., Uemura, Y. J., Hart, H. R., and Lay, K. W. 1990. Search for anomalous internal magnetic fields in high-T_c superconductors as evidence for broken time-reversal symmetry. *Phys. Rev. Lett.* 64: 2082–2085.

Kittel, C. 1951. Optical rotation by ferromagnet. *Phys. Rev.* 83: 208.

Kliger, D. S., Lewis, J. W., and Randall, C. E. 1990. Polarized Light in Optics and Spectroscopy. Academic Press, San Diego.

Landau, L. D. and Lifshitz, E. M. 1960. Electrodynamics of Continuous Media. Pergamon Press, Elmsford, N.Y.

Maxwell, J. C. 1892. A Treatise on Electricity and Magnetism, Vol. II, 3rd ed., Chapter XXI, Article 811-812, pp. 454–455. Oxford University Press, Oxford.

Moog, E. R. and Bader, S. D. 1985. SMOKE signals from ferromagnetic monolayers: $p(1\times1)$ Fe/Au(100). *Superlattices Microstructures* 1:543–552.

Moog, E. R., Bader, S. D., and Zak, J. 1990. Role of the substrate in enhancing the magneto-optic response of ultrathin films: Fe on Au. *Appl. Phys. Lett.* 56:2687–2689.

Qiu, Z. Q., Pearson, J., and Bader, S. D. 1992. Magneto-optic Kerr ellipticity of epitaxial grown Co/Cu overlayers and superlattices. *Phys. Rev. B* 46:8195–8200.

Shen, Y. R. 1964. Faraday rotation of rare-earth ions. I. Theory. *Phys. Rev.* 133:A511–515.

Spielman, S., Fesler, K., Eom, C. B., Geballe, T. H., Fejer, J. J., and Kapitulnik, A. 1990. Test for nonreciprocal circular birefringence in $YBa_2Cu_3O_7$ thin films as evidence for broken time-reversal symmetry. *Phys. Rev. Lett.* 65:123–126.

Voigt, W. 1908. Magneto- und Elektrooptic. B.G. Teuner, Leipzig.

Weaver, J. H. and Frederikse, M. P. R. 1999. Optical properties of metal and semiconductors. *In* CRC Handbook of Chemistry and Physics, 80th ed. (D.L. Lide, ed.). Sect. 12, p. 129. CRC Press, Boca Raton, Fla.

Wilczek, F. 1991. Anyons. *Sci. Am.* May:58–65.

Zak, J., Moog, E. R., Liu, C., and Bader, S. D. 1990. Universal approach to magneto-optics. *J. Magn. Magn. Mater.* 89:107–123.

Zak, J., Moog, E. R., Liu, C., and Bader, S. D. 1991. Magneto-optics of multilayers with arbitrary magnetization directions. *Phys. Rev. B* 43:6423–6429.

KEY REFERENCES

Bennemann, K. H. (ed.). 1998. Nonlinear Optics in Metals. Clarendon Press, Oxford.

Covers the physics of linear as well as nonlinear magneto-optics.

Dillon, 1971. See above.

Provides an excellent background to the magneto-optic Kerr effect.

Freeman, A. J. and Bader, S. D. (eds.). 1999. Magnetism Beyond 2000. North Holland, Amsterdam.

Has 46 review articles that cover cutting-edge issues and topics in magnetism, many of which are being addressed using magneto-optic techniques.

Kliger, D. S., Lewis, J. W., and Randall, C. E. 1990. Polarized Light in Optics and Spectroscopy. Academic, Boston.

Covers instrumentation.

APPENDIX A: THE MEDIUM-BOUNDARY AND MEDIUM-PROPAGATION MATRICES

The first matrix, denoted $\boldsymbol{A}$, is the 4×4 medium-boundary matrix. It relates the components of the electric and magentic fields with the s and p components of the electric field, which are, respectively, perpendicular to and parallel to the plane of incidence. We define the incident (i) and reflected (r) waves at each boundary between two layers as in Figure 7. Then, we obtain the relation between the x, y components of **E** and **H** and the the s, p components of the electric field as

$$\begin{pmatrix} E_x \\ E_y \\ H_x \\ H_y \end{pmatrix} = A \begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ E_{\mathrm{s}}^{\mathrm{r}} \\ E_{\mathrm{p}}^{\mathrm{r}} \end{pmatrix} \tag{10}$$

where

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ \frac{i}{2}[-Q_y \tan\theta(1+\cos^2\theta) + Q_z \sin^2\theta] & \cos\theta + iQ_x \sin\theta \\ \frac{in}{2}(Q_y \sin\theta + Q_z \cos\theta) & -n \\ n\cos\theta & \frac{\mathrm{in}}{2}(Q_y \tan\theta + Q_z) \\ 1 & 0 \\ \frac{i}{2}[Q_y \tan\theta(1+\cos^2\theta) + Q_z \sin^2\theta] & -\cos\theta + iQ_x \sin\theta \\ \frac{in}{2}(Q_y \sin\theta - Q_z \cos\theta) & -n \\ -n\cos\theta & -\frac{in}{2}(Q_y \tan\theta - Q_z) \end{pmatrix} \tag{11}$$

The second matrix, denoted D, is the 4 × 4 medium-propagation matrix. It relates the s and p components of the electric field at the two surfaces (1 and 2) of a film of thickness d. This relation can then be expressed by the matrix product

$$\begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ E_{\mathrm{s}}^{\mathrm{r}} \\ E_{\mathrm{p}}^{\mathrm{r}} \end{pmatrix} = \boldsymbol{D} \begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ E_{\mathrm{s}}^{\mathrm{r}} \\ E_{\mathrm{p}}^{\mathrm{r}} \end{pmatrix}_2 \tag{12}$$

where

$$\boldsymbol{D} = \begin{pmatrix} U\cos\delta_i & U\sin\delta_i & 0 & 0 \\ -U\sin\delta_i & U\cos\delta_i & 0 & 0 \\ 0 & 0 & U^{-1}\cos\delta_{\mathrm{r}} & -U^{-1}\sin\delta_{\mathrm{r}} \\ 0 & 0 & U^{-1}\sin\delta_{\mathrm{r}} & U^{-1}\cos\delta_{\mathrm{r}} \end{pmatrix} \tag{13}$$

with

$$\begin{aligned} U &= \exp(-ikd\cos\theta) \\ \delta_i &= \frac{kd}{2}(Q_y \tan\theta + Q_z) \\ \delta_{\mathrm{r}} &= \frac{kd}{2}(Q_y \tan\theta - Q_z) \end{aligned} \tag{14}$$

Here, $k = 2\pi/\lambda$ with λ being the wavelength of the light.

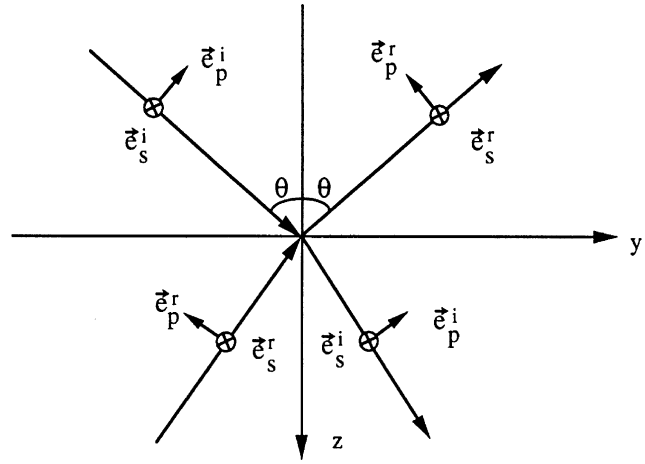


Figure 7. Definitions of the s and p directions for the incidence and reflection waves at the boundary between two media.

With the matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, we can calculate the magneto-optic effect under any conditions. Consider a multilayer structure that consists of N individual layers and a beam of light impinging on the top of the structure from the initial medium (i). After multiple reflections, there will be a reflected beam backscattered into the initial medium and a transmitted beam that emerges from the bottom layer into the final medium (f). The electric fields in the initial and final media can then be expressed as

$$\boldsymbol{P}_{\mathrm{i}} = \begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ E_{\mathrm{s}}^{\mathrm{r}} \\ E_{\mathrm{p}}^{\mathrm{r}} \end{pmatrix}_{\mathrm{i}} = \begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ r_{\mathrm{ss}}E_{\mathrm{s}}^{\mathrm{i}} + r_{\mathrm{sp}}E_{\mathrm{p}}^{\mathrm{i}} \\ r_{\mathrm{ps}}E_{\mathrm{s}}^{\mathrm{i}} + r_{\mathrm{pp}}E_{\mathrm{p}}^{\mathrm{i}} \end{pmatrix} \tag{15}$$

and

$$\boldsymbol{P}_{\mathrm{f}} = \begin{pmatrix} E_{\mathrm{s}}^{\mathrm{i}} \\ E_{\mathrm{p}}^{\mathrm{i}} \\ 0 \\ 0 \end{pmatrix}_{\mathrm{f}} = \begin{pmatrix} t_{\mathrm{ss}}E_{\mathrm{s}}^{\mathrm{i}} + t_{\mathrm{sp}}E_{\mathrm{p}}^{\mathrm{i}} \\ t_{\mathrm{ps}}E_{\mathrm{s}}^{\mathrm{i}} + t_{\mathrm{pp}}E_{\mathrm{p}}^{\mathrm{i}} \\ 0 \\ 0 \end{pmatrix}$$

where r and t are reflection and transmission coefficients of the corresponding components. If $\mathbf{P}_m$ is the field component at the bottom surface of the mth layer, then the field component at the top surface of the mth layer will be $\boldsymbol{D}\mathbf{P}_m$. Because E_x, E_y, H_x, and H_y are related to **P** by the matrix $\boldsymbol{A}$, the boundary condition—E_x, E_y, H_x, and H_y are continuous—at the interface between the mth layer and the $(m+1)$th layer is

$$\boldsymbol{A}_m\boldsymbol{P}_m = \boldsymbol{A}_{m+1}\boldsymbol{D}_{m+1}\boldsymbol{P}_{m+1} \tag{16}$$

Then the relation between $\mathbf{P}_{\mathrm{i}}$ and $\mathbf{P}_{\mathrm{f}}$ can be derived as

$$\begin{aligned} \boldsymbol{A}_i\boldsymbol{P}_i &= \boldsymbol{A}_1\boldsymbol{D}_1\boldsymbol{P}_1 = \boldsymbol{A}_1\boldsymbol{D}_1\boldsymbol{A}_1^{-1}\boldsymbol{A}_1P = \boldsymbol{A}_1\boldsymbol{D}_1\boldsymbol{A}_1^{-1}\boldsymbol{A}_2\boldsymbol{D}_2\boldsymbol{P}_2 \\ &= \cdots = \prod_{\mathrm{m}=1}^{\mathrm{N}} (\boldsymbol{A}_m\boldsymbol{D}_m\boldsymbol{A}_m^{-1})\boldsymbol{A}_{\mathrm{f}}\boldsymbol{P}_{\mathrm{f}} \end{aligned} \tag{17}$$

If this expression is put in the form $\mathbf{P}_i = T\mathbf{P}_f$, where

$$T = \boldsymbol{A}_i^{-1} \prod_{m=1}^{N} (\boldsymbol{A}_m \boldsymbol{D}_m \boldsymbol{A}_m^{-1}) \boldsymbol{A}_f \equiv \begin{pmatrix} G & H \\ I & J \end{pmatrix} \tag{18}$$

then the 2×2 matrices of $\boldsymbol{G}$ and $\boldsymbol{I}$ can used to obtain the Fresnel reflection and transmission coefficients:

$$\boldsymbol{G}^{-1} = \begin{pmatrix} t_{ss} & t_{sp} \\ t_{ps} & t_{pp} \end{pmatrix} \quad \text{and} \quad \boldsymbol{I}\boldsymbol{G}^{-1} = \begin{pmatrix} r_{ss} & r_{sp} \\ r_{ps} & r_{pp} \end{pmatrix} \tag{19}$$

The Kerr rotation ϕ' and ellipticity ϕ'' for p- and s-polarized light are then given by

$$\phi_s = \phi_s' + i\phi_s'' = \frac{r_{ps}}{r_{ss}} \quad \text{and} \quad \phi_p = \phi_p' + i\phi_p'' = \frac{r_{sp}}{r_{pp}} \tag{20}$$

APPENDIX B: THE ULTRATHIN LIMIT

In the ultrathin limit the magneto-optic expressions simplify further. For ultrathin films the total optical thickness of the film is much less than the wavelength of the light, $\sum_i n_i d_i \ll \lambda$. In this limit the $\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{-1}$ matrix can be simplified to

$$\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 & -\frac{i2\pi d}{\lambda} \\ \frac{2\pi d}{\lambda} n Q_y \sin\theta & 1 + \frac{2\pi d}{\lambda} n Q_x \sin\theta & \frac{i2\pi d}{\lambda} \cos^2\theta & 0 \\ \frac{2\pi d}{\lambda} n^2 Q_z & \frac{i2\pi d}{\lambda} n^2 & 1 - \frac{2\pi d}{\lambda} n Q_x \sin\theta & 0 \\ -\frac{i2\pi d}{\lambda} n^2 \cos^2\theta & \frac{2\pi d}{\lambda} n^2 Q_z & -\frac{2\pi d}{\lambda} n Q_y \sin\theta & 1 \end{pmatrix} \tag{21}$$

If the initial and final media are nonmagnetic, then the 2×2 matrices of $\boldsymbol{G}$ and $\boldsymbol{I}$ in Equation 19 yield the reflection coefficients

$$\begin{aligned} r_{ss} &= \frac{n_i \cos\theta_i - n_f \cos\theta_f}{n_i \cos\theta_i + n_f \cos\theta_f} \\ r_{pp} &= \frac{n_f \cos\theta_i - n_i \cos\theta_f}{n_f \cos\theta_i + n_i \cos\theta_f} \\ r_{ps} &= -\frac{4\pi}{\lambda} \frac{n_i \cos\theta_i}{(n_i \cos\theta_i + n_f \cos\theta_f)(n_f \cos\theta_f)} \\ &\quad \left(\cos\theta_f \sum_m d_m n_m^2 Q_z^{(m)} - n_f n_i \sin\theta_i \sum_m d_m Q_y^{(m)} \right) \\ r_{sp} &= -\frac{4\pi}{\lambda} \frac{n_i \cos\theta_i}{(n_i \cos\theta_i + n_f \cos\theta_f)(n_f \cos\theta_i + n_i \cos\theta_{f)}} \\ &\quad \left(\cos\theta_f \sum_m d_m n_m^2 Q_z^{(m)} + n_f n_i \sin\theta_i \sum_m d_m Q_y^{(m)} \right) \end{aligned} \tag{22}$$

Here n_i, θ_i, and n_f, θ_f are the refraction indices and the incident angles of the initial and the final media, respectively. Equation 22 provides a basis for the additivity law for multilayers in the ultrathin limit, which states that the total Kerr signal is a simple summation of the Kerr signals from each magnetic layer and is independent of the nonmagnetic spacer layers in the multilayer structure. This additivity law is true only in the limit where the total optical thickness of the layered structure is much less than the wavelength of the incident beam. For thick film, it is obvious that the additivity law must break down because the light attenuates and will not penetrate to the deeper layers of the structure.

Z. Q. QIU
University of California at Berkeley,
Berkeley, California
S. D. BADER
Argonne National Laboratory
Argonne, Illinois

This page intentionally left blank

ELECTROCHEMICAL TECHNIQUES

INTRODUCTION

Electrochemistry, in contrast to most materials characterization disciplines, contains a bifurcated methodology having applications both in materials synthesis and as a physical characterization technique. On the one hand, a variety of functional surfaces are prepared electrochemically, for example, the anodization of aluminum, electroetching of semiconductors, and electroplating of metallic films. Similarly, the electrosynthesis of bulk materials such as conducting organic polymers has taken on importance in recent years. On the other hand, electrochemistry has been demonstrated to be of use as a physical characterization technique for the quantification of conducting and semiconducting surfaces. The corrosion properties of metallic surfaces and semiconductor electronic properties are two key areas where electrochemical characterization has had a high impact. In keeping with the central analytical theme of this work, this chapter limits its focus to the class of electrochemical techniques specifically aimed at characterizing materials and their interfaces (as opposed to synthetic electrochemistry). In some cases, the line between materials synthesis and measurement can become hazy. The field of sensors based on chemically modified electrodes (CMEs) is a good example of this confusion. Often sensors of this type are electrosynthesized and the synthesis parameters represent an important portion of the systems characterization. Cases such as this will be included in the topical coverage of this chapter.

Electrochemistry differs from other analytical techniques in a second important way. Unlike most instrumental characterization techniques, the sample under study is made into part of the measuring circuitry, and thus, inappropriate sample preparation can effectively lead to a malfunction of the instrument. More importantly, this synergism demands that the user have some knowledge of the detailed circuitry associated with the instrument being employed. Most instruments can be treated as "black boxes"; the user need only understand the rudiments of the instrument's functions. However, if this approach is employed in the execution of an electrochemical experiment, the result is often an artifactual instrument response that is easily misinterpreted as the chemical response of the sample. The electrochemical literature repeatedly testifies to the preeminence of this problem.

Modern electroanalytical techniques have become viable characterization tools because of our abilities both to precisely measure small currents and to mathematically model complex heterogeneous charge transfer processes. Modeling of electrode-based charge transfer processes is only possible using digital simulation methods. The mathematics and computer science related to this underpinning of electrochemistry is sophisticated and beyond the scope of this work. In most cases, an incomplete understanding of this aspect of electroanalysis will not affect the quality of the experimental outcome. In cases where the experimenter wishes to obtain a more detailed understanding of the physical chemical basis of electrochemistry a variety of texts are available. Two volumes, the first by Bard and Faulkner (1980) and the second by Gileadi (1993) are particularly well suited to a more detailed understanding of electrochemistry and electroanalysis. The latter volume is specifically aimed at the materials scientist while the former work is one of the central texts in physical electrochemistry. A third text edited by Kissinger and Heineman (1996) is an excellent source of experimental details and analytical technologies. All of these texts will be most valuable as supplements to this work, providing more chemical details but less materials-characterization emphasis than found here.

Electrochemical measurements can be divided into two categories based on the required instrumentation. Potentiometric measurements utilize a sensitive voltmeter or electrometer and measure the potential of a sample against a standard of known potential. Voltammetric measurements utilize a potentiostat to apply a specified potential waveform to a sample and monitor the induced current response. Amperometric measurements, in which the potential is held constant and the current monitored as a function of time, are included in this latter category. Potentiometric measurements form the historical basis of electrochemistry and are of present utility as a monitoring technique for solution-based processes (i.e., pH monitoring or selective ion monitoring). However, potentiometry is limited with regard to materials characterization. Thus, this chapter focuses on voltammetric measurements and the use of the potentiostat. The potentiostat is fundamentally a feedback circuit that monitors the potential of the test electrode (referred to as the working electrode) versus a reference half-cell (typically called a reference electrode). If the potential of the working electrode drifts from a prescribed offset potential versus the reference electrode, a correcting potential is applied. A third electrode, the counterelectrode (or auxiliary electrode) is present in the associated electrochemical cell to complete the current pathway. The potentiostat typically contains a current-following circuit associated with the auxiliary electrode, which allows a precise determination of the current, which is reported to a recording device as a proportional potential. The rudimentary operation of the potentiostat is covered in this text. However, for many applications a more detailed analysis of potentiostat electronics is desirable. To this end, the reader is directed to Gileadi's excellent monograph (Gileadi et al., 1975).

Potentiostats are available from a series of vendors and range in price from less than $1000 to approximately

$30,000. Variations in price have to do with whether or not digital circuits are employed, the size of the power supply utilized, and the dynamic range and stability of the circuitry. A typical research grade potentiostat presently costs about $20,000 and thus is accessible as a standard piece of laboratory instrumentation. For certain specific applications, less expensive equipment is available (in the approximately $1,000 to $2,000 range) with top-end machines costing around $50,000.

Electrochemical investigations involve making the sample of interest into one electrode of an electrochemical cell. As such, it is immediately obvious that this technique is limited to materials having excellent to good conductivity. Metallic samples and conducting organic polymers therefore, immediately jump to mind as appropriate samples. What may be less obvious is the application of electrochemical techniques to the characterization of semiconductors; however, this application is historically central to the development of electronic materials and continues to play a key role in semiconductor development to date.

The listing above points to obvious overlaps with other materials-characterization techniques and suggests complimentary strategies that may be of utility. In particular, the characterization of electronic systems using electrical circuit responses (see ELECTRICAL AND ELECTRONIC MEASUREMENT) or electron spectroscopy (see ELECTRON TECHNIQUES) is often combined with electrochemical characterization to provide a complete picture of semiconducting systems. Thus, for example, one might employ solid-state electronics to determine the doping level or (n-type versus p-type character) of a semiconducting sample prior to evaluation of the sample in an electrochemical cell. Likewise, electron spectroscopy might be used to evaluate band-edge energetics or the nature of surface states in collaboration with electrochemical studies that determine interfacial energetics and kinetics. In passing, it is interesting to note that the first "transistor systems" reported by Bell Laboratories were silicon- and germanium-based three-electrode electrochemical cells (Brattain and Garrett, 1955). While these cells never met commercial requirements forcing the development of the solid-state transistor, these studies were critical in the development of our present understanding of semiconductor interfaces.

This chapter considers the most commonly utilized techniques in modern electrochemistry: polarography (i.e., currents induced by a slowly varying linear potential sweep), cyclic voltammetry (i.e., currents induced by a triangular potential waveform), and AC impedance spectroscopy (i.e., the real and imaginary current components generated in response to an AC potential of variable frequency). The characterization of semiconducting materials, the corrosion of metallic interfaces, and the general behavior of redox systems are considered in light of these techniques. In addition, the relatively new use of electrochemistry as a scanning probe microscopy technique for the visualization of chemical processes at conducting and semiconducting surfaces is also discussed.

LITERATURE CITED

Bard, A. J. and L. R. Faulkner. 1980. Electrochemical Methods, Fundamentals and Applications. John Wiley & Sons, New York.

Brattain, W. H. and Garrett, C. G. B. 1955. *Bell System Tech. J.* 34:129.

Gileadi, E. 1993. Electrode Kinetics for Chemists, Chemical Engineers, and Materials Scientists. VCH Publishers, New York.

Gileadi, E., Kirowa-Eisner, E., et al. 1975. Interfacial Electrochemistry: An Experimental Approach. Addison-Wesley, London.

Kissinger, P. T. and Heineman, W. R. (eds.). 1996. Laboratory Techniques in Electroanalytical Chemistry. Marcel Dekker, New York.

INTERNET RESOURCES

http://electrochem.cwru.edu/estir/

Web site of the Electrochemical Science and Technology Information Resource (ESTIR), established by Zoltan Nagy at Case Western Reserve. It serves as the unofficial site of the electrochemistry community.

http://www.electrochem.org/

Web site of The Electrochemical Society.

http://www.eggpar.com/index.html

Web site of Princeton Applied Research (a major supplier of electrochemical instrumentation. This site provides a series of "Application Notes."

http://seac.tufts.edu

Web site of the Society of Electroanalytical Chemistry, providing a full compliment of links to electrochemical sites.

ANDREW B. BOCARSLY

CYCLIC VOLTAMMETRY

INTRODUCTION

A system is completely characterized from an electrochemical point of view if its behavior in the three-dimensional (3D) space composed of current, potential, and time is fully specified. In theory, from such phenomenologic data one can determine all of the system's controlling thermodynamic and kinetic parameters, including the mechanism of charge transfer, rate constants, diffusion coefficients, standard redox potentials, electron stoichiometry, and reactant concentrations. A hypothetical $i(E,t)$ (current as a function of potential and time) data set is shown in Figure 1. This particular mathematical surface has been synthesized using the assumption that the process of interest involves a reversible one-electron charge transfer. The term "reversible" is used here in its electrochemical sense, to indicate that charge transfer between the electrode and the redox-active species both is thermodynamically reversible—i.e., the molecule(s) of interest can be both oxidized and reduced at potentials near the standard redox

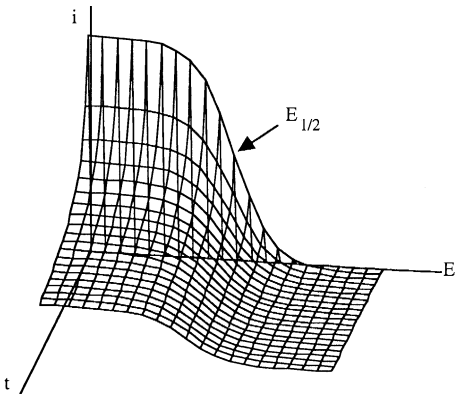


Figure 1. Theoretically constructed current-potential-time [$i(E,t)$] relation for an ideal, reversible, one-electron charge-transfer reaction taking place at an electrode surface. The half-wave potential, $E_{1/2}$, is the redox potential of the system under conditions where reactant and product diffusion coefficents are similar and activities can be ignored.

potential—and occurs at a rate of reaction sufficiently rapid that for all potentials where the reaction occurs, the process is never limited by charge-transfer kinetics. In such a case, the process is completely described by the standard redox potential of the electrochemical couple under investigation, the concentration(s) of the couple components, and the diffusion coefficients of the redox-active species.

On the other hand, one can imagine a redox reaction that is totally controlled by the kinetics of charge transfer between the electrode and the redox couple. In this case, a different 3D current-potential-time surface is obtained that depends on the charge-transfer rate constant, the symmetry of the activation barrier, and the reactant concentrations. Of course, a variety of mechanisms in between these two extremes can also be obtained, each leading to a somewhat different 3D representation. In addition, purely chemical processes can be coupled to charge-transfer events, providing more complex reaction dynamics that will be reflected in the current-potential-time surface.

The enormous information content and complexity of electrochemical dynamics in the current-potential-time domain introduces two major experimental complications associated with electrochemical data. First is the pragmatic consideration of how much time is necessary to obtain a sufficiently complete data set to allow for a useful analysis. The second, far more serious complication is that different charge-transfer mechanisms often translate into subtle changes in the $i(E,t)$ response. Thus, even after one has access to a complete data set, visual inspection is typically an insufficient means to determine the mechanism of charge transfer. Furthermore, knowledge of this mechanism is critical to determining the correct approach to data analysis. Thus, given the requisite data set, one must rely on a series of large calculations to move from the data to chemically meaningful parameters.

The approach described so far is of limited value as a survey tool, or as a global mechanistic probe of electrochemical phenomena. It is, however, a powerful technique for obtaining precise thermodynamic and kinetic parameters once the reaction mechanism is known. The mathematical resolution to the problem posed here is to consider the projection of the $i(E,t)$ surface onto a plane that is not parallel to the i-E-t axis system. This is the basis of cyclic voltammetry. Not surprisingly, a single projection is not sufficient to completely characterize a system; however, if judiciously chosen, a limited set of projections (on the order of a half dozen) will provide sufficient information to determine the reaction mechanism and obtain the key reaction parameters to relatively high precision. Perhaps more importantly, reduction of the data to two dimensions allows one to determine the mechanism based on pattern recognition, and thus no calculational effort is required. Practically, the projection of interest is obtained by imposing a time-dependent triangular waveform on the electrode as given by Equation 1:

$$E(t) = \begin{cases} E_i + \omega t & 0 \quad \text{for} \quad \leq t \leq t_\lambda \\ E_i + \omega t_\lambda - \omega t & \text{for} \quad t_\lambda \leq t \end{cases} \qquad (1)$$

where E_i is the initial potential of the scan, ω is the scan rate (in mV/s), and τ_λ is the switching time of the triangular waveform as shown in Figure 2A. The induced current is then monitored as a function of the electrode potential as shown schematically in Figure 2B. The wave shape of the i-E plot is diagnostic of the mechanism. Different projection planes are simply sampled by varying the value of

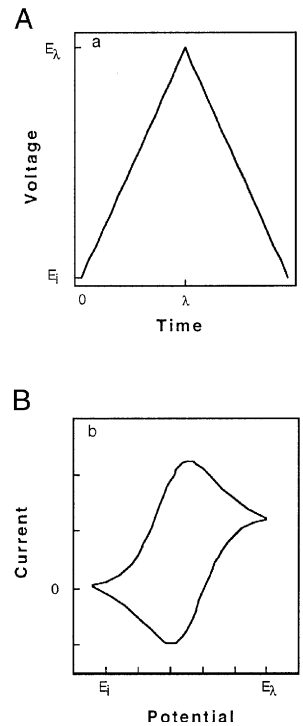


Figure 2. (**A**) Triangular $E(t)$ signal applied to the working electrode during a cyclic voltammetric scan. E_λ is the switching potential where the scan direction is reversed. (**B**) Cyclic voltammetric current versus potential response for an ideal, reversible one-electron redox couple under the $E(t)$ waveform described in (**A**).

ω for a series of cyclic voltammetric scans. Although the 2D traces obtained offer a major simplification over the 3D plots discussed above, their chief power, as identified initially by Nicholson and Shain (Nicholson and Shain, 1964), is that they are quantified from a pattern-recognition point of view by two easily obtained scan-rate-dependent parameters. Nicholson and Shain showed that these diagnostic parameters, peak current(s) and peak-to-peak potential separation, produce typically unique scan-rate dependencies as long as they are viewed over a scan-rate range that traverses at least 3 orders of magnitude. The details of this analysis are provided later. First, one needs to consider the basic experiment and the impact of laboratory variables on the results obtained.

The qualitative aspect of cyclic voltammetry has made it a general tool for the evaluation of electrochemical systems. With regard to materials specifically, the technique has found use as a method for investigating the electrosynthesis of materials (including conducting polymers and redox-produced crystalline inorganics), the mechanism of corrosion processes, the electrocatalytic nature of various metallic systems, and the photoelectrochemical properties of semiconducting junctions. Here a general description of the cyclic voltammetric technique is provided along with some examples focusing on materials science applications.

PRINCIPLES OF THE METHOD

Reaction Reversibility

The Totally Reversible Reaction. The simplest charge-transfer mechanism to consider is given by Equation 2, a reversible redox process:

$$\mathrm{Ox} + n\mathrm{e}^- \Leftrightarrow \mathrm{Red} \tag{2}$$

For this process, the concentration of the reactant and products at the electrode surface is given by the Nernst equation (Equation 3) independent of reaction time or scan rate:

$$E = E_R - (2.303)\frac{RT}{nF}\log\frac{[\mathrm{Red}]}{[\mathrm{Ox}]} \tag{3}$$

where E is the electrode potential, E_R is the standard redox potential of the redox couple, [Red] and [Ox] are the concentrations of the reduced and oxidized species, respectively, R is the gas constant, T is the temperature (K), F is Faraday's constant, and n is defined by Equation 2. [At room temperature, 298 K, 2.303 $(RT/nF) = 59$ mV/n.] Equation 3 indicates that the electrode potential, E, is equal to the redox couple's redox potential, E_R, when the concentrations of the oxidized species and reduced species at the electrode/electrolyte interface are equal. It can be shown that this situation is approximately obtained when the electrode is at a potential that is halfway between the anodic and cathodic peak potentials. This potential is referred to as the cyclic voltammetric half-wave potential, $E_{1/2}$. More precisely it can be demonstrated that the half-wave potential and the standard redox potential are related by Equation 4:

$$E_{1/2} = E_R - \frac{RT}{nF}\ln\left(\frac{\gamma_\mathrm{Ox}}{\gamma_\mathrm{Red}}\left[\frac{D_\mathrm{Red}}{D_\mathrm{Ox}}\right]^{1/2}\right) \tag{4}$$

where D_Ox and D_Red are the diffusion coefficients of the oxidized and reduced species, respectively, and the γ values represent the activity coefficients of the two halves of the couple. While this result is intuitively appealing, a solid mathematical proof of the relationship is complex and beyond the scope of this unit. The mathematical details are well presented in a variety of textbooks, however (Gileadi et al., 1975; Bard and Faulkner, 1980; Gosser, 1993).

By its very nature, the reversible redox reaction cannot cause a substantial change in the connectivity or shape of a molecular system. As a result, the diffusion coefficients of the oxidized and reduced species are expected to be similar, as are the activity coefficents—in which case Equation 4 reduces to $E_{1/2} = E_\mathrm{R}$. Even if there is some variation in the diffusion coefficients, the square root functionality invariably produces a ratio close to 1, and thus the second term in Equation 4 can safely be ignored. Likewise, at a low concentration of electroactive species (as is typically employed in cyclic voltammetry), the activity coefficients can safely be ignored. Once a system has been demonstrated to be reversible (or quasireversible, as discussed later), the redox potential can then be directly read off the cyclic voltammogram.

For the reversible mechanism, the idealized cyclic voltammetric response is illustrated in Figure 2B. Within the context of the Nicholson and Shain diagnostic criteria, this cyclic voltammetric response provides a coupled oxidation and reduction wave separated by $\sim 60/n$ mV independent of the scan rate. The exact theoretical value for the peak-to-potential separation is not critical since this value is based on a Monte Carlo approximation and is somewhat dependent on a variety of factors that are often not controlled. In a real electrochemical cell the ideal value is typically not achieved. The peak height of the anodic and cathodic waves should be equivalent, and the current function (for either the anodic or cathodic waves) is expected to be invariant with scan rate. Under such conditions the peak current is given by Equation 5 (Bard and Faulkner, 1980):

$$i_\mathrm{p} = (2.69 \cdot 10^5)n^{3/2}AD_\mathrm{o}^{1/2}\omega^{1/2}C_\mathrm{o} \tag{5}$$

where n is as defined by Equation 3, A is the area of the working electrode in cm^2, D_o is the diffusion coefficient of the electroactive species having units of cm^2/s, ω is the scan rate in mV/sec, and C_o is the bulk electrolyte concentration in moles/cm^3. (Note that these are not the standard units of concentration.)

One important caveat must be noted here: the peak-to-peak separation is not solely dependent on the charge-transfer mechanism; cell resistances, for example, will increase the peak-to-peak separation beyond the anticipated $60/n$ mV. Thus, the practical application of the diagnostic

is a constant small peak-to-peak separation ($\sim 100/n$ mV) with scan rate. It is also important to realize that the impact of cell resistance on potential ($V = iR$) is scan-rate dependent since the magnitude of the observed current increases with scan rate. Thus, under conditions of high cell resistance (for example, when a nonaqueous electrolyte is used), a reversible couple may yield a scan rate–dependent peak-to-peak potential variation significantly greater than the ideal ~60-mV shift. It is therefore critical to evaluate all three of the diagnostics over a reasonable range of scan rates before reaching a mechanistic conclusion.

The Totally Irreversible Reaction. The opposite extreme of a reversible reaction is the totally irreversible reaction. As illustrated by Equation 6, such a reaction is kinetically sluggish in one direction, producing a charge-transfer–limited current. The reaction is assigned a rate constant k; however, it is important to note that k is dependent on the overpotential, η, where η is the difference between the redox potential of the couple and the potential at which the reaction is being observed ($E_R - E_{electrode}$). It is convenient to define a heterogeneous charge transfer rate constant, k_s, which is the value of k when $\eta = 0$.

$$\mathrm{Ox} \xrightarrow{ne^-} \mathrm{Red} \tag{6}$$

For this case, a mass transport–limited current is only achieved at large overpotentials, since the rate constant for charge transfer is small for reasonable values of the electrode potential. The concentrations of redox species near the electrode are never in Nernstian equilibrium. Thus, one cannot determine either the redox potential or the diffusion coefficients associated with this system. However, careful fitting of the cyclic voltammetric data can provide the heterogeneous charge-transfer rate constant, k_s, and activation-barrier symmetry factor as initially discussed by Nicholson and Shain (1964) and reviewed by Bard and Faulkner (Bard and Faulkner, 1980; Gosser, 1993). In the case of a very small value of k_s, the irreversible cyclic voltammogram is easily identified, since it consists of only one peak independent of the scan rate employed. This peak will shift to higher potential as the scan rate is increased. For moderate but still limiting values of k_s one will observe a large peak-to-peak separation that is scan-rate dependent. The $E_{1/2}$ value will also vary with scan rate. In this case, it is important to rule out unexpected cell resistance as the source of the potential dependence, before concluding that the reaction is irreversible. Often modern digital potentiostats are provided with software that allows the direct measurement (and real time correction) of cell resistance via a current-interrupt scheme in which the circuit is opened for a short (millisecond) time period and the cell voltage is measured. Analog potentiostats often have an iR compensation circuit (Kissinger and Heineman, 1996).

The Quasireversible Reaction. Although arbitrary, it is practically useful to define reversible charge-transfer systems as those having a heterogeneous charge-transfer rate constant in excess of $0.3\omega^{1/2}$, while totally irreversible systems are those exhibiting rate constants less than $2 \times 10^{-5}\omega^{1/2}$. (Note that this definition is quite surprising in that the requisite minimum rate constant for a reversible reaction depends on the potentiostat being employed. As the slew rate of the system increases, the ability to see large charge-transfer rate constants is enhanced. This is unfortunate, in that it clouds the distinction between thermodynamics and kinetics.) This definition produces a large number of reactions having intermediate rate constants ($0.3\omega^{1/2} \geq k_s \geq 2 \times 10^{-5}\omega^{1/2}$), which are referred to as quasireversible systems. These systems will appear reversible or irreversible depending on the scan rate employed. For sufficiently large scan rates, the rate of interfacial charge transfer will be limiting and the system will appear irreversible. For slower scan rates, the system response time will allow the Nernst equation to control the interfacial concentrations and the system will appear reversible. Depending on the scan rate employed one can determine systemic thermodynamic parameters (redox potential, n, and diffusion coefficient) or kinetic parameters. As in the reversible case, the current function for quasireversible systems tends to be scan rate independent. The peak-to-peak potential dependence is also often an inconclusive indicator. However, the potential of the peak current(s) as a function of scan rate is an excellent diagnostic. At low scan rates, the peak potential can approach the theoretical $\sim 30/n$ mV separation from the half-wave potential in a scan rate–independent manner, however, as the scan rate is increased and the system enters the irreversible region and the peak potential shifts with $\log \omega$. This dependence is given by Equation 7 (Gileadi, 1993):

$$E_p^{Irr} = E_{1/2} - \sigma\left[0.52 + \log\left(\frac{D\omega}{\sigma}\right)^{1/2} - \log k_s\right] \tag{7}$$

where σ is the Tafel slope, a measure of the kinetic barrier, and the other terms are as previously defined. Note that the slope of E_p versus $-0.5\log\omega$ allows one to determine (σ/D), while D and $E_{1/2}$ can be obtained from cyclic voltammograms taken in the reversible region allowing one to determine $k(E_{1/2})$ (see Fig. 5C for an example of this type of behavior).

Nonreversible Charge-Transfer Reactions. In contrast to mechanistically "irreversible reactions," which indicate a kinetic barrier to charge transfer, a mechanistically "nonreversible reaction" refers to a complex reaction mechanism in which one or more chemical steps are coupled to the charge-transfer reaction (Nicholson and Shain, 1965; Polcyn and Shain, 1966a,b; Saveant, 1967a,b; Brown and Large, 1971; Andrieux et al., 1980; Bard and Faulkner, 1980; Gosser, 1993; Rieger, 1994). For example, consider the generic chemical reaction shown as Equation 8:

$$\begin{gathered}\mathrm{Ox} + ne^- \Leftrightarrow \mathrm{Red} \\ \mathrm{Red} \xrightarrow{k} \mathrm{Product}\end{gathered} \tag{8}$$

where k represents the rate constant for a nonelectrochemical transformation such as the formation or destruction of

a chemical bond. This reaction couples a follow-up chemical step to a reversible charge-transfer process and is thus referred to as an EC process (electrochemical step followed by a chemical step). Consider the effect of this coupling on the cyclic voltammetric response of the Ox/Red system. At very fast scan rates, Ox can be converted to Red and Red back to Ox before any appreciable Product is formed. Under these conditions, the Nicholson and Shain diagnostics will appear reversible. However, as the scan rate is slowed down, the redox couple will spend a longer time period in the Red state and thus the formation of Product will diminish the amount of Red below the "reversible" level. Therefore, at slow scan rates the ratio of peak currents (i_a/i_c) will increase above unity (the reversible value). On the other hand, the current function for the cathodic current will decrease below the reversible level as Red is consumed and thus becomes unavailable for back-conversion to Ox. The peak-to-peak potential is not expected to change for the mechanistic case presented. This set of diagnostics is unique and can be used to demonstrate the presence of an EC mechanism. Likewise, one can consider mechanisms where a preliminary chemical step is coupled to a follow-up electrochemical step (a CE mechanism) or where multiple chemical and electrochemical steps are coupled. In many of the cases that have been considered to date, the Nicholson and Shain diagnostics provide for a unique identification of the mechanism if viewed over a sufficiently large scan-rate window. An excellent selection of mechanisms and their expected Nicholson and Shain diagnostics are presented in Brown and Large (1971).

PRACTICAL ASPECTS OF THE METHOD

Since a cyclic voltammetric experiment is fundamentally a kinetic experiment, the presence of a well-defined internal "clock" is essential. That is, the time dependence is provided experimentally by the selected scan rate since the input parameter $E(t)$ is implicitly a time parameter. However, in order for this implicit time dependence to be of utility it must be calibrated with a chemically relevant parameter. The parameter used for this purpose is the diffusion of an electroactive species in the electrolyte. Thus, the cyclic voltammetric experiment is fundamentally a "quiet" experiment in which convective components must be eliminated and the diffusion condition well defined. This is done by considering the geometry of the electrochemical cell, the shape of the electrode under investigation, and the time scale of the experiment.

Additionally, from the essence of the experiment, the application of an electrode potential combined with the monitoring of the induced current, it is important to know to high precision the time-dependent electrode potential function, $E(t)$, for all values of t. This capability is established by using a potentiostat to control an electrochemical cell having a three-electrode cell configuration. Successful interpretation of the electrochemical experiment can only be achieved if the experimenter has a good knowledge of the characteristics and limitations of the potentiostat and cell configuration employed.

Electrochemical Cells

Cyclic voltammetric experiments employ a "three-electrode" cell containing a working electrode, counterelectrode (or auxiliary electrode), and reference electrode. The working electrode is the electrode of interest; this electrode has a well-defined potential for all values of $E(t)$. The counterelectrode is simply the second electrode that is requisite for a complete circuit. The reference electrode is in fact not an electrode at all, but rather an electrochemical half cell used to establish a well-defined potential against which the working electrode potential can be measured. Typically, a saturated calomel electrode (SCE—a mercury-mercurous couple) or a silver–silver chloride electrode is utilized for this purpose. Both of these reference half cells are commercially available from standard chemical supply houses. While the consideration of reference half cells is important to any electrochemical experiment, it is beyond the scope of this unit; the reader is referred to the literature for a consideration of this topic (Gileadi, 1993; Gosser, 1993; Rieger, 1994). The exact geometry of the three electrodes, along with the shape and size of the working electrode, will determine the internal cell resistance and capacitance. These electrical parameters will provide an upper limit for both the current flow (via Ohm's law: $i = V/R$, where V is voltage and R is resistance) and the cell response time (via the capacitive time constant of the cell). The contact between the reference half cell and the remainder of the electrochemical cell is typically supplied through a high-impedance frit, ceramic junction, or capillary. A high-impedance junction is employed for two purposes: to eliminate chemical contamination between the contents of the reference electrode electrolyte and the electrolyte under investigation and to guarantee a minimal current flow between the reference and working electrodes; the voltage drop associated with this current, referred to as iR drop (Ohm's law), is uncompensated by standard potentiostat circuitry. If this voltage drop becomes substantial (see Equation 9 and discussion below), the data will be highly distorted and not representative of the chemical thermodynamics/kinetics under investigation.

Potentiostats and Three-Electrode Electrochemical Cells

At its heart, the potentiostat is a negative-feedback device that monitors the potential drop between the working electrode and the reference electrode. If this value deviates from a preselected value, a bias is applied between the working and counterelectrodes. The working/counterelectrode voltage drop is increased until the measured working electrode versus reference electrode potential returns to the preset value (Gileadi et al., 1975; Gileadi, 1993). In order to carry out a cyclic voltammetric experiment, a potential waveform generator must be added to the potentiostat. The waveform generator produces the potential stimuli presented by Equation 1. Depending on the instrument being utilized, the waveform generator may be internal to the instrumentation or provided as a separate unit.

The cyclic voltammetric "figure of merit" of a potentiostat will depend on the size of the power supply utilized

and the rise time (or slew rate) of the power supply. The rise time determines the maximum scan rates that can be used in fashioning the $E(t)$ waveform. In addition to the power-supply slew rate, the cell requirements of a nonconvective system coupled with practical limits at which the potential of the working electrode can be varied (due to the resistance-capacitance (RC) time constant associated with a metal/electrolyte interface) provide upper and lower limits for the accessible scan-rate range. Unless the cell is carefully insulated from the environment, thermal and mechanical convective currents will set in for scan rates much below ~1 to 2 mV/s; this establishes a practical lower limit for the scan rate. The upper limit is dependent on the size of the electrode and the resistance of the cell. For typical cyclic voltammetric systems (electrode size ~1 mm^2), scan rates above ~10 V/s tend to introduce complications associated with the cell RC time constant. However, scan rates as high as 100,000 V/s have been reported in specially engineered cells employing ultramicroelectrodes and small junction potentials. More realistically, careful limitation of cell resistance allows one to achieve maximum scan rates in the 10 to 100-V/s range.

The size of the power supply determines the potentiostat compliance voltage (the largest voltage that can be applied between the working and counterelectrodes). This voltage determines the potentiostat's ability to control the potential of the working electrode. If one is working in a single-compartment cell with an aqueous electrolyte containing a relatively high salt concentration (≥0.1 molar salt), then a relatively modest power supply (~10 V) will provide access to all reasonable working electrode potentials. However, in order to carry out cyclic voltammetric studies in high-resistance cells (i.e., those with low salt concentrations, multiple electrochemical compartments, and/or nonaqueous electrolytes) a compliance voltage on the order of 100 V may be required. It is extremely important to note when using a high-compliance voltage potentiostat that the potential reported by the potentiostat is the voltage drop between the reference electrode and the working electrode. Although this value will never exceed ~3 V, in order to achieve this condition, the potentiostat may have to apply ~100 V between the counter and working electrodes. *Since the leads to these electrodes are typically exposed, the researcher must use utmost care to avoid touching the leads, even when the potentiostat is reporting a low working electrode potential. If the experimenter is accidentally inserted into the circuit between the working and counterelectrodes, the full output voltage of the power supply may run through the experimenter's body.*

In order to determine the potential of the working electrode with regard to the reference electrode, the potentiostat employs an electrometer. The value measured by the electrometer is used both to control the potentiostat feedback loop and to produce the potential axis in the cyclic voltammogram. As such it is assumed that the electrometer reading accurately reflects the potential of the working electrode, ϕ. In fact, the electrometer reading, E_{obs}, is better represented by Equation 9:

$$E_{obs} = \phi + iR \tag{9}$$

where iR is the uncompensated voltage drop between the working and reference electrodes. In general, application of a voltage between two electrodes (in this case, the working and reference) may cause a current to flow. This would result in a large value for the iR term (and be deleterious to the reference half cell). This problem is circumvented by employing a high-impedance junction between the working electrode and reference electrode, as noted earlier. This junction ensures that the value of i will be quite small and thus iR will have a small value. In this case $E_{obs} \approx \phi$ and the potentiostat reading is reliable. However, if R is allowed to become excessively large, then even for a small value of i, the iR term cannot be ignored, and an error is introduced into the cyclic voltammogram. High scan rates produce large peak currents, not only exacerbating the iR drop but introducing a phase-lag problem associated with the cell's RC time constant. Both these effects can severly distort the cyclic voltammogram. If this occurs, it can be remedied by switching to a low-impedance reference electrode. Decreasing the cell R improves both the iR and RC responses at the cost of destabilizing the potential of the reference half cell.

The Working Electrode

The working electrode may be composed of any conducting material. It must be recognized that the shape, area, and internal resistance of this electrode affect the resulting current (Nicholson and Shain, 1964; Bard and Faulkner, 1980; Kissinger and Heineman, 1996), and therefore these parameters must be controlled if the cyclic voltammetric data is to be used analytically. Almost all of the kinetic modeling that has been carried out for cyclic voltammetric conditions assumes a semi-infinite linear-diffusion paradigm. In order to meet this condition, one needs a relatively small electrode (~1 mm^2) that is planar. The small size assures low currents. This both limits complications associated with iR drop and provides well-defined diffusion behavior. If the above conditions cannot be met, then a correction factor can be employed prior to data analysis (Kissinger and Heineman, 1996).

In cases where cyclic voltammetric studies are applied to solution species, typical working electrode materials include platinum, gold, mercury, and various carbon materials ranging from carbon pastes and cloths to glassy carbon and pyrolytic graphite. These materials are selected because they are inert with respect to corrosion processes under typical electrochemical conditions. A second selection criteria is the electrode material's electrocatalytic nature (or lack thereof). Platinum, broadly speaking, presents a catalytic interface. This is particularly true of reactions involving the hydrogen couple. As a result, this material is most often utilized as the counterelectrode, thereby ensuring that this electrode does not kinetically limit the observed current. For the same reasons, platinum tends to be the primary choice for the working electrode. The other materials noted are employed as working electrodes because they either are electrocatalytic for a specific redox couple of interest, provide exceptional corrosion inertness in a specific electrolyte of interest, or present a high overpotential with respect to interfering redox couples. With

respect to this latter attribute, carbon and mercury are of interest, since both have a high overpotential for proton reduction. As such, one can access potentials significantly negative of the water redox potential in aqueous electrolyte using a carbon or mercury working electrode. Typically, potentials that are either more negative than the electrolyte reduction potential or more positive than the electrolyte oxidation potential are not accessible, since the large current associated with the electrolysis of the electrolyte masks any other currents. Carbon also presents an interesting interface for oxidation processes in aqueous electrolytes since it also has a high overpotential for water oxidation. Mercury, on the other hand, is not useful at positive potentials since it is not inert in this region, oxidizing to Hg^{2+}. In addition to standard metal-based electrodes, a variety of cyclic voltammetric studies have been reported for conducting polymer electrodes, semiconducting electrodes, and high-T_c superconducting electrodes.

An alternate basis for selecting an electrode material is mechanical properties. For example, the mercury electrode is a liquid electrode obtained by causing mercury to flow through a glass capillary. The active electrode area is at the outlet of the capillary, where a droplet of mercury forms, expands, and eventually drops off. This has two major electrochemical consequences. First, the electrode surface is renewed periodically (which can be helpful if surface poisoning by solutions species is an issue); second, the area of the electrode increases in a periodic manner. One can control (as a function of time) the electrode area and lifetime by controlling the pressure applied to the capillary. Obviously, this type of control is not available when using a solid electrode.

Similarly, certain solid materials offer unique mechanical characteristics providing enhanced electrode properties. For example, carbon electrodes are available in a variety of crystalline and bulk morphological forms. Unlike many materials carbon is available as a fiber, woven cloth, or paper. The availability of carbon as a paper has recently become important in the search for new electrode materials using combinatorial materials techniques. Large-area carbon-paper electrodes can be used to hold a well-defined library of potential new electrode materials made by constructing a stoichiometric distribution of alloys of two or more materials as described by Mallouk et al. (Reddington et al., 1998) Each alloy electrode is prepared as a small dot on a carbon-paper electrode. Each carbon-paper electrode holds a grid of >100 dots that provide a gradient of stoichiometries over the chemical phase-space being investigated. A survey marker is then needed to determine which dots are catalytic when the whole carbon sheet is used as an electrode. (In Mallouk's case a fluorescent pH indicator was used to identify electrodes that experienced a large pH shift in the electrolyte at modest potentials.) Potential new electrocatalyst formulations can then be cut out of the carbon grid and mounted as individual electrodes, for which a complete electroanalysis can then be carried out. In this manner a wide expanse of compositional phase-space can be searched for new electrocatalytic materials in a minimum amount of time. As a result, complex alloys containing three or more constituents can be evaluated as potential new electrode materials.

Electrolytes

The basic requirements of an electrolyte are the abilities to support the flow of current and to solubilize the redox couple of interest. In the context of a cyclic voltammetric experiment solubility becomes central, both because of the resistance limitation expressed by Equation 9 and because of the need to maintain a diffusion-limited current. It is most useful to consider the electrolyte as composed of three components, a solvent system, a supporting electrolyte, and an electroactive species. The electroactive species is the redox couple of interest. In the case of a species in solution, the species is typically present in millimolar concentrations. Below ~0.1 mM, the current associated with this species is below the sensitivity of the cyclic voltammetric experiment. At concentrations above ~10 mM it is difficult to maintain a diffusion-limited system.

The low concentration of electroactive material employed in the cyclic voltammetric experiment precludes the presence of sufficient charge carriers to support the anticipated current. Therefore, a salt must be added to the system to lower the electrolyte resistivity. This salt, which is typically present in the 0.1 to 1 M concentration range, is referred to as the supporting electrolyte. The supporting electrolyte cannot be electroactive over the potential range under investigation. Since this substance is present in significantly higher concentrations than the electroactive species, any redox activity associated with the supporting electrolyte will mask the chemistry of interest.

Water, because of its high dielectric constant and solubilizing properties, is the electrolyte solvent of choice. However, organic electrolytes are often needed to solubilize specific electroactive materials or provide a chemically nonreactive environment for the redox couple under investigation. Because of its good dielectric and solvent properties, acetonitrile is a primary organic electrolyte for cyclic voltammetry. Acetonitrile will dissociatively dissolve a variety of salts, most typically perchlorates and hexafluorophosphate-based systems. Tetraalkylammonium salts are often the supporting electrolytes of choice in this solvent, with tetra-*n*-butylammonium perchlorate—which dissolves well in acetonitrile, is electrochemically inert over a large potential window, and is commercially available at high purity—being a popular option. One important limitation of the acetonitrile systems is their affinity for water. Acetonitrile is extremely hydroscopic, and thus water will be present in the electrolyte unless the solvent is initially dried by reflux in the presence of a drying agent and then stored under an inert gas. Extremely dry acetonitrile can be prepared using a procedure published by Billon (1959). A variety of other solvents have been employed for specific electrochemical applications. The types of solvents available, compatible supporting electrolytes, solvent drying procedures, and solvent electrochemistry are nicely summarized by Mann (1969).

DATA ANALYSIS AND INITIAL INTERPRETATION

As noted earlier, one of the major attributes of cyclic voltammetric experiments is the ability to determine the

charge-transfer mechanism with relatively little data analysis using a pattern-recognition approach. In accord with the procedures established by Nicholson and Shain, initial data analysis is achieved by considering the response of current and peak-to-peak potential shifts as a function of the potential scan rate (Nicholson and Shain, 1964; Bard and Faulkner, 1980; Gosser, 1993). Three diagnostic plots were suggested by these authors:

A plot of current function, $(i_p/\omega^{1/2})$;
A plot of peak-to-peak potential separation versus scan rate;
A plot of the ratio of the anodic peak current to cathodic peak current versus scan rate.

These three diagnostic plots have become the cornerstones of cyclic voltammetric analysis. The current function required for the first plot is the peak current divided by the square root of the scan rate. This functional form is related to the fact that in the absence of kinetic limitations, the rate of charge transfer is diffusion controlled. Fick's second law of diffusion introduces a $\omega^{1/2}$ (where ω is the scan rate) dependence into the current response. Thus, division of the peak current by $\omega^{1/2}$ removes the diffusional dependence from the cyclic voltammogram's current response.

Both the first and third diagnostic require measurement of peak currents. This represents a special challenge in cyclic voltammetry due to the unusual nature of the baseline. This is most easily seen by the representation in Figure 3, which shows an idealized cyclic voltammogram plotted both in its normal *i-E* form and as i versus t. Recall that both forms are mathematically equivalent since E and t are linearly related by Equation 1. Note that the first wave observed starting at a potential of E_i has a peak current that is easily obtained (line 1), since the baseline is well established. However, it is important to realize that the peak current for the return wave is not given by line 2. This is easily seen by looking at the time-dependent plot. Here it can be seen that the return wave has a baseline that is very different from the initial wave, and thus the measurement of peak current is complex. In addition, it has been noted that the baseline established for the return wave is a function of the switching potentials. Several approaches have been suggested to resolve this baseline problem. Nicholson (1966) has suggested a mathematical solution that produces the true peak current based on the values of lines 1, 2, and 3 in Figure 3A. Using this approach the ratio of peak currents is obtained as Equation 10:

$$\frac{i_c}{i_a} = \frac{\text{line 1}}{\text{line 2}} + 0.485\frac{\text{line 3}}{\text{line 4}} + 0.086 \qquad (10)$$

where i_c and i_a are the cathodic and anodic peak currents respectively.

A critical, and often-forgotten, aspect of the Nicholson and Shain analysis is that it is necessary to observe the diagnostics over 3 orders of magnitude in scan rate in order to come to a definitive conclusion. This is often difficult to do. While in many cases one can still reach reasonable mechanistic conclusions, the uncertainty in the conclusions is increased as the scan rate range is limited. Certainly, determination of a charge transfer mechanism using less than 2 orders of magnitude in scan rate cannot be justified. Since a system which is chemically reactive may change with time, it is useful to randomize the order of scan rates employed. This way, the time dependence of the system will not be confused with the scan rate dependence.

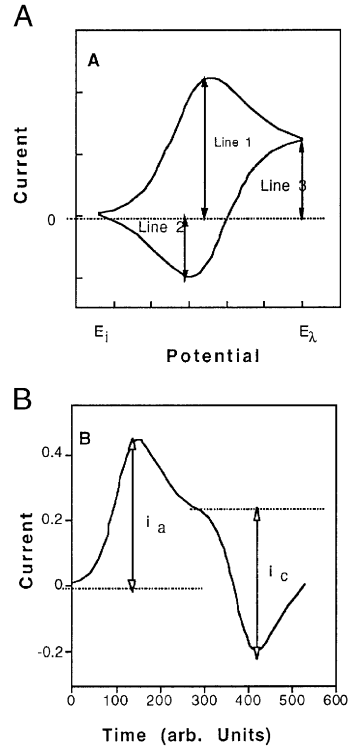


Figure 3. (**A**) Ideal one-electron reversible cyclic voltammogram showing measurements from an $i = 0$ baseline. (**B**) Same as (**A**) but showing the time dependence of the data and identifying the true baselines for the anodic and cathodic cyclic voltammetric peaks.

An Example

One mechanism that has received much attention in recent years is the electrocatalytic mechanism. In its simplest form it can be expressed as in Equation 11:

$$\begin{aligned} &\text{Ox} + n\text{e}^- \Leftrightarrow \text{Red} \\ &\text{Red} + Z \xrightarrow{k\Leftrightarrow} \text{Ox} + \text{Product} \end{aligned} \qquad (11)$$

The electrocatalytic mechanism involves an initial charge transfer from the electrode to Ox followed by a bimolecular solution charge transfer between Red and a dissolved reactant (Z) having a rate constant k. This follow up step regenerates Ox, so that the Ox/Red couple is never consumed. This mechanism is often referred to as

"mediated charge transfer." It allows one to circumvent an interfacial activation barrier for the direct oxidation of Z at the electrode. In addition, it tends to endow the reaction with a degree of specificity, since the interaction of Z with Red can be tailored by one's choice of molecular system. As such, this mechanism forms the basis for many modern electrochemical sensors. One example of this application is given in Figure 4, which demonstrates a sensor for the detection of the amino acid sarcosine as a clinical marker for kidney function. Sarcosine is unreactive at solid metal electrodes. The experiment presented here utilizes a graphite electrode and a phosphate buffered (pH = 7.0) electrolyte containing 6×10^{-6} molar sarcosine oxidase (SOX), a naturally occurring enzyme that selectively oxidizes sarcosine according to the reactions shown in Equation 12. The redox mediator in this case is $[Mo(CN)_8]^{4-}$ that has been bonded to the electrode surface. The sensor mechanism is a variation on the mechanism shown in Equation 11. Note that in this case, the oxidized form of the mediator is the active species.

$$\begin{aligned} [Mo(CN)_8]^{4-} &\Leftrightarrow [Mo(CN)_8]^{3-} + e^- \\ [Mo(CN)_8]^{3-} + SOX &\rightarrow [Mo(CN)_8]^{4-} + SOX^+ \\ SOX^+ + Sarcosine &\rightarrow SOX + Sarcosine^+ \\ Sarcosine^+ &\rightarrow \text{chemical products} \end{aligned} \qquad (12)$$

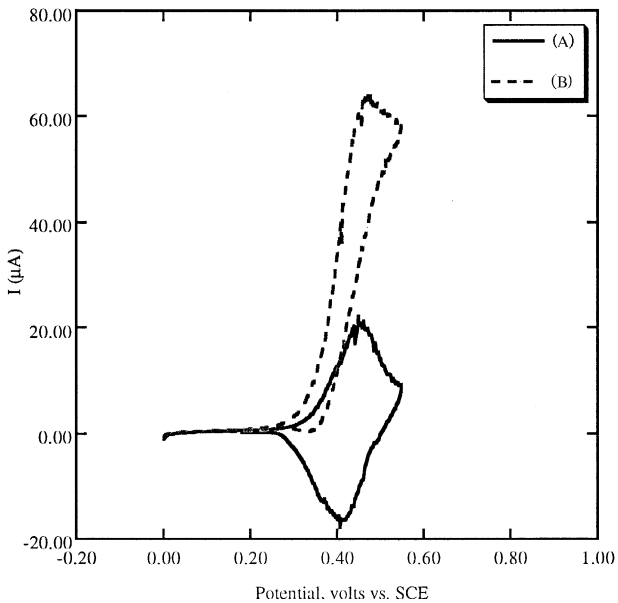


Figure 4. Cyclic voltammograms of a graphite electrode coated with a $Mo(CN)_8^{4-}$-containing polymer in an aqueous electrolyte containing 0.05 M phosphate buffer (pH = 7), 0.35 M KCl supporting electrolyte, and 6 μM sarcosine oxidase. A scan rate of 2 mV/s was employed, with the scan being initiated at 0.00 V versus SCE and proceeding to positive potentials. Scan (**A**) shows a quasireversible wave for the one-electron oxidation of the Mo complex. Scan (**B**) is the same as (**A**) with the addition of 50 mM sarcosine. The increase in the anodic wave and decrease in the cathodic wave are indicative of a mediated charge-transfer mechanism. This electrode system has been used to analyze for creatinine, a species involved in kidney function.

Scan (a) in Figure 4 shows the cyclic voltammetric response of the $[Mo(CN)_8]^{4-}$ system. Because of its low concentration, the addition of sarcosine oxidase has no perceptible effect on the cyclic voltammetric response. The electrocatalytic chain shown in Equation 12 is activated by the addition of the analyte, sarcosine, as shown in scan (b). The response is highly selective for the presence of sarcosine due to the selectivity of both the $[Mo(CN)_8]^{4-}$ mediator and the enzyme. In addition the system is very sensitive to low concentrations of sarcosine. The shift from a reversible system, scan (a), to an electrocatalytic system, scan (b), is trivially monitored by cyclic voltammetry, nicely illustrating the qualitative utility of this technique in determining reaction mechanisms. Note that both the disappearance of the cathodic wave and the enhancement in the anodic wave are requisite for the mediation mechanism (Equation 11).

Finally, it should be noted that the exact scan-rate dependence observed will be a function of k, the chemical rate constant for all coupled processes. Hence, one can obtain this quantity from a quantitative analysis of the scan-rate dependence. Thus, for chemically coupled electrochemical systems, cyclic voltammetry can be utilized as a kinetic tool for determination of the nonelectrochemical rate constant. Classically, this analysis has involved curve fitting data sets to the numerically derived functional forms of the cyclic voltammetric wave shape. This approach has been well documented and discussed from both theoretical and applied viewpoints by Gosser (1993). A more powerful kinetic analysis is available by comparison of data sets to digitally simulated cyclic voltammograms. Until recently this approach required extensive programming skills. However, digital simulation programs are now commercially available that allow one to import a cyclic voltammetric data set and correlate it with a simulated voltammogram based on a proposed mechanism (Feldberg, 1969; Speiser, 1996). Both Princeton Applied Research (PAR) EG&G and BioAnalytical Systems (BAS) offer digital simulators that interface with their data-collection software or operate on a freestanding PC data file. Such programs run on personal computers and bring high-powered kinetic and mechanistic cyclic voltammetric analysis to all users. The power of cyclic voltammetry, however, still resides in the fact that mechanistic and kinetic/thermodynamic data for a wide variety of systems can be directly obtained with little effort from raw data sets via the pattern-matching processes highlighted in this review.

Another Example

The cyclic voltammetric response of a real system is shown in Figure 5. The inorganic compound under study is shown in Figure 5A. The molecular system consists of two iron sites sandwiching a central Pt(IV). Each of the iron sites is composed of a ferrocyanide unit, $Fe(CN)_6^{4-}$, a well-studied, one-electron charge-transfer couple. The compound is held together by two bridging cyanide (CN^-) ligands. The study presented here involved an aqueous electrolyte composed of 0.1 M $NaNO_3$ and 2.5 mM electroactive complex. A platinum working electrode and an SCE reference

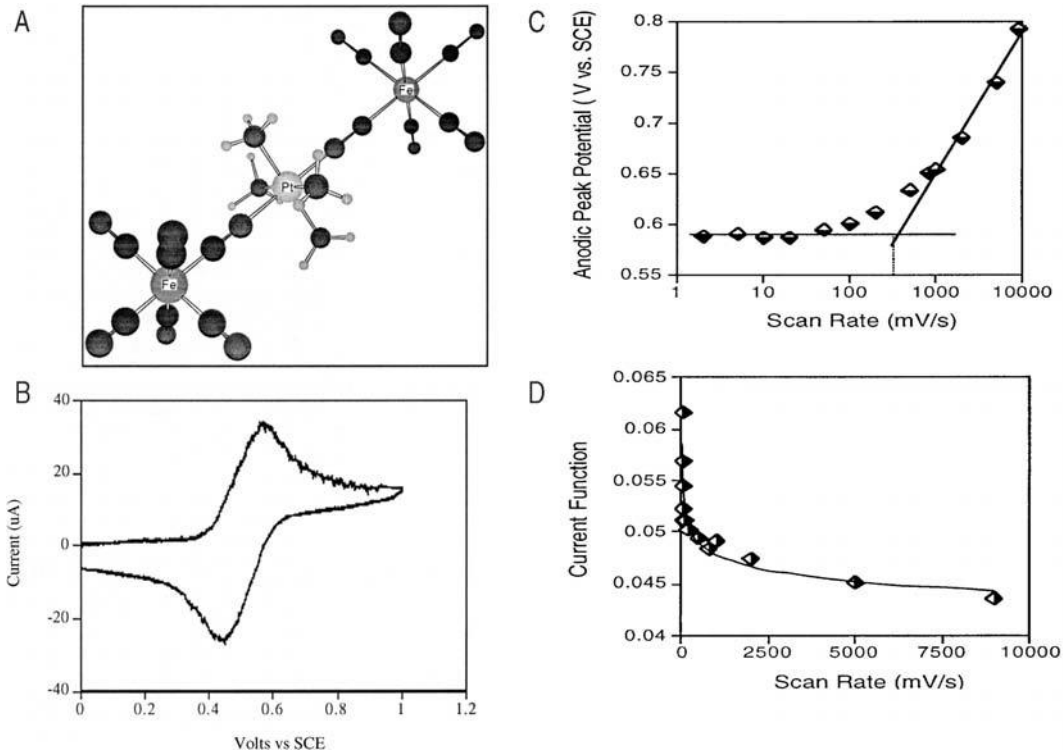


Figure 5. (**A**) Model of $\{[Fe(II)(CN)_6]-[Pt(IV)(NH_3)_4]-[Fe(II)(CN)_6]\}^{4-}$ (**B**) A cyclic voltammogram of the complex in part (**A**) taken at 100 mV/s using a platinum working electrode. The scan initiates at 0.0V versus SCE and proceeds in the positive direction. (**C**) Shift in anodic peak potential as scan rate is increased, showing reversible behavior <300 mV/s (no scan rate dependence) and overall quasireversible behavior. From this data a heterogeneous charge-transfer rate constant at $E_{1/2}$ of 4.6×10^{-3} cm/s is determined. (**D**) Current function for the data set showing the pattern for a catalytic charge-transfer process (see Equation 11).

half cell were employed. Initially, a cyclic voltammetric scan at a rate of 100 mV/s was obtained (shown in Fig. 5B). The presence of a symmetric set of waves yielding identical peak currents (once the baseline correction was applied) yielded the preliminary conclusion that this may be a reversible or quasireversible system. The symmetric nature of the cyclic voltammogram allowed determination of a half wave value, $E_{1/2} = 0.51$ V versus SCE and assignment of this value as the standard redox potential for the couple under investigation. The presence of only one pair of cyclic voltammetric waves along with the observed redox potential indicated that the signal is associated with one or more iron centers, and therefore that the Pt(IV) is electrochemically silent in this system. The value of n was determined by adding an equimolar concentration of hexaammineruthenium(II), an ideally reversible one-electron couple, to the cell. The peak current for the Fe/Pt complex was 2.8 times that observed for the hexaammineruthenium(II) complex (Zhou et al., 1990). Based on Equation 7, a value of $n = 2$ was deduced and the cyclic voltammetric wave was assigned to the simultaneous redox reaction of both irons. Note that in theory one could directly derive n from either the peak-to-peak potential separation or the peak current (Equation 7). However, the peak-to-peak potential separation at 100 mV/s is 120 mV, suggesting the system is not purely reversible. While Equation 7 might still be employed, a knowledge of the diffusion coefficent is required. Thus, the use of an internal standard represents the best method for obtaining n. Note, the standard must be chosen so that $D^{1/2}$ of the standard and the unknown are comparable.

To further develop the analysis, a plot of E_p for the anodic wave versus logω (Fig. 5C) was undertaken, clearly indicating a quasireversible system with reversible behavior exhibited below a scan rate of 300 mV/s. Construction of the current-function diagnostic indicates a further complexity in the reaction mechanism (Fig. 5D). This behavior is predicted for an electrocatalytic mechanism. In the present case, it has been shown that the observed mechanism is associated with the oxidation of tetraaminoplatinum(II) cations by the oxidized complex to regenerate the reduced complex and form a coordination polymer. The peak-to-peak potential diagnostic similarly supports this conclusion.

Based on this analysis one predicts that a coordination polymer can be synthesized as a thin layer on an electrode by holding the electrode potential at a value positive with respect to the anodic peak potential in the presence of an electrolyte containing a high concentration of the Fe/Pt complex and $[Pt(NH_3)_4]^{2+}$. This conclusion has been

experimentally verified (Pfennig and Bocarsly, 1991). The product obtained is photoactive and can thus be patterned (Wu et al., 1995). It has potential applications as an electrochemical sensor material for alkali cations (Bocarsly et al., 1988; Coon et al., 1998).

SAMPLE PREPARATION

The working electrode can be composed of any material that has a satisfactory conductivity. Thus, electrodes have been fabricated from metals, inorganic semiconductors (both small and large band gap), and conducting organic polymers. Materials in these classes have been electrochemically examined in a variety of bulk forms, including single crystals, polycrystalline solids, or particulates in a paste or pressed pellet. Conducting liquids such as mercury and gallium are also good working electrode candidates. In certain instances nonconducting materials can also be examined electrochemically using the chemically modified electrode approach. For example, the coordination polymer mentioned previously (see Another Example under Data Analysis) can be formed as an ultrathin layer (<1 μm thick) on an electrode surface using electrosynthesis techniques. The redox properties of this insulating solid can be studied under these conditions. Likewise, various nonconducting crystalline materials having redox activity can be studied by first electrocrystallizing microcrystallites of the material onto the surface of an appropriate conducting substrate.

In some cases, it is advantageous to acquire the optical properties of a material while it is undergoing oxidation or reduction. In this case, reflectivity studies can be undertaken by utilizing an optically polished working electrode. Alternatively, the material can be deposited as a thin film on an optically transparent conductor such as tin oxide, ITO (indium tin oxide) or a fine metal mesh. In this case, the optical transmission properties of the material can be evaluated. This latter approach also allows for spectroelectrochemical studies of nonconducting redox thin films. These approaches have found utility in the characterization of semiconductors and conducting organic polymers.

Independently of the material being studied, the size and shape of the working electrode must be taken into account. The cyclic voltammetric response can be distorted if too much current is passed, since this will introduce an unacceptably large iR drop (see Equation 9). Limitation of the electrode area is one of the fundamental methods for limiting current without introducing new resistances. In addition, the capacitance of the electrode-electrolyte interface is proportional to electrode area. This capacitance introduces a phase shift in the current-potential response, which distorts the cyclic voltammogram. Thus, it is critical to minimize electrode area. Pragmatically, the area A should be in the range of 10^{-3} cm$^2 \leq A < 0.1$ cm^2). In cases where the electrode material is fairly resistive (i.e., semiconductors), areas of ~ 1 cm^2 can be successfully employed. In all cases, it is very important to have a counterelectrode that has a substantially larger area than the working electrode. If this condition is not maintained, it is possible for the counterelectrode to become the current limiting electrode, thereby invalidating the cyclic voltammetric diagnostic results.

The mathematically modeled response of the electrode to a triangular potential waveform, which forms the basis for cyclic voltammetric mechanistic studies, has been developed for a spherical electrode. Further, because of the system's dependence on diffusion through the electrolyte, the current-potential response is dependent on the shape of the electrode. Fortunately, the dependence is not strong, and thus can often be ignored. Electrodes that are discs or cylinders can be employed with little negative effect. Often, square planar electrodes are also acceptable; however, in this case, care must be taken that edge effects are not perturbing the current signal. If the electrode is mounted in an insulating material so that a portion of the sample is exposed to electrolyte, it is important that the electrode/insulator interface be totally impervious to electrolyte. If a small amount of electrolyte is held between the electrode and the mount, an enhanced current will be achieved in this area, which will dramatically perturb the electrochemical response. Finally, it should be noted that electrode shape does become critical as the electrode size decreases, thus for electrodes in the 1 to 100-μm length regime, shape and size become major parameters in determining the dynamic response of the system.

SPECIMEN MODIFICATION

In some cases, the cyclic voltammetric properties of a system are dependent on the mechanical conditions of the surface. This is most evident when analyzing single-crystal semiconducting materials, which tend to show crystal-plane dependent charge-transfer properties. This effect has also been reported for single-crystal metal electrodes. Specimen surface modification can be accomplished using standard metallurgical techniques such as abrasive (or chemical) polishing, chemical etching, or electrochemical etching. Since surface damage (either inadvertent or deliberate) can modify the electrochemical response, the condition of the working electrode surface must be monitored. Standard chemical etches are often employed to remove damage during sample preparation and electrode mounting. Chemical etches can also be used to remove native oxides from a sample's surface. Such oxides are often semiconducting or insulating; thus, they modify the sample's electrochemical response.

PROBLEMS

Perhaps the biggest error made in utilizing cyclic voltammetry is believing that this technique is the ideal method for determining the electrochemical properties of a redox system. Cyclic voltammetry is a powerful technique when applied to carefully chosen systems. However, little useful data will be acquired if the system of interest has excessive charge-transfer activation barriers or is heavily coupled to a complex chemical process. It is important to recall that

for an irreversible system, cyclic voltammetry provides no information about the redox potential—which is usually a central piece of information.

Aside from this "philosophical" problem, the experimenter needs also to be aware of several pragmatic pitfalls. Unlike in almost all other analytical characterization techniques, the "sample" is an active component of the instrument in an electrochemical reaction. Thus, the effective capacitances and resistances associated with the electrochemical cell actually modify the system response of the potentiostat. As such, artifactual signals or artifactual changes in signal are easily introduced into the study. Such signals can be mistakenly identified with chemical processes. It is therefore critical to know something about the resistive and capacitive response of the electrochemical cell under investigation, and it is important to minimize both of these elements. Usually the shape of the cyclic voltammogram or the behavior of the Nicholson and Shain diagnostics will give a clue that resistive components are affecting the observed signal. However, since the charge-transfer process in an irreversible system is identical to a physical resistance, artifactual resistances may not become obvious until they are large. Thus, cell design is of critical experimental importance. Excessively high impedance between the working and reference electrodes can lead to the total loss of peaks in the cyclic voltammogram. Such uncompensated resistance can arise from the use of high-resistance electrolytes; the placement of a junction separator between the working electrode and the reference electrode (i.e., in a multicompartment cell, the working and reference electrodes must be in the same compartment); junction potentials due to the presence of a different electrolyte in the reference half cell than employed in the cell; or a partial plugging of the junction separator. Thus, care must be taken in the selection of cell geometry and the choice of reference-electrode systems. The reader is directed to Gileadi (Gileadi et al., 1975; Gileadi, 1993) with regard to these details. In cases where a high uncompensated resistance cannot be avoided, various strategies can be implemented (Kissinger and Heineman, 1996). These include the use of a potentiostat with an iR compensation circuit and/or the addition of a Luggins capillary to the electrochemical cell. This latter device is a J-shaped capillary tube that holds the reference electrode and is placed as close as is possible to the working electrode (Gileadi et al., 1975; Gileadi, 1993).

In analyzing the response of a cyclic voltammetric cell, it is important to recognize that cyclic voltammetry is intrinsically a low-noise method. If the data is presenting a poor signal-to-noise ratio, then an instrumental error is involved. Two types of error need to be considered: (1) an actual failure in the potentiostat circuitry, and (2) an excessive (or variable) reference-electrode impedance. Noise is most often due to this second source. The impedance is induced by a partial blockage of the reference electrode junction. An alternate process that may introduce a noise-like condition is adsorption or desorption of reactants or products on the electrode surface. Surface adsorption can strongly distort cyclic voltammetric data. In certain cases, this process does not present itself as noise, but rather as an asymmetry in one of the cyclic voltammetric peaks. Variations in electrode response upon multiple scans of the potential window are also often due to adsorption processes. In cases where this occurs, the best solution is to find a solvent system or electrode material that does not exhibit this behavior. However, in some cases it is worth noting that adsorption can give rise to interesting materials properties. This phenomena has led to the development of the field of chemically modified electrodes and self-assembled surfaces, in which the surface properties of a material are tailored to provide specific chemical or physical responses to an external event. For example, highly selective and sensitive sensors can be developed by surface confining a molecular species to an electrode surface that interacts in a well-defined and unique manner with a target analyte. This type of device is illustrated by the sarcosine oxidase example provided earlier.

Finally, one needs to be aware that an artifactual response can be induced by instrumentation external to the electrochemical cell. When operating above ~500 mV/s, it is important to assure that electronics external to the cell are not limiting the potential slew rate. In particular, some potentiostat power supplies cannot be slewed sufficiently fast to achieve these rates. In addition, the recording device employed may not be able to keep up with high scan rates and may therefore produce artifactual results. This is particularly likely to be encountered if a mechanical x-y recorder is employed. However, limitations can also inadvertently be encountered with an analog-to-digital converter that interfaces the potentiostat output to a computer. For extremely high-speed experiments (>10 V/s), it is often necessary to observe the output on an oscilloscope.

LITERATURE CITED

Andrieux, C. P., Blockman, C., et al. 1980. Homogeneous redox catalysis of electrochemical reactions. Part V. Cyclic voltammetry. *J. Electroanal. Chem.* 113:19–40.

Bard, A. J. and Faulkner, L. R. 1980. Electrochemical Methods, Fundamentals and Applications. John Wiley & Sons, New York.

Billon, J. P. 1959. Êlectrochimie dans l'acétonitrile. *J. Electroanal. Chem.* 1:486–501.

Bocarsly, A. B., Amos, L. J., et al. 1988. Morphological variation at the $[NiFe(CN)_6]^{2-/1-}$ derivatized nickel electrode: A technique for the evaluation of alkali cation containing solutions. *Anal. Chem.* 60:245–249.

Brown, E. R. and Large, R. F. 1971. Cyclic voltammetry, AC polarography, and related techniques. *In* Physical Methods of Chemistry, Part IIA, Electrochemical Methods, Vol. 1. (A. Weissberger and B. W. Rossiter, eds.). pp. 423–530. John Wiley & Sons, New York.

Coon, D. R., Amos, L. J., et al. 1998. Analytical applications of cooperative interactions associated with charge transfer in cyanometalate electrodes: Analysis of sodium and potassium in human whole blood. *Anal. Chem.* 70:3137–3145.

Feldberg, S. W. 1969. Digital simulation: A general method for solving electrochemical diffusion-kinetic problems. *Electroanal. Chem.* 3:199–296.

Gileadi, E. 1993. Electrode Kinetics for Chemists, Chemical Engineers, and Materials Scientists. VCH Publishers, New York.

Gileadi, E., Kirowa-Eisner, E., et al. 1975. Interfacial Electrochemistry, An Experimental Approach. Addison-Wesley, Reading, Mass.

Gosser, D. K. 1993. Cyclic Voltammetry, Simulation and Analysis of Reaction Mechanisms. VCH Publishers, New York.

Kissinger, P. T. and Heineman, W. R. (eds.). 1996. Laboratory Techniques in Electroanalytical Chemistry. Marcel Dekker, New York.

Mann, C. K. 1969. Nonaqueous solvents for electrochemical use. *Electroanal. Chem.* 3:57–133.

Nicholson, R. S. 1966. Semiempirical procedure for measuring with stationary electrode polarography rates of chemical reactions involving the product of electron transfer. *Anal. Chem.* 38:1406.

Nicholson, R. S. and Shain, I. 1964. Theory of stationary electrode polarography. Single scan and cyclic methods applied to reversible, irreversible, and kinetic systems. *Anal. Chem.* 36:706–723.

Nicholson, R. S. and Shain, I. 1965. Theory of stationary electrode polarography for a chemical reaction coupled between two charge transfers. *Anal. Chem.* 37:178–190.

Pfennig, B. W. and Bocarsly, A. B. 1991. Surface attached $[(NC)_5Fe\text{-}CN\text{-}Pt(NH_3)_4.NC\text{-}Fe(CN)_{5\ D}]^{4-}$: A study in the electrochemical and photochemical control of surface morphology. *Inorg. Chem.* 30:666–672.

Polcyn, D. S. and Shain, I. 1966a. Multistep charge transfers in stationary electrode polarography. *Anal. Chem.* 38:370–375.

Polcyn, D. S. and Shain, I. 1966b. Theory of stationary electrode polarography for a multistep charge transfer with catalytic (cyclic) regeneration of the reactant. *Anal. Chem.* 38:376–382.

Reddington, E., Sapienza, A., et al. 1998. Combinatorial electrochemistry: A highly parallel, optical screening method for discovery of better electrocatalysts. *Science* 280:1735–1737.

Rieger, P. H. 1994. Electrochemistry. Chapman & Hall, New York.

Saveant, J. M. 1967a. Cyclic voltammetry with asymmetrical potential scan: A simple approach to mechanisms involving moderately fast chemical reactions. *Electrochim. Acta* 12: 999–1030.

Saveant, J. M. 1967b. ECE mechanisms as studied by polarography and linear sweep voltammetry. *Electrochim. Acta* 12: 753–766.

Speiser, B. 1996. Numerical simulation of electroanalytical experiments: Recent advances in methodology. *Electroanal. Chem.* 19:2–109.

Wu, Y., Pfennig, B. W., et al. 1995. Development of redox-active optical mesostructures at chemically modified electrode interfaces. *Inorg. Chem.* 34:4262–4267.

Zhou, M., Pfennig, B. W., et al. 1990. Multielectron transfer and single crystal X-ray structure of a trinuclear cyanide-bridged platinum iron species. *Inorg. Chem.* 29:2456–2460.

KEY REFERENCES

Bard and Faulkner, 1980. See above.

This publication is considered by many to be the central text covering modern electrochemistry. It provides very few experimental details; however, it does an excellent job on developing the physical chemistry of electrochemistry. While a wide variety of electrochemical techniques are considered, the book focuses on controlled potential techniques including cyclic voltammetry. This work is excellent at developing the thermodynamics and chemical kinetics (mechanism) electrochemical processes and the relationship between cyclic voltammetric response and chemical mechanism.

Gileadi, 1993. See above.

The second half of this text provides an excellent discussion of the experimental details necessary to carry out electrochemical studies via a series of teaching experiments. Topics include reference electrodes, potentiostats, and cyclic voltammetry.

Gosser, 1993. See above.

A dedicated monograph that discusses cyclic voltammetry as a tool for investigating chemical mechanisms (i.e., reaction kinetics). This work includes discussions related to the determination of redox potentials, activation processes, rate constants, and complex chemical mechanisms. Pragmatic experimental details are also discussed.

Kissinger and Heineman, 1996. See above.

An excellent source of information on how to carry out laboratory electrochemical studies in general. This text covers an extensive list of electrochemical techniques including cyclic voltammetry. It provides both theoretical and experimental details.

ANDREW B. BOCARSLY
Princeton University
Princeton, New Jersey

ELECTROCHEMICAL TECHNIQUES FOR CORROSION QUANTIFICATION

INTRODUCTION

Corrosion is the deterioration of a solid body upon interaction with atoms or molecules that originate outside the solid body. In the majority of cases, the solid body is a metal or alloy. Metallic structures and components exposed to atmospheres can undergo atmospheric corrosion, while those that are used to carry out specific processes will be subject to various forms of attack depending upon the temperature and the chemical composition of the medium. In addition, velocity effects of fluids and erosive effects of particulates can exert additional influence on the corrosion process (see TRIBOLOGICAL AND WEAR TESTING). Thus corrosion phenomena of one form or another exist in the petrochemical, automotive, electric power generation, water supply and treatment, semiconductor, transportation, and space exploration industries.

Corrosion is so widespread because most metallic materials of practical significance are chemically metastable. Therefore, nature provides the inherent tendency and driving force for the material to revert to its most stable thermodynamic state, which could be, e.g., an oxide, a hydroxide, a sulfide, a carbonate, or a sulfate. Thus the corrosion scientist and engineer must constantly battle with nature to preserve the metastability of the metallic material for long duration. Herein lies their major challenge.

The direct economic impact of corrosion has been variously estimated. A 1976 study put the annual economic cost of corrosion in the United States at 70 billion dollars. While one may take issue with specific figures, it is clear that the costs are substantial. In addition to these direct costs, indirect costs are also associated with corrosion,

resulting from factors such as loss of production or the need for alternative procedures during the replacement of a corroded component or structure, gradual decrease in process efficiency resulting from corrosion product buildup and its impact on process flow and heat transfer, the need to overdesign certain components because of uncertainty due to corrosion, and loss of property and lives.

Accurate measurement of corrosion is essential to predict the life of critical metallic components in service. Corrosion measurement also becomes very important in evaluating the effectiveness of corrosion control strategies. The techniques used to measure corrosion vary depending upon whether one is dealing with gas phase corrosion or liquid phase corrosion. Electrochemical techniques, which are the subject of the present chapter, are applicable specifically to corrosion occurring in liquid media. Corrosion can be either general or localized. Much of the discussion in this chapter will focus on the phenomenon of general atom loss from metallic surfaces. When a metal corrodes in a liquid, the corrosion can involve a direct dissolution of the metal into the liquid, the conversion of the metal into an inorganic corrosion product layer on the surface, or a combination of both reaction layer formation and dissolution. The most widely spread corrosion processes in liquids occur in aqueous environments. To a more limited extent, corrosion is an issue in molten salt environments, e.g., in molten carbonate fuel cells or in the molten salt corrosion of turbine blades. Corrosion in liquid hydrocarbons is of interest to the petrochemical industry; however, in view of the nonconducting nature of these liquids, the applicability of electrochemical techniques to such systems is questionable.

In the following, three electrochemical techniques are described for the quantification of corrosion processes in liquid systems. The first two are direct current (dc) techniques, the first of which requires the use of relatively high applied potentials, while the second one uses applied voltages in the vicinity of the corrosion potential. The third technique, which uses alternating current (ac) voltages and currents, has the advantage of independently measuring different frequency-dependent processes in corrosion.

TAFEL TECHNIQUE

Principles of the Method

A metal in contact with an aqueous electrolyte solution acquires a potential. This potential is a mixed potential generated by a combination of anodic reactions and cathodic reactions. The classical paper on mixed potential theory was written by Wagner and Traud (1938). The development of such a mixed potential is schematically described in Figure 1, where the y axis is the electrode potential and the x axis is the current involved in the electron exchange reaction. In the lower portion of the graph, two straight lines are shown corresponding to the metal oxidation reaction, $M = M^+ + e^-$, and the reverse reaction, metal reduction. The intersection of these lines gives the equilibrium electrode potential of the charge-transfer reaction involving the metal, M. At the metal–aqueous solution interface, other charge transfer reactions are possible, for example, $H^+ + e^- = 1/2H_2$. The oxidation and reduction lines for this reaction are shown in the upper portion of the graph in Figure 1.

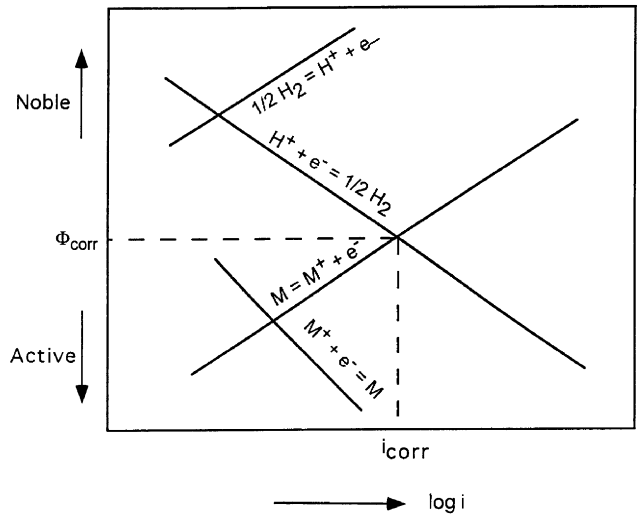


Figure 1. Schematic of corrosion potential and corrosion current.

The potential corresponding to the equilibrium of this charge transfer reaction can be read off on the y axis. The Butler-Volmer equation (Bockris and Reddy, 1970) represents in a general sense the current resulting from an applied potential for a specific charge-transfer reaction. If $\Phi - \Phi_0$ is the difference between the applied potential and the equilibrium potential for the charge-transfer reaction, then

$$i = i_0\left[\exp\left(\frac{(1-\alpha)F(\Phi-\Phi_0)}{RT}\right) - \exp\left(\frac{-\alpha F(\Phi-\Phi_0)}{RT}\right)\right] \tag{1}$$

where the potential Φ is the inner, or Galvani, potential of a metal in contact with an electrolyte; the potential Φ_0, represents the situation where the forward and backward rates of the charge-transfer reaction responsible for the buildup of potential Φ are equal (i.e., the specific charge-transfer reaction is in equilibrium); i_0 is the exchange current characteristic of the equilibrium of the electron-transfer reaction; $(1-\alpha)(\Phi-\Phi_0)$ represents the potential through which the metal has to be activated to effect ionization and $\alpha(\Phi-\Phi_0)$ the activation barrier for the metal ions in solution to revert back to the metallic state; and F is the Faraday constant, R the universal gas constant, and T the absolute temperature. Equation 1 can be rewritten for the case of a metal charge-transfer reaction in the form

$$i_M = i_{0,M}\left[\exp\left(\frac{\Phi-\Phi_{0,M}}{\beta_M^a}\right) - \exp\left(\frac{-(\Phi-\Phi_{0,M})}{\beta_M^c}\right)\right] \tag{2}$$

The constants in the exponential of Equation1 have been combined into the term β in Equation 2, the superscripts a and c indicating the anodic and cathodic reactions,

respectively, and the subscript M the metal charge-transfer reaction. Taking the hydrogen charge-transfer reaction as the second important reaction occurring on the metal surface, one can express the Butler-Volmer equation in the form

$$i_{\mathrm{H}} = i_{0,\mathrm{H}}\left[\exp\left(\frac{\Phi - \Phi_{0,\mathrm{H}}}{\beta_{\mathrm{H}}^{\mathrm{a}}}\right) - \exp\left(\frac{-(\Phi - \Phi_{0,\mathrm{H}})}{\beta_{\mathrm{H}}^{\mathrm{c}}}\right)\right] \quad (3)$$

In practice, however, the potential attained by a metal in solution corresponds neither to the equilibrium of the metal charge-transfer reaction nor to the equilibrium of the hydrogen charge-transfer reaction but falls somewhere in between. This is explained by the mixed-potential theory (Wagner and Traud, 1938). In a typical situation, such as indicated in Figure 1, the anodic part of the metal charge-transfer reaction and the cathodic part of the hydrogen charge-transfer reaction would dominate. The potential corresponding to this situation is the corrosion potential Φ_{corr}, which is given by the intersection of the anodic limb of the metal charge-transfer reaction and the cathodic limb of the hydrogen charge-transfer reaction. The corrosion current i_{corr} can be read off on the x axis corresponding to this intersection.

According to the mixed-potential theory, the total current i at any potential is given by the sum of the partial currents corresponding to the various reactions on the metal surface. Taking the anodic part of the metal charge-transfer reaction and the cathodic part of the hydrogen charge-transfer reaction to dominate,

$$i = i_{0,\mathrm{M}} \exp\left(\frac{\Phi - \Phi_{0,\mathrm{M}}}{\beta_{\mathrm{M}}^{\mathrm{a}}}\right) - i_{0,\mathrm{H}} \exp\left(\frac{-(\Phi - \Phi_{0,\mathrm{H}})}{\beta_{\mathrm{H}}^{\mathrm{c}}}\right) \quad (4)$$

At the corrosion potential, the anodic current due to the metal charge-transfer reaction and the cathodic current due to the hydrogen charge-transfer reaction are equal and is given by the corrosion current i_{corr}. Thus,

$$i_{\mathrm{corr}} = i_{0,\mathrm{M}} \exp\left(\frac{\Phi_{\mathrm{corr}} - \Phi_{0,\mathrm{M}}}{\beta_{\mathrm{M}}^{\mathrm{a}}}\right) - i_{0,\mathrm{H}} \exp\left(\frac{-(\Phi_{\mathrm{corr}} - \Phi_{0,\mathrm{H}})}{\beta_{\mathrm{H}}^{\mathrm{c}}}\right) \quad (5)$$

On combining Equations 4 and 5,

$$i = i_{\mathrm{corr}}\left[\exp\left(\frac{\Phi - \Phi_{\mathrm{corr}}}{\beta_{\mathrm{M}}^{\mathrm{a}}}\right) - \exp\left(\frac{-(\Phi - \Phi_{\mathrm{corr}})}{\beta_{\mathrm{H}}^{\mathrm{c}}}\right)\right] \quad (6)$$

The Tafel approach (1904; Bockris and Reddy, 1970; Fontana and Greene, 1978; Uhlig and Revie, 1985; Kaesche, 1985) involves the determination of the intersection point of the dominant anodic reaction curve and the dominant cathodic reaction curve in order to determine the corrosion current i_{corr} in accordance with Equation 5. The β's are temperature-dependent constants for the specific corrosion system and are known as Tafel constants. The terms $i_{0,\mathrm{M}}$ and $i_{0,\mathrm{H}}$ are the exchange current densities of the anodic and cathodic reactions, respectively. In other words, $i_{0,\mathrm{M}}$ represents the current corresponding to the equilibrium of the charge-transfer reactions involving metal dissolution and metal deposition and $i_{0,\mathrm{H}}$ is similarly defined. At high anodic potentials,

$$i = i_{0,\mathrm{M}} \exp\left(\frac{\Phi - \Phi_{0,\mathrm{M}}}{\beta_{\mathrm{M}}^{\mathrm{a}}}\right) \quad (7)$$

Equation 4 then reduces to

$$\log i = \log i_{0,\mathrm{M}} + \frac{\Phi - \Phi_{0,\mathrm{M}}}{2.303\beta_{\mathrm{M}}^{\mathrm{a}}} \quad (8)$$

which is an expression of the Tafel law (Tafel, 1904). According to this law, a plot of log i vs. the potential must be a straight line. The anodic Tafel slope β_{a} equals $2.303\beta_{\mathrm{M}}^{\mathrm{a}}$. Similarly, at high cathodic voltages, a straight-line behavior is expected with the cathodic Tafel constant β_c, which equals $2.303\beta_{\mathrm{H}}^{\mathrm{c}}$. If one extrapolates the straight lines corresponding to the anodic reaction and the cathodic reaction, respectively, their intersection corresponds to the corrosion current and the corrosion potential. From the measured corrosion current, a corrosion rate, r, in moles per second, can be determined using Faraday's law:

$$r = \frac{i_{\mathrm{corr}}}{nF} \quad (9)$$

where n represents the valence of the corroding metal (number of equivalents). An engineering unit of corrosion rate prevalent in the United States is mils per year (mpy), where one mil equals one-thousandth of an inch. The corrosion rate R in mpy is given by

$$R = 0.129\frac{i_{\mathrm{corr}}}{nD} \quad (10)$$

where D is the density of the corroding metal.

Practical Aspects of the Method

A typical electrochemical cell that can be used to measure the corrosion rate of a metal in an aqueous medium is shown in Figure 2. The cell consists of a working electrode that is the metal undergoing corrosion. A reference electrode (e.g., calomel) provides a stable reference with respect to which the potential of the working electrode can be measured. When the potential of the metal is changed, the resulting current flow is measured between the working electrode and the counter electrode. An electronic potentiostat in conjunction with the electrochemical cell is used to carry out the electrochemical measurements. Excellent potentiostats are available from manufacturers of electrochemical research equipment such as Princeton Applied Research or Solatron Instrumentation Group. The potentiostat has connections for the working electrode, counter electrode, and reference electrode. In the Tafel experiment, the potential of the working electrode is increased in steps in the anodic direction and the corresponding current measured. Both application of potential and the measurement of current are accomplished using the electronic potentiostat. A typical voltage step is 10 mV. The voltage range of the Tafel experiment can be 100 to

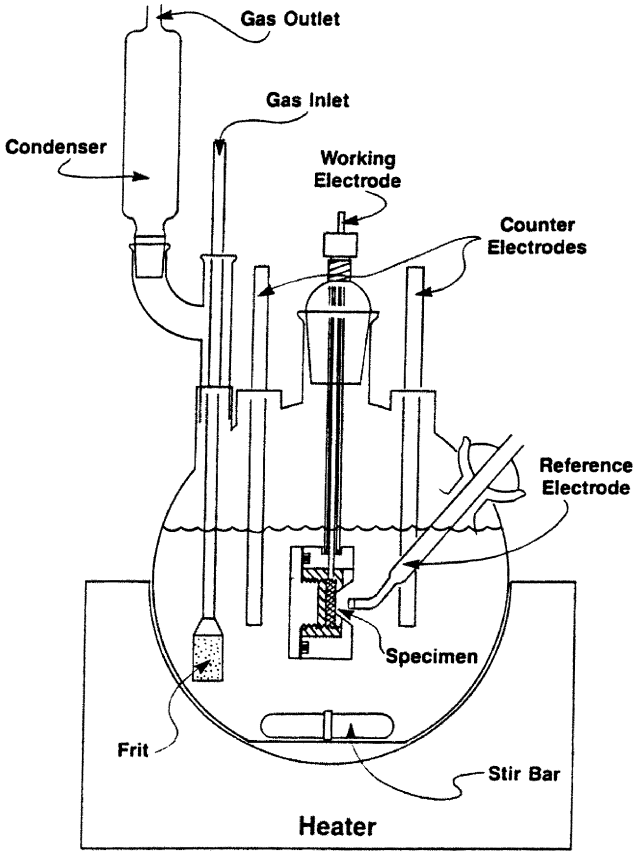


Figure 2. Experimental electrochemical cell arrangement for corrosion studies.

200 mV. Similar to the anodic case, a voltage-current curve is generated for potentials applied in the cathodic direction. Data acquisition from the potentiostat can be carried out using a computer. Corrosion measurement software packages are available from electrochemical equipment manufacturers such as Princeton Applied Research that allow voltage steps and voltage ranges to be predetermined for a corrosion experiment.

Figure 3 represents a Tafel study (Barnartt, 1971) on iron in 1 M $NaHSO_4$ solution at 25°C. In the plot shown, $\Delta\varepsilon$ is the potential difference from the corrosion potential in the anodic or cathodic direction. The authors describe a three-point method applicable in any potential range to get accurate Tafel slopes. The three points correspond to three voltage-current data sets: the first at a selected $\Delta\varepsilon$ value, the second at 2 $\Delta\varepsilon$, and the third at -2 $\Delta\varepsilon$. The mathematical analysis to derive Tafel slopes from these points is described by Barnartt (1971). Figure 4 presents an example from a study by Bandy and Jones (1976) that demonstates the difficulty in determining Tafel slopes. This figure shows that while the cathodic Tafel slope is easy to determine, the anodic Tafel slope is less evident. In this case, the authors use a procedure described by Stern and Roth (1957) to calculate the anodic curve in the vicinity of the corrosion potential and thereby determine the actual corrosion rate. The anodic currents are derived from cathodic data as the sum of the net current and

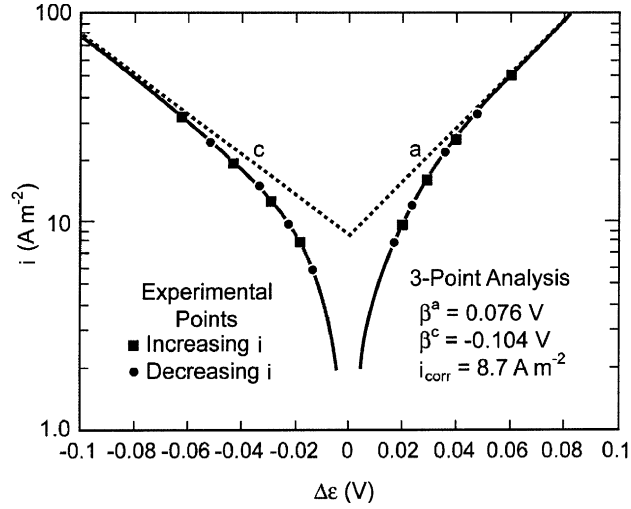


Figure 3. Polarization curves for Ferrovac E iron in 1 M $NaHSO_4$ at 25°C.

cathodic current. The approach is useful in cases where the cathodic Tafel line is well defined.

Figure 5 shows an example from the author's laboratory (Ramanarayanan and Smith, 1990) involving corrosion rate measurements at a temperature of 225°C in a 1% sodium chloride solution saturated with 90:10 Ar-H_2S at a pressure of 2000 psi. Because of the high temperatures involved, a nickel wire reference electrode is used. The wire forms a nonstoichiometric nickel sulfide film on its surface and functions as a pseudoreference as long as the sulfur content in the aqueous solution is fixed. During corrosion, a film of iron-deficient pyrrhotite, $Fe_{1-x}S$, grows on the surface of the working electrode, which is 4130

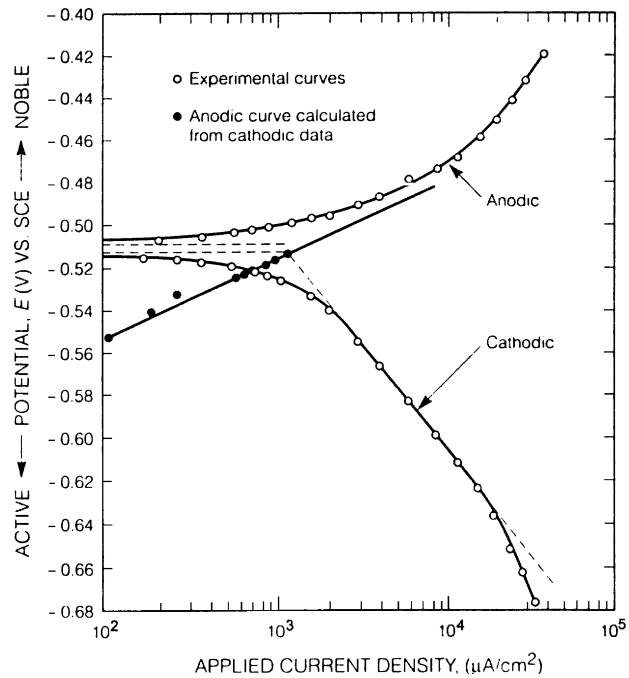


Figure 4. Experimental polarization curves for 1080 steel in deaerated 1 N H_2SO_4.

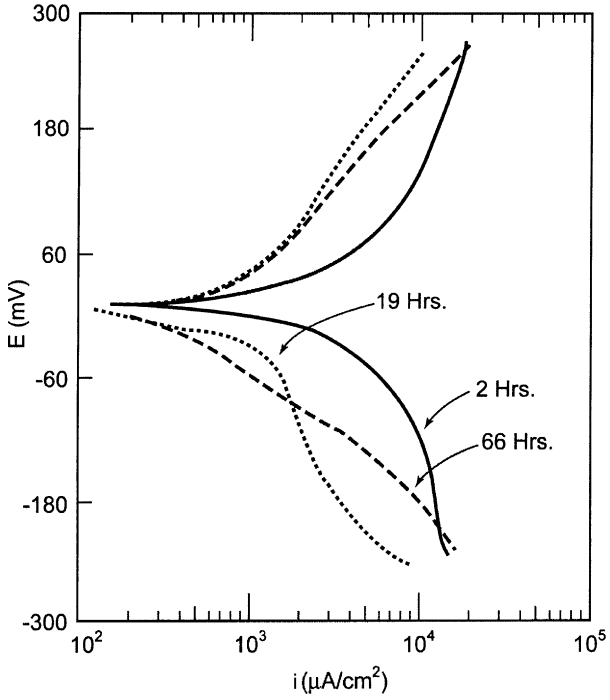


Figure 5. Polarization curves as a function of time for the corrosion of 4130 carbon steel in 1% NaCl solution saturated with Ar–10% H_2S at 2000 psi and 220°C.

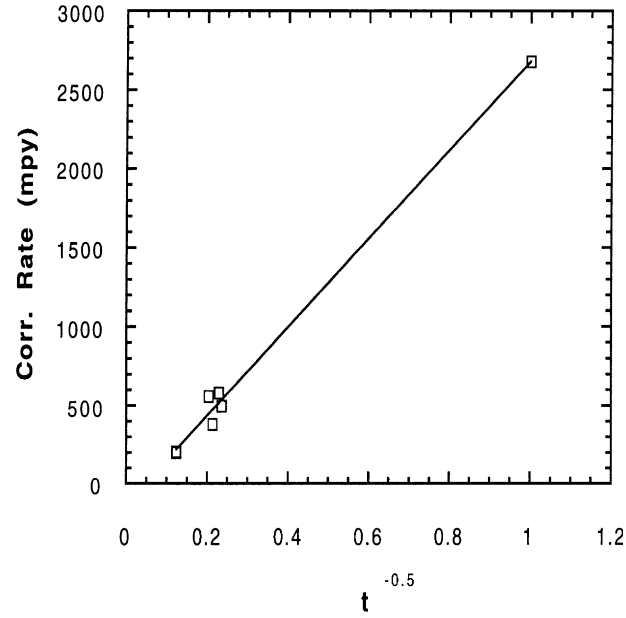


Figure 6. Demonstration of diffusion-controlled kinetics during the corrosion of 4130 steel in 1% NaCl saturated with Ar–10% H_2S at 220°C and 2000 psi.

carbon steel. It is assumed as an approximation that, within the duration of a Tafel run, the growth of the sulfide film is negligible. Figure 5 shows Tafel plots obtained at different times. In this case, the anodic charge-transfer reaction occurs at the metal-FeS interface, and the cathodic charge-transfer reaction occurs at the sulfide film–solution interface. Under the applied voltage, the potential-driven diffusion of iron through FeS is rapid enough so that charge transfer is still the rate-limiting step under the conditions of the Tafel experiment. On the other hand, by analyzing the rates obtained at different times, one obtains a parabolic law for corrosion kinetics as shown in Figure 6. This example illustrates that with films in which diffusion is relatively fast, under an applied potential, one can create conditions where the overall rate under the conditions of the experiment is charge-transfer limited. However, the overall time dependence of a number of Tafel experiments at different times enables the researcher to determine the actual rate-limiting step in corrosion in the absence of an applied potential.

Problems

The presence of passive films on the surface will interfere with the correct use of the Tafel approach, especially in the study of corrosion at low temperatures. Ideally, the Tafel equation describes the direct dissolution of metal into the liquid phase with charge transfer as the rate-limiting step.

Another error in the use of the Tafel technique comes from solution resistance (Scully, 1995). If the solution resistance is significant, it can make a significant contribution to the applied potential. Approaches to minimize the solution resistance involve the placement of the reference electrode as close as practicable to the working electrode using a Luggin-Haber capillary. Many reviews (Britz, 1978; Hayes and Kuhn, 1977-78; Mansfeld, 1982; Scribner and Taylor, 1990) are available on minimizing the solution resistance contribution.

The Tafel approach assumes pure charge transfer control. In many systems, either the reactant (cathodic step) or product (anodic step) can lead to a concentration depletion or buildup in the vicinity of the working electrode (corroding sample). In this case, the corrosion rate becomes controlled by mass transfer (Jones, 1992). Mass transfer effects can be minimized with adequate stirring. A rotating-disk technique is frequently used to minimize mass transfer effects and to simulate turbulent liquid streams (Poulson, 1983).

Finally, the high potentials applied in the Tafel approach to determine corrosion rates can introduce changes in the nature of the surface because of metal dissolution. Such changes can affect the surface area of the corroding metal, for example.

LINEAR POLARIZATION

Principles of the Method

Wagner and Traud (1938), Stern and Geary (1957), and later Oldham and Mansfeld (1971) have considered other ways of treating Equation 4 to get at the corrosion current. Differentiating Equation 4 with respect to the potential yields

$$\frac{\partial i}{\partial \Phi} = \frac{i_{0,\mathrm{M}}}{\beta_{\mathrm{M}}^{\mathrm{a}}} \exp\left(\frac{\Phi - \Phi_{0,\mathrm{M}}}{\beta_{\mathrm{M}}^{\mathrm{a}}}\right) + \frac{i_{0,\mathrm{H}}}{\beta_{\mathrm{H}}^{\mathrm{c}}} \exp\left(\frac{-(\Phi - \Phi_{0,\mathrm{H}})}{\beta_{\mathrm{H}}^{\mathrm{c}}}\right) \quad (11)$$

At the corrosion potential Φ_{corr}, the above derivative can be expressed in the following form by combining Equations 5 and 7:

$$\left(\frac{\partial i}{\partial \Phi}\right)_{\Phi_{corr}} = i_{corr}\left(\frac{1}{\beta_M^a} + \frac{1}{\beta_H^c}\right) \quad (12)$$

Stern and Geary (1957) expressed Equation 12 in the form

$$i_{corr} = \frac{\Delta i}{2.303\ \Delta\Phi}\left(\frac{\beta_a \beta_c}{\beta_a + \beta_c}\right) \quad (13)$$

where β_a and β_c are the Tafel constants. The ratio $\Delta\Phi/\Delta I$ is the polarization resistance R_p. For the Stern-Geary approximation to be valid, the applied potential must be within ~10 mV of the corrosion potential in practice. This is a rather general statement; the maximum permissible applied potential would depend on the specific corrosion system being investigated. For a proper evaluation of the corrosion current, the Tafel constants β_a and β_c must be known. The results of polarization measurements on type 430 stainless steel in 1 N H_2SO_4 from eight different laboratories (ASTM standard G 59) are summarized in Figure 7. The mean result is represented by curve 1 while curves 2 and 3 indicate 95% confidence limits (Scully, 1995).

For Equation 13 to apply, the current-potential curve must be linear in the vicinity of the corrosion potential. However, such perfect linearity is not always observed in practice. This has been pointed out by several authors (Barnartt, 1969; Stern, 1958; Antropov, Gerasimenko, and Gerasimenko, 1966). As described by Mansfeld (1976a), the second derivative of Equation 4 with respect to Φ reads

$$\frac{\partial^2 i}{\partial \Phi^2} = \frac{i_{0,M}}{(\beta_M^a)^2}\exp\left(\frac{\Phi - \Phi_{0,M}}{\beta_M^a}\right) - \frac{i_{0,H}}{(\beta_H^c)^2}\exp\left(\frac{\Phi_{0,H} - \Phi}{\beta_H^c}\right) \quad (14)$$

In view of Equation 5, Equation 14 can be rewritten to give

$$\left(\frac{\partial^2 i}{\partial \Phi^2}\right)_{\Phi_{corr}} = i_{corr}\left(\frac{1}{(\beta_M^a)^2} - \frac{1}{(\beta_H^c)^2}\right) \quad (15)$$

The linearity condition of the current-potential curve requires the second derivative in Equation 15 to be zero, a condition that is satisfied only when the Tafel constants for the anodic and the cathodic curves are the same. Oldham and Mansfeld (1972) have shown that, in a general sense, linearity occurs only at one potential, Φ_L, given by

$$\Phi_L = \Phi_{corr} + \frac{2\beta_M^a \beta_H^c}{\beta_M^a + \beta_H^c}\ln\left(\frac{\beta_M^a}{\beta_H^c}\right) \quad (16)$$

It can be seen that when the Tafel slopes are equal, $\Phi_L = \Phi_{corr}$. According to Oldham and Mansfeld, for an iron-acid system with $\beta_M^a = 13$ mV and $\beta_H^c = 52$ mV, the potential at which linear behavior occurs is displaced by ~30 mV from Φ_{corr}.

Mansfeld (1974) has further examined the question of nonlinearity, especially to estimate errors arising from measurements using corrosion rate meters. In some of these meters, a value of $\Phi - \Phi_{corr}$ of 10 mV is applied and the corresponding corrosion currents Δi_+ and Δi_- measured. Assuming linear behavior, the corrosion current is estimated from the slope, $\Delta i/\Delta\Phi$. Let the corrosion current measured in this manner be given by

$$i'_{corr} = B\frac{\Delta i}{\Delta \Phi} \quad (17)$$

It can be shown that the error ε is given by

$$\varepsilon = 1 - \frac{i'_{corr}}{i_{corr}} = 1 - \frac{B}{\Delta\Phi}\left[\exp\left(\frac{\Delta\Phi}{\beta_M^a}\right) - \exp\left(-\frac{\Delta\Phi}{\beta_H^c}\right)\right] \quad (18)$$

where i_{corr} is the true corrosion current.

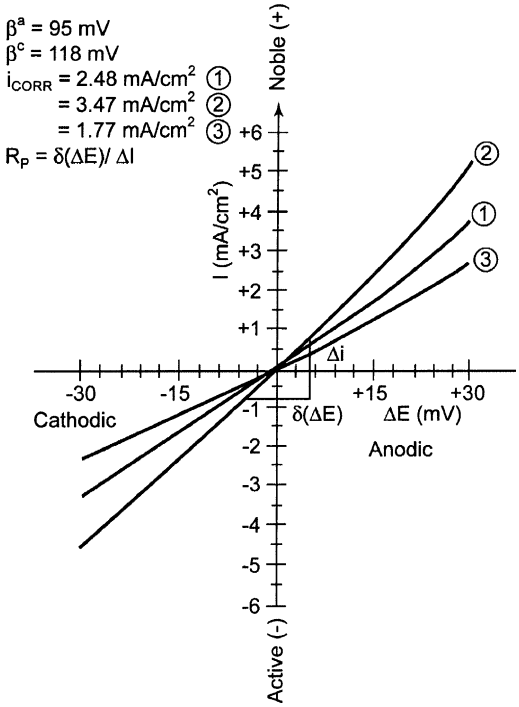


Figure 7. ASTM G 59 polarization curves for type 430 stainless steel in 1 N H_2SO_4 based on results from eight independent laboratories: curve 1, mean; curves 2 and 3, 95% confidence limits.

Practical Aspects of the Method

The experimental setup for carrying out linear polarization experiments is similar to that described under Tafel Technique above. The main difference is in the voltage range selected. The applied voltages used generally do not extend beyond ±10 mV of the corrosion potential. In this range, in many instances, the potential- current curve is linear. Similar to the Tafel method, software packages are now available from vendors such as Princeton Applied Research to enable the investigator to carry out measurements within a specified potential range using specified voltage steps.

The linear polarization technique has been widely used, and there are commercial monitoring probes available. The results shown in Figure 7 for the corrosion of 410 stainless steel in sulfuric acid provide a good example of

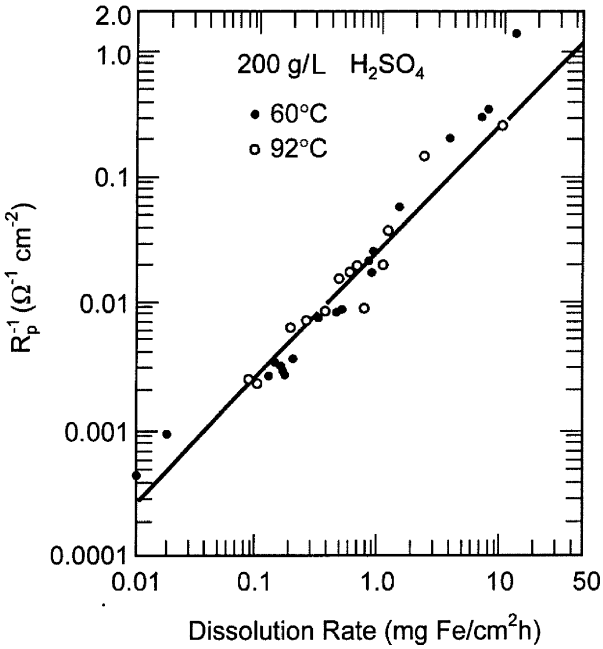


Figure 8. Inverse polarization resistance as a function of dissolution rate of iron for various inhibitor concentrations.

the application of this technique. Another example is given here from the study by Mennenoh and Engell (1962). This work was published within a few years of the exposition of the technique by Stern and Geary (1957). Polarization resistance measurements were used to investigate the effectiveness of corrosion inhibitors in pickling baths for steel. The proportionality of the polarization resistance R_p to corrosion rate was established. This proportionality is shown for steel in sulfuric acid medium containing various inhibitors in Figure 8. Investigations were carried out at two temperatures, 62° and 92°C; however, the influence of temperature does not appear to be significant. From this same study, a curve of the effect of inhibitor concentration on the inverse of the polarization resistance is shown in Figure 9. Other examples are available in detailed reviews (e.g., Mansfeld, 1976a).

Problems

The technique assumes that the corrosion rate is limited by charge-transfer reactions on the entire surface of the metal. Corrosion product films are presumed to be absent. Further, it is assumed there are no contributions from ohmic voltage drops.

Equations have been derived in the literature that take into account the influence of iR drop arising from the electrolyte resistance and the resistance of surface films (Mansfeld, 1976b). The use of iR drop compensation in the measurement circuit has also been employed by several investigators (Jones, 1968; Wilde, 1967; Walker and France, 1969; Cessna, 1971; Booman and Holbrook, 1963; Lauer and Osteryoung, 1966; Sawyer and Roberts, 1974). The influence of iR drop is evident from Figure 10, which shows iR compensated and uncompensated curves for Fe in 0.01 N HCl/EtOH.

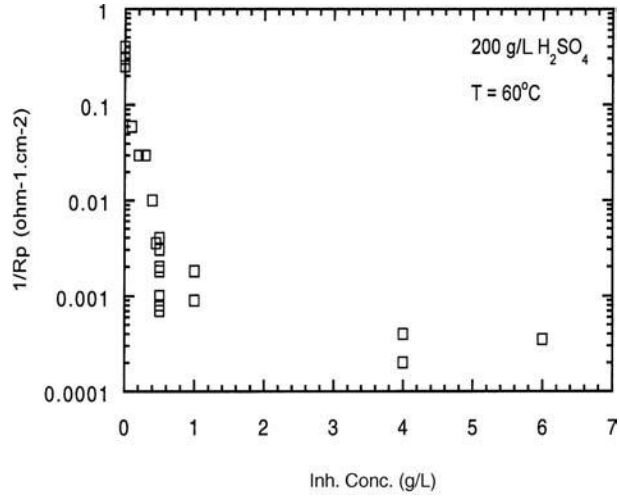


Figure 9. Dependence of polarization resistance on inhibitor concentration.

Another important limitation is that the Tafel slopes must be known in order to apply Equation 13 to determine the corrosion current. An independent measurement of the Tafel slopes using high applied potentials in the cathodic and anodic regimes can be done. However, one can introduce errors due to surface area–structure changes arising from the large current densities involved in the high-potential region. In fact, one advantage of the linear polarization technique is that the applied potentials are low and the corroding sample is minimally disturbed.

The need for independent measurement of Tafel slopes can be eliminated using the curve-fitting approach proposed by Mansfeld (1973). Equation 6 can be rewritten in terms of the Tafel slopes β_a and β_c as

$$i = i_{corr}\left[\exp\left(\frac{2.303\ \Delta\Phi}{\beta_a}\right) - \exp\left(-\frac{2.303\ \Delta\Phi)}{\beta_c}\right)\right] \qquad (19)$$

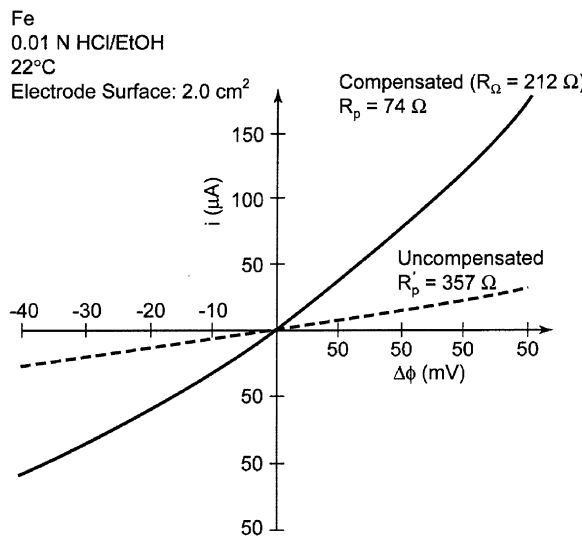


Figure 10. Experimental polarizarion curves for Fe–EtOH + HCl showing the effect of uncompensated iR drop.

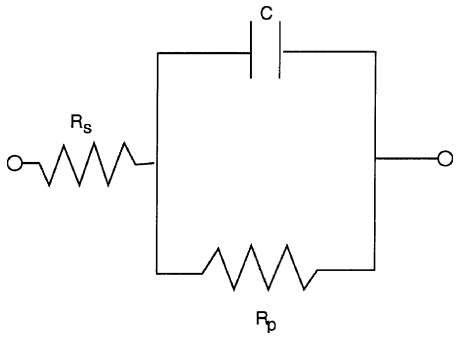


Figure 11. Simple equivalent circuit representation of an electrochemical cell.

Combining Equation 19 with the Stern-Geary relationship represented by Equation 13 and noting that $\Delta\Phi/\Delta i$ equals the polarization resistance R_p, one has

$$2.303R_p i_{corr} = \frac{\beta_a\beta_c}{\beta_a+\beta_c}\left[\exp\left(\frac{2.303\ \Delta\Phi}{\beta_a}\right) - \exp\left(-\frac{2.303\ \Delta\Phi}{\beta_c}\right)\right] \quad (20)$$

Since the right-hand side of Equation 20 essentially depends on Tafel slopes, various combinations can be tried until the best fit to the experimental curve is obtained.

ELECTROCHEMICAL IMPEDANCE SPECTROSCOPY

Principles of the Method

Electrochemical impedance spectroscopy (EIS) is an excellent approach in the determination of R_p described under Linear Polarization above. The technique, however, is not limited to charge-transfer reactions and is capable of simultaneously measuring different steps in a corrosion process. Such a step can include, for example, diffusional transport through a corrosion product film. Herein lies the power of this technique. It consists of applying a small-amplitude voltage perturbation, $V(\omega)$, to the working electrode at a number of different frequencies. At each frequency, the current response $i(\omega)$ will have a similar waveform to the voltage perturbation but will be out of phase with it by the phase angle Φ. The frequency-dependent electrochemical impedance $Z(\omega)$ is the proportionality constant that links the voltage signal to the current response. Thus,

$$Z(\omega) = \frac{V(\omega)}{i(\omega)} = |Z|\exp(j\phi) \quad (21)$$

In Equation 21, $j = \sqrt{-1}$.

Equation 21 can be rewritten in the form

$$Z(\omega) = Z(\mathrm{R}) + jZ(\mathrm{I}) \quad (22)$$

where $Z(\mathrm{R})$ and $Z(\mathrm{I})$ represent the real and imaginary components of the electrochemical impedance. Consider the equivalent circuit represented in Figure 11.

This is the simplest form of a corroding system, in which R_s represents the solution resistance, C is the capacitance of the double layer at the working electrode–solution interface, and R_p is the polarization resistance. The impedance diagram for such an electrochemical cell can be represented either by the Nyquist plot shown in Figure 12A or the Bode plot shown in Figure 12B. It is clear from the Nyquist representation that

$$|Z|^2 = Z(\mathrm{R})^2 + Z(\mathrm{I})^2 \quad (23)$$

Furthermore,

$$Z(\mathrm{R}) = |Z|\cos\phi \quad \text{and} \quad Z(\mathrm{I}) = |Z|\sin\phi \quad (24)$$

The electrochemical impedance $Z(\omega)$ is a fundamental electrochemical parameter. In EIS, a given electrochemical cell is described by a corresponding equivalent circuit. For the equivalent circuit shown in Figure 11,

$$Z(\omega)_{\omega\to 0} = R_s + R_p \quad (25)$$

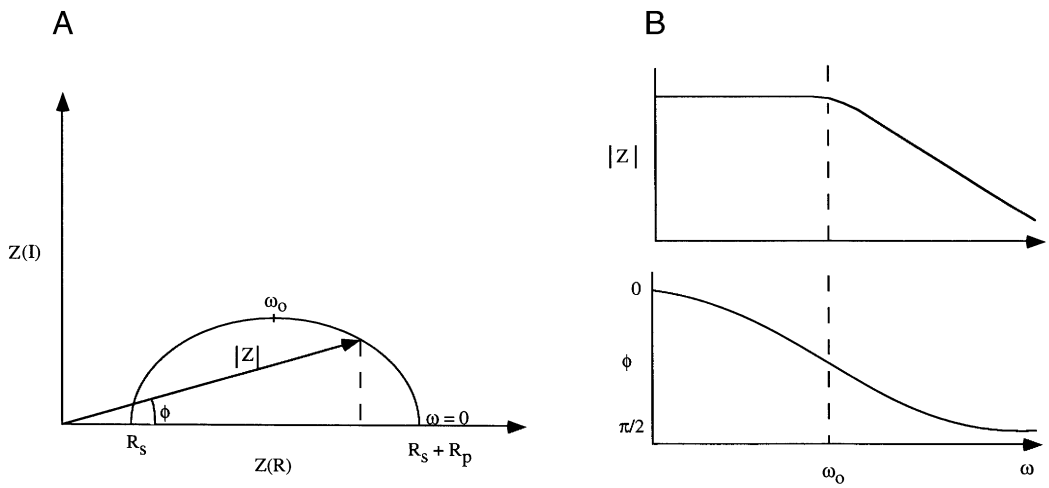


Figure 12. Nyquist (**A**) and Bode (**B**) representations of the equivalent's circuit in Figure 11.

As described before, the inverse of the polarization resistance R_p is proportional to the corrosion rate.

The polarization resistance is also related to the term ω_0 in the Nyquist plot. Thus

$$\omega_0 = 1/R_pC \tag{26}$$

The Bode representation is a combination of two plots, one showing the frequency dependence of the magnitude of the electrochemical impedance and the other the frequency dependence of the phase angle. The Nyquist and Bode representations are equivalent and either can be used. Excellent review articles on the use of EIS are available in the literature (Macdonald and McKubre, 1981; McKubre and Macdonald, 1977; Macdonald, 1977; Sluyters-Rehbach and Sluyters, 1970).

Practical Aspects of the Method

The electrochemical cell arrangement used in EIS is similar to that described for the other electrochemical techniques. The working, counter and reference electrodes are connected to an electronic potentiostat or electrochemical interface (obtainable from Princeton Applied Research or Solatron Instrumentation Group). A function generator is used to feed a wide range of potential frequencies to the potentiostat that applies the signal to the working electrode. To measure the in-phase and out-of-phase components of the cell response to the perturbating voltage signal, an appropriate detection or analyzer system must be used in conjunction with the potentiostat. A phase-sensitive detector or lock-in-amplifier allows the measurement of impedances with frequencies above ~1 Hz (de Levie and Husovsky, 1969; Armstrong and Henderson, 1972). Digital transfer function analyzers can be used for measurements in the range 0.1 to 10^4 Hz (Gabrielli and Keddam, 1974). As opposed to a frequency-by-frequency method, fast Fourier transform (FFT) allows a broad band of frequencies to be simultaneously applied. The acquisition, storage, analysis, and display of data are facilitated by the use of a computer. Software packages are available to carry out EIS measurements for corrosion quantification (Princeton Applied Research; Solatron Instrumentation Group).

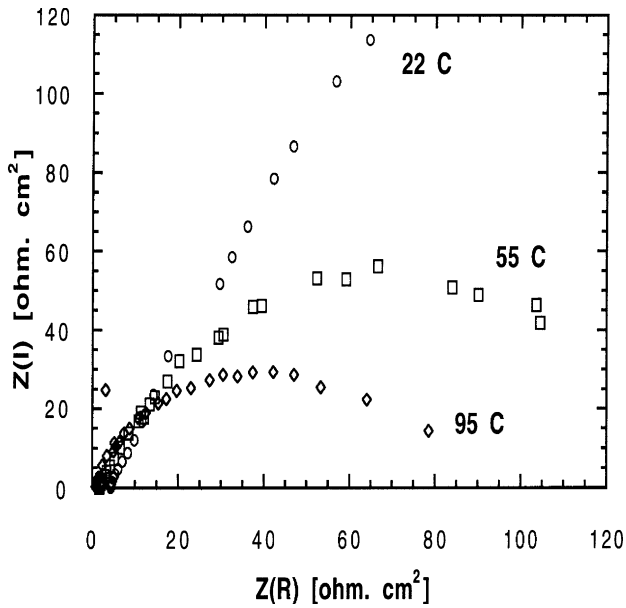


Figure 13. Temperature dependence of impedance data for the corrosion of 4130 steel in aqueous H_2S containing 3% NaCl at 0.08 h.

Table 1. Temperature dependence of impedance parameters

Temperature (°C)	ω_{max}(Hz)	R_D(Ω cm^2)	R_t(Ω cm^2)
22	0.3502	468	14.4
55	0.7007	147	8.8
75	0.8505	118	6.5
95	0.9009	79	5.0

Some examples are now described. The first is a study from the author's laboratory (Vedage et al., 1993). It deals with the corrosion of carbon steel (4130 as an example) in a 3% sodium chloride solution saturated with an 80:20 Ar-H_2S gas mixture. Electrochemical impedance measurements were carried out over the temperature range 22 to 95°C. A typical electrochemical cell used in this study is shown in Figure 2. A frequency range from 10^4 to 2.5×10^{-3} Hz was used in the measurements. The ac excitation potential used was 10 mV. A continuous flow of the 80:20 Ar-H_2S mixture was maintained through the stirred aqueous chloride solution during the measurements.

The impedance results as a function of temperature at a corrosion time of 0.08 h are summarized in Figure 13. A typical semicircular form is exhibited in the Nyquist representation of the data. At intermediate to low frequencies, <10 Hz, a second semicircle of diameter ~5 Ω cm^2 at 95°C could be delineated. The diameters of the larger and the smaller semicircles as well as the value of ω_0 as a function of temperature are shown in Table 1. Electron microscopic studies of the surface of carbon steel after corrosion showed a continuous layer of iron-deficient iron sulfide (pyrrhotite) on the surface. Experimentally, it was observed that the thickness of this layer increased as a function of the square root of time, in accordance with the well-known parabolic rate law (Wagner, 1933), which indicates that the corrosion process is controlled by solid-state diffusion through the iron sulfide reaction product.

The resistances corresponding to the diameter of the larger semicircle, R_D, can be traced to solid-state diffusion of iron atoms through iron-deficient iron sulfide. Those corresponding to the diameters of the smaller semicircles, R_p, are consistent with a charge transfer step at the steel-FeS interface. Thus, different steps in a corrosion process can be independently investigated.

The fact that R_D corresponds to diffusion through the FeS layer can be easily established as follows. According to the parabolic rate law,

$$X^2 = k_Dt \tag{27}$$

where X in the present example denotes the thickness of the FeS corrosion product and k_D is a diffusion-related constant. Differentiating Equation 27 with respect to time,

$$2X\frac{dX}{dt} = k_D \tag{28}$$

Since the corrosion process is supposed to be sulfide growth controlled, dX/dt is a measure of the corrosion rate or the corrosion current and is inversely proportional to R_D. If this proportionality constant is denoted by k, then

$$2\frac{kX}{R_D} = k_D \tag{29}$$

Combining Equations 27 and 29 yields

$$R_D^2 = \frac{4k^2}{k_D}t \tag{30}$$

Thus a plot of R_D^2 as a function of time should be linear, passing through the origin. This is demonstrated in Figure 14. Furthermore, within the small temperature range of the present study, the activation energy for iron diffusion through iron sulfide could be calculated using the results from impedance measurements. This gives a value of ~15 kcal/mol. The value quoted in the literature from high-tem perature diffusion measurements in iron sulfide falls in the range 13 to 30 kcal/mol.

Electrochemical impedance spectroscopy is also a powerful technique for the study of the effect of inhibitors on corrosion. In other studies, corrosion inhibitors were added to the H_2S containing sodium chloride solution (Ramanarayanan and Chiang, 1992). Results are presented as a plot of $1/R_D$ as a function of time in Figure 15.

Under steady-state conditions, the addition of sodium meta-vanadate at a concentration level of 0.005% provides

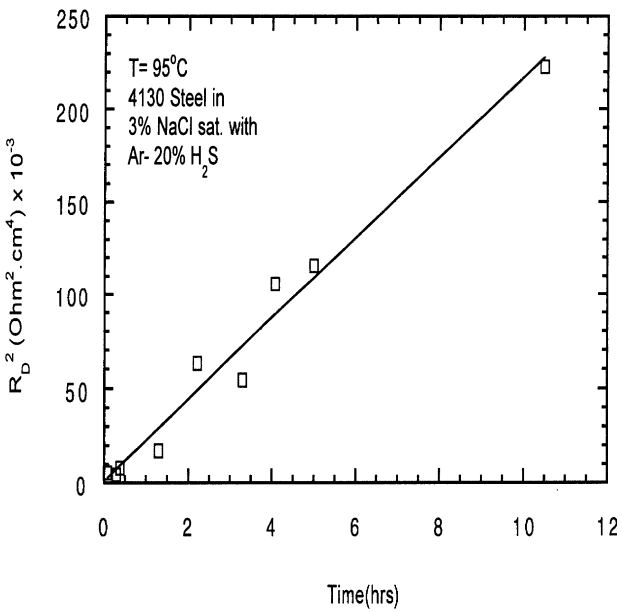


Figure 14. Sulfide growth-controlled corrosion of carbon steel in aqueous NaCl saturated with H_2S.

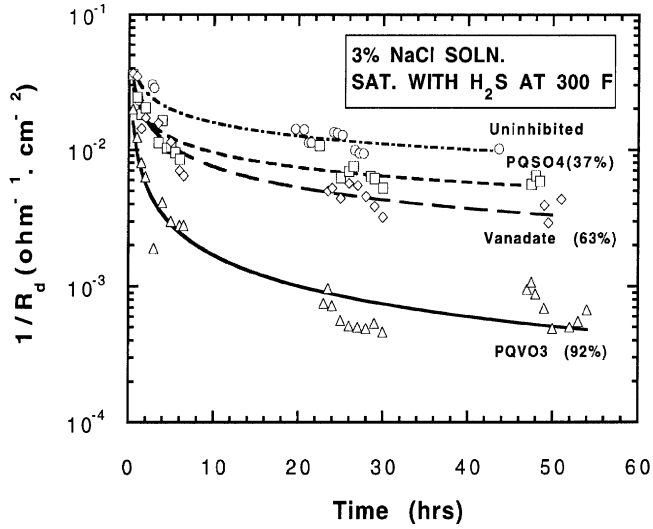


Figure 15. Effect of inhibitors on H_2S corrosion.

~60% corrosion inhibition. Another inhibitor, polyquinoline vanadate, leads to almost 90% inhibition. In these studies, the efficiency of the inhibitor (IE) given in terms of percentage inhibition can be expressed as

$$\text{IE} = \frac{(1/R_D)_{\text{uninh.}} - (1/R_D)_{\text{inh.}}}{(1/R_D)_{\text{uninh.}}} \times 100 \tag{31}$$

It is conceivable that the interaction of inhibitors with metal surfaces can be significantly affected by applied potentials. Thus, EIS with a minimal perturbation of ~10 mV is particularly suited to such studies.

Electrochemical impedance spectroscopy is a very popular technique, and investigations have been carried out by a number of authors on different corroding systems (Keddam et al., 1981; Epelboin et al., 1972, 1975; McKubre and Macdonald, 1981; Armstrong and Edmondson, 1973; de Levie, 1964, 1967).

Another example is presented that deals with the hot corrosion of metals. Hot corrosion is a high-temperature corrosion phenomenon in which the protective tendency of a surface oxide film on a metal-alloy surface is severely compromised by the presence of a molten salt that tends to dissolve the oxide film (Goebel et al., 1973; Bornstein and DeCrescente, 1971; Luthra, 1982; Luthra and Shores, 1980; Rapp, 1986). Hot-corrosion problems have been of major concern to the gas turbine industry, where high-temperature alloys in service have been attacked by a molten sodium sulfate film. A schematic of the hot-corrosion process is shown in Figure 16 (Rapp and Goto, 1981). On the surface of the metal is a thin oxide film, MO, in which the transport of ionic and electronic charger carriers can occur during corrosion.

According to Rapp and Goto, a negative solubility gradient for the surface oxide in the molten salt at the oxide-salt interface is a condition for the sustained hot corrosion of a metal. This condition may be written as

$$\left(\frac{dN_{MO}}{dX}\right)_{x=0} < 0 \tag{32}$$

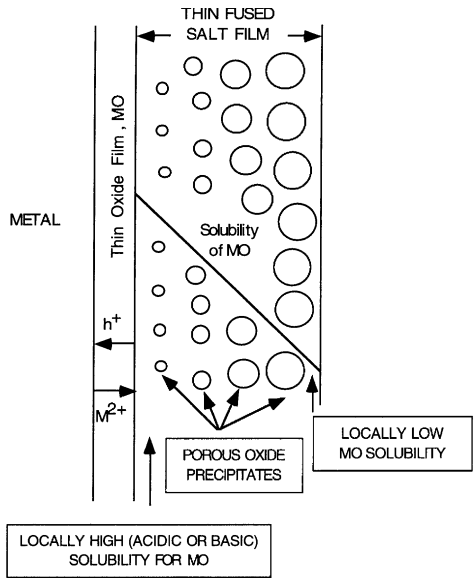


Figure 16. Schematic of the mechanism of hot corrosion of metals.

where N_{MO} is the solubility of the oxide MO in the salt film and $x = 0$ represents the oxide film–salt film interface. This condition allows the oxide to dissolve in the salt at $x = 0$ where the solubility is high and reprecipitate away from the interface where the solubility is low. What is the rationale for the existence of a negative solubility gradient? To answer this, let us look at the hot corrosion of Ni having a surface NiO film by a sodium sulfate film with a certain partial pressure of SO_3 gas at its outer surface. In principle, NiO can dissolve in sodium sulfate by two mechanisms: acid dissolution and basic dissolution. These two modes of dissolution are represented by the reactions

$$\underset{\text{(acidic solution)}}{NiO \rightarrow Ni^{2+}(\text{in salt}) + O^{2-}(\text{in salt})} \quad (33)$$

$$\underset{\text{(basic dissolution)}}{NiO + O^{2-}(\text{in salt}) \rightarrow NiO_2^{2-}(\text{in salt})} \quad (34)$$

In the case of NiO dissolution in sodium sulfate, basic dissolution is the preferred mode for hot corrosion. A measure of the basicity of the salt at any location is the prevailing concentration of oxide ions, O^{2-}. Thus NiO solubility will be higher where the concentration of O^{2-} is higher. The concentration of O^{2-} ions in the sodium sulfate melt is fixed by the equilibrium

$$SO_3 + O^{2-} = SO_4^{2-} \quad (35)$$

In view of the equilibrium of reaction 35, at the outer surface of the salt film, where the effective SO_3 partial pressure is high, the oxide ion concentration and therefore the basicity are low. At the oxide-salt interface, the SO_3 pressure is the lowest and therefore the basicity is the highest. This will give rise to a negative gradient in basicity and therefore NiO solubility at the oxide film–salt film interface.

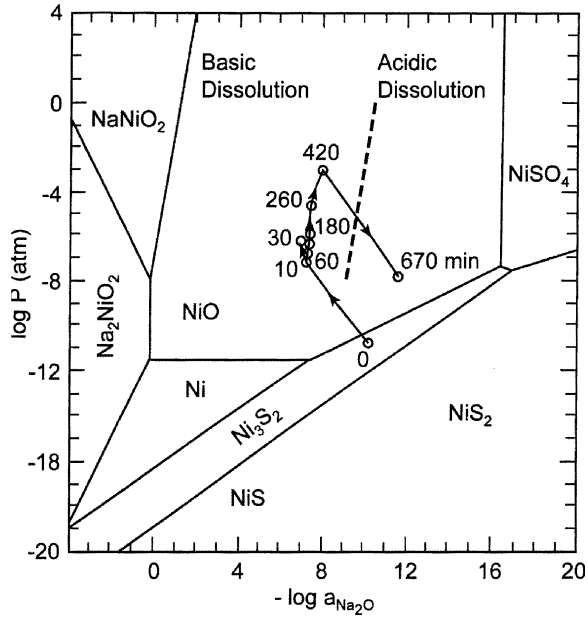


Figure 17. Trace of salt chemistry during the hot corrosion of Ni at 1200 K.

A limited number of studies have used EIS to investigate hot corrosion (Farrell et al., 1985; Gao et al., 1990; Wu and Rapp, 1991). Wu and Rapp applied this technique to study the hot corrosion of Ni in sodium sulfate melt at 1200 K. Amplitudes <10 mV were used in the frequency range 10^{-3} to 10^5 Hz. A nickel wire wound around a stabilized zirconia tube served as the working electrode. A sodium sulfate film of ~2 mm thickness, in equilibrium with a $O_2 + SO_2 + SO_3$ gas mixture, separated the Ni working electrode from a platinum foil counterelectrode. Two sets of reference electrodes, O_2-stabilized zirconia and Ag-Ag_2SO_4-fused silica, permitted the measurement of the oxygen partial pressure and the sodium oxide activity (basicity) at the corroding interface. A stability phase diagram of log P_{O_2} vs. log a_{Na_2O} at 1200 K is shown in Figure 17. The diagram delineates regimes of stability of different compounds. Also shown are the regimes of acid dissolution and basic dissolution. The trajectory from 0 to 670 min represents the path traversed during one impedance spectroscopy study of Ni precovered with a NiO film of ~0.8 μm thickness. Initially (0 min), the Ni surface is in the Ni_3S_2 stability regime. After 10 min of exposure, the metal surface shifts to a basic dissolution regime. The Nyquist plot corresponding to this condition is shown in Figure 18A. A lot of scatter can be observed in the low-frequency region. At 30 min, the surface is still in the basic dissolution regime (Fig. 18B). At 670 min, the corroding surface shifts to an acid dissolution regime (Fig. 18C). This is accompanied by a substantial decrease in the corrosion rate, as is suggested by the impedance curve corresponding to 670 min.

Problems

Powerful as the technique is, it is not always easy to use. Some of the problems and limitations have been summarized by Macdonald (1990). One limitation is the difficulty

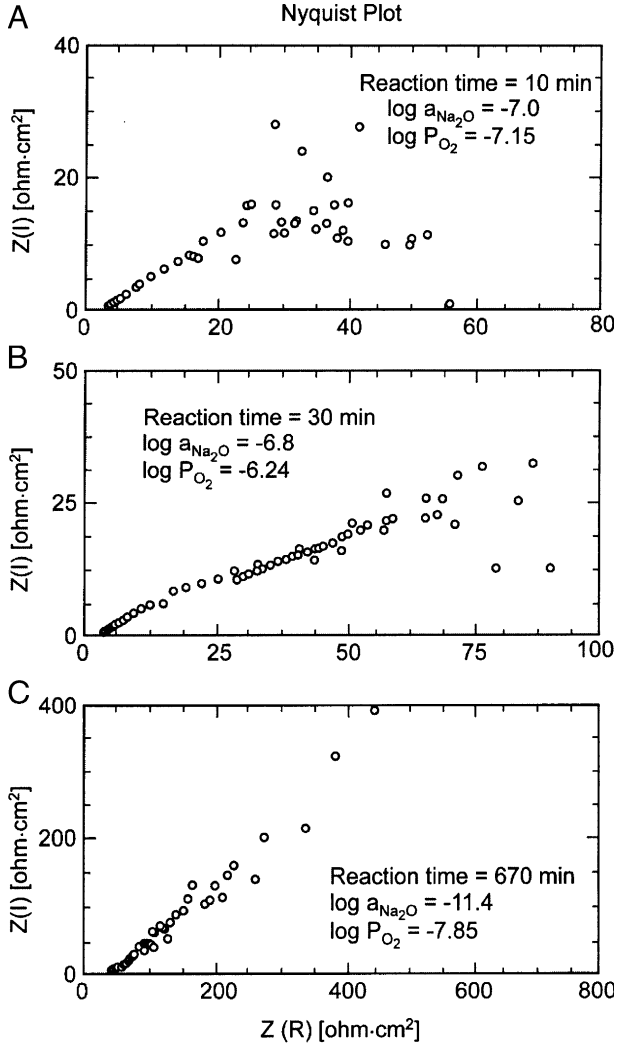


Figure 18. Impedance data for the hot corrosion of Ni in molten Na_2SO_4 after preoxidation in air at 1200 K for 5 min.

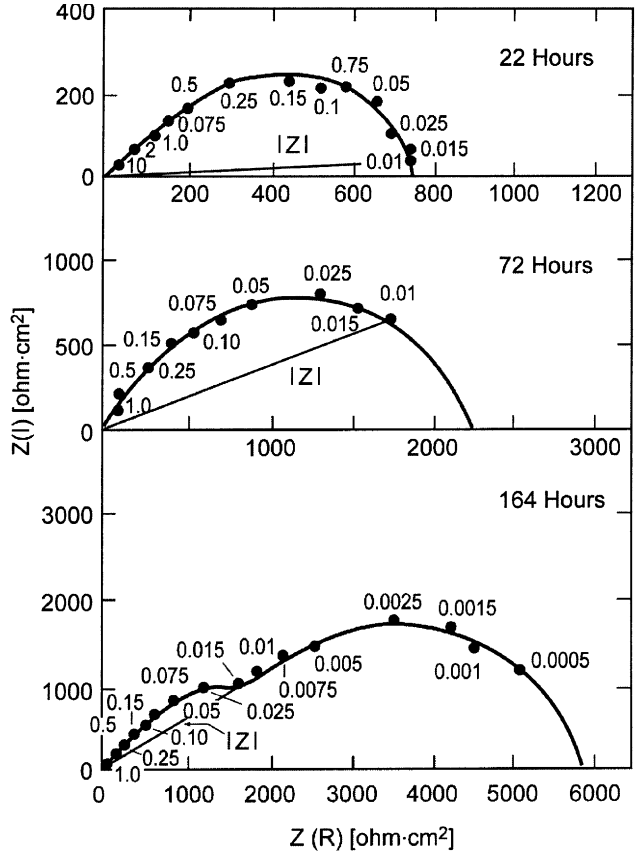


Figure 19. Impedance data for 90:10 Cu-Ni alloy in flowing sea water as a function of exposure time.

involved in acquiring a sufficient number of low-frequency data points in order to make a correct estimation of R_p. This problem can be demonstrated using data for a 90:10 Cu-Ni alloy (Syrett and Macdonald, 1979; Macdonald et al., 1978) undergoing corrosion in flowing sea water (Fig. 19). It can be seen that for an exposure time of 22 h, the semicircle is rather well defined. But at longer exposure times, even frequencies as low as 0.0005 Hz are insufficient to obtain a complete quantification of the interfacial impedance.

Acquisition of impedance data requires a finite amount of time depending on the system. In corroding systems where the initial rate of corrosion is very high, as is true in many cases, great care must be exercised in integrating $1/R_p$ over time in order to get an engineering value for the corrosion rate.

Porosity in surface films is another source of error in impedance measurements. Corrosion product films are often of variable porosity and of different sizes and shapes, and their contributions to the impedance spectrum are difficult to evaluate.

The accurate analysis of impedance data requires good correspondence between the corroding system and an appropriate electrical equivalent circuit. A complete correspondence is often difficult to achieve in practice.

LITERATURE CITED

Antropov, L. I., Gerasimenko, M. A., and Gerasimenko, Yu. S. 1966. *Prot. Met.* 2:98.

Armstrong, R. D. and Edmondson, K. 1973. *Electrochim Acta* 18:937.

Armstrong, R. D. and Henderson, M. 1972. *J. Electroanal. Chem.* 40:121.

Bandy, R. and Jones, D. A. 1976. *Corrosion* 32:126.

Barnartt, S. 1969. *Corrosion Sci.* 9:45.

Barnartt, S. 1971. *Corrosion* 27:467.

Bockris, J. O'M. and Reddy, A. K. N. 1970. Modern Electrochemistry, Vol. 2. Plenum, New York.

Booman, G. L. and Holbrook, W. B. 1963. *Anal. Chem.* 34:1793.

Bornstein, N. S. and DeCrescente, M. A. 1971. *Metall. Trans.* 2:2875.

Britz, D. 1978. *J. Electrochem. Soc.* 88:309.

Cessna, J. C. 1971. *Corrosion* 27:244.

Epelboin, I., Gabrielli, C., Keddam, M., Lestrade, J-C., and Takenouti, H. 1972. *J. Electrochem. Soc.* 119:1632.

Epelboin, I., Gabrielli, C., Keddam, M., and Takenouti, H. 1975. *Electrochim. Acta.* 20:913.

Farrell, D. M., Cox, W. M., Stott, F. H., Eden, D. A., Dawson, J. L., and Wood, G. C. 1985. *High Temp.Technol.* 3:15.

Fontana, M. G. and Greene, N. D. 1978. Corrosion Engineering. McGraw- Hill, New York.

Gao, G., Stott, F. H., Dawson, J. L., and Farrell, D. M. 1990. *Oxidation of Metals*. 33:79.

Gabrielli, C. and Keddam, M. 1974. *Electrochim. Acta*. 19:355.

Goebel, J. A., Pettit, F. S., and Goward, G. W. 1973. *Metall. Trans*. 4:261.

Hayes, M. and Kuhn, J. 1977–78. *J. Power Sources* 2:121.

Jones, D. A. 1968. *Corrosion Sci*. 8:19.

Jones, D. A. 1992. Principles and Prevention of Corrosion. Macmillan, New York.

Kaesche, H. 1985. Metallic Corrosion. NACE, Houston.

Keddam, M., Matto, O. R., and Takenouti, H. 1981. *J. Electrochem. Soc*. 128:257.

Lauer, G. and Osteryoung, R. A. 1966. *Anal. Chem*. 38:1106.

de Levie, R. 1964. *Electrochim. Acta*. 9:1231.

de Levie, R. 1967. *Adv. Electrochem. Electrochem. Eng*. 6:329.

de Levie, R., Husovsky, A. A. 1969. *J. Electroanal. Chem*. 20:181.

Luthra, K. L. 1982. *Metall. Trans*. 13A:1853 and 1943.

Luthra, K. L. and Shores, D. A. 1980. *J. Electrochem. Soc*. 127:2202.

Macdonald, D. D. 1977. Transient Techniques in Electrochemistry. Plenum, New York.

Macdonald, D. D. 1990. *Corrosion* 46:229.

Macdonald, D. D. and McKubre, M. C. H. 1981. Electrochemical impedance techniques in corrosion science. *In* Electrochemical Corrosion Testing, STP 272 (F. Mansfeld and U. Bertocci, eds.) pp. 110-149. American Society for Testing and Materials, Philadelphia.

Macdonald, D. D., Syrett, B. C., and Wing, S. S. 1978. *Corrosion* 34:289.

Mansfeld, F. 1973. *J. Electrochem. Soc*. 120:515.

Mansfeld, F. 1974. *Corrosion* 30:92.

Mansfeld, F. 1976a. The polarization resistance technique for measuring corrosion currents. *In* Advances in Corrosion Science and Technology (M. G. Fontana and R. W. Staehle, eds.) pp. 163–262. Plenum, New York.

Mansfeld, F. 1976b. *Corrosion* 32:143.

Mansfeld, F. 1982. *Corrosion* 38:556.

Mansfeld, F. and Oldham, K. B. 1971. *Corrosion Sci*. 11:787.

McKubre, M. C. H. and Macdonald, D. D. 1977. Electronic instrumentation for electrochemical studies. *In* A Comprehensive Treatrise of Electrochemistry (J. O'M. Bockris, B. E. Conway, and E. Yeager, eds.). Plenum, New York.

McKubre, M. C. H. and Macdonald, D. D. 1981. *J. Electrochem. Soc*. 128:524.

Mennenoh, S. and Engell, H. J. 1962. *Stahl Eisen.* 82:1796.

Oldham, K. B. and Mansfeld, F. 1971. *Corrosion* 27:434.

Oldham, K. B. and Mansfeld, F. 1972. *Corrosion* 28:180.

Poulson, B. 1983. *Corrosion Sci*. 23: 391.

Ramanarayanan, T. A. and Chiang, L. Y. 1992. U. S. Patent No. 5,158,693.

Ramanarayanan, T. A. and Smith, S. N. 1990. *Corrosion* 46:66.

Rapp, R. A. 1986. *Corrosion* 42:568.

Rapp, R. A. and Goto, K. S. 1981. *In* Proceedings of the Second International Symposium on Molten Salts (J. Braunstein and J. R. Selman, eds.) pp. 159–177. Electrochemical Society, Pennington, N.J.

Sawyer, D. T. and Roberts, Jr., J. L. 1974. Experimental Electrochemistry for Chemists. John Wiley & Sons, New York.

Scribner, L. L. and Taylor, S. R. 1990. The measurement and correction of electrolyte resistance in electrochemical tests, ASTM STP 1056. American Society for Testing and Materials, Philadelphia.

Scully, J. R. 1995. *In* Corrosion Tests and Standards (R. Baboian, ed.). American Society for Testing and Materials, Philadelphia.

Sluyters-Rehbach, M. and Sluyters, J. H. 1970. *In* Electroanalytical Chemistry, Vol. 4 (A. J. Bard, ed.) pp. 3–128. Marcel Dekker, New York.

Stern, M. 1958. *Corrosion* 14:440.

Stern, M. and Geary, A. L. 1957. *J. Electrochem. Soc.* 104:56.

Stern, M. and Roth, R. M. 1957. *J. Electrochem. Soc*. 104:390.

Syrett, B. C. and Macdonald, D. D. 1979. *Corrosion* 35:505.

Tafel, Z. 1904. *Z. Phys. Chem*. 50:641.

Uhlig, H. H. and Revie, R. W. 1985. Corrosion and Corrosion Control. John Wiley & Sons, New York.

Vedage, H., Ramanarayanan, T. A., Mumford, J. D., and Smith, S. N. 1993. *Corrosion* 49:114.

Wagner, C. 1933. *Z. Phys. Chem*. 21B:25.

Wagner, C. and Traud, W. 1938. *Z. Elektrochem.* 44:391.

Walker, M. S. and France Jr., W. D. 1969. *Mater. Protect*. 8:47.

Wilde, B. E. 1967. *Corrosion* 23:379.

Wu, Y. M. and Rapp, R. A. 1991. *J. Electrochem. Soc*. 138: 2683.

KEY REFERENCES

Mansfeld, 1976a. See above.

An excellent review of the polarization resistance approach for measuring corrosion currents. A good treatment of historical development, theory, and specific literature studies.

McKubre et al., 1977. See above.

A thorough discussion of electronic instrumentation and circuitry for electrochemical studies including corrosion measurements. It is particularly useful to the researcher who wants to make instrumentation changes to suit specific needs.

Rapp, 1986. See above.

Provides an excellent basis for the electrochemical aspects of high-temperature molten salt corrosion (hot corrosion).

Uhlig and Revie, 1985. See above.

A comprehensive treatment of corrosion and its control from both scientific and engineering perspectives that gives the reader an excellent overall perspective. The serious student will find this broad coverage to be very instructive and good intitial preparation for deeper focus on a specific aspect.

Wagner and Traud, 1938. See above.

Mixed potential theory is outlined. It is clarified that there need not exist spatially separated anodic sites and cathodic sites for corrosion to occur. Anodic and cathodic sites can statistically interchange, and both types of reactions can occur simultaneously. All that is needed for corrosion to occur is that the potential difference across the metal–electrolyte interface be between the equilibrium potentials for the anodic reaction and the cathodic reaction.

T. A. RAMANARAYANAN
Exxon Research and Engineering Company
Annandale, New Jersey

SEMICONDUCTOR PHOTOELECTROCHEMISTRY

INTRODUCTION

This unit discusses methods and experimental protocols in semiconductor electrochemistry. We first discuss the basic principles that govern the energetics and kinetics of charge flow at a semiconductor-liquid contact. The principal electrochemical techniques of photocurrent and photovoltage measurements used to obtain important interfacial energetic and kinetic quantities of such contacts are then described in detail.

After this basic description of concepts and methods in semiconductor electrochemistry, we describe methods for characterizing the optical, electrical, and chemical properties of semiconductors through use of the electrochemical properties of semiconductor-liquid interfaces. The latter part of this unit focuses on methods that provide information primarily on the properties of the semiconductor surface and on the semiconductor-liquid junction.

In some cases, the semiconductor-liquid junction provides a convenient method for measuring properties of the bulk semiconductor that can only be accessed with great difficulty through other techniques; in other cases, the semiconductor-liquid contact enables measurement of properties that cannot be determined using other methods. Due to the extensive amount of background material and the interdisciplinary nature of work in this field, the discussion is not intended to be exhaustive, and the references cited in the various protocols should be consulted for further information.

This unit will cover the following methods in semiconductor electrochemistry:

- Photoconductor/photovoltage measurements
- Measurement of semiconductor band gaps using semiconductor-liquid interfaces
- Diffusion length determination using semiconductor-liquid contacts
- Differential capacitance measurements of semiconductor-liquid contacts
- Transient decay dynamics of semiconductor-liquid contacts
- Measurement of surface recombination velocity using time-resolved microwave conductivity
- Electrochemical photocapacitance spectroscopy
- Laser spot scanning methods at semiconductor-liquid contacts
- Flat-band potential measurements of semiconductor-liquid interface
- Time-resolved photoluminescence spectroscopy to determine interfacial charge transfer kinetic parameters
- Steady-state *J-E* data to determine kinetic properties of semiconductor-liquid interfaces.

PHOTOCURRENT/PHOTOVOLTAGE MEASUREMENTS

Principles of the Method

Thermodynamics of Semiconductor-Liquid Junctions

Energetics of Semiconductor-Liquid Contacts. When a semiconductor is placed in contact with a liquid, interfacial charge transfer occurs until equilibrium is reached. The direction and magnitude of this charge flow are dictated by the relevant energetic properties of the semiconductor-liquid contact. As shown in Figure 1, the key energetic quantities of the semiconductor are the energy of the bottom of the conduction band, E_{cb}, the energy of the top of the valence band, E_{vb}, the Fermi level, E_F, and the band gap energy, E_g ($= E_{vb} - E_{cb}$).

The key energetic quantity of the liquid is its electrochemical potential, $E(A/A^-)$, defined by the redox couple formed from an electroactive electron acceptor species, A, and an electroactive donor species, A^-, present in the electrolyte phase. The electrochemical potential of the phase containing specific concentrations of A and A^- is related to the formal electrochemical potential $E^{0\prime}(A/A^-)$ of this redox system by the Nernst equation:

$$E(A/A^-) = E^{0\prime}(A/A^-) + \frac{kT}{n}\ln\frac{[A]}{[A^-]} \qquad (1)$$

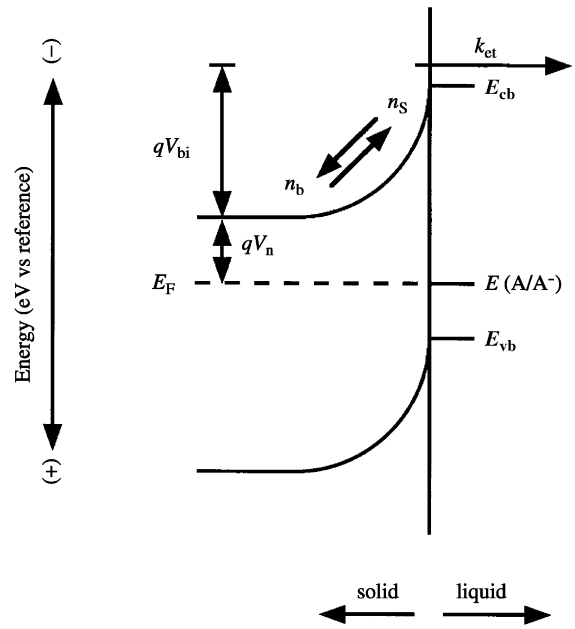


Figure 1. Energy of an *n*-type semiconductor-liquid junction under equilibrium conditions. At equilibrium the Fermi level of the semiconductor, E_F, is equal to the electrochemical potential of the solution. The surface electron concentration n_s is proportional to the bulk concentration of electrons n_b and the equilibrium built-in voltage V_{bi}. The energy difference between the Fermi level and the conduction band in the bulk is constant and equal to qV_n. The rate of charge transfer from the semiconductor to solution species is governed by the interfacial electron transfer rate constant k_{et}. Note the standard electrochemical sign convention, with positive energies representing more tightly bound electrons, is used throughout this unit.

where n is the number of electrons transferred, k is Boltzmann's constant, T is the absolute temperature, [A] is the concentration of acceptor species, and [A^-] is the concentration of donor species.

The doping level of the semiconductor is also important in determining the degree of interfacial charge transfer, because the dopant density determines the position of E_F in the semiconductor before contact with the electrolyte. For an n-type semiconductor and nondegenerate doping, the dopant density N_d is given by the Boltzmann-type relationship

$$N_d = N_c \exp(E_{cb} - E_F)kT \tag{2}$$

where N_c is the effective density of states in the conduction band of the semiconductor.

Equilibrium Charge Density, Electric Field, and Electric Potential Distributions at Semiconductor-Liquid Contacts. After contact between the semiconductor and the liquid, net charge will flow until the Fermi level is equal everywhere in both phases. At equilibrium, this charge flow will produce a spatially nonuniform charge density in the semiconductor and in the liquid. This nonzero charge density in both phases will, in turn, produce an electric field and an electric potential in the vicinity of the semiconductor-liquid contact.

For a given difference between E_{cb} and $E(A/A^-)$ at an n-type semiconductor-liquid contact, solution of Poisson's equation leads to the following well-known expressions for the charge density, Q, and the magnitudes of the electric field $\boldsymbol{E}$ and electric potential V as a function of distance x into the semiconductor:

$$Q = qN_d \qquad 0 \le x \le W \tag{3a}$$

$$Q = 0 \qquad x > W \tag{3b}$$

$$\mathbf{E}(x) = \frac{qN_d}{\varepsilon_s}(W - x) \qquad 0 \le x \le W \tag{4a}$$

$$\mathbf{E}(x) = 0 \qquad x > W \tag{4b}$$

$$V(x) = -\frac{qN_d}{2\varepsilon_s}(W - x)^2 \qquad 0 \le x \le W \tag{5}$$

$$W = \sqrt{\frac{2\varepsilon_s[E_{cb} + qV_n - E(A/A^-)]}{N_d q^2}} \tag{6}$$

Here, q is the absolute value of the charge on an electron, ε_s is the static dielectric constant of the semiconductor, W is the depletion width, and V_n is the potential difference between the Fermi level and the conduction band level in the bulk of the semiconductor (Fig. 1).

Equation 3 is reasonable because the dopants are present in relatively low concentration in the solid (perhaps 1 ppm or less), so essentially all of the dopants are ionized until a sufficient distance has been reached that the required electrical charge has been transferred across the solid-liquid interface. Equations 4 and 5 then follow directly from the charge density profile of Equation 3, once the value of W is known from the amount of charge transferred (Equation 6). Analogous equations for p-type semiconductor-liquid contacts can be obtained to relate the energetics of the valence band, E_{vb}, to $E(A/A^-)$. For a typical difference between $E(A/A^-)$ and E_{cb} of 1 eV, Equation 4 shows that the electric field near the surface of the semiconductor is $\approx 10^5$ V cm^{-1}.

This junction formation leads to diode-like behavior, in which charge carriers experience a large barrier to current flow in one direction across the interface but display an exponentially increasing current density as a voltage is applied to the system in order to produce a current flow in the opposite direction across the contact. Understanding the microscopic origin of this rectification and interpreting its properties in terms of the chemistry of the semiconductor-liquid contact of concern is the topic covered next.

Charge Transfer at Equilibrium. Interfacial charge transfer at a semiconductor-liquid interface can be represented by the following balanced chemical equation:

$$\text{Electron in solid} + \text{acceptor in solution} \rightleftharpoons \text{electron vacancy in solid} + \text{donor in solution} \tag{7}$$

When Equation 7 is obeyed, the current should depend linearly on the concentration of electrons near the semiconductor surface and on the concentration of acceptor ions that are available to capture charges at the semiconductor surface (Fig. 1). In a solution containing a redox couple A/A^-, the rate of direct electron transfer from an n-type semiconductor to the acceptor species A can therefore be expressed as

$$\text{Rate of electron injection into solution} = k_{et} n_s [A]_s \tag{8}$$

where k_{et} is the rate constant for the electron transfer, n_s is the surface concentration of electrons, and $[A]_s$ is the concentration of acceptors in the interfacial region near the semiconductor-liquid contact. The units of k_{et} are centimeters to the fourth power per second, because the rate of charge flow represents a flux of charges crossing the interface, with units of reciprocal centimeters squared per second, and the concentrations n_s and $[A]_s$ are expressed in units of reciprocal cubic centimeters.

The electron flux in the opposite direction, i.e., from the electron donors in the solution to the empty states in the conduction band of the semiconductor, can be described as

$$\text{Rate of electron transfer from solution} = k'_{et}[A^-]_s \tag{9}$$

In Equation 9, k'_{et} is the reverse reaction rate constant and $[A^-]_s$ is the concentration of donors in the interfacial region near the semiconductor-liquid contact. In this expression, the concentration of empty states in the conduction band of the semiconductor has been incorporated implicitly into the value of k'_{et}.

At equilibrium, the rates of Equations 8 and 9 must be equal. Denoting the equilibrium electron concentration at the semiconductor surface by the quantity n_{s0}, we obtain

$$k_{et} n_{s0} [A]_s = k'_{et} [A^-]_s \tag{10}$$

Away from equilibrium, the net rate of electron transfer into solution ($-dn/dt$) is simply the forward rate minus the reverse rate. From Equations 8 to 10, we then obtain the general relationship

$$-\frac{dn}{dt} = k_{et}[A]_s n_s - k'_{et}[A^-]_s \tag{11}$$

or

$$-\frac{dn}{dt} = k_{et}[A]_s(n_s - n_{s0}) \tag{12}$$

Dark Current–Potential Characteristics of Semiconductor-Liquid Junctions. Because the current density (i.e., the current I divided by the exposed area of the electrode A_s) is merely the electron transfer rate multiplied by the charge on an electron, the interfacial electron transfer current density can be written as

$$J = -q\left(\frac{dn}{dt}\right) = -C(n_s - n_{s0}) \tag{13}$$

where the constant $C = qk_{et}[A]_s$. The current density J is defined to be negative when a reduction occurs at the electrode surface. Therefore, when $n_s > n_{s0}$, a negative (reduction) current will flow, because the electrode will tend to donate electrons to the solution.

A useful form of this equation is

$$J = -Cn_{s0}\left(\frac{n_s}{n_{s0}} - 1\right) \tag{14}$$

To obtain explicitly the dependence of the current density on the potential applied across the solid-liquid interface, we must relate the electron concentration at the surface of a semiconductor to the electron concentration in the bulk. The surface electron concentration at equilibrium is given by

$$n_{s0} = n_b \exp\left(\frac{-qV_{bi}}{kT}\right) \tag{15}$$

where n_b is the concentration of electrons in the bulk of the semiconductor and V_{bi} is the built-in voltage, i.e., the potential dropped across the semiconductor at equilibrium. Similarly, when a potential E is applied to the semiconductor relative to the situation at equilibrium, the total voltage drop in the semiconductor depletion region is $V_{bi} + E$, so we obtain an analogous Boltzmann relationship away from equilibrium:

$$n_s = n_b \exp\left(\frac{-q(V_{bi} + E)}{kT}\right) \tag{16}$$

These equations represent the physical situation that the electron concentration at the semiconductor surface can be either increased or decreased through the use of an additional voltage. This applied potential controls the surface carrier concentration in the same fashion as the built-in voltage, so the same Boltzmann relationship applies.

Equations 15 and 16 lead to a simple expression for the variation in the surface electron concentration as a function of the applied potential:

$$\frac{n_s}{n_{s0}} = \exp\left(\frac{-qE}{kT}\right) \tag{17}$$

This makes sense, because any change in the voltage dropped across the solid should exponentially change the electron concentration at the semiconductor surface relative to its value at equilibrium.

Substituting Equation 17 into Equation 14, we obtain the desired relationship between the current density and the potential of a semiconductor-liquid junction:

$$J = -Cn_{s0}\left[\exp\left(\frac{-qE}{kT}\right) - 1\right] \tag{18}$$

This equation is the simple rate equation 14, which has been rewritten to emphasize the explicit dependence of the current density on E.

Equation 18 is often written with only one constant:

$$J = -J_0\left[\exp\left(\frac{-qE}{kT}\right) - 1\right] \tag{19}$$

where $J_0 = Cn_{s0}$. The parameter J_0 is called the exchange current density, because it is the value of the current density that is present at equilibrium. For convenience, J_0 is defined as a positive quantity. The parameter J_0 is clearly dependent on the value of the equilibrium surface electron concentration, because a smaller exchange current should flow at equilibrium if there are fewer electrons available to exchange with a particular solution.

The current density–potential (or J-E) characteristic described by Equations 18 and 19, where the current can flow predominantly in only one direction under an applied potential, is called "rectification." The rectification characteristic is typical of electrical diodes. Equations that have the form of Equations 18 and 19 are therefore generally called "diode equations."

Factors that Affect Current Density–Potential Properties of Semiconductor-Liquid Contacts. The dependence of the charge transfer rate on the solution redox potential is perhaps the most important experimental property of semiconductor electrodes. Regardless of the value of the redox potential of the solution, $E(A/A^-)$, the diode behavior of Equation 19 will be obeyed. Changes in $E(A/A^-)$, however, will produce different values of J_0, because J_0 depends on n_{s0}. These different exchange currents will produce a measurable change in the J-E behavior of the semiconductor-liquid contact.

For an n-type semiconductor, more positive redox potentials will yield smaller values of J_0 and will produce highly rectifying diode behavior. For p-type semiconductors, the

opposite behavior is expected, so that negative redox potentials should produce highly rectifying contacts while positive redox potentials should produce poorly rectifying contacts. Rectifying *J-E* behavior is required for efficient photoelectrochemical devices that use either *n*- or *p*-type semiconductors; thus, one goal in constructing semiconductor-liquid junctions is to ensure that chemical control is maintained over the *J-E* properties of the semiconductor-liquid contact.

Variations in redox potentials also give rise to changes in an experimental parameter known as the barrier height. The barrier height, ϕ_b, for an *n*-type semiconductor-liquid contact is the potential difference between the redox potential of the solution and the conduction band edge. The value of $q\phi_b$ additionally reflects the free energy associated with interfacial electron transfer. Using this parameter, the expression for n_{s0} can be rewritten as

$$n_{s0} = N_c \exp\left(\frac{-q\phi_b}{kT}\right) \tag{20}$$

The value of N_c is known for most semiconductors and is generally $\approx 10^{19}$ cm^{-3} (Sze, 1981).

We can now explicitly incorporate the barrier height into our diode expressions by substituting Equation 20 into the expression for J_0:

$$J_0 = CN_c \exp\left(\frac{-q\phi_b}{kT}\right) \tag{21}$$

Equation 21 indicates that a value of J_0 can be predicted if the values of both ϕ_b and C are known.

Although we have derived the diode behavior of a semiconductor-liquid junction by assuming that electron transfer is the important charge flow process across the interface, the diode equation is generally applicable to semiconductor-liquid devices even when other processes are rate limiting. The *J–E* relationships for other possible charge flow mechanisms, such as recombination of carriers at the surface and/or in the bulk of the semiconductor, almost all adopt the form of Equation 19 (Fonash, 1981). The major difference between the various mechanisms is the value of J_0 for each system. Mechanistic studies of semiconductor-liquid junctions therefore generally reduce to investigations of the factors that control J_0. Such studies also involve quantitative comparisons of the magnitude of J_0 with the value expected for a specific charge transport mechanism. These types of investigations have yielded a detailed level of understanding of many semiconductor-liquid interfaces. Recent reviews describing more details of this work have been written by Koval and Howard (1992), Lewis (1990), and Tan et al. (1994b).

Basic J-E Equations for Illuminated Semiconductor-Liquid Junctions. Photoelectrochemical experiments also often deal with the behavior of semiconductor electrodes under illumination. Illumination of a semiconductor with light above its band gap energy produces excess electron-hole pairs, and movement of these charge carriers produces a photocurrent and a photovoltage at the semiconductor-liquid contact.

Current Components at Illuminated Semiconductor-Liquid Junctions. The effects of illumination are relatively simple to incorporate into the *J-E* behavior of a semiconductor-liquid contact. The total current in such a system can be conceptually partitioned into two components: one that originates from majority carriers and one from minority carriers. Absorption of photons creates both majority carriers and minority carriers; therefore, increases in both current components are expected under illumination.

Majority Carrier Currents. The concentration of majority carriers generated by absorption of moderate-intensity light is usually small compared to the concentration of majority carriers that is obtained from thermal ionization of dopant atoms in the solid. This implies that such levels of illumination do not significantly perturb the majority carrier behavior either in the semiconductor or at the semiconductor-liquid interface. Because the majority carrier concentrations are essentially unchanged, the rate equations that govern majority carrier charge flow also are unchanged. Majority carriers should thus exhibit a *J-E* characteristic that is well described by the diode equation, regardless of whether the semiconductor is in the dark or is exposed to moderate levels of illumination.

Minority Carrier Currents. Unlike the situation for majority carriers, illumination generally effects a substantial change in the concentration of minority carriers. Calculation of the minority carrier current is greatly simplified by considering the effects of the electric field at the semiconductor-liquid junction. For most semiconductor-liquid junctions in depletion, the electric field is so strong that essentially all of the photogenerated minority carriers are separated from the photogenerated majority carriers and then collected. Using this approximation, the photogenerated minority carrier current density J_{ph} is simply equal to the photon flux absorbed by the semiconductor multiplied by the charge on an electron, q.

Total Current Under Illumination. The total current density–potential characteristics of an illuminated semiconductor electrode can thus be obtained by adding together, with the appropriate sign, the majority and minority carrier components of the current density. The majority carrier current density obeys the diode equation, while the minority carrier photocurrent density is related to the absorbed light intensity. The expression for the total current density is therefore

$$J = J_{ph} - J_0\left[\exp\left(\frac{-qE}{kT}\right) - 1\right] \tag{22}$$

The sign of the minority carrier current (photocurrent) density is opposite to that of the majority carrier current density, because holes crossing the interface lead to an oxidation current, while electrons crossing the interface lead to a reduction current. Equation 22 is obviously just the diode curve of Equation 19 offset by a constant amount J_{ph} over the voltage range of interest (Fig. 2).

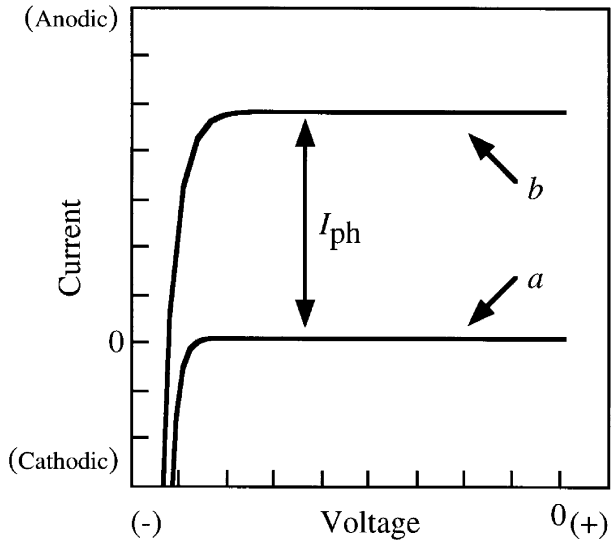


Figure 2. Ideal current-voltage behavior of a semiconductor-liquid junction (*a*) in the dark and (*b*) under illumination. The current observed under illumination is offset from the current observed in the dark by the value of the photocurrent, I_{ph}.

Properties of Photocurrent Density–Potential Behavior of Semiconductor-Liquid Junctions. For $J_{ph} > J_0$, as is generally the case, the 1 in Equation 22 can be neglected. We then obtain

$$J \approx J_{ph} - J_0\left[\exp\left(\frac{-qE}{kT}\right)\right] \tag{23}$$

We can then define the open-circuit voltage V_{oc} as the absolute value of the voltage present when no net current flows and obtain

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{J_{ph}}{J_0}\right) \tag{24}$$

This voltage is significant in the field of solar energy conversion, as it represents the maximum free energy that can be extracted from a semiconductor-liquid interface.

Equation 24 brings out several important features of the open-circuit voltage. First, V_{oc} increases logarithmically with the light intensity, because J_{ph} is linearly proportional to the absorbed photon flux. Second, the open-circuit voltage of a system increases (logarithmically) as J_0 decreases. Chemically, such behavior is reasonable, because J_0 represents the tendency for the system to return to charge transfer equilibrium. Third, Equation 24 emphasizes that a mechanistic understanding of J_0 is crucial to controlling V_{oc}. Only through changes in J_0 can systematic, chemical control of V_{oc} be established for different types of semiconductor-liquid junctions.

Another parameter that is often used to describe illuminated semiconductor-liquid junctions is the short-circuit photocurrent density J_{sc}. Short-circuit conditions imply $V = 0$. From Equation 22, the net current density at short circuit (J_{sc}) equals J_{ph}. The short-circuit current density provides a measure of the collection efficiency of photogenerated carriers in a particular photoelectrochemical cell. The reader is referred to earlier reviews for a more extensive discussion of how these parameters are relevant in solar energy research using semiconductor-based photoelectrochemical cells (Lewis, 1990; Tan et al., 1994b).

Practical Aspects of the Method

Basic Electrochemical Cell Design. Current density–potential data for semiconductor electrodes are typically obtained using a potentiostat (Fig. 3). This instrument ensures that the measured *J-E* properties are characteristic of the semiconductor-liquid interface and not of the counter electrode–liquid contact that is needed to complete the electrical circuit in the electrochemical cell. A three-electrode arrangement consisting of a semiconductor (working) electrode, a counter electrode, and a reference electrode is typically used to acquire data. The potentiostat uses feedback circuitry and applies the voltage needed between the working electrode and counter electrode to obtain the desired potential difference between the working and reference electrodes. The potentiostat then records the current flowing through the working electrode–counter electrode circuit at this specific applied potential. Nominally, no current flows through the reference electrode, which only acts as a point of reference for the system. The scan rate should be 50 mV s^{-1} or slower in order to minimize hysteresis arising from diffusion of species to the electrode surface during the *J-E* scan.

The electrochemical data are collected directly as the current vs. the applied potential. Electrode areas are, of course, needed to obtain current densities from the measured values of the current. The projected geometric area of the electrode is usually obtained by photographing the electrode and a microruler simultaneously under a microscope and digitally integrating the area defined by the exposed semiconductor surface.

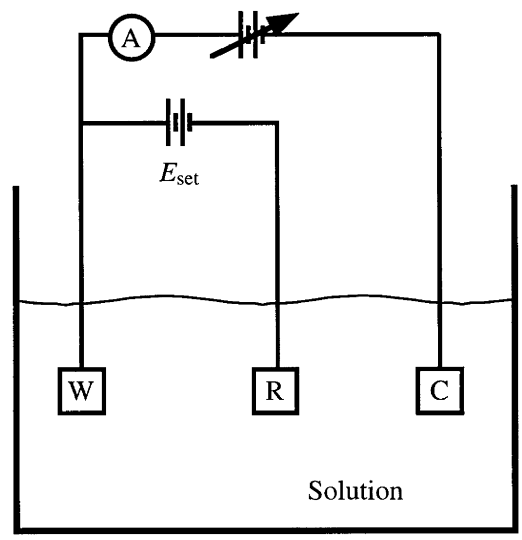


Figure 3. Circuit consisting of a simple potentiostat and an electrochemical cell. A potential is set between the working and reference electrode, and the current flow from the counter electrode to the working electrode is measured.

Reference Electrodes. Reference electrodes are constructed according to conventional electrochemical protocols. For example, two types of reference electrodes are an aqueous (or nonaqueous) saturated calomel electrode (SCE) and a nonaqueous ferrocenium-ferrocene electrode.

A simple SCE can be constructed by first sealing a platinum wire through one leg of an H-shaped hollow glass structure. The platinum wire is then covered with mercury, and a ground mixture of approximately equal amounts of mercury and calomel (Hg_2Cl_2) dispersed into a small amount of saturated potassium chloride solution is then placed on top of the mercury. The remainder of the tube is filled with saturated potassium chloride solution, and the other leg of the structure, which contacts the solution, is capped with a fritted plug. Prior to use, the nonaqueous SCE should be calibrated against a reference electrode with a known potential, such as an aqueous SCE prepared in the same fashion.

For work in nonaqueous solvents, a convenient reference electrode is the ferrocenium-ferrocene reference. This electrode consists of a glass tube with a fritted plug at the bottom. The tube is filled with a ferrocene-ferrocenium-electrolyte solution made using the same solvent and electrolyte that is to be used in the electrochemical experiment. A platinum wire is inserted into the top of the solution to provide for a stable reference potential measurement.

When both forms of the redox couple are present in the electrochemical cell, an even simpler procedure can be used to construct a reference electrode. A platinum wire can be inserted into the electrolyte–redox couple solution or into a Luggin capillary that is filled with the electrolyte–redox couple solution (see Luggin capillaries, below). This wire then provides a stable reference potential that is equal to the Nernstian potential of the electrochemical cell. At any convenient time, the potential of this reference can be determined vs. another reference electrode, such as an SCE, through insertion of the SCE into the cell. This approach not only is convenient but also is useful when water and air exposure is to be minimized, as is the case for reactive semiconductor surfaces in contact with deoxygenated nonaqueous solvents.

Luggin Capillaries. The cell resistance can be reduced by minimizing the distance between the working and reference electrodes. These small distances can be achieved through the use of a Luggin capillary as a reference electrode. The orifice diameter of the capillary should generally be $\sim$0.1 mm. A convenient method to form such a structure is to pull a disposable laboratory pipette under a flame and then to use a caliper to measure and then break the pipette glass at the point that corresponds to the desired orifice radius. The pipette is then filled with the reference electrode solution of interest, and the flow of electrolyte out of the pipette is minimized by capping the top of the pipette with a rubber septum. The contact wire is then inserted through the septum and into the electrolyte. Under some conditions, a syringe needle connected to an empty syringe can be inserted through the septum to facilitate manipulation of the pressure in the head space of the pipette. This procedure can be used to minimize mixing between the solution in the pipette and the solution in the electrochemical cell.

Illumination of Semiconductor-Liquid Contacts

Monochromatic Illumination. Low-intensity monochromatic illumination can be obtained readily from a white light source and a monochromator. This is useful for obtaining spectral response data to measure the diffusion length or the optical properties of the semiconductor electrode, as described in more detail under Measurement of Semiconductor Band Gaps Using Semiconductor-Liquid Interfaces.

Laser illumination can also be used to provide monochromatic illumination. However, care should be taken to diffuse the beam such that the entire electrode surface is as uniformly illuminated as possible. Because the photovoltage is a property of the incident light intensity, careful measurement of the photovoltage requires maintaining a uniform light intensity across the entire electrode surface. This protocol has not been adhered to in numerous measurements of the *J-E* properties of semiconductor electrodes, and the photovoltages quoted in such investigations are therefore approximate values at best. To control the light intensity from the laser, neutral density filters can be used to attenuate the incident beam before it strikes the electrode surface.

Regardless of whether the monochromatic light is obtained from a white light–monochromator combination or from a laser, measurement of the incident photon power is readily achieved with pyranometers, photodiodes, thermopiles, or other photon detectors that are calibrated in their response at the wavelengths of interest.

Polychromatic Illumination. For polychromatic illumination, solar simulators provide the most reproducible laboratory method for measuring *J-E* properties under standard, "solar-simulated" illumination. A less expensive method is to use tungsten-halogen ELH-type projector bulb lamps. However, their intensity-wavelength profile, like that of almost any laboratory light source, is not very well matched to the solar spectrum observed at the surface of the earth. Calibration of the light intensity produced by this type of source should not be done with a spectrally flat device such as a thermopile. Since laboratory sources typically produce more photons in the visible spectral region than does the sun at the same total illumination power, maintaining a constant power from both illumination sources tends to yield higher photocurrents, thus producing overestimates in efficiency of photoelectrochemical cells in the laboratory, relative to their true performance under an actual solar spectral distribution in the field.

An acceptable measurement method instead involves calibration of the effective incident power produced by the laboratory source through use of a photodetector whose spectral response characteristics are very similar to that of the photoelectrochemical cell of concern. Preferably, the response properties of the photodetector are linear with light intensity and the absolute response of the detector is known under a standard solar spectral

distribution and illumination power. The absolute detector response can be obtained either by measurements performed under a calibrated solar simulator or by measurement of the output of the detector in actual sunlight. If sunlight is used, another calibration is then required to determine the actual solar power striking the plane of the detector at the time of the measurement. Useful primary or secondary reference detectors for this purpose are silicon cells that have been calibrated on balloon flights by NASA. Spectrally flat radiometers, such as those produced by Eppley for the purpose of measuring the absolute spectral power striking a specific location on the surface of the earth, are also useful for determining the solar power under conditions used to calibrate the detector to be used in the laboratory. If these primary or secondary reference detectors are routinely available, it is of course also possible to determine the *J-E* properties of the photoelectrochemical cell directly in sunlight, as opposed to having to establish a reference detector measurement and then recalibrate a laboratory light source to produce the equivalent effective spectral power for the purposes of the measurement.

Effects of Cell Configuration. Under potentiostatic control, the concentrations of both forms of the redox species need not be as high as might be required to sustain identical performance in an actual, field-operating photoelectrochemical cell. This occurs because a two-electrode photovoltaic-type cell configuration requires sufficiently high concentrations of both forms of the redox couple dissolved in the solution to suppress mass transport limitations without mechanical stirring of the electrolyte. In a three-electrode cell with an *n*-type semiconductor electrode, the primary consideration is that sufficient redox donor be present such that the anodic current is limited by the light intensity and not by the mass transport of donor to the electrode surface. A high concentration of redox acceptor is not required to achieve electrode stability and often is undesirable when the oxidized form of the redox material absorbs significantly in the visible region of the spectrum. The concentration overpotential that results from a low concentration of electron acceptor in the electrolyte can be assessed and corrected for analytically using Equation 26 below.

In contrast, the performance of an actual energy conversion device using a two-electrode cell configuration is so dependent on the properties of the working electrode, the counter electrode, the electrolyte, the cell thickness, and the properties of the various optical interfaces in the device that many design trade-offs are involved and are unique to a particular cell configuration used in the device assessment. Emphasis here has been placed on determining the properties of the semiconductor electrode in "isolation," using a potentiostat, so that a comparison from electrode to electrode can be performed without considering the details of the device configuration used in each measurement.

Data Analysis and Initial Interpretation

Typical Raw Current–Potential Data. A representative example of a current-potential curve is shown in Figure 4,

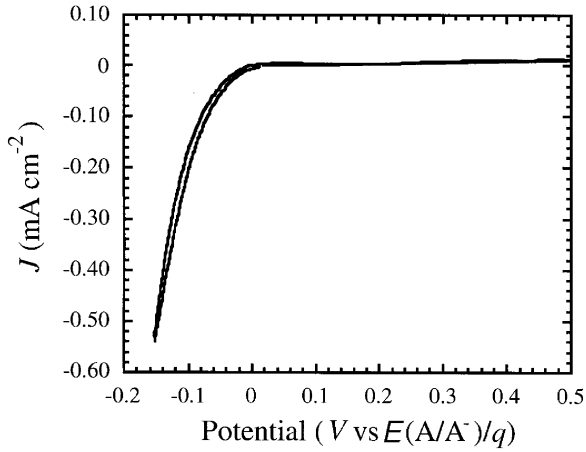


Figure 4. Representative example of current-voltage data for a semiconductor-liquid interface. The system consists of a silicon electrode in contact with a methanol solution containing lithium chloride and oxidized and reduced forms of benzyl viologen. In this example, the current has been divided by the surface area of the electrode, yielding a current density as the ordinate. The curve has not been corrected for cell resistance or concentration overpotential.

which displays data for an *n*-type silicon electrode in contact with a redox-active methanol solution. In this figure, the current has been divided by the surface area of the electrode to allow for quantitative analysis of the data. To extract meaningful results from the current-potential curve, it is also necessary to perform corrections for concentration overpotential and solution resistance, as discussed below.

Corrections to J-E Behavior: Concentration Overpotentials. Attention to electrochemical cell design is critical to minimize concentration overpotentials, mass transport restrictions on the available current density, and uncompensated resistance drops between the working and reference electrodes. Even with good cell design, in nonaqueous solvents the *J-E* curves must generally be corrected for concentration overpotential losses as well as for uncompensated ohmic resistance losses to obtain the inherent behavior of the semiconductor-liquid contact.

To minimize mass transport limitations on the current, the electrolyte should be vigorously stirred during the *J-E* measurement. For a given redox solution, the limiting anodic current density $J_{l,a}$ and the limiting cathodic current density $J_{l,c}$ should be determined using a platinum foil electrode placed in exactly the same configuration as the semiconductor working electrode. The areas of the two electrodes should also be comparable. If the redox couple is known to be electrochemically reversible, the platinum-electrode data can then be used to obtain the cell parameters needed to perform the necessary corrections to the *J-E* data of the semiconductor electrode (see Steady-State *J-E* Data to Determine Kinetic Properties of Semiconductor-Liquid Interfaces).

Alternatively, the semiconductor electrode can be fabricated into a disk configuration, and can be rotated in the electrolyte. Under these conditions, the mass transport

parameters E can be determined analytically, and the limiting current density (Bard et al., 1980) is

$$J_{1,c} = 0.620nFD_0^{2/3}\omega_{rde}^{1/2}\upsilon^{-1/6}[A]_b \qquad (25)$$

where F is Faraday's constant, D_0 is the diffusion coefficient, ω_{rde} is the angular velocity of the electrode, υ is the kinematic velocity of the solution ($\approx$0.01 $cm^2\ s^{-1}$ for dilute aqueous solutions near 20°C), and $[A]_b$ is the bulk concentration of oxidized acceptor species. A similar equation yields the limiting anodic current density based on the parameters for the reduced form of the redox species. This procedure allows control over the limiting current densities instead of merely measuring their values in a mechanically stirred electrolyte solution.

Laminar flow typically ceases to exist above Reynolds numbers (defined as the product of ω and the disk radius of the electrode divided by υ) of 2×10^5 (Bard et al., 1980), so for electrode radii of 1 mm, this corresponds to a typical upper limit on the rotation velocity of 1 to 2×10^7 rpm. Beyond this limit, Equation 25 does not describe the mass transport to the electrode. Smaller electrodes can increase this limit on ω, but use of smaller electrodes is generally not advisable, because edge effects become important and can distort the measured electrochemical properties of the solid-liquid contact by hindering diffusion of minority carriers and allowing recombination at the edges of the semiconductor crystal.

Once the limiting current densities and the *J-E* data are collected for a reversible redox system at a metal electrode, the concentration overpotential η_{conc} can be determined (Bard et al., 1980):

$$\eta_{conc} = \frac{kT}{nq}\left[\ln\left(\frac{J_{1,a}}{-J_{1,c}}\right) - \ln\left(\frac{J_{1,a}-J}{J-J_{1,c}}\right)\right] \qquad (26)$$

These values can then be used to correct the data at a semiconductor electrode to yield the proper *J-E* dependence of the solid-liquid contact in the absence of such concentration overpotentials.

Corrections to J-E Behavior: Series Resistance Overpotentials. Even with good cell design, measurement of the cell resistance R_{soln} is required to perform another correction to the *J-E* data. Values for R_{soln} can be extracted from the real component of the impedance in the high-frequency limits of Nyquist plots ELECTROCHEMICAL TECHNIQUES FOR CORROSION QUANTIFICATION for the semiconductor electrode or can be determined from steady-state measurements of the ohmic polarization losses of a known redox couple at the platinum electrode. In the former method, R_{soln} is simply taken as the real part of the impedance in the high-frequency limit of the Nyquist plot. In the latter method, the current-potential properties of a platinum electrode are determined under conditions where the platinum electrode is in an identical location to that of the semiconductor electrode. After correction of the data for concentration polarization, R_{soln} can be obtained from the inverse slope of the platinum current-potential data near the equilibrium potential of the solution.

The final corrected potential E_{corr} is then calculated from E, the concentration overpotential η_{conc}, R_{soln}, and the current I using (Bard et al., 1980)

$$E_{corr} = E - \eta_{conc} - IR_{soln} \qquad (27)$$

The measured value of I is divided by the projected geometric area of the electrode and plotted vs. E_{corr} to obtain a plot of the *J-E* behavior of the desired semiconductor-liquid contact.

Measurement of J_{sc} and V_{oc} of Illuminated Semiconductor-Liquid Contacts. The open-circuit photovoltage and short-circuit photocurrent should be measured directly using four-digit voltmeters connected to the photoelectrochemical cell as opposed to estimating their values from a time-dependent scan of the *J-E* data. This steady-state measurement eliminates any bias that might arise due to the presence of hysteresis in the current-potential behavior. Also, in some cases, the light-limited photocurrent is not reached at short circuit; in this case, both the light-limited photocurrent value and the short-circuit photocurrent value are of experimental interest and should be measured separately.

Sample Preparation

Electrodes for photoelectrochemical measurements should be constructed to allow exposure of the front face of the semiconductor to the solution while providing concealment of the back contact and of the edges of the electrode. This is readily accomplished using an insulating material that is inert toward both the etchant and the working solution of interest. The area of the electrode should be large enough to allow ready measurement of the bulk surface area but should be small enough to limit the total current flowing through the electrochemical cell (because larger currents require larger corrections for the cell resistance). Because of these trade-offs, electrode areas are typically 0.1 to 1 cm^2.

Ohmic contacts vary widely between semiconductors, and several books are available for identifying the ohmic contact of choice for a given semiconductor (Willardson et al., 1981; Pleskov et al., 1986; Finklea, 1988). Although most ohmic contacts are prepared by evaporating or sputtering a metal on the back surface of the semiconductor, some semiconductors are amenable to more convenient methods such as using a scribe to rub a gallium-indium eutectic on the back surface of the solid. This latter procedure is commonly used to make an ohmic contact to *n*-type silicon. The quality of an ohmic contact can be verified by making two contacts, separated by a contact-free region, on one side of an electrode and confirming that there is only a slight resistance between these contacts as measured by a *J-E* curve collected between these contact points.

The proper choice of a chemical etch depends on the semiconductor, its orientation, and the desired surface properties. Generally, an ideal etchant produces an atomically smooth surface with no electrical surface defects. Fluoride-based etches are most commonly used with

silicon: a 40% (w/w) ammonium fluoride solution is well suited for (111)-oriented Si and a solution of HF is appropriate for (100)-oriented Si (Higashi et al., 1991, 1993). For many III to V semiconductors, an etch in 0.05% (v/v) Br_2 followed by a rinse in a solution of NH_4OH produces abrupt discontinuities in the dielectric at the solid-air interface (Aspnes et al., 1981). An exhaustive literature provides information on additional etches for these and other semiconductors, and the reader is referred to these references for further information (Wilson et al., 1979; Aspnes et al., 1981; Higashi et al., 1993). There are also several published reports of effective dry etching methods (Higashi et al., 1993; Gillis et al., 1997).

Because many semiconductors are reactive in aerobic environments, it is often necessary to carry out experiments in anaerobic conditions using nonaqueous solvents. Air-sensitive experiments can be performed in specialized glassware that is continuously purged with an inert gas or in an inert atmosphere glove box specially modified for electrical throughputs. Although outlined here for current-voltage measurements, the electrode preparation techniques are applicable not only to these measurements but also to most other techniques discussed in this unit.

MEASUREMENT OF SEMICONDUCTOR BAND GAPS USING SEMICONDUCTOR-LIQUID INTERFACES

Principles of the Method

The basic photoelectrochemical techniques discussed above can be used to determine many important properties of the semiconductor and of the semiconductor-liquid contact. Monitoring the wavelength dependence of the photocurrent produced at a semiconductor-liquid contact can provide a nondestructive, routine method for determining some important optical properties of a semiconductor. Specifically, the value of the band gap energy and whether the electronic transition is optically allowed or forbidden can be obtained from measurement of the spectral response of the photocurrent at a semiconductor-liquid contact.

Two types of optical transitions are commonly observed for semiconductors: direct gaps and indirect gaps. Near the absorption edge, the absorption coefficient α can be expressed as (Pankove, 1975; Sze, 1981; Schroder, 1990)

$$\alpha \sim (h\nu - E_g)^{\beta} \tag{28}$$

where h is Planck's constant, ν is the frequency of the light incident onto the semiconductor, and β is the coefficient for optical transitions. The absorption coefficient is obtained from Beer's law, in which the ratio of transmitted, Γ, to incident, Γ_0, photon flux for a sample of thickness d is (Pankove, 1975)

$$\Gamma/\Gamma_0 = \exp(-\alpha d) \tag{29}$$

For optically allowed, direct gap transitions, $\beta = \frac{1}{2}$, whereas for indirect, optically forbidden transitions, $\beta = 2$ (Sze, 1981).

Typically, α is determined through measurements of the extinction of an optical beam through a known thickness of the semiconductor sample. Beer's law is applied to solve for α from the measured ratio of the incident and transmitted intensities of light through the sample at each wavelength of interest. For a direct band gap semiconductor, a plot of α^2 vs. photon energy will give the band gap as the x intercept. If the semiconductor has an indirect band gap, a plot of $\alpha^{1/2}$ vs. photon energy will have two linear regions, one corresponding to absorption with phonon emission and one corresponding to absorption with phonon capture. The average of the two intercepts is the energy of the indirect band gap. One-half the difference between the two intercepts is the phonon energy emitted or captured during the band gap excitation. An alternative method for determining the absorption coefficient vs. the wavelength is to determine the real and imaginary parts of the refractive index from reflectance, transmittance, or ellipsometric data and then to use the Kramers-Kronig relationship to determine the absorption coefficient of the solid over the wavelength region of interest (Sze, 1981; Lewis et al., 1989; Adachi, 1992).

Practical Aspects of the Method

The photocurrent response at a semiconductor-liquid interface can also be used to determine α as a function of wavelength and thus to determine E_g and β for that semiconductor. This method avoids having to make tedious optical transmission measurements of a solid sample in a carefully defined optical configuration. According to the Gärtner equation, the photocurrent is given as (Sze, 1981; Lewis et al., 1989; Schroder, 1990)

$$I_{ph} = q\Gamma_0(1 - R^*)\left(1 - \frac{\exp(-\alpha W)}{1 + \alpha L}\right) \tag{30}$$

where L is the minority carrier diffusion length and R^* is the optical reflectivity of the solid. In a semiconductor sample with a very short minority carrier diffusion length and with $\alpha W \ll 1$, this equation simplifies to (Sze, 1981; Schroder, 1990)

$$I_{ph} = q\Gamma_0(1 - R^*)(\alpha W) \quad \text{for } \alpha L \ll 1 \tag{31}$$

Under these conditions, α can be measured directly from the photocurrent at each wavelength. These values can then be plotted against the photon energy to determine the band gap energy and transition profile, direct or indirect, for the semiconductor under study. The only parameter that needs to be controlled experimentally is Γ_0 at the various wavelengths of concern. Methods for determining and controlling Γ_0, which are common to this method and to other methods that use illuminated semiconductor-liquid contacts, are described in detail under Diffusion Length Determination Using Semiconductor-Liquid Contacts.

Data Analysis and Initial Interpretation

Figure 5 shows the spectral dependence of the photocurrent at an n-type MoS_2 electrode (Tributsch et al., 1977). The principal analysis step in determining the band gap energy or any other intrinsic parameter, such as the

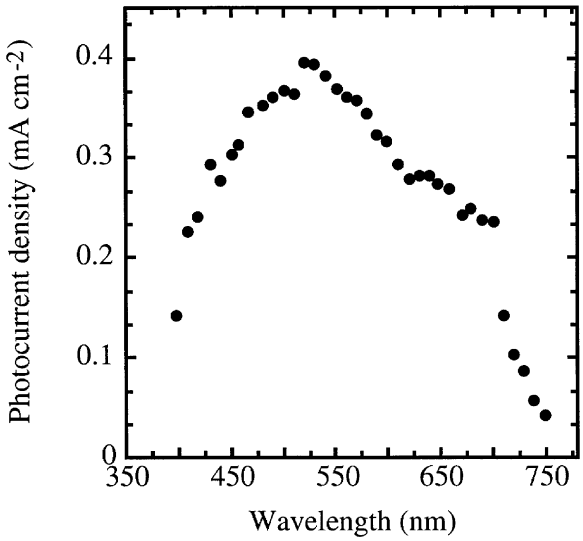


Figure 5. Spectral response of MoS_2 photocurrents (n type) in the anodic saturation region. (Reprinted with permission from Tributsch et al., 1977.)

diffusion length, from this spectrum depends on accurately transforming the wavelength-dependent photocurrent information to a corresponding absorption coefficient. With a table of absorption coefficients, photocurrent densities, and wavelengths, one can plot the absorption coefficient vs. photon energy and, by application of the methods described above, extract the band gap energy. The use of photocurrents to determine minority carrier diffusion lengths is discussed under Diffusion Length Determination Using Semiconductor-Liquid Contacts, below, and the subtleties of transforming this photocurrent data into absorption coefficients are illustrated.

DIFFUSION LENGTH DETERMINATION USING SEMICONDUCTOR-LIQUID CONTACTS

Principles of the Method

The minority carrier diffusion length L is an extremely important parameter of a semiconductor sample. This quantity describes the mean length over which photogenerated minority carriers can diffuse in the bulk of the solid before they recombine with majority carriers. The value of L affects the bulk diffusion-recombination limited V_{oc} and the spectral response properties of the solid.

The ASTM (American Society for Testing and Materials) method of choice for measurement of diffusion length is the surface photovoltage method. A conceptually similar methodology can, however, be used when a liquid provides the electrical contact to the semiconductor. Use of a semiconductor-liquid contact has the advantage of allowing a reproducible analysis of the surface condition as well as control over the potential across the semiconductor during the experiment. In either mode, the method works well only for silicon and other indirect gap materials.

We assume that the semiconductor is n type, so we are therefore interested in measuring the diffusion length of holes, L_p. Simplification of the Gärtner equation, assuming that $\alpha W \ll 1$, yields the following expression for the wavelength dependence of the open-circuit photovoltage (Sze, 1981; Lewis et al., 1989; Schroder, 1990):

$$V_{oc} \sim [1 + (L_p\alpha)^{-1}]^{-1} \tag{32}$$

Thus, provided that $L_p \gg W$ and $\alpha^{-1} \gg W$, a plot of $(V_{oc})^{-1}$ vs. α^{-1} will yield a value of the inverse of the slope that is equal to L_p. An additional check on the data is that the x intercept of the plot should equal $-L_p$ (Schroder, 1990).

Measurement of L_p using the surface photovoltage method in air requires that the semiconductor be capacitively coupled through an insulating dielectric such as mica to an optically transparent conducting electrode. However, for a semiconductor-liquid contact, either the photovoltage or photocurrent can be determined as a function of wavelength λ, and in this implementation the method is both nondestructive and convenient.

Practical Aspects of the Method

Both the surface photovoltage method and the determination of the optical properties of the semiconductor require accurate measurement of the wavelength dependence of the quantum yield for carrier collection at the semiconductor-liquid contact. The quantum yield is the ratio of the rate at which a specimen forces electrons through an external circuit, $I(\lambda)/q$, to the rate at which photons are incident upon its surface:

$$\Phi(\lambda) = \frac{I(\lambda)/q}{\Gamma_0(\lambda)A_s} \tag{33}$$

In this equation, $\Gamma_0(\lambda)$ represents the flux of monochromatic light, which is assumed to be constant over the area of the specimen, A_s.

Typically, the quantum yield is measured at short circuit. Commercial silicon photodiodes have very stable quantum yields of >0.7 throughout most of the visible spectrum, making them a nearly ideal choice for a calibration reference. The quantum yield of the experimental specimen can be calculated as follows:

$$\Phi_{cell}(\lambda) = \Phi_{ref}(\lambda)\left(\frac{I_{cell}(\lambda)A_{s,ref}}{I_{ref}(\lambda)A_{s,cell}}\right) \tag{34}$$

Experimentally, the excitation monochromator is scanned to record $I_{cell}(\lambda)$; then the experimental specimen is replaced with the reference and $I_{ref}(\lambda)$ is recorded. A significant source of error in this method, however, is the drift in the intensity of the light source over time, which can affect both $I_{cell}(\lambda)$ and $I_{ref}(\lambda)$.

A preferred experimental setup is shown in Figure 6. In this arrangement, the ratio of the photocurrent response of the experimental specimen to that of an uncalibrated photodiode, $I_{cell}(\lambda)/I_{uncal}(\lambda)$, is recorded as the experimental variable. It is not necessary to know the quantum yield or area of the uncalibrated photodiode, which merely acts to calibrate the light intensity at each wavelength. The geometry of the uncalibrated photodiode with respect to the light source and the pick-off plate should be arranged such that the surface of the uncalibrated diode is

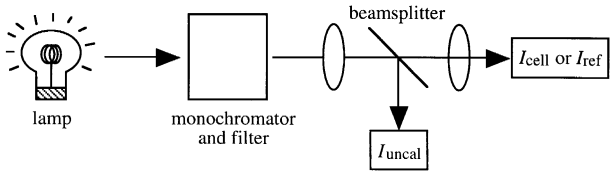


Figure 6. Spectral response measurement system consisting of a white light source, a monochromator, a beam splitter, and a calibrated photodiode.

illuminated by a small fraction of the light that is incident on the main specimen. If $t(\lambda)$ is the ratio of the light diverted to the uncalibrated photodiode relative to that which reaches the main specimen, then

$$I_{uncal}(\lambda) = q\Phi_{uncal}(\lambda)t(\lambda)\Gamma_0(\lambda)A_{s,uncal} \tag{35}$$

The photocurrent response of the calibrated photodiode relative to that of the same uncalibrated photodiode, $I_{ref}(\lambda)/I_{uncal}(\lambda)$, must also be determined. When $I_{cell}(\lambda)$ and $I_{ref}(\lambda)$ in Equation 34 are replaced with the ratios $I_{cell}(\lambda)/ I_{uncal}(\lambda)$ and $I_{ref}(\lambda)/I_{uncal}(\lambda)$, respectively, the unknown terms Φ_{uncal} and $A_{s,uncal}$ divide out.

In the foregoing discussion, we have assumed that the excitation light is continuous, e.g., furnished by a xenon arc lamp or a tungsten filament source coupled to a monochromator. Pulsed light sources can also be used, and the most convenient method is to have the excitation beam interrupted by a mechanical chopper, with a lock-in amplifier used to monitor the photocurrent. If the response time of the specimen is short compared to the chopping period, then this method yields exactly the same information as the DC experiment. However, lock-in detection offers a substantial advantage in the signal-to-noise ratio, which can be important at low light and/or low current levels. Also, for experiments that are not carried out at short circuit, but at some applied bias, this method provides for an automatic subtraction of the dark current. In addition, when a lock-in detection method is used, it is possible to measure the differential quantum yield of a signal that is superimposed on top of a DC illumination source. This information is very useful for materials that have quantum yields that are dependent on the incident light intensity. Finally, when the response time of the system is on the same order as the chopping period, variations in the photocurrent with the chopping period and analysis of the data in the frequency domain can yield detailed information about the kinetic process occurring in the semiconductor bulk and at the semiconductor-liquid interface (Peter, 1990).

Data Analysis and Initial Interpretation

To determine the diffusion length of a solid using photoelectrochemical methods, it is important to have an accurate relationship between the excitation wavelength and the absorption coefficient of the solid, because any error in this relationship leads to the calculation of incorrect values for the diffusion length in the analysis of the experimental data.

Although various equations have been proposed for the relationship between α and λ for silicon, the ASTM standard recommends the following for a stress-relieved, polished silicon wafer (Schroder, 1990):

$$\alpha = 5263.67 - 114425\lambda^{-1} + 5853.68\lambda^{-2} + 399.58\lambda^{-3} \tag{36a}$$

where the wavelength is in units of micrometers and the absorption coefficient is in units of reciprocal centimeters. This relationship is valid for the wavelength range of 0.7 to 1.1 μm, which is typically the range used for determining diffusion lengths. For a non−stress-relieved silicon wafer, the relationship is

$$\alpha = 10696.4 - 33498.2\lambda^{-1} + 36164.9\lambda^{-2} + 13483.1\lambda^{-3} \tag{36b}$$

Recent expressions have also been developed that accurately fit published absorption data for GaAs and InP. These expressions are

$$\alpha = (286.5\lambda^{-1} - 237.13)^2 \tag{36c}$$

for GaAs in the 0.75- to 0.87-μm wavelength range and

$$\alpha = (252.1\lambda^{-1} - 163.2)^2 \tag{36d}$$

for InP in the 0.8- to 0.9-μm wavelength range.

For quantitative results, the wavelength dependence of the optical reflectivity of the surface, R^*, must be known. Silicon has a weak dependence of reflectance on wavelength, which is given by the following empirical fit to the data:

$$1 - R^* = 0.6786 + 0.03565\lambda^{-1} - 0.03149\lambda^{-2} \tag{37}$$

Figure 7 illustrates the procedure for determining the diffusion length from a plot of Φ vs. $1/\alpha$ for silicon (Schroder, 1990). The minority carrier diffusion length is readily obtained by extrapolating the quantum yield data to the x intercept and taking the value of $1/\alpha$ when $\Phi = 0$. Since the quantum yield is measured at different wavelengths, the photon flux is adjusted to assure a constant photovoltage at each Φ measurement.

Other Methods for Determination of Diffusion Length. Many other methods for the measurement of minority carrier diffusion length involve determination of the minority carrier lifetime τ and the minority carrier diffusion constant D and application of the relationship $L = (D\tau)^{1/2}$ to determine the minority carrier diffusion length. One can, for example, monitor the bulk band-to-band radiative luminescence lifetime under conditions where the surface nonradiative recombination processes are negligible (Yablonovitch et al., 1987). Another technique uses the absorption of microwave radiation by free carriers. In this method, pulsed excitation of a semiconductor sample inside a microwave cavity produces a transient increase

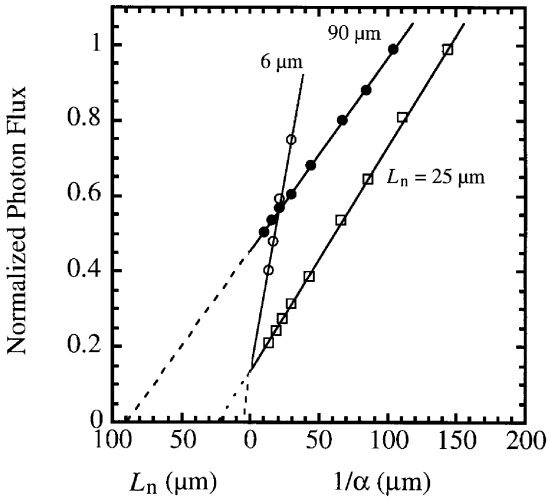


Figure 7. Plot of Φ vs. the inverse absorption coefficient for three Si diodes with different diffusion lengths. The minority carrier diffusion lengths are obtained from the value of $|1/\alpha|$ when the quantum yield is zero. (Reprinted with permission from Schroder, 1990.)

in microwave absorption, and the absorption decay yields the carrier lifetime (Yablonovitch et al., 1986). The analysis of the data to extract a value for τ, and thus to determine L_p, is similar in all cases, and details can be found, e.g., in the book by Many et al. (1965). Semiconductor-liquid contacts are very useful in certain circumstances, as in, e.g., the iodine-treated silicon surface in methanol, for which the surface recombination velocity is so low that the observed recombination is almost always dominated by nonradiative decay processes in the bulk of the silicon sample (Msaad et al., 1994). In these cases, measurement of the photogenerated carrier decay rate by any convenient method directly yields the minority carrier lifetime, and thus the minority carrier diffusion length, of the sample of concern. These values can also be spatially profiled across the semiconductor-liquid contact to obtain important information concerning the uniformity of the material properties of the bulk solid of interest.

DIFFERENTIAL CAPACITANCE MEASUREMENTS OF SEMICONDUCTOR-LIQUID CONTACTS

Principles of the Method

Differential capacitance measurements of semiconductor-liquid contacts are very useful in obtaining values for the dopant density of the bulk semiconductor that forms the semiconductor-liquid contact. In addition, such measurements have been found to be of great use in determining doping profiles of heterojunctions (Seabaugh et al., 1989) and of epitaxial layers of semiconductors (Leong et al., 1985) fabricated for use in light-emitting diodes, transistors, solar cells, and other optoelectronic devices.

To obtain an expression for the differential capacitance vs. potential properties of a semiconductor-liquid contact, we refer again to Equations 3 to 6, which describe the basic electrostatic equilibrium conditions at a semiconductor-liquid interface. Because all of the dopants in the depletion region are assumed to be ionized, the charge in the depletion region at a potential E can be expressed as

$$Q = \left[q\varepsilon\varepsilon_0 N_d\left(V_{bi} + E - \frac{kT}{q}\right)\right]^{1/2} \tag{38}$$

Taking the derivative of Q with respect to E yields an expression for the differential capacitance C_d of the semiconductor:

$$C_d = \frac{dQ}{dE} = \left[\frac{q\varepsilon\varepsilon_0 N_d}{2(V_{bi} + E - kT/q)}\right]^{1/2} \tag{39}$$

where ε is the relative permittivity of the semiconductor and ε_0 the permittivity of free space. A plot of C_d^{-2} vs. E (i.e., a Mott-Schottky plot) should thus yield a straight line that has a slope of $2(q\varepsilon_s N_d)^{-1}$ (with $\varepsilon_s = \varepsilon\varepsilon_0$ and an x intercept of $kT/q - V_{bi}$.

Equivalent Circuit Model for Measuring Differential Capacitance of Semiconductor-Liquid Contact. The most common procedure for obtaining differential capacitance vs. potential data is to apply a small AC (sinusoidal) voltage at the DC potential of interest. An analysis of the resulting cell impedance and phase angle in response to this sinusoidal perturbation yields the value of the semiconductor differential capacitance C_d. This method, of course, requires that C_d can be accessed from the experimentally measured electrical impedance behavior of a semiconductor-liquid contact. In general, an equivalent circuit is required to relate the measured impedance values to physical properties of the semiconductor-liquid contact.

The conventional description of the equivalent circuit for a semiconductor-liquid junction can be represented as in Figure 8A (Gerischer, 1975). The subscript s refers to

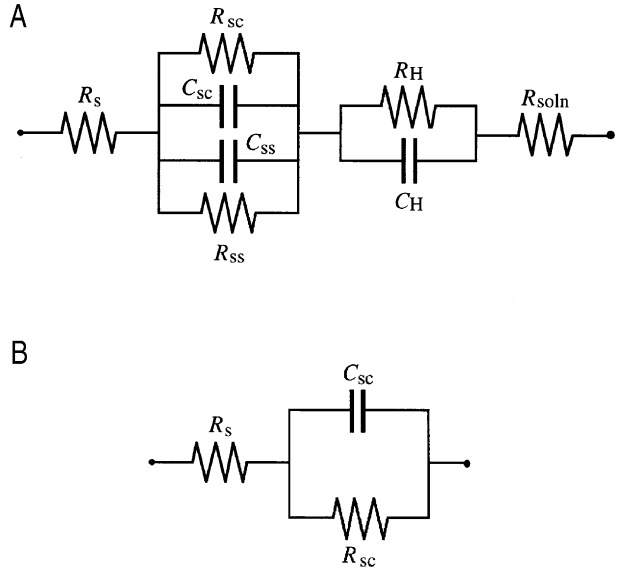


Figure 8. (**A**) Circuit of a semiconductor-liquid junction and (**B**) a simplification of that circuit.

the semiconductor bulk, sc to the space charge, ss to surface states, H to the Helmholtz layer, and soln to solution. This circuit is reasonable because there will clearly be a capacitance and resistance for the space-charge region of the semiconductor. Because surface states represent an alternate pathway for current flow across the interface, C_{ss} and R_{ss} are in parallel with the elements for the space-charge region. The Helmholtz elements are in series with both the space-charge and surface-states components, since current flows through the Helmholtz layer regardless of the pathway of current flow through the semiconductor. Finally, the series resistance of the electrode and contact and the solution series resistance are included as distinct resistive components in this circuit.

A possible simplification that can often be justified combines the resistances of the solution and of the electrode into one series resistor (Figure 8B). If the Helmholtz layer resistance is small, the RC portion of the Helmholtz layer circuit impedance is dominated by C_H. Furthermore, C_H is usually much larger than C_{sc}. Therefore, C_H can typically be neglected, because the smallest capacitance value for a set of capacitors connected in series dominates the total capacitance of the circuit. Additionally, if the AC impedance measurement is performed at a high frequency, such that surface states do not accept or donate charge as rapidly as the space-charge region, then the elements associated with surface state charging-discharging processes can also be neglected. Alternatively, if the surface state density of the semiconductor is low enough, the contributions of C_{ss} and R_{ss} are negligible.

At sufficiently high frequencies, the impedance of this simplified circuit can be written in terms of $G = (R_s)^{-1}$ and $B = (\omega C_{sc} R_s)^{-2}$, where G is the conductance and B is the susceptance (both in units of siemens). This yields the desired determination of C_d vs. E from the impedance response of the system. Once C_d is determined, Equation 39 is used to obtain a value for N_d. When using this method, C_d^{-2} vs. E should be determined at several different AC frequencies ω to verify that the slope and intercept do not change as the frequency is varied.

Practical Aspects of the Method

In a typical Mott-Schottky experiment, the DC bias to the electrochemical cell is delivered by a potentiostat, and the AC voltage signal is supplied by an impedance analyzer that is interfaced to the potentiostat. The DC potential applied to the semiconductor junction, which is usually in the range of 0 to 1 V vs. $E(A/A^-)$, is always in the direction of reverse bias, since forward-bias conditions induce Faradaic current to flow and effectively force the cell to respond resistively instead of capacitively. The input AC signal, usually 5 to 10 mV in amplitude, must be small to ensure that the output signal behaves linearly around the DC bias point and is not convoluted with the electric field in the semiconductor (Morrison, 1980).

The quantities that are measured by the impedance analyzer are the total impedance Z_{tot} and the phase angle θ between Z_{tot} and the input AC signal. From these data, the real component of the impedance, Z_{re}, which is defined by $Z_{re} = Z_{tot} \cos \theta$, and the imaginary component of the impedance, Z_{im}, which is defined by $Z_{im} = Z_{tot} \sin \theta$, can be calculated. For the simplified three-element equivalent circuit illustrated in Figure 8, the dependencies of Z_{tot} vs. AC signal frequency (Bode plot) and Z_{im} vs. Z_{re} (Nyquist plot) ideally obey patterns that not only permit extraction of the electrical elements in the circuit but also indicate the frequency range over which the cell acts capacitively. Of course, deviations from these ideal patterns can allow evaluation of the appropriateness of the theoretical equivalent circuit to the description of the experimental cell.

A typical Bode plot for a silicon electrode in contact with a redox-active solution is shown in Figure 9. At high frequencies f of the input AC voltage signal, the capacitive reactance χ_c [i.e., the effective impedance of the capacitor, given by $\chi_c = (2\pi f C_{sc})^{-1}$] is small relative to R_{sc}, so most of the current flows through the C_{sc} pathway, and the observed impedance is therefore simply R_s. At low frequencies, χ_c is high relative to R_{sc}, so most of the current flows through the R_{sc} pathway, and the observed impedance is $R_s + R_{sc}$. In the intermediate-frequency range, as is shown in Figure 9, increments in frequency translate directly into changes in χ_c, so the magnitude of Z_{tot} is dictated by C_{sc}. The Mott-Schottky measurements should be performed in this capacitive frequency regime, which ideally has a slope of -1 in a Bode plot.

A Nyquist plot is distinguished from the Bode plot in that it illustrates explicitly the branching between Z_{re} and Z_{im} as a function of frequency. A Nyqyist plot for a silicon electrode in contact with a redox-active solution is

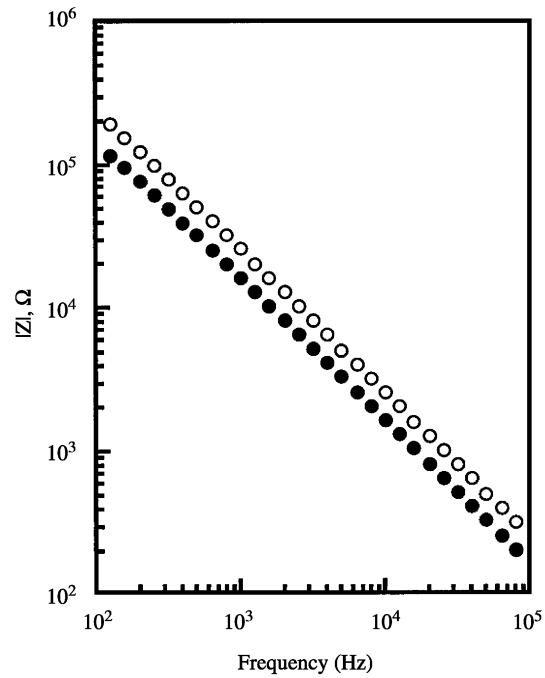


Figure 9. Typical Bode plots taken at intermediate frequencies where the impedance is dominated by capacitive circuit elements. The electrode-solution system is identical to that used to obtain the data in Figure 4. The open circles represent data taken at +0.2 V vs. $E(A/A^-)$ and the filled circles represent data taken at +0.8 V vs. $E(A/A^-)$.

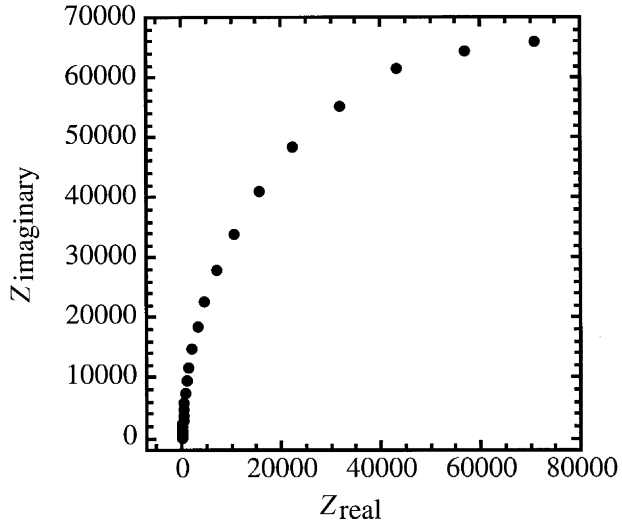


Figure 10. Nyquist plot for the system described in Figure 4. The data were collected at an applied DC bias of 0.45 V.

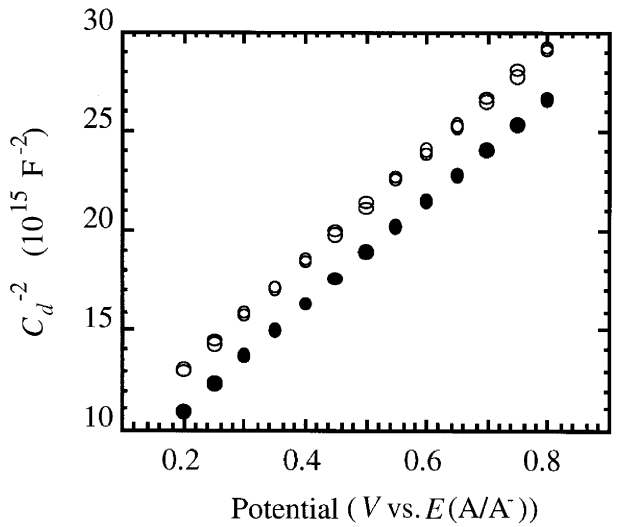


Figure 11. Mott-Schottky plots for the system described in Figure 4. The data shown are for two different concentrations of acceptor species and for two different acquisition frequencies, 100 and 2.5 kHz.

shown in Figure 10. At high frequency, $Z_{re} \cong R_s$. As the frequency is incrementally decreased, the contribution of Z_{im} rises until the magnitude of the capacitive reactance is identical to the resistance of R_{sc}. At this point, equal current flows through the circuit elements C_{sc} and R_{sc}. As the frequency is decreased further, the system behaves increasingly resistively, until the impedance is entirely dictated by R_{sc}. In practice, since the frequency may not reach a point where the system is predominantly resistive, it is generally easier to extract the value of R_{sc} from a Nyquist plot than from a Bode plot. Circle-fitting algorithms can provide reasonable estimates of R_{sc} if the Nyquist plot possesses a sufficient arc over the frequency range explored experimentally.

Data Analysis and Initial Interpretation

Assuming the three-element equivalent circuit depicted in Figure 8B,

$$C_{sc} = \frac{1 + \sqrt{1 - 4Z_{im}^2/R_{sc}^2}}{2\omega Z_{im}} \qquad (40)$$

where the angular frequency $\omega = 2\pi f$. It is also assumed that $R_s \ll R_{sc}$, so that R_s contributes negligibly to the measured impedance. Typically, since R_s is on the order of 10^2 Ω and R_{sc} is on the order of 10^5 to $10^7\Omega$, Equation 40 gives an accurate value for C_{sc}. Therefore, measurement of Z_{im} at various frequencies and DC biases, as well as extraction of the R_{sc} value at each DC bias from the circular fit of the Nyquist plot at that potential, allows calculation of C_{sc}. The subsequent compilation of C_{sc}^{-2} vs. E for each frequency yields the Mott-Schottky plot, as shown in Figure 11 for a typical silicon-solution interface. Linear regression of these data points and extrapolation of these linear fits to the x intercepts of $C_{sc}^{-2} = 0$ gives the value of V_{bi} for each frequency.

Differential capacitance measurements can readily be performed with a potentiostat in conjunction with an impedance analyzer, both of which are set up as described above, yielding direct values of the imaginary and real components of the impedance. Alternatively, AC voltametry can be performed using a potentiostat in conjunction with a waveform generator and a lock-in amplifier. In this technique, the magnitude of the AC component of the current (i.e., Z_{tot}) and the phase angle with respect to the input AC signal are measured at a given frequency. The real and imaginary parts of the impedance can then be extracted as described above. This powerful technique therefore requires no additional capabilities beyond those that are required for basic measurements of the behavior of semiconductor photoelectrodes.

In aqueous solutions, useful electrolytes are those that yield anodic dissolution of the semiconductor but do not etch the solid or produce a passivating oxide layer in the absence of such illumination. For Si, the appropriate electrolyte is NaF-H_2SO_4 (Sharpe et al., 1980), whereas for GaAs, suitable electrolytes are Tiron (dihydroxybenzene-3,5-disulfonic acid disodium salt) and ethylenediaminetetraacetic acid (EDTA)–0.2 M NaOH (Blood, 1986). A variety of nonaqueous solutions can also be used to perform differential capacitance measurements, although more attention needs to be paid to series resistance effects in such electrolytes.

Either solid-state or solid-liquid junctions can be utilized to measure the dopant density of the bulk semiconductor from C^{-2}-vs.-E plots, as described above. Solid-liquid contacts offer the additional opportunity to conveniently obtain a depth profile of the semiconductor doping level in the same experimental setup. A more comprehensive review of the technique of electrochemical profiling is given by Blood (1986).

To characterize the dopant profile through the solid, an anodic potential is applied to the semiconductor electrode such that the surface is partially dissolved. The thickness of material dissolved from the electrode, W_d, is given by the

time integral of the current density passed, J (Ambridge et al., 1975):

$$W_d = \frac{M}{n + F\rho}\int J\, dt \tag{41}$$

where M is the molecular weight of the semiconductor, n^+ is the number of holes required to oxidize one atom of the electrode material to a solution-phase ion, and ρ is the density of the solid. A separate Mott-Schottky plot is then obtained at each depth of etching through the material. The depth of the dopant density measurement is given by $W' = W + W_d$. Thus, periodic dissolutions and Mott-Schottky determinations can yield a complete profile of N_d as a function of distance into the solid.

Problems

The most prevalent problem in Mott-Schottky determinations is frequency dispersion in the x intercepts of the linear regressions. Nonlinearity in these plots may originate from various experimental artifacts (Fajardo et al., 1997). If the area of the semiconductor is too small, edge effects may dominate the observed impedance. The area of the counter electrode must also be large relative to the working electrode (as a rule of thumb, by a factor of 10), to ensure that any electrochemical process occurring at the counter electrode does not introduce additional electrical elements into the equivalent circuit. Additionally, the redox component that is oxidized or reduced at the counter electrode must have a sufficient concentration in solution (roughly 5 mM), or this redox process may be manifested as an additional resistance that is a major component of the total cell impedance. Finally, the measuring resistor within the potentiostat should be adjusted in response to the magnitude of the currents passing through the cell in order to obtain the most accurate impedance value possible.

TRANSIENT DECAY DYNAMICS OF SEMICONDUCTOR-LIQUID CONTACTS

Principles of the Method

Solid-liquid junctions are very useful in determining the electrical properties of the semiconductor surface. Principal methods in this endeavor involve measurement of the transient carrier decay dynamics, electrochemical photocapacitance spectroscopy (EPS), and laser spot scanning (LSS) techniques.

Under conditions where bulk recombination is slow relative to surface recombination, measurements at the solid-liquid interface can provide valuable information regarding the concentration and energetics of electrical trap sites at the semiconductor surface. Furthermore, the presence of a liquid allows, in principle, manipulation of these trap levels through chemical reactions induced by addition of reagents to the liquid phase. Methods for determining the surface recombination velocity of solid-liquid interfaces are therefore described here.

The rate of nonradiative recombination mediated through surface states can be described under many experimental conditions by the steady-state Shockley-Read-Hall (SRH) rate equation (Hall, 1952; Shockley et al., 1952; Schroder, 1990; Ahrenkiel et al., 1993):

$$R_{SRH}(\text{surface}) = \frac{n_s p_s - n_i^2}{(n_s + n_{1,s})/N_{t,s}k_{p,s} + (p_s + p_{1,s})/N_{t,s}k_{n,s}} \tag{42}$$

where n_i is the intrinsic electron concentration for the semiconductor, $N_{t,s}$ defines the density of surface states (in units of reciprocal centimeters squared), p_s is the surface hole concentration, and $k_{n,s}$ and $k_{p,s}$ are the capture coefficients for electrons and holes, respectively, by surface states. Each capture coefficient is related to the product of the thermal velocity v of the carrier in the solid and the capture cross-section σ for each kinetic event such that (Many et al., 1965; Krüger et al., 1994a)

$$k_{p,s} = \nu_p\sigma_p \quad \text{and} \quad k_{n,s} = \nu_n\sigma_n \tag{43}$$

The symbols $n_{1,s}$ and $p_{1,s}$ in Equation 42 represent the surface concentrations of electrons in the conduction band and holes in the valence band, respectively, when the Fermi level is at the trap energy. Values for $n_{1,s}$ and $p_{1,s}$ can be obtained through use of the principle of detailed balance, which, when applied to a system at equilibrium in the dark, yields (Many et al., 1965; Blakemore, 1987)

$$n_{1,s} = N_c\exp[(E_{cb} - E_t)/kT] \tag{44a}$$

and

$$p_{1,s} = N_v\exp[(E_t - E_{vb})/kT] \tag{44b}$$

where E_t is the energy of the surface trapping state.

When other recombination processes are minimal and the initial carrier concentration profiles are uniform throughout the sample, surface recombination should dominate the entire decay properties of the sample. Under these conditions, the observed minority carrier decay dynamics are given by (Ahrenkiel et al., 1993)

$$\text{Rate} = -\frac{d\Delta p}{dt} = R_{SRH} \approx \frac{\Delta p}{\tau_{s,1}} \tag{45}$$

Furthermore, under such conditions, the fundamental filament decay lifetime $\tau_{s,1}$ is given by (Schroder, 1990; Ahrenkiel et al., 1993)

$$\tau_{s,1} = d/S_{low} \quad S_{low} \equiv S_p = N_{t,s}k_{p,s} \tag{46}$$

where d is the thickness of the sample and S_{low} is the surface recombination velocity (in units of centimeters per second) at low-level injection ($\Delta n < N_d$ and $\Delta p < N_d$).

The other limiting situation is obtained under uniform illumination at an intensity high enough to produce high-level injection conditions (i.e., $\Delta n \gg N_d$ and $\Delta p \gg N_d$). By using Equation 42 with injected carrier concentrations $\Delta n \gg N_d$ and $\Delta p \gg N_d$ and, because equal numbers of

electrons and holes are created by the optical injection pulse, taking $\Delta n = \Delta p$, the recombination rate is

$$\text{Rate} = \frac{d\,\Delta p}{dt} = R_{\text{SRH}} \approx \frac{\Delta p}{\tau_{\text{s,h}}} \tag{47}$$

Now, however, the filament decay lifetime is (Blakemore, 1987; Schroder, 1990)

$$\tau_{\text{s,h}} = \frac{d}{S_{\text{high}}} \qquad S_{\text{high}} = \frac{S_{\text{p}} S_{\text{n}}}{S_{\text{p}} + S_{\text{n}}} \tag{48}$$

For conditions of $k_{\text{p,s}} = k_{\text{n,s}}$, $S_{\text{high}} = \frac{1}{2} S_{\text{low}}$, and the surface recombination decay lifetime under high-level injection, $\tau_{\text{s,h}}$, is equal to $2\tau_{\text{s,l}}$.

Practical Aspects of the Method

The carrier concentration decays can be monitored using a number of methods. Luminescence is a convenient and popular probe for direct band gap semiconductors, while for materials like silicon, conductivity methods are often employed. The two methods are complementary in that the photoluminescence signal decays when either the minority or majority carrier is captured, whereas conductivity signals are weighted by the individual mobilities of the carrier types. Thus, if all of the holes in silicon were to be trapped, the band gap luminescence signal would vanish, whereas the conductivity signal would still retain 75% of its initial amplitude (because of the carrier mobilities for silicon, $\mu_{\text{n}} \simeq e\mu_{\text{p}}$). Either method can be used to probe the carrier concentration dynamics, and both are sometimes used on the same sample to ensure internal consistency of the methodology. A somewhat more specialized method of monitoring the carrier concentrations has also been recently developed using ohmic-selective contacts on silicon, where the photovoltage developed between these contacts yields a probe of the carrier concentration decay dynamics in the sample of concern (Tan et al., 1994a).

Measurements of Surface Recombination Velocity Using Time-Resolved Photoluminescence Methods. One of the most reliable methods for monitoring minority carrier decay dynamics is the time-resolved photoluminescence (TRPL) method. The use of time-correlated single-photon counting has added further sensitivity to the TRPL technique and has allowed for sub-picosecond lifetime measurements.

To perform time-correlated single-photon TRPL, one starts with pulsed, short-time monochromatic laser light tuned to an energy greater than the band gap energy of the material under study. This laser source can be produced by the output of a pulse-compressed Nd-yttrium-aluminum garnet (YAG) pumped dye laser, if picosecond timing resolution suffices, or from the fundamental mode of a solid-state Ti-sapphire laser, if even shorter time resolution is desired. The timing characteristics of the laser pulse can be determined through use of an autocorrelator. Any chirping in the light source must be removed by passing the light through a few meters of optical fiber.

Prior to reaching the sample surface, the laser light is split by means of a beam splitter (Fig. 12). The less intense optical beam emanating from the beam splitter is directed into a photodiode, while the more intense beam is directed onto the semiconductor sample. The photodiode and sample must be positioned at equal distances from the beam splitter to ensure correct timing in the experiment. The voltage output of the photodiode is directly connected to the START input of a time-to-amplitude converter (TAC). Thus, with each laser pulse, the sample is illuminated and the TAC is triggered on (start). Both the beam diameter and the polarization of the incident light are experimental parameters used for controlling the light intensity, and thus the injection level, of the semiconductor.

The light emitted by the sample due to radiative recombination processes is collected and focused onto a spectrometer that is tuned to the wavelength of the transition to be monitored. The light output from this spectrometer is then directed onto a single-photon detector. When a single photon is detected, the resulting voltage output produced by the detector is used to trigger the STOP input of the TAC. Once triggered into the off position, the TAC will in turn produce a voltage pulse whose magnitude is linearly dependent on the time duration between the START and STOP signals. The distribution of voltage-pulse magnitudes is thus a distribution in time of the photons emitted from the sample after initial illumination. By allowing a computer-controlled multichannel analyzer (MCA) to display a histogram, with respect to magnitude, of the distribution of the voltage pulses that are produced by the TAC, a single-photon TRPL spectrum is obtained (Ahrenkiel et al., 1993; Kenyon et al., 1993).

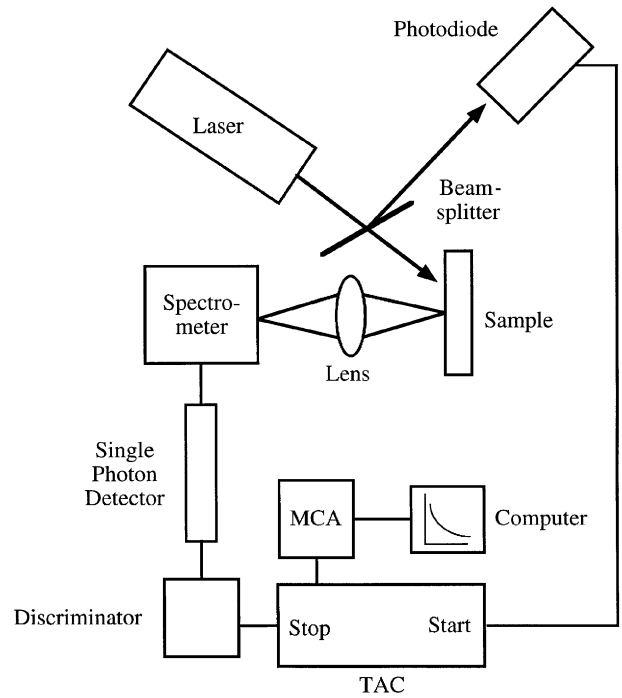


Figure 12. Time-resolved single-photon photoluminescence spectrometer.

To enhance the signal-to-noise ratio in a TRPL spectrum, it is necessary to place a pulse- height discriminator between the single-photon detector and the STOP input of the TAC. This discriminator will ensure that low-voltage pulses that arise from either electronic or thermal noise and/or high-voltage pulses that can be produced by multiphoton events do not accidentally trigger the STOP signal on the TAC. To obtain the true TRPL spectrum for minority carriers, this experimentally obtained TRPL spectrum must be deconvoluted with respect to the system response function of the entire apparatus. The system response function must be determined independently for the setup of concern and is typically measured by exciting a diffuse reflector or scattering solution instead of the semiconductor. A variety of numerical methods allow generation of the corrected photoluminescence decay curve from the observed TRPL decay and the system response function (Love et al., 1984). Deconvolution is especially important if the TRPL decay time is comparable to the system response time.

The success of the single-photon counting method is due to the development of detectors with sufficient gain and time response that a single photon produced by an electron-hole recombination event can be detected within a system response time of 30 ps. Two varieties of these detectors are photomultiplier tubes (PMTs) and microchannel plates (MCPs). A limitation of these detectors, however, is that only photon energies 1.1 eV can be detected. Recently, advancements in the timing resolution of single-photon avalanche photodiodes (SPADs) have allowed for detection of photons in the near-infrared region and have raised the possibility of performing TRPL with materials having a band gap of <1.1 eV, so that additional semiconductor materials can be explored using this approach.

Data Analysis and Initial Interpretation

Once a TRPL spectrum is recorded and the effective lifetime τ_{pl} has been measured, it is necessary to extract the lifetime for surface recombination, or any other dominant recombination mechanism. The rate of radiative band-to-band emission R_r from a semiconductor is given by (Pankove, 1975; Ahrenkiel et al., 1993; Krüger et al., 1994b)

$$R_r = k_r(np - n_i^2) \tag{49}$$

where k_r is the rate constant for radiative recombination. For low-level injection, Equation 49 reduces to

$$R_r = K_r[(N_d + \Delta n)(p_0 + \Delta p) - n_i^2] \approx k_r N_d \Delta p \tag{50}$$

or

$$R_r = -d(\Delta p)/\Delta p = k_r N_d\, dt \tag{51}$$

where p_0 is the equilibrium minority carrier concentration in the dark.

When nonradiative recombination mechanisms such as surface recombination are active in parallel with radiative recombination, the measured photoluminescence signal $I_{pl}(t)$ for a volume element in the solid containing a spatially uniform density of excess minority carriers is given by (Ahrenkiel et al., 1993; Krüger et al., 1994a)

$$I_{pl}(t) = \frac{\Delta p_0 \exp(-\tau/\tau_{pl})}{\tau_{r,l}} \quad (\text{photons cm}^{-3}\,\text{s}^{-1}) \tag{52}$$

where Δp_0 is the initial density of injected minority species after the laser excitation. The quantity τ_{pl} is the measured effective total lifetime for the system, which is determined from the experimental TRPL spectrum (Fig. 13) and is expressed as

$$\frac{1}{\tau_{pl}} = \frac{1}{\tau_s} + \frac{1}{\tau_r} \tag{53}$$

where τ_r is the radiative lifetime and τ_s is the surface recombination lifetime. Depending on the injection level, τ_r is defined as either (Many et al., 1965; Schroder, 1990)

$$\tau_{r,l} = \frac{1}{k_r N_d} \quad \text{low injection} \tag{54a}$$

or

$$\tau_{r,h} = \frac{1}{k_r\, \Delta p_0} \quad \text{high injection} \tag{54b}$$

Since k_r, N_d, and Δp_0 are all measurable parameters, the decay time for a luminescence spectrum can be easily deconvoluted through application of Equation 53 to retrieve the surface recombination lifetime and the resulting surface recombination velocity for a semiconductor of known thickness.

One problem in TRPL experiments involves the fraction of surface area of the semiconductor that is illuminated by the laser beam. The illumination should always be uniform across the surface. Frequently, the injection level is

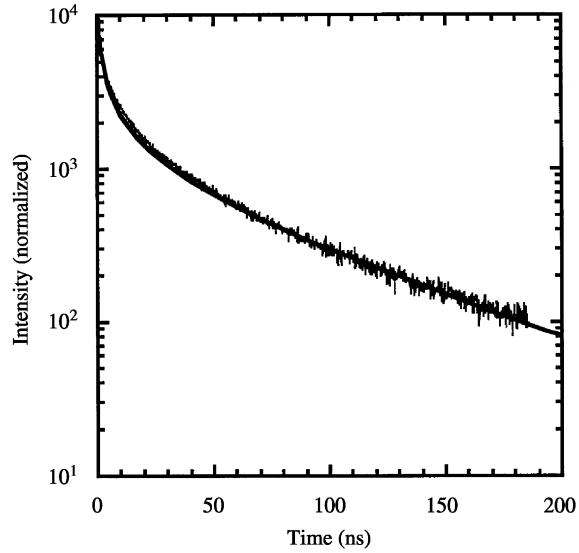


Figure 13. Fitting example of a time-resolved photoluminescence decay.

controlled by controlling the diameter to which the laser light is focused. However, carrier recombination kinetics in the dark differ from those in the light. With a nonuniformly illuminated surface, photogenerated carriers can diffuse into the nonilluminated regions of the semiconductor. The net recombination lifetime measured will thus be a convolution of lifetimes for processes occurring in the illuminated and nonilluminated regions of the semiconductor. This is especially of concern with beam spots $\sim 10\,\mu m$ in diameter. In such situations one cannot make a quantitative assessment for the carrier dynamic lifetimes of an illuminated semiconductor because the majority of the recombination occurs in the unilluminated regions of the sample. In essence, the desired lifetime is affected by "background" recombination lifetimes occurring in the dark.

The repetition rate of modern lasers is also of concern, as is their intensity. The repetition rate should always be set such that all carrier concentrations return to their equilibrium dark values prior to the subsequent laser pulse. Otherwise, the injection levels and hence recombination lifetimes will vary from pulse to pulse. For a semiconductor in accumulation, the time to restore equilibrium can be as large as a few milliseconds. Thus the repetition rate should be regulated as the bias across the semiconductor is varied. One method of increasing the time between subsequent pulses is to use a Kerr shutter as a pulse gate.

Caution should also be taken to ensure that the laser intensity does not induce sample damage. The following protocols should always be followed to assure that lifetimes derived from TRPL experiments are reliable. First, Mott-Schottky measurements should be performed to extract the flat-band potential and barrier height. Second, *J-E* curves should be acquired both in the dark and under the same light intensity conditions that will be used in the TRPL experiments. Finally, after performing the TRPL experiments, the Mott-Schottky and *J-E* curves should be redetermined to verify that the integrity of the sample has not been compromised during the TRPL measurement.

MEASUREMENT OF SURFACE RECOMBINATION VELOCITY USING TIME-RESOLVED MICROWAVE CONDUCTIVITY

Principles of the Method

Another contactless method for measuring the lifetime of minority carriers after an initial excitation pulse is time-resolved microwave conductivity (TRMC). The change in reflected microwave power, $\Delta R_M(t)$, is proportional to the change in the conductivity of the semiconductor. This conductivity change is, to first order, a measure of the time dependence of excess minority carriers in the semiconductor (Naber et al., 1969; Chen, 1988; Forbes et al., 1990; Schroder, 1990; Ramakrishna et al., 1995):

$$\begin{aligned} &\Delta R_M(t) \simeq \Delta\sigma(t) = q(\mu_n \Delta n + \mu_p \Delta p) \\ &\Delta R_M(t) \simeq q(\mu_n + \mu_p)\Delta P \text{ (under low-level inection)} \end{aligned} \tag{55}$$

Practical Aspects of the Method

To implement the method, either the semiconductor under study is connected to a microwave source through a radio frequency (RF) coupler, which consists of a series of coils connected to the base of the sample mount, or the sample is connected to the source using a waveguide. The waveguide system offers numerous advantages. Since the sample is mounted inside the waveguide and is illuminated colinearly with the microwave radiation, better protection from electromagnetic interference and better-defined and more uniform microwave radiation is available.

A microwave experimental setup consists of four primary components (Fig. 14): (1) the source, which consists of a microwave generator, an isolator, and an attenuator; (2) the sample, which is mounted onto the end of a microwave waveguide; (3) the detector, which by rectification creates a voltage across a load resistor, with the voltage being proportional to the amplitude of the reflected microwave power; and (4) the power meter (a short-circuit plate consisting of silver-plated brass), which is used to measure the power from the source and is mounted on a separate waveguide at an equal distance from the source and the sample. These four components are positioned around a microwave circulator (Schroder, 1990).

Microwaves from the source, usually $\kappa\alpha$ radiation from a Gunn diode operating at $\sim$10 GHz, are directed via the circulator to the waveguide system that contains the sample. The reflected microwaves are in turn directed by the circulator to the detector. Most commercially available circulators allow for timing resolutions in the sub-nanosecond regime. To measure the source power, the circulator directs microwaves to the short-circuit plate rather than to the sample. During the experiment, the sample is initially illuminated by a laser pulse or strobe light, which also triggers an oscilloscope. The reflected microwaves are amplified and measured, and the corresponding signal is displayed on the oscilloscope as a function of time from this illumination. The reflected microwaves not only emanate only from the surface of the semiconductor, but

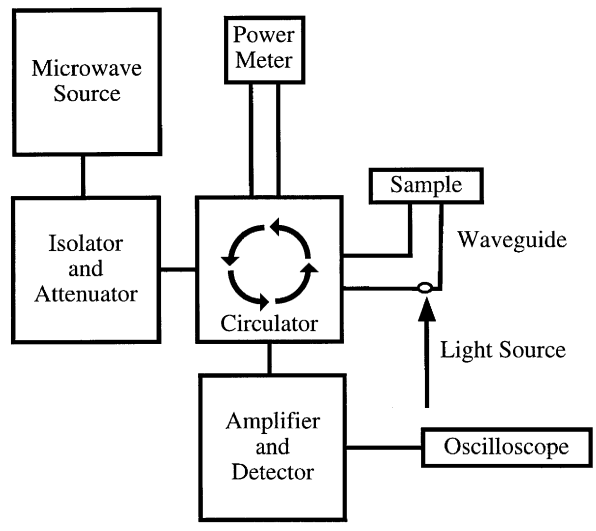


Figure 14. Microwave reflectance spectrometer.

actually penetrate a skin depth into the sample. For silicon at 10 GHz, this skin depth is 350 μm for a resistivity of 0.5 Ω-cm and is 2200 μm for a resistivity of 10 Ω-cm. Once the microwave spectrum is collected by a digital oscilloscope, the lifetime can be measured directly.

Data Analysis and Initial Interpretation

Figure 15 displays the microwave conductivity decay for *n*-Si (solid line) in contact with concentrated sulfuric acid under high-level injection conditions (10^{15} injected carriers per square centimeter in a 190-μm-thick sample) (Forbes et al., 1990). The data can be fit using a simulation package that takes into account all aspects of drift-diffusion, generation, recombination, and charge transfer for photogenerated carries in a semiconductor. Since most bulk parameters for common semiconductors are known, simulations are performed by parametrically varing an unknown parameter, such as the surface recombination velocity, until a fit with a minimal residual is obtained. Such a simulated fit is shown in Figure 15 (dotted line) for a surface recombination velocity of 25 cm s^{-1} and a bulk lifetime of 2 ms.

If no simulation package is available, a spectrum can be fit with a biexponential of the form

$$I(t) = A_1 \exp(-t/\tau_1) + A_2 \exp(-t/\tau_2) \qquad (56)$$

The fit is performed from the time of peak conductivity intensity to the time at which the conductivity signal has decayed to 1% of its peak value. Subsequently, the resultant fit can be described with an effective lifetime value, $\langle\tau\rangle$, computed using the relationship

$$\langle\tau\rangle = \frac{A_1\tau_1 + A_2\tau_2}{A_1 + A_2} \qquad (57)$$

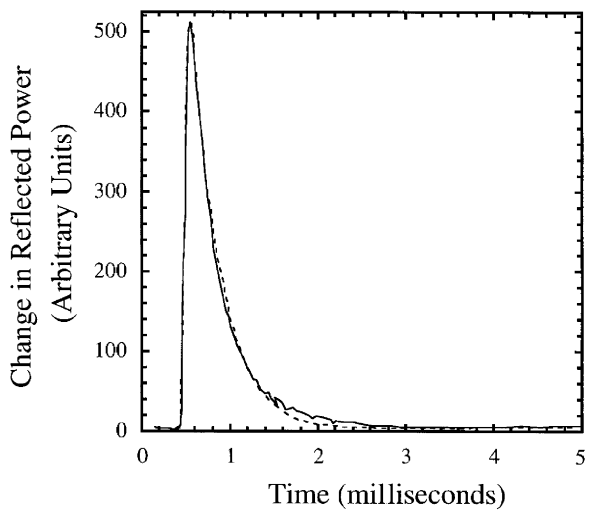


Figure 15. Microwave conductivity decay (solid line) under high injection (10^{15} carriers/cm^2) for a 190-μm thick sample of *n*-Si in contact with concentrated sulfuric acid. A simulation with a surface recombination velocity of 25 cm/s^1 and a bulk lifetime of 2 ms is shown (dotted line) for comparison. (Reprinted with permission from Forbes et al., 1990.)

where A_1 and A_2 are the preexponential factors and τ_1 and τ_2 are the exponential decay times of the fit (Equation 56). This effective lifetime has terms that are linear rather than quadratic with respect to the decay times τ_n. Equation 57 hence does not bias a $\langle\tau\rangle$ value toward the longer of the two decay times derived from the fitting procedure. The above-mentioned fitting procedure is also applicable to all time-dependent decay phenomena, such as radiative recombination, discussed under Transient Decay Dynamics of Semiconductor-Liquid Contacts, above.

Sample Preparation

Measurements of bulk or surface parameters require two different sample preparation procedures. If one is interested in bulk recombination lifetimes, then the samples should be made progressively thicker, on the order of 1 mm, as the bulk lifetime to be measured becomes longer, and the surface recombination velocity should be made as small as possible. For measurement of the surface recombination velocity, the sample should be of the thickness typically used in wafers, or on the order of 0.1 mm. The sample should also be polished on both ends to ensure consistent transmission and reflectivity of the microwaves. The electrolyte must also not be too conductive or the radiation signal will be greatly attenuated.

Microwave conductivity experiments permit a great deal of versatility as to the sample geometry that can be used. Of concern, however, is what percentage of the microwave decay signal recorded is indeed that of the semiconductor and not that of the measurement apparatus. When the geometry of a sample/apparatus is off resonance, it has been found that the system response is very fast, while an on-resonance cavity results in a large increase in the system fall time. One method of determining the extent of resonance in microwave experiments is to use etched silicon wafers as calibration templates. It has been determined that Si wafers etched in $HF:HNO_3:CH_3COOH$ solutions result in a surface recombination velocity as large as 10^5 cm s^{-1}. Conversely, it has been shown that immersing the sample in HF during decay measurements gives a measured surface recombination velocity as low as 0.25 cm s^{-1}. Both these protocols are useful when calibrating one's instrument and/or samples.

ELECTROCHEMICAL PHOTOCAPACITANCE SPECTROSCOPY

Principles of the Method

Photocapacitance spectroscopy has been used to characterize deep-level states in semiconductors since the mid-1960s. The use of EPS, however, offers many advantages relative to solid-state photocapacitance measurements. The presence of an electrolyte in EPS measurements allows leakage currents to be reduced by adjusting the availability and reorganization energy of electronic states in the electrolyte. Also, EPS can be used for depth profiling of trap states because layers of the semiconductor can be etched between subsequent EPS measurements. The depth-profiling procedure allows for detailed examination

of the influence of surface oxide layers or surface passivation layers on the surface recombination kinetics. If the etching is performed electrochemically, then the entire depth-profiling and EPS experiment can be conducted in a single apparatus (Haak et al., 1982).

Basic Principles of EPS. The principles behind EPS are relatively straightforward. As described in Equations 38 to 40, the differential capacitance of the semiconductor-liquid contact is dominated by the differential capacitance of the space-charge layer of the semiconductor. Any additional charge introduced by injection of carriers into the solid, either optically or thermally, will therefore either populate or depopulate states in the bulk or surface of the semiconductor. This change in charge density in the solid will affect the thickness of the space-charge layer. In turn, the change in space-charge layer thickness will affect the capacitance of the semiconductor-liquid junction. Measurement of the capacitance as a function of the wavelength (energy) of sub–band-gap light incident onto the semiconductor-liquid junction can therefore provide information on the energies, time constants, physical location, and character (i.e., either acceptor or donor) of the traps in the semiconducting electrode.

A typical EPS spectrum will exhibit plateaus and/or peaks (Fig. 16). A plateau is observed in the spectrum when a transition involves one of the bands of the semiconductor, because once a threshold energy for the transition has been reached, transitions either to or from a band can proceed over a wide range of photon energies. In contrast, the EPS spectrum exhibits a peak whenever localized state or impurity/defect atoms participate in the charge injection process. These injections require two-level transitions and thus are only observable over a smaller range of illumination wavelengths (energy) (Anderson, 1982; Haak et al., 1982; Goodman et al., 1984; Haak et al., 1984).

The energy of a localized state relative to a band edge energy is given by the onset energy of the peak in the EPS spectrum. When all of the states at a given illumination energy have been saturated, populated, or depopulated, the spectrum will saturate with respect to capacitance as the illumination intensity is increased. The density of such states can thus be measured from the magnitude of the capacitance at saturation. Furthermore, by fixing the intensity of the light and scanning a range of energies around a plateau or peak, the optical absorption cross-section for trap states can be measured from the rate of change of the capacitance with time. The corresponding emission rate can then be determined from the rate of decay of the capacitance with time after the interruption of the incident illumination (Haak et al., 1984).

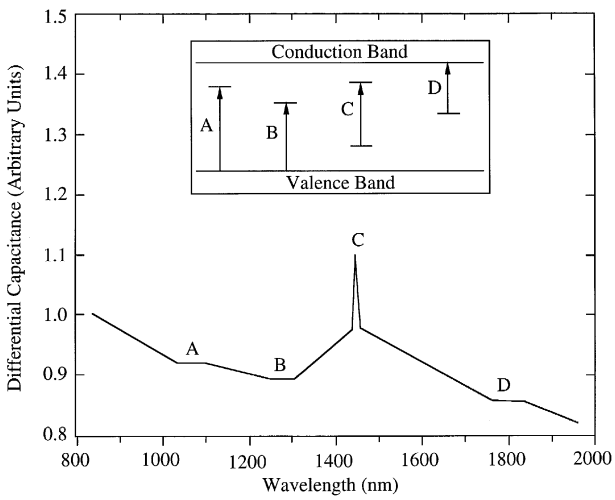


Figure 16. Electrochemical photocapacitance spectrum. Displayed is the measured differential capacitance as a function of the wavelength of the incident light. The inset illustrates various types of electro-optical transitions that can be associated with the observed plateaus and peaks in the EPS spectrum.

With EPS, contributions to the measured capacitance from bulk and surface states can be distinguished by taking EPS measurements at constant illumination for different bias voltages. The bias voltage determines the thickness of the space-charge layer and thus allows measurement of the density of bulk states constituting the space-charge layer. The bias voltage has, however, little influence on the potential drop at the surface or in the Helmholtz layer. The rate of change of the EPS capacitance at a fixed illumination energy and intensity as a function of sample bias thus provides a measure of the density of bulk trap states as a function of the sample thickness.

Quantifying Electrochemical Photocapacitance Spectra

Bulk Capacitance Measurements. The density of bulk states (in reciprocal cubic centimeters) can be readily derived from the Mott-Schottky relations of Equations 38 to 40. Upon algebraic manipulation, the density of optically active states at a particular illumination energy is given by (Anderson, 1982; Haak et al., 1982, 1984; Goodman et al., 1984)

$$N_t(\lambda) = \pm \frac{8\pi[E - E(A/A^-)]}{\varepsilon_{sc} q} (C_0^2 - C_d^2) \qquad (58)$$

where C_d is the differential capacitance measured at this particular illumination energy (and corresponds to photons of energy just above the threshold energy for trap states of interest) and C_0 is the capacitance measured just prior to reaching the threshold energy for the trap states of interest. At an illumination intensity sufficient to saturate all of the trap states at the energy under study, the measured plateau or peak capacitance, $C_{d,sat}$, corresponds to the total density of bulk trap states, N_t, optically excitable at this particular illumination energy. The $\pm$ in Equation 58 takes into consideration the fact that $N_t(\lambda)$ can either increase or decrease as the space-charge layer is either populated or depleted of charge carriers. Changes in $N_t(\lambda)$ also are related to the character of the state, with acceptor states becoming more positive when emptied and donor states becoming more negative when emptied.

Surface Capacitance Measurements. The density of surface states (in reciprocal centimeters squared) can similarly

be expressed from the Mott-Schottky relationships. In a derivation analogous to the bulk state treatment of Equation 39, the density of surface states at a particular illumination energy can be shown to be (Anderson, 1982; Haak et al., 1982, 1984; Goodman et al., 1984)

$$N_{t,s}(\lambda) = \pm \frac{\varepsilon_{sc} N_d C_H}{8\pi} \left(\frac{1}{C_0^2} - \frac{1}{C_d^2} \right) \quad (59)$$

where C_d is the differential capacitance measured at the threshold energy for which the surface states under study are ionized (populated or depopulated) and C_0 is the capacitance measured at an illumination energy just prior to influencing the surface states under study. As for the case of bulk states, the measured capacitance at saturation is the total number of surface states present at a particular illumination energy, $N_{t,s}$.

Kinetic Rates. The kinetics of filling and emptying optically active states can also be deduced using EPS. Kinetic equations can be derived for both surface and bulk states, but only surface kinetic processes will be discussed in detail. To perform these measurements, the sample is subjected to an intense initial illumination pulse capable of saturating the states of interest. The capacitance is then measured continuously after the illumination is terminated. The time rate of change in the measured capacitance parallels the time rate of change for the concentration of filled surface states and is given by (Haak et al., 1984)

$$\frac{dN_{t,s}}{dt} = \kappa_s [N_{t,s} - N_{t,s}(\lambda)] - \kappa_n n_s \quad (60)$$

where κ_s and κ_n are the ionization (capture) and the neutralization (emission) rate constants, respectively.

Both rate constants are functions of both optical and thermal emission processes. For surface states that become more positive when emptied (acceptor states), these rate constants are given as (Haak et al., 1984)

$$\kappa_s = c_{n,t} + c_{n,o} + k_{p,s} p_s \quad (61a)$$

$$\kappa_n = c_{p,t} + c_{p,o} + k_{n,s} n_s \quad (61b)$$

For surface states whose character becomes more negative when emptied (donor states),

$$\kappa_s = c_{p,t} + c_{p,o} + k_{n,s} n_s \quad (62a)$$

$$\kappa_n = c_{n,t} + c_{n,o} + k_{p,s} p_s \quad (62b)$$

In defining these rate constants, $c_{p,t}$ and $c_{p,o}$ are the thermal and optical emission rate constants for holes, respectively, and $c_{n,t}$ and $c_{n,o}$ are the thermal and optical emission rate constants for electrons, respectively. The quantities $k_{p,s}$ and $k_{n,s}$ are the capture coefficients for electrons and holes defined in Equation 43. It is important to note that while an increase in positive charge produces an increase in capacitance for an n-type semiconductor, the opposite effect is observed for p-type semiconductors.

Extracting the Surface Recombination Capture Coefficient. When the semiconductor is a wide-band-gap material with deep-level states, one can measure the time dependence of the differential capacitance after interruption of the illumination, C_t, and plot

$$\ln \left(\frac{C_0^2 - C_t^2}{C_0^2 - C_{t=0}^2} \right) \quad \text{for bulk states} \quad (63a)$$

or

$$\ln \left(\frac{C_0^2 - C_t^2}{1/C_0^2 - 1/C_{t=0}^2} \right) \quad \text{for bulk states} \quad (63b)$$

vs. time, where $C_t = 0$ is the differential capacitance measured at the time of the interruption of the illumination. For an n-type material with deep electron trap states, $c_{n,t}$ and $k_{p,s}$ are zero, and the resulting slope for these plots equals $c_{p,t} + k_{n,s} n_s$. If the surface or bulk states that give rise to the plotted capacitance are sufficiently above the valance band and hence not thermally ionizable by holes, then $c_{p,t} = 0$, and the product of $k_{n,s}$ and n_s can be calculated directly for surface states by EPS measurements (Haak et al., 1984).

Practical Aspects of the Method

Electrolyte Selection. Ideally, the electrolyte should be transparent to light ranging from the band gap energy down to $\sim E_g$. This range permits the entire band gap region to be probed. Additionally, the solvent/electrolyte should be a good electrical conductor. The electrolyte must also be chosen such that the surface is stable both in the dark and under illumination (Haak et al., 1984).

Aqueous systems appear to provide the most stability for oxygen-passivated semiconductor surfaces. Studies have shown that for n-CdSe, the interface states associated with oxygen adsorption remain unchanged in aqueous solution relative to the situation for these surfaces in contact with a vacuum (Goodman et al., 1984; Haak et al., 1984).

Measurement Frequency. As with the Mott-Schottky experiments, the frequency of the perturbation voltage can be adjusted in EPS to maximize the signal-to-noise ratio. The voltage perturbation frequency used in an EPS experiment should be chosen to yield a relatively large value for the phase shift between the voltage perturbation and the current response. Generally, a phase shift of between 70° and 90° is acceptable. However, to minimize errors in the EPS spectrum, the semiconductor-solution system should show a limited range of frequency dispersion. A large frequency dispersion is usually an indication of poor crystalline quality in the semiconductor and will generally lead to irreproducible EPS spectra (Haak et al., 1984).

Electrode Bias Voltage. The EPS spectral features associated with bulk states generally vary linearly with the space-charge depletion width, whereas those associated with surface or interface states are relatively insensitive

to the electrode potential. Thus, variation of the electrode potential is a powerful method to distinguish between bulk and surface states. The electrode potential is also the experimental parameter that determines the depth into the semiconductor that is probed by the EPS experiment.

When measuring the effect of passivation or varying the concentration of electrolyte on surface state density, the bias voltage should be chosen such that the semiconductor is strongly reverse biased. Since bulk states are unchanged in response to surface modification, the extent of surface state perturbation induced by surface passivation or electrolyte dissolution is deduced by the residuals between two EPS spectra taken at large reverse biases prior to and just after the surface modification procedure (Haak et al., 1984).

Experimental Setup. The experimental setup is analogous to that used in Mott-Schottky experiments described under Differential Capacitance Measurements of Semiconductor-Liquid Contacts, Principles of the Method. A significant difference is that the working electrode is illuminated with monochromatic light provided by a xenon lamp or tungsten halogen lamp and a monochromator. The differential capacitance is measured in a fashion analogous to that described for Mott-Schottky measurements and the data are plotted, at a fixed frequency, as a function of the wavelength of this monochromatic light. At each wavelength, the perturbation voltage is applied at a rate determined by the chosen fixed frequency. During each cycle time (1/frequency), the real and the imaginary components of the electrode impedance are determined and amplified by a differential amplifier. The resulting differential capacitance values measured at each of these cycles are then averaged by a frequency response analyzer. For a thorough analysis of trap states in a semiconductor, various spectra should be collected, each varying with respect to the degree of applied bias. In most EPS experiments, changes in the capacitance as small as 1 part in 10^4 can be measured (Haak et al., 1984).

Prior to any EPS measurement, it is necessary to record an optical transmission spectrum of the solution in a cell with a path length equivalent to the distance between the EPS cell's optical window and the working electrode to be illuminated. This procedure is necessary to determine which features in the bias-independent part of an EPS spectrum might result from solvent absorption (Haak et al., 1984).

Problems

Other concerns with EPS solutions are surface degradation and film formation. While for most experimental EPS measurements, the effects of surface degradation and film formation caused by the accumulation of photogenerated minority carriers at the solid-liquid interface is minimal, it does become a concern when a sample is to undergo measurements over the course of hours or days. To help counter the effect of surface degradation of film formation caused by the accumulation of photogenerated minority carriers at the solid-liquid interface, an additional redox species can be added to the electrolyte in order to scavenge carriers that arrive at the semiconductor-liquid contact. This process can, however, have a detrimental effect if the redox species is present in high concentration, because interfacial charge transfer between the redox species and the semiconductor may produce an overall dark current that can degrade the sensitivity of the EPS experiment (Haak et al., 1984).

LASER SPOT SCANNING METHODS AT SEMICONDUCTOR-LIQUID CONTACTS

Principles of the Method

Laser spot scanning (LSS) involves recording the photocurrent while a highly focused laser spot is scanned over a semiconductor surface. The technique has seen broadened appeal since the early 1990s, when higher laser fluences and spot sizes on the order of 3 μm became available through use of fiber optics. Laser spot scanning offers the opportunity to monitor surface recombination lifetimes as a function of varying treatments to the semiconductor surface as well as other local parameters of semiconductor-liquid contacts (Mathian et al., 1985; Carlsson and Hom-ström, 1988; Carlsson et al., 1988; Eriksson et al., 1991).

Four steps are thought to be involved in producing LSS spectra. Assuming a circular laser spot of radius r, the first step of an LSS experiment involves the generation of minority carriers by the laser spot (Eriksson et al., 1991). The flux of photogenerated carriers is equal to (Eriksson et al., 1991)

$$F(\lambda) = G_0(r)2\pi r\,dr \tag{64}$$

where $G_0(r)$ is the Beer-Lambert generation function. Carriers are thus generated in a cylinder whose radius is equal to the radius of the laser spot and whose depth is α^{-1}, where α is the absorption coefficient. Generally, the illumination wavelength is chosen such that α^{-1} is larger than the depletion width.

Due to the strong electric fields present in the depletion region, minority carriers generated in the depletion region should migrate rapidly, with little sideways diffusion, to the solid-liquid interface. However, minority carriers that are created below the depletion layer will diffuse laterally and either contribute a background photocurrent or undergo nonradiative recombination. The extent to which these carriers contribute to the detected photocurrent depends on their lifetime in the semiconductor bulk. The shorter the lifetime, the less prominent their diffusion and hence the smaller the detected photocurrent.

The minority carriers generated in the depletion region are thus the primary source of the detected photocurrent. The equations for the diffusion currents of these carriers are (Eriksson et al., 1991)

$$\text{Difusion into spot} = J_D(r)2\pi r\,dr \tag{65}$$

$$\text{Diffusion out of spot} = -J_D(r+dr)2\pi r\,dr \tag{66}$$

where $J_D(r)$ is the diffusion flux of minority carriers along the surface. Through Fick's second law, one obtains

$$J_D(r) - J_D(r+dr) = \frac{Dr^2P_s(r)}{dr^2} \tag{67}$$

where D now defines the minority (hole for an n-type semiconductor) carrier diffusion coefficient. The value of D is related to the minority carrier mobility μ_p by $\mu \equiv Dq/kT$, and $p_s(r)$ is the radial dependence of the minority carrier concentration measured from the center of the laser spot (Sze, 1981; Carlsson and Homström, 1988; Eriksson et al., 1991).

Experimentally, LSS spectra record the change in photocurrent as a function of position on the sample. The recorded photocurrent becomes lower in regions where surface recombination is rapid. Thus, the LSS current is expected to decrease in the vicinity of step edges or grain boundaries, regions of polycrystalline growth, or other areas of rapid surface recombination. This lateral profile of surface recombination is not available directly in a traditional spectral response experiment, which does not use a tightly focused source of illumination (Carlsson and Homström, 1988; Carlsson et al., 1988; Eriksson et al., 1991).

Unlike EPS, the lifetime of minority carriers in LSS is determined by two processes, surface recombination and charge transfer:

$$-\tau_{s,l}^{-1} p_s(r) 2\pi r\, dr \qquad \text{surface recombination} \tag{68}$$

$$-k_{ht} p_s(r) 2\pi r\, dr \qquad \text{charge transfer} \tag{69}$$

where k_{ht} is the minority carrier charge transfer rate constant. Recent modeling has shown that the radial dependence for the concentration of minority carriers relative to their concentration at the laser spot's perimeter, r_0, is (Eriksson et al., 1991)

$$\frac{p_s(r)}{p_s(r_0)} = \exp\left[-(1 - k_{ht}\tau_s)^{1/2} \frac{\Delta r}{L}\right] \tag{70}$$

where $\Delta r = r - r_0$ and L is the diffusion length, which, when only surface recombination processes are considered, is defined as $L \equiv (\tau_s D)^{1/2}$ (Many et al., 1965). If we further define r_{eff} as the radial distance where the minority carrier concentration is $1/e$ of its value at the laser spot's perimeter, it can be shown that

$$\frac{r_{eff}}{L} = (1 + k_{ht}\tau_s)^{-1/2} \tag{71}$$

Thus, when the surface recombination lifetime is short, or the charge transfer to the electrolyte is slow, $r_{eff} = L$. In contrast, when either surface recombination or the charge transfer process is fast, $r_{eff} = 0$. This latter situation corresponds to the limit in which essentially all of the minority carriers have been contained and consumed within the laser spot area (Eriksson et al., 1991).

Clearly, the deficiency of LSS is that one cannot distinguish the contribution of surface recombination from the effect of charge transfer kinetics on the induced photocurrent. Thus, to obtain quantifiable surface recombination lifetimes, either the interfacial charge transfer rate must be slow or the exact magnitude of the charge transfer rate constant must be available independently. Nevertheless, recent experiments with LSS have shown the extent to which layers and step edges influence the combined surface lifetime of minority carriers. For an InSe sample, LSS has shown that a step edge, which presents a barrier against lateral diffusion and hence allows carriers to locally accumulate on a surface, can increase the rate of surface recombination by nearly 20-fold (Mathian et al., 1985; Eriksson et al., 1991). Further experiments on polycrystalline silicon have shown that the recorded change in light-induced photocurrent as a function of distance from a step or grain boundary, l, can be approximated as (Mathian et al., 1985; Eriksson et al., 1991)

$$I_{ph}(\lambda) = k_{LSS} \frac{L}{2D} \left(\frac{1}{1 + k_{ht}\tau_s L/2D}\right) \exp\left(-\frac{l}{L}\right) \tag{72}$$

where K_{LSS} is a constant related to the effective circuit used in the LSS experiments (Mathian et al., 1985; Eriksson et al., 1991).

Practical Aspects of the Method

One aspect of the appeal of LSS is the degree of simplicity associated with the experimental setup. The sample or electrochemical cell is placed on a precision micromanipulator that is equipped with motorized control of the sample position. For greater control of the sample position at the expense of the magnitude of the motion, the sample can be placed onto a piezoelectric mount. Alternatively, as with scanning probe microscopy techniques, the sample can be held in a fixed position and a fiber-optic light source can be mounted on a piezoelectric drive to provide a means of rastering the light source over the sample (Mathian et al., 1985; Carlsson et al., 1988; Carlsson et al., 1988; Eriksson et al., 1991).

The mounted sample requires only an ohmic back contact such that reverse-bias conditions can be achieved and photocurrents measured. The light source in LSS must be monochromatic, but inexpensive light sources such as a He-Ne laser suffice for experiments with most semiconductors. The only excitation requirement is that the penetration depth α^{-1} at the excitation wavelength has to be larger than the depletion width of the semiconductor under reverse bias.

To obtain the LSS spectra, the laser light is passed into an optical fiber and focused onto the sample surface. In electrochemical cells, the fiber optic can be embedded through the cell case. The photocurrent data are then collected as a function of laser spot position for either various ambient atmospheres or electrolyte concentrations (Carlsson et al., 1988; Carlsson et al., 1988; Eriksson et al., 1991).

In a fashion analogous to displaying scanning tunneling microscopy SCANNING TUNNELING MICROSCOPY or atomic force microscopy (AFM) data, the measured photocurrents obtained while rastering the laser spot across the sample can be represented by shades of color. Usually, a linear shading of 256 colors is associated with the range of measured photocurrents. If, as a function of sample position, the measured photocurrent is represented by its color shades, then one will obtain an LSS topograph spectrum of a sample surface. To increase the signal-to-noise ratio of an LSS spectrum or topography scan, the laser light can be chopped prior to entering the fiber optic, and a

lock-in amplifier can then be used to record and average the photocurrent that is synchronous with the chopping rate. With respect to apparatus, LSS experimentation is therefore very similar to spectral response equipment (Mathian et al., 1985; Carlsson et al., 1988; Eriksson et al., 1991).

When high-resolution LSS spectra or topographies are required, the intensity of the light prior to entering the fiber optics is measured and regulated by means of a beamsplitter, a PMT, and a series of computer-controlled light attenuators. While the spatial resolution of most LSS surface topography is limited by the spatial profile of the laser spot, the contrast or sensitivity of the method to topography variations is determined by variations in the intensity of the laser beam.

By bundling a series of fiber optics around the central light-emitting optical fiber, one can measure the intensity of light reflected from the sample surface. This in turn provides another channel of data on the reflective characteristics, and hence electronic characteristics, of the surface as a function of the photovoltage or the surface recombination processes (Eriksson et al., 1991).

Problems

As stated above, the deficiency of LSS is that one cannot distinguish the contribution of surface recombination from the effect of charge transfer kinetics on the induced photocurrent. Thus, the interfacial charge transfer rate for the system under study must be slow or the exact magnitude of the charge transfer rate constant must be available independently. One, mechanical, concern is that the ohmic contact needs to remain intact during LSS experiments. Hence, corrosive systems such as Si-HF, which are accessible with time-resolved methods, are not practical for LSS experiments. Another concern is assuring that the sample's dopant density is sufficiently small that the solution-induced depletion width is $<\alpha^{-1}$ of the rastered light beam. Randomly increasing the dopant density can, however, lead to confounding problems. Since under the LSS method one cannot determine if a variation in the photocurrent results from surface defects or from defects buried within the depletion region, changes in the photocurrent arising from the dopant atoms and/or buried defect sites can easily be attributed to poor surface qualities. One method of deconvoluting buried effects from surface effects is to overlay an LSS spectrum onto a conventional topographic image of the surface, such as those obtainable from AFM or transmission electron microscopy (TEM), and to correlate variations in the LSS signal with visible topographic features. For moderately doped semiconductors, however, the influences of buried defects are minimal.

FLAT-BAND POTENTIAL MEASUREMENTS OF SEMICONDUCTOR-LIQUID INTERFACE

Principles of the Method

We now describe methods that can be used to probe the properties of the semiconductor-liquid interface. Energetic and kinetic processes are equally important, and both can be probed using the methods described here.

As described in Equations 3 to 6, the barrier height is a key energetic property of a semiconductor-liquid contact. The conventional method for measuring this property is to measure the C^2-E behavior of the solid-liquid contact and to use the intercept of the Mott-Schottky plot, Equation 38, to yield the built-in voltage of the semiconductor-liquid interface. We describe this method in somewhat more detail below.

In an ideal semiconductor-liquid junction, for a wide range of values of $E(\mathrm{A/A^-})$, E_{cb} and E_{vb} will remain constant vs. a fixed reference potential. This can be verified by measurements of ϕ_b in solutions of various redox potentials by plotting ϕ_b vs. $E(\mathrm{A/A^-})$. The plot should be linear with a slope of unity, and for an n-type semiconductor the intercept yields the value of E_{cb} in the solvent of concern.

There are, however, several cases in which the plots of ϕ_b vs. $E(\mathrm{A/A^-})$ do not have a slope of unity. For an n-type semiconductor, at very positive redox potentials, the band edges will appear to shift as $E(\mathrm{A/A^-})$ is changed. This will occur because at some point the electrochemical potential will become sufficiently positive that ionization of atoms in the valence band of the semiconductor will be thermodynamically feasible. At this point, the dopant atoms need not be ionized to produce charge transfer equilibrium in response to an increase in the electrochemical potential of the solution phase. Instead, the required charge can be obtained from ionization of electrons associated with the lattice atoms of the solid. Under these conditions, the depletion width is essentially constant, and thus no change in the barrier height is observed as $E(\mathrm{A/A^-})$ is increased. Such conditions are called carrier inversion, because they occur whenever the Fermi level of an n-type semiconductor becomes sufficiently positive that the minority carrier concentration at the surface of the solid exceeds the majority carrier concentration in the bulk of the semiconductor.

Another source of a lack of ideality in a plot of ϕ_b vs. $E(\mathrm{A/A^-})$ is due to surface states. If a sufficient density of surface states is present at the solid-liquid interface, the charge required to respond to a change in $E(\mathrm{A/A^-})$ can come from ionization of surface state localized charge instead of ionization of dopant atoms. This situation also produces little or no change in W and thus again produces little or no change in ϕ_b as $E(\mathrm{A/A^-})$ is varied. This situation is called Fermi level pinning, because the Fermi level of the solid appears to be locked into a fixed, pinned position relative to the value of E_{cb}. Under such conditions, E_{cb} will appear to move vs. a fixed reference potential as $E(\mathrm{A/A^-})$ is varied, so again a lack of ideality in a plot of ϕ_b vs. $E(\mathrm{A/A^-})$ is observed.

Unlike carrier inversion, which is a thermodynamic property of the semiconductor-liquid contact, Fermi level pinning arises from the presence of surface states at a particular semiconductor-liquid junction. For a given surface state density, the amount of potential that can be accommodated by emptying or filling surface states is readily calculated if the dielectric constant and thickness of the Helmholtz layer are known. Assuming values for these

constants of $4\varepsilon_0$ and 3×10^{-8} cm, respectively, one obtains, for the maximum voltage drop,

$$\Delta V = (1.357 \times 10^{-14} \text{V} \cdot \text{cm}^2) N_{t,s} \quad (73)$$

For smaller values of $N_{t,s}$, part of the charge transferred in response to a change in $E(\text{A/A}^-)$ is removed from surface states and part is removed from ionization of dopant atoms. In this case, a plot of ϕ_b vs. $E(\text{A/A}^-)$ is often linear but has a slope of < 1.0. The value of $N_{t,s}$ at different energies can often be determined by plotting ϕ_b vs. $E(\text{A/A}^-)$ and observing the region over which the slope is not unity. Data on the density of surface states vs. energy in the band gap are thus complementary to those obtained by EPS.

A final region in which a plot of ϕ_b vs. $E(\text{A/A}^-)$ is not linear is when the Boltzmann relationship between the electron concentration at the semiconductor surface and the electron concentration in the bulk of the solid, Equation 16, is no longer valid. This occurs when the barrier height is so low that $E(\text{A/A}^-) \approx E_{cb}$. Under these conditions, $n_s \geq n_b$, and Fermi-Dirac statistics must be used to relate n_s to n_b, as opposed to the simplification possible when the occupancy of empty states in the conduction band is so low that the Boltzmann approximation of Equation 16 is valid. This situation, which produces a very large value of J_0, is readily diagnosed because the *J-E* behavior of the semiconductor-liquid interface resembles that of an ohmic contact, with no rectification of current through the solid-liquid contact except at very large values of the applied potential.

Practical Aspects of the Method

The protocols for collection of differential capacitance vs. potential plots have been discussed in depth under Differential Capacitance Measurements of Semiconductor-Liquid Contacts, Principles of the Method. To determine the barrier height the flat-band potential of the solid-liquid contact, the intercept of the Mott-Schottky plot should yield the value of V_{bi} after inclusion of a small correction term, kT/q, which equals 0.0259 V at room temperature. The value of V_{bi} can be related to the barrier height as follows:

$$\phi_b = V_{bi} + \frac{kT}{q} \ln\left(\frac{N_c}{N_d}\right) \quad (74)$$

The value for V_{bi} should be independent of the measurement frequency, and the slope of the plot of C^{-2} vs. E should be in accord with that expected from the value of N_d and the electrode area A_s, according to Equation 39. In addition, a useful check on the validity of the V_{bi} value obtained from the data is to ensure that the experimentally determined value of V_{bi} shifts 59 mV for every decade that N_d is increased. This dependence occurs because the barrier height ϕ_b is independent of the dopant density of the semiconductor, so the value of V_{bi} must change, as given by Equation 74, in response to a variation in N_d. If the majority carrier mobility is known, the value of N_d can be determined independently by four-point probe measurements of the sample resistivity, or it can be determined by Hall effect measurements DEEP-LEVEL TRANSIENT SPECTROSCOPY, which yield both N_d and μ_n directly from the measurement on an *n*-type semiconductor sample (Sze, 1981).

Data Analysis and Initial Interpretation

Ideal Mott-Schottky plots should be nondispersive across all measured voltages and frequencies. Figure 11 shows a representative Mott-Schottky plot for the *n*-Si/CH_3OH system, which forms ideal junctions with numerous redox compounds (Fajardo et al., 1995, 1997). As stated, the difference between the closed and open circles is an order-of-magnitude increase in the acceptor concentration, 10 vs. 100 mM, respectively. At each concentration, data points are plotted at input AC voltage frequencies of 100 and 2.5 kHz. This figure illustrates the quality of data required to achieve quantitative results from Mott-Schottky plots. The data are not only nondispersive across the frequency ranges used but also somewhat nondispersive across an order-of-magnitude change in acceptor concentration.

After extrapolation, the open-circle data points had x intercepts of 0.028 and 0.023 V, respectively, a variance of <5 mV. After addition of the correction term to these values to yield corresponding V_{bi} values, Equation 74 can be applied and will yield barrier heights of 518 and 513 mV, respectively. The variance in the barrier height between the closed and open circles was measured to be less than 63 mV. It is critical to show the nondispersive (linear) nature of the Mott-Schottky plots for a wide range of frequencies. Should the Mott-Schottky plots become dispersive, the data set should be re-collected using longer acquisition times. It is also crucial to have short cables, all of a consistent length, and solid well-fitted connectors on all the interconnects between the cell and potentiostat.

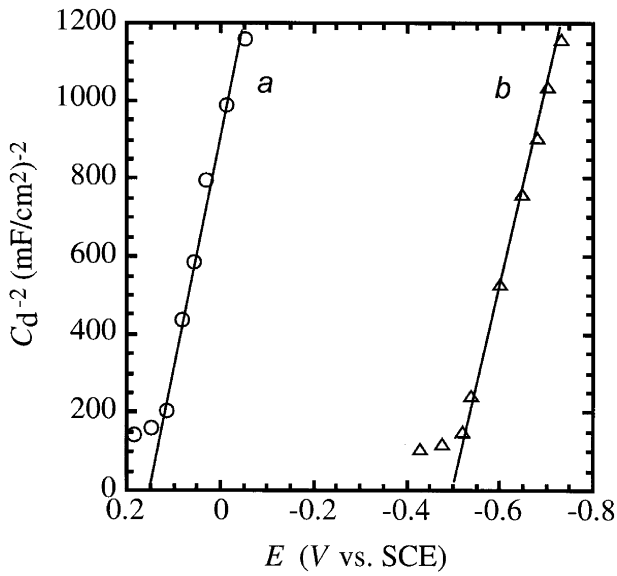


Figure 17. Mott-Schottky plots of *p*-Si/CH_3CN with tetrabutylammonium phosphate (TBAP) electrolyte (*a*) in the absence of redox couples and (*b*) with the redox couples $PhNO_2$ (30 mM) and $PhNO_2^-$ (10 mM). (Reprinted with permission from Nagasubramanian et al., 1982.)

To help prevent current and/or ground loops from developing, the cables must be kept straight. Figure 17 shows somewhat nonideal Mott-Schottky plots for the *p*-Si/CH_3CN-tetra(*n*-butyl)ammonium perchlorate contact, taken at 5 kHz both in the absence of redox couples (open circles) and in the presence of $PhNO_2$ (30 mM) and $PhNo_2^-$ (10 mM) (open triangles) (Nagasubramanian et al., 1982). The slight nonlinearity of these plots makes determination of the flat-band potential to within better than ~0.1 V problematic.

TIME-RESOLVED PHOTOLUMINESCENCE SPECTROSCOPY TO DETERMINE INTERFACIAL CHARGE TRANSFER KINETIC PARAMETERS

Principles of the Method

Measurement of the decay dynamics of photogenerated charge carriers has been discussed under Measurement of Surface Recombination Velocity Using Time-Resolved Microwave Conductivity. In the presence of redox donors and acceptors in the solution phase, measurement of the carrier decay dynamics in the solid can also yield information on the rates of interfacial charge transfer from carriers in the semiconductor to redox species in the solution. Use of this method to determine these important interfacial kinetic parameters is described below.

The rate of disappearance of charge carriers at a semiconductor-liquid contact is given by the SRH model of Equation 42, with additional terms required to account for the interfacial charge transfer processes. Thus the boundary conditions that must be used in evaluating the carrier concentration decay dynamics are

$$j_p/q = k_{ht}[A^-]_s(p_s - p_{0,s}) + k_{p,s}N_{t,s}[f_{t,s}p_s - (1 - f_{t,s})p_{1,s}] \tag{75}$$

and

$$-j_n/q = k_{et}[A]_s(n_s - n_{0,s}) + k_{n,s}N_{t,s}[(1 - f_{t,s})n_s - f_{t,s}n_{1,s})] \tag{76}$$

The values of j_p and j_n represent the currents densities to and from the valence and conduction band, respectively, and $f_{t,s}$ is the fraction of surface states occupied with electrons. Measurement of the decay dynamics in the absence of, and then in the presence of, increasing concentrations of redox species can therefore allow determination of k_{et} and k_{ht} for the semiconductor-liquid contacts of interest.

Practical Aspects of the Method

As described under Transient Decay Dynamics of Semiconductor-Liquid Contacts, Practical Aspects of the Method, the TRPL method is very useful to determine charge carrier capture dynamics to adsorbed species or to species that very rapidly capture charges from the semiconducting electrode. Use of the method is, however, limited to a rather narrow range of potentials, light intensities, and interfacial charge transfer rate constants. In fact, recent simulations show that the photoluminescence decays are generally insensitive to the value of the minority carrier charge transfer rate constant k_{ht} for an *n*-type semiconductor. Instead, diffusional spreading and drift-induced separation of photogenerated carriers in the space-charge layer of the semiconductor dominate the time decay of the observed luminescence signal under most experimentally accessible conditions (Krüger et al., 1994a,b).

This behavior can be readily understood. The laser excitation pulse produces an exponential concentration profile of injected carriers, so the luminescence is initially very intense because it involves bimolecular recombination between electrons and holes that are both created in the same region of space in the solid (Many et al., 1965; Ahrenkiel et al., 1993). Diffusional motion of carriers tends to alleviate this carrier concentration gradient and therefore will eventually eliminate the quadratic recombination to instead produce a simple, exponentially decaying, photoluminescence signal. In the limit of negligible nonradiative recombination, the long-lifetime photoluminescence decay asymptote is thus dictated by the carrier mobilities, the optical penetration depth of the light, and the carrier injection level used in the experiment.

The short-lifetime photoluminescence decay asymptote can also be readily understood. When carriers are removed nonradiatively from the solid as fast as they arrive at the solid-liquid interface, either by surface recombination or by interfacial charge transfer processes (or a combination thereof), the photoluminescence decay will reach a short-lifetime asymptote. Any additional increases in k_{ht} or $N_{t,s}$ cannot affect the photoluminescence decay dynamics because the carrier quenching rate is already limited by diffusion of carriers to the surface. Thus, like the long-lifetime photoluminescence decay asymptote, the short-lifetime photoluminescence decay asymptote is dictated only by the carrier mobilities, the optical penetration depth of the light, and the carrier injection level used in the experiment.

Due to these competing processes, generally, values of k_{ht} and of the minority carrier low-level surface recombination velocity S_p can be obtained from an analysis of the photoluminescence decays only when the following restricted sets of conditions are satisfied simultaneously: 10^1 cm $s^{-1} \leq S_p \leq 10^4$ cm s^{-1}, 10^{-18} cm^4 $s^{-1} \leq k_{ht} \leq 10^{-15}$ cm^4 s^{-1}, and the electrode potential E is in the region $0 < E < +0.15$ V relative to the flat-band potential of the *n*-type semiconductor-liquid interface. The simulations further show that it is not possible to extract a "field dependence" of the charge transfer rate constant when the semiconductor-liquid contact is maintained in reverse bias ($E \geq +0.15$ V vs. the flat-band potential) and is subjected to light pulses that produce low or moderate carrier injection levels, because under such conditions the photoluminescence decay dynamics are dominated by drift-induced charge separation in the space-charge layer of the semiconductor. Under high-level injection conditions, no field dependence can be observed because the majority of the photoluminescence decay dynamics occur near the flat-band condition, so the value of the band bending in the semiconductor under dark, equilibrium conditions has negligible influence on the luminescence transients produced by a high-intensity laser pulse.

A full discussion of the regimes in which the interfacial charge transfer kinetics can be accessed experimentally using this method is beyond the scope of this work, and the reader is referred to the recent literature for an extensive evaluation of the limitations of this method for such purposes. However, it is useful for measurement of the surface recombination velocity of solid-liquid contacts in certain specific cases.

Problems

Details of the method have been described above. The only additional concern for interfacial kinetic measurements is to ensure that electrode corrosion, passivation, adsorption, and other surface-related processes are not significant for the system of concern, because otherwise interpretation of the carrier decay dynamics is problematic. The rate constants should be first order in the acceptor concentration, or there is no proof that the kinetic process of interest is rate limiting. One method for minimally establishing whether surface damage has occurred or not is to collect *J-E* data for the system in the electrolyte of concern and in another electrolyte solution of interest both before and after the photoluminescence decay experiments to ensure that no surface chemistry changes are present that would affect the electrical behavior of the solid-liquid contact. This is necessary, but not sufficient, to secure a robust rate constant determination and is often helpful in ruling out measurements on clearly unstable electrode surfaces. The TRPL data should always be buttressed by measurements at varying concentrations of acceptor in the solution in order to establish whether or not the expected rate law has been observed in the system.

STEADY-STATE *J-E* DATA TO DETERMINE KINETIC PROPERTIES OF SEMICONDUCTOR-LIQUID INTERFACES

Principles of the Method

The final method to be discussed involves the use of steady-state *J-E* measurements of the semiconductor-liquid contact. This method has been described in detail under Photocurrent-Photovoltage Measurements. Here, however, we specifically address the use of steady-state *J-E* data to measure interfacial charge transfer rate constants at semiconductor-liquid contacts.

The governing kinetic equations are Equations 8 and 19. The transfer of charge from the conduction band in the semiconductor to the acceptor species in the solution should be first order in the electron concentration at the surface of the solid, n_s, and first order in the acceptor concentration in the solution, $[A]_s$. Any measurement of k_{et} requires, at minimum, proof of this rate law to meet the conventional benchmarks for the experimental determination of the rate constant of a chemical process. In addition, it is important to establish that one is not simply working in a linear region of an adsorption isotherm, such that the electron capture event is entirely proceeding to adsorbed species, with the concentration of the adsorbate being linearly related to the concentration of the redox species in the solution phase.

The first-order dependence of the rate on $[A]_s$ is readily investigated through variation in the concentration of redox acceptors that are dissolved in the electrolyte. The first- order dependence of the rate on n_s is investigated through the dependence of J on E. According to Equation 19, a plot of ln J vs. E should have a slope of $-q/kT$ if the rate law of Equation 8 is obeyed.

Although this dependence of J on $[A]_s$ is conceptually simple, it has rarely been observed experimentally. Including surface state capture as an intermediate pathway for electron transfer from the semiconductor to the acceptor in solution (Fig. 18) produces the following rate law:

$$\text{Rate} = k_{et}[A]_s n_s + k_{sol}[A]_s N_{t,s} f_{t,s} \qquad (77)$$

where k_{sol} is the electron transfer rate constant from a surface trap to a solution acceptor. At steady state, $df_{t,s}/dt = 0$, producing the following rate expression:

$$\text{Rate} = k_{et}[A]_s n_s + k_{sol}[A]N_{t,s}\left(\frac{k_{n,s}n_s}{k_{n,s}n_s + k_{sol}[A]_s}\right) \qquad (78)$$

where $k_{n,s}$, the capture coefficient for electrons by traps, is the rate constant for transfer of an electron from the conduction band into a surface state. If an electron slowly fills a surface trap and proceeds quickly to the solution, such

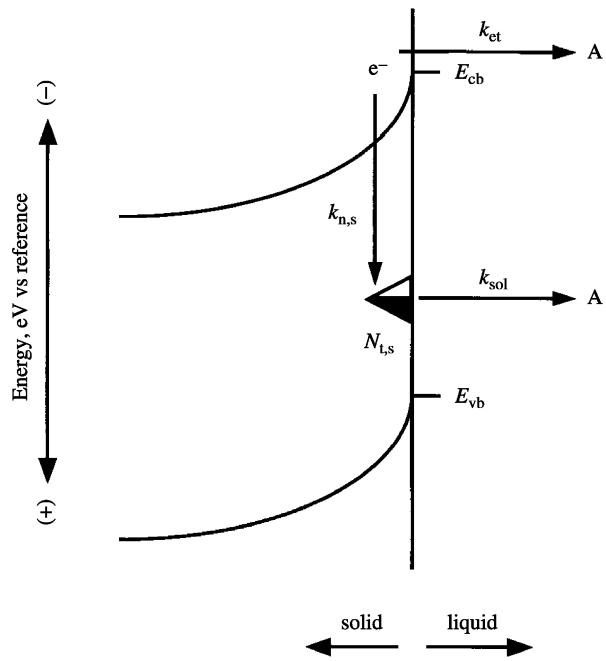


Figure 18. Two possible mechanisms for interfacial electron transfer. The upper pathway shows the direct transfer of electrons from the conduction band edge of the semiconductor to the acceptor in solution, with the rate constant k_{et}. The lower pathway depicts the surface state mediation of interfacial electron transfer. Here, $N_{t,s}$ is the density of surface states, $k_{n,s}$ is the rate constant for electron capture from the conduction band into surface states, and k_{sol} is the rate constant for electron injection from the surface states into solution.

that $k_{sol}[A]_s \gg k_{n,s}n_s$, and $k_{n,s}$ is interpreted as a collisional event through Equation 43, then Equation 78 becomes

$$\text{Rate} = k_{et}[A]_s n_s + (\sigma \nu_n) N_{t,s} n_s \quad (79)$$

The thermal velocity of electrons is well known for n-Si to be 10^7 cm s^{-1}. Equation 79 indicates that, in the scenario of slow electron trapping by surface states and fast electron ejection into solution, a large $N_{t,s}$ will eliminate the dependence of J on $[A]_s$. In the converse case of fast electron trapping and a rate-determining step of electron ejection into solution, such that $k_{n,s}n_s \gg k_{sol}[A]_s$, Equation 78 reduces to

$$\text{Rate} = k_{et}[A]_s n_s + k_{sol}[A]_s N_{t,s} \quad (80)$$

For the rate in Equation 80, a large $N_{t,s}$ will eliminate the dependence of J on n_s. Regardless of which surface state mediation step dictates the rate, a significant density of electrical traps will at best hamper and at worst prevent steady-state measurements of k_{et}. Gerischer (1991) has suggested that redox species couple more strongly to surface states than to conduction band states, and hence even a small $N_{t,s}$ would generally overwhelm the steady-state, electron transfer current flow under practical experimental conditions.

Practical Aspects of the Method

Significant effort must go into preparation of nearly defect-free surfaces in order to extract values for k_{et} from steady-state J-E data. Recent results have shown that this is possible for n-Si and n-InP semiconductor-liquid contacts (Fajardo et al., 1997; Pomykal et al., 1997), and other systems are currently under investigation as well. Special care should be taken in determining the kinetics of semiconductor electrodes according to Equation 8. Rate constants that do not meet these criteria are often quoted in units of centimeters to the fourth power per second (Meier et al., 1997), and this is clearly not in accord with conventional chemical kinetic protocols.

To establish the desired kinetic behavior experimentally, the concentration of the acceptor must be varied (Rosenbluth et al., 1989). However, in doing so, the electrochemical potential of the solution will change if the concentration of the other redox partner is held constant (Equation 1). One approach is to dilute the solution, thereby not varying $E(A/A^-)$ while changing the concentration of the desired redox species. This is useful but often precludes simultaneous differential capacitance measurements, which can require high concentrations of both forms of the redox couple in the electrolyte to avoid spurious results due to concentration polarization at the counter electrode of the system. Another method is to change the concentration of only one form of the redox species. However, care must then be taken to ensure that the band edge positions of the semiconductor do not shift as the redox potential is changed. If this is not the case, interpretation of the data is difficult and problematic.

Once the correct kinetic dependencies on $[A]_s$ and n_s have been established, it is straightforward to evaluate k_{et} from the measured value of J at a given potential. To do this, however, requires an independent measurement of the value of n_s at this potential. Traditionally, C^{-2}-E methods are used for this purpose, and this experimental protocol has been discussed in detail under Differential Capacitance Measurements of Semiconductor-Liquid Contacts, Practical Aspects of the Method. Care should be taken to ensure that the band edge positions do not move versus a fixed reference potential, because then the data are problematic to interpret. Otherwise, simple arithmetic manipulation of Equation 8 yields the desired value of k_{et} if n_s, $[A]_s$, and J are known.

Data Analysis and Initial Interpretation

Figure 19 shows plots of $\ln(-J)$-E for an n-Si electrode in contact with solutions having varying ratios of $[A]/[A^-]$. In this example, an ideal first-order concentration dependence is evident from a +59-mV shift in the $\ln(-J)$-E curve for a factor of 10 increase in $[A]/[A^-]$. A first-order dependence on the surface electron concentration can be verified by fitting $\ln(-J)$-E curves to a standard diode equation:

$$J = -J_0\left[\exp\left(\frac{-qE}{\gamma kT}\right)\right] \quad (81)$$

Note that this expression is a simplified form of Equation 19 and is valid when the electrode is sufficiently reverse biased. The diode quality factor γ is a floating parameter that should yield a value of 1.0 for an electron-transfer-limited system (compare Equations 81 and 19).

Prior to verifying these dependencies, all J-E curves should be corrected for cell resistance and concentration overpotential, as discussed under Photocurrent-Photovoltage Measurements, Data Analysis and Initial Interpretation. Assuming a diode quality factor of unity, the value of the rate constant can be extracted at any given potential. However, in practice, it is best to obtain the value of the rate constant using values of the current for which resistance effects are still minimal and for which the simplified form of the diode equation is still valid.

Problems

As mentioned above under Principles of the Method, to obtain reliable values of the heterogeneous rate constant,

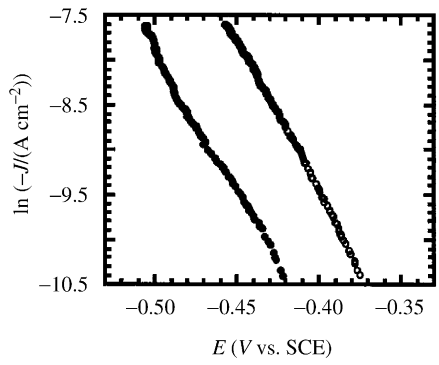


Figure 19. Plots of $\ln(-J)$ versus E for the system described in Figure 4 for two different ratios of $[A]/[A^-]$. The data on the right had a ratio of $[A]/[A^-]$ 10-fold higher than that for the curve on the left.

the semiconductor-liquid system of interest must exhibit second-order kinetics, with a first-order dependence on both the concentration of majority carriers at the surface and the concentration of redox species. However, most systems do not exhibit either one or both of these dependencies. A lack of concentration dependence is more likely in systems that employ redox species or electrolytes that can adsorb on the surface of the electrode and with electrodes that have high surface recombination velocities. Such systems are often dominated by alternate recombination mechanisms precluding a quantitative kinetic analysis of desired rate process. In addition, it is often difficult to obtain sufficiently well-behaved Mott-Schottky plots so that band edge movement can be assessed to within a few millivolts when the redox potential is only varied over a small range of potential. Thus, behavior generally needs to be explored, and verified, for several redox couples that have $E(A/A^-)$ varying over a significant potential range to ensure confidence in the results of the kinetic measurements.

ACKNOWLEDGMENTS

We acknowledge the Department of Energy, Office of Basic Energy Sciences, and the National Science Foundation for their generous support of semiconductor electrochemistry that is incorporated partly in this work. In addition, we acknowledge helpful discussions with G. Coia, G. Walker, and O. Krüger during the preparation of this unit.

LITERATURE CITED

Adachi, S. 1992. Physical Properties of III-V Semiconductor Compounds. John Wiley & Sons, New York.

Ahrenkiel, R. K. and Lundstrom, M. S. 1993. Minority Carriers in III-V Semiconductors: Physics and Applications. *In* Semiconductors and Semimetals (R. K. Willardson, A. C. Beer, and E. R. Weber, eds.) pp. 40–150. Academic Press, San Diego.

Ambridge, T. and Faktor, M. M. 1975. An Automatic Carrier Concentration Profile Plotter Using An Electrochemical Technique. *J. Appl. Electrochem.* 5:319–328.

Anderson, J. C. 1982. Theory of Photocapacitance in Amorphous Silicon MIS Structures. *Philos. Mag. B* 46:151–161.

Aspnes, D. E. and Studna, A. A. 1981. Chemical Etching and Cleaning Procedures for Si, Ge and Some III-V Compound Semiconductors. *Appl. Phys. Lett.* 39:316–318.

Bard, A. J. and Faulkner, L. R. 1980. Electrochemical Methods: Fundamentals and Applications. John Wiley & Sons, New York.

Blakemore, J. S. 1987. Semiconductor Statistics. Dover Publications, New York.

Blood, P. 1986. Capacitance-Voltage Profiling and the Characterization of III-V Semiconductors Using Electrolyte Barriers. *Semicond. Sci. Technol.* 1:7–27.

Carlsson, P. and Homström, B. 1988. Laser Spot Scanning for Studies of the Photoelectrochemical Properties of InSe. *Finn. Chem. Lett.* 121:52–53.

Carlsson, P., Holmström, B., Uosaki, K., and Kita, H. 1988. Fiber Optic Laser Spot Microscope: A New Concept for Photoelectrochemical Characterization of Semiconductor Electrodes. *Appl. Phys . Lett.* 53:965–972.

Chen, M. C. 1988. Photoconductivity Lifetime Measurements on HgCdTe Using a Contactless Microwave Technique. *J. Appl. Phys.* 64:945–947.

Eriksson, S., Carlsson, P., and Holmström, B. 1991. Laser Spot Scanning in Photoelectrochemical Systems, Relation between Spot Size and Spatial Resolution of the Photocurrent. *J. Appl. Phys.* 69:2324–2327.

Fajardo, A. M. and Lewis, N. S. 1997. Free-Energy Dependence of Electron-Transfer Rate Constants at Si/Liquid Interfaces. *J. Phys. Chem. B* 101:11136–11151.

Fajardo, A. M., Karp, C. D., Kenyon, C. N., Pomykal, K. E., Shreve, G. A., Tan, M. X., and Lewis, N.S. 1995. New Approaches to Solar Energy Conversion Using Si/Liquid Junctions. *Solar Energy Mater. Solar Cells* 38:279–303.

Finklea, H. O. 1988. Semiconductor Electrodes. Elsevier, New York.

Fonash, S. J. 1981. Solar Cell Device Physics. Academic Press, New York.

Forbes, M. D. E. and Lewis, N. S. 1990. Real-Time Measurements of Interfacial Charge Transfer Rates at Silicon/Liquid Junctions. *J. Am. Chem. Soc.* 112:3682–3683.

Gerischer, H. 1975. Electrochemical Photo and Solar Cells: Principles and Some Experiments. *J. Electroanal. Chem.* 58:263–274.

Gerischer, H. 1991. Electron-Transfer Kinetics of Redox Reactions at the Semiconductor/Electrolyte Contact: A New Approach. *J. Phys. Chem.* 95:1356–1359.

Gillis, H. P., Choutov, D.A., Martin, K. P., Bremser, M. D., and Davis, R.F. 1997. Highly Anisotropic, Ultra-Smooth Patterning of GaN/SiC By Low-Energy-Electron Enhanced Etching in DC Plasma. *J. Electron. Mater.* 26:301–305.

Goodman, C. E., Wessels, B. W., and Ang, P. G. P. 1984. Photocapacitance Spectroscopy of Surface States on Indium Phosphide Photoelectrodes. *Appl. Phys. Lett.* 45:442–444.

Haak, R. and Tench, R. 1984. Electrochemical Photocapacitance Spectroscopy Method for Characterization of Deep Levels and Interface States in Semiconductor Materials. *J. Electrochem. Soc.* 131:275–283.

Haak, R., Ogden, C., and Tench, D. 1982. Electrochemical Photocapacitance Spectroscopy: A New Method for Characterization of Deep Levels in Semiconductors. *J. Electrochem. Soc.* 129: 891–893.

Hall, R. N. 1952. Electron-Hole Recombination in Germanium. *Phys. Rev.* 87:387–388.

Higashi, G. S., Becker, R. S., Chabal, Y. J., and Becker, A. J. 1991. Comparison of Si(111) Surface Prepared Using Aqueous- Solutions of NH_4F Versus HF. *Appl. Phys. Lett.* 58:1656–1658.

Higashi, G. S., Irene, E. A., and Ohmi, T. 1993. Surface Chemical Cleaning and Passivation for Semiconductor Processing. Materials Research Society, Pittsburgh, Pa.

Kenyon, C. N., Ryba, G. N., and Lewis, N. S. 1993. Analysis of Time-Resolved Photocurrent Transients at Semiconductor-Liquid Interfaces. *J. Phys. Chem.* 97:12928–12936.

Koval, C. A. and Howard, J.N. 1992. Electron Transfer at Semiconductor Electrode-Liquid Electrolyte Interfaces. *Chem. Rev.* 92:411–433.

Krüger, O., Jung, C., and Gajewski, H. 1994a. Computer Simulation of the Photoluminescence Decay at the GaAs-Electrolyte Junction. 1. The Influence of the Excitation Intensity at the Flat-Band Condition. *J. Phys. Chem.* 98:12653–12662.

Krüger, O., Jung, C., and Gajewski, H. 1994b. Computer Simulation of the Photoluminescence Decay at the GaAs-Electrolyte Junction. 2. The Influence of the Excitation Intensity under Depletion Layer Conditions. *J. Phys. Chem.* 98:12663–12669.

Leong, W. Y., Kubiak, R. A., and Parker, E. H. C. 1985. Dopant Profiling of Si-MBE Material Using the Electrochemical CV Technique. Paper presented at the First International Symposium on Silicon MBE, Toronto, Ontario, Electrochemical Society.

Lewis, N. S. 1990. Mechanistic Studies of Light-Induced Charge Separation at Semiconductor/Liquid Interfaces. *Acc. Chem. Res.* 23:176–183.

Lewis, N. S. and Rosenbluth, M. 1989. Theory of Semiconductor Materials. *In* Photocatalysis: Fundamentals and Applications (N. Serpone and E. Pelizzetti, eds.) pp. 45–98. John Wiley & Sons, New York.

Love, J. C. and Demas, J. N. 1984. Phase Plane Method for Deconvolution of Luminescence Decay Data with a Scattered-Light Component. *Anal. Chem.* 56:82–85.

Many, A., Goldstein, Y., and Grover, N. B. 1965. Semiconductor Surfaces. North-Holland Publishing , New York.

Mathian, G., Pasquinelli, M., and Martinuzzi, S. 1985. Photoconductance Laser Spot Scanning Applied to the Study of Polysilicon Defect Passivation. *Physica* 129B:229–233.

Meier, A., Kocha, S. S., Hanna, M. C., Nozik, A. J., Siemoneit, K., Reineke-Koch, R., and Memming, R. 1997. Electron-Transfer Rate Constants for Majority Electrons at GaAs and GaInP2 Semiconductor/Liquid Interfaces. *J. Phys. Chem. B* 101: 7038–7042.

Morrison, S. R. 1980. Electrochemistry at Semiconductor and Oxidized Metal Electrodes. Plenum, New York.

Msaad, H., Michel, J., Lappe, J. J., and Kimerling, L. C. 1994. Electronic Passivation of Silicon Surfaces by Halogens. *J. Electron. Mater.* 23:487–491.

Naber, J. A. and Snowden, D. P. 1969. Application of Microwave Reflection Technique to the Measurement of Transient and Quiescent Electrical Conductivity of Silicon. *Rev. Sci. Instrum.* 40:1137–1141.

Nagasubramanian, G., Wheeler, B. L., Fan, F.-R. F., and Bard, A. J. 1982. Semiconductor Electrodes. XLII. Evidence for Fermi Level Pinning from Shifts in the Flatband Potential of *p*-Type Silicon in Acetonitrile Solutions with Different Redox Couples. *J. Eletrochem. Soc.* 129:1742–1745.

Pankove, J. I. 1975. Optical Processes in Semiconductors. Dover Publications, New York.

Peter, L. M. 1990. Dynamic Aspects of Semiconductor Photoelectrochemistry. *Chem. Rev.* 90:753–769.

Pleskov, Y. V. and Guervich, Y. Y. 1986. Semiconductor Photoelectrochemistry. Consultants Bureau, New York.

Pomykal, K. E. and Lewis, N. S. 1997. Measurement of Interfacial Charge-Transfer Rate Constants at *n*-Type InP/CH3OH Junctions. *J. Phys. Chem. B* 101:2476–2484.

Ramakrishna, S. and Rangarajan, S. K. 1995. Time-Resolved Photoluminescence and Microwave Conductivity at Semiconductor Electrodes: Depletion Layer Effects. *J. Phys. Chem.* 99:12631–12639.

Rosenbluth, M. L. and Lewis, N. S. 1989. "Ideal" Behavior of the Open Circuit Voltage of Semiconductor/Liquid Junctions. *J. Phys. Chem.* 93:3735–3740.

Schroder, D. K. 1990. Semiconductor Material and Device Characterization. John Wiley & Sons, New York.

Seabaugh, A. C., Frensley, W. R., Matyi, R. J., and Cabaniss, G. E. 1989. Electrochemical C-V Profiling of Heterojunction Device Structures. *IEEE Trans. Electron Dev.* ED-36:309–313.

Sharpe, C. D. and Lilley, P. 1980. The Electrolyte-Silicon Interface: Anodic Dissolution and Carrier Concentration Profiling. *J. Electrochem. Soc.* 127:1918–1922.

Shockley, W. and Read, W. T. 1952. Statistics of the Recombination of Holes and Electrons. *Phys. Rev.* 87:835–8342.

Sze, S. M. 1981. The Physics of Semiconductor Devices. John Wiley & Sons, New York.

Tan, M. X., Kenyon, C. N., and Lewis, N. S. 1994a. Experimental Measurement of Quasi-Fermi Levels at an Illuminated Semiconductor/Liquid Contact. *J. Phys. Chem.* 98:4959–4962.

Tan, M. X., Laibinis, P. E., Nguyen, S. T., Kesselman, J. M., Stanton, C. E., and Lewis, N. S. 1994b. Principles and Applications of Semiconductor Photoelectrochemistry. *Prog. Inorg. Chem.* 41:21–144.

Tributsch, H. and Bennett, J. C. 1977. Electrochemistry and Photochemistry of MoS_2 Layer Crystals. I. *J. Electroanal. Chem.* 81:97–111.

Willardson, R. K. and Beer, A. C. 1981. Semiconductors and Semimetals: Contacts Junctions, Emitters. Academic Press, New York.

Wilson, R. H., Harris, L. A., and Gerstner, M. E. 1979. Characterized Semiconductor Electrodes. *J. Electrochem. Soc.* 126:844–850.

Yablonovitch, E., Sandroff, C. J., Bhat, R., and Gmitter, T. 1987. Nearly Ideal Electronic Properties of Sulfide Coated GaAs Surfaces. *Appl. Phys. Lett.* 51:439–441.

Yablonovitch, E., Swanson, R. M., Eades, W. E., and Weinberger, B. R. 1986. Electron-Hole Recombination at the Si-SiO_2 Interface. *Appl. Phys. Lett.* 48:245–247.

KEY REFERENCES

Bard and Faulkner, 1980. See above.

General electrochemistry text with a good introduction to impedance spectroscopy.

Schroder, 1990. See above.

Describes in more detail many of the techniques discussed here, including photoluminescence and photoconductivity decay.

Shockley and Read, 1952. See above.

Classic reference for carrier recombination.

Tan et al., 1984b.

A detailed introduction to the thermodynamcis and kinetics of semiconductor-liquid interface.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

A	Oxidized acceptor species
$A_1(A_2)$	Preexponential floating variables of a biexponential fit to a decay curve
A^-	Reduced acceptor species (donor)
A_s	Surface area of electrode
$[A]_b$	Bulk concentration of oxidized acceptor species
$[A]_s$	Acceptor concentration at semiconductor surface
$[A^-]_s$	Donor concentration at semiconductor surface
B	Circuit susceptance
C	Constant equal to $qk_{et}[A]_s$
C_0	Differential capacitance measured at base of EPS peak
C_d	Experimentally measured differential capacitance of a semiconductor
$C_{d,sat}$	Plateau or peak capacitance measured under high illumination intensity
C_H	Differential capacitance of Helmholtz layer

$c_{n,o}$	Optical emission rate constant for electrons
$c_{n,t}$	Thermal emission rate constant for electrons
$c_{p,o}$	Optical emission rate constant for holes
$c_{p,t}$	Thermal emission rate constant for holes
C_{sc}	Differential capacitance of space-charge layer
C_{ss}	Differential capacitance of semiconductor surface
C_t	Time-dependent differential capacitance
d	Semiconductor sample thickness
D	Carrier diffusion constant
D_0	Diffusion coefficient
$-dn/dt$	Net rate of electron transfer into solution
$\boldsymbol{E}$	Electric field
E	Applied potential
$E(\mathrm{A/A^-})$	Redox potential
$E(\mathrm{A/A^-})$	Electrochemical potential of contacting solution
E_{cb}	Energy of conduction band edge
E_{corr}	Potential corrected for resistance and concentration overpotentials
E_F	Fermi level of semiconductor
E_g	Band gap energy
$E^0(\mathrm{A/A^-})$	Formal electrochemical potential of contacting solution
E_t	Energy of a trap state
E_{vb}	Energy of valence band edge
f	Frequency of AC input voltage signal
F	Faraday's constant
$F(\lambda)$	Flux of photogenerated carriers
$f_{t,s}$	Fraction of surface states filled by electrons
G	Circuit conductance
$G_0(r)$	Beer-Lambert generation function
h	Planck's constant
I	Current
I_{ph}	Photocurrent
I_{pl}	Photoluminescence intensity
J	Current density
J_0	Exchange current density
$J_D(r)$	Diffusion flux of minority carriers along surface
$J_{l,a}$	Limiting anodic current density
$J_{l,c}$	Limiting cathodic current density
j_n (j_p)	Charge transfer current to and from conduction (valence) band
J_{ph}	Photocurrent density
J_{sc}	Short-circuit current density
k	Boltzmann's constant
k_{et}	Interfacial electron transfer rate constant
k_{et}^-	Reverse interfacial electron transfer rate constant
k_{ht}	Interfacial hole transfer rate constant
K_{LSS}	Effective circuit constant in an LSS experiment
k_n (k_p)	Electron (hole) capture coefficient for traps
k_r	Radiative recombination rate constant
k_{soln}	Electron-transfer rate constant from a surface trap to a solution acceptor
l	Step or grain boundary
L	Minority carrier diffusion length
L_p	hole diffusion length
M	Molecular weight of semiconductor
n	Number of electrons transferred
n^+	Number of holes required to oxidize one atom of electrode material
$n_{1,s}$ ($p_{1,s}$)	Concentration of electrons (holes) in conduction (valence) band when trap energy is at Fermi energy
n_b	Electron concentration in bulk of semiconductor
N_c	Effective density of states in conduction band
N_d	Dopant density
n_i	Intrinsic electron concentration
n_s (p_s)	Surface electron (hole) concentration
n_{s0}	Equilibrium surface electron concentration
N_t	Concentration of bulk states (in $\mathrm{cm^{-3}}$)
$N_t(\lambda)$	Density of optically active bulk states in a semiconductor at a particular illumination wavelength
$N_{t,s}$	Concentration of surface states (in $\mathrm{cm^{-2}}$)
$N_{t,s}(\lambda)$	Density of optically active surface states in a semiconductor at a particular illumination wavelength
N_v	Effective density of states in valence band
p_0	Equilibrium minority carrier concentration in the dark
$p_s(r)$	Radial dependence of minority carriers about an LSS laser spot
q	Electronic charge
Q	Charge density
r	Radius of laser spot
R	Gas constant
R^*	Optical reflectivity of solid
r_0	Perimeter of laser spot
r_{eff}	Distance from center of laser spot where concentration of minority carriers is $1/e$ of its value at perimeter of laser spot
R_H	Resistance of Helmholtz layer
R_r	Rate of radiative band-to-band emission
R_s	Resistance of bulk semiconductor
R_{sc}	Resistance of space-charge layer
R_{soln}	Solution resistance
R_{SRH}	Rate of Shockley-Read-Hall (SRH) nonradiative recombination in semiconductor
R_{ss}	Resistance due to semiconductor surface states
S_{high}	High-level surface recombination velocity (in $\mathrm{cm\ s^{-1}}$)
S_{low}	Low-level surface recombination velocity for electrons (holes)
T	Temperature
V	Electric potential
V_{bi}	Built-in voltage
v_n (v_p)	Thermal velocity of electrons (holes) in a solid
V_n	Potential difference between Fermi level and conduction band in bulk
V_{oc}	Open-circuit voltage
W	Depletion width
W'	Depth of dopant density measurement ($W + W_d$)
W_d	Thickness of material dissolved from an electrode
x	Distance into semiconductor
Z_{im}	Imaginary component of impedance
Z_{re}	Real component of impedance

Z_{tot}	Total impedance
α	Absorption coefficient
β	Coefficient of optical transitions
γ	Diode quality factor
Γ	Transmitted photon flux
Γ_0	Incident photon flux
Δp (Δn)	Optically induced excess minority (majority) carrier concentration
Δp_0	Optically induced excess minority carrier concentration directly after excitation ($t = 0$)
$\Delta R_m(t)$	Reflected microwave power
ε	Relative permittivity
ε_0	Permittivity of free space
ε_s	Static dielectric constant of semiconductor
η_{conc}	Concentration overpotential
θ	Phase angle
κ_n	Neutralization (emission) rate constant
κ_s	Ionization (capture) rate constant
λ	Excitation wavelength
μ_n (μ_p)	Mobility of electrons (holes) in semi- conductor
ν	Frequency of incident radiation
ρ	Density of solid
σ_n (σ_p)	Capture cross-section of electrons (holes) in a trap
τ	Minority carrier lifetime
$\langle\tau\rangle$	Weighted lifetime parameter for fit to a decay curve
τ_1 (τ_2)	Lifetimes of biexponential fit to a decay curve
τ_{pl}	Effective total lifetime for photoluminescence spectrum
$\tau_{r,h}$	High-level injection radiative lifetime
$\tau_{r,l}$	Low-level injection radiative lifetime
$\tau_{s,h}$	High-level SRH lifetime associated with surface recombination
$\tau_{s,l}$	Low-level SRH lifetime associated with surface recombination
υ	Kinematic velocity of a solution
Φ	Quantum yield
ϕ	Barrier height
χ_c	Capacitive reactance
ω	Angular frequency of AC input signal
ω_{rde}	Angular velocity of rotating disk electrode

SAMIR J. ANZ
ARNEL M. FAJARDO
WILLIAM J. ROYEA
NATHAN S. LEWIS
California Institute of Technology
Pasadena, California

SCANNING ELECTROCHEMICAL MICROSCOPY

INTRODUCTION

Scanning electrochemical microscopy (SECM) is one of a number of scanning probe microscopy (SPM) techniques that arose out of the development of the scanning tunneling (Binnig and Rohrer, 1982) and atomic force microscopes (Binnig et al., 1986). Scanning probe microscopes operate by scanning, or "rastering," a small probe tip over the surface to be imaged. The SECM tip is electrochemically active, and imaging occurs in an electrolyte solution. In most cases, the SECM tip is an ultramicroelectrode (UME) (Wightman, 1988; Wightman and Wipf, 1989), and the tip signal is a Faradaic current from electrolysis of solution species. Some SECM experiments use an ion-selective electrode (ISE) as a tip. In this case, the tip signal is usually a voltage proportional to the logarithm of the ion activity in solution. The use of an electrochemically active tip allows an extremely versatile set of experiments, with chemical sensitivity to processes occurring at a substrate surface as an essential aspect.

A requirement of SPM techniques is that the signal from the tip must be perturbed in some reproducible fashion by the presence of the surface. One of the two methods used in SECM to provide this signal change is known as the "feedback" mode (Kwak and Bard, 1989; Bard et al., 1991, 1994). The feedback mode employs a Faradaic current that flows from the electrolysis of a mediator species [e.g., ferrocene or $Ru(NH_3)_6^{3+}$ at a UME tip as a probe of a substrate surface. As the tip is brought close to the surface, the electrolysis current is perturbed in a manner that is characteristic for electronically insulating or conducting surfaces. "Positive" feedback occurs when the mediator is restored to its original oxidation state at the substrate by an electrochemical, chemical, or enzymatic reaction (Fig. 1A). The regeneration of the mediator in the tip-substrate gap increases the tip current as the gap width decreases. "Negative" feedback occurs when the surface physically blocks the diffusion of mediator molecules to the tip electrode, producing a drop in tip current with a decrease in gap width (Fig. 1B). Thus, feedback can provide topographic images of either electronically insulating or conducting surfaces. Examples from many types of surfaces have been obtained, such as electrodes (Bard et al., 1989; Lee et al., 1990b; Wittstock et al., 1994a) or polymer films (Kwak et al., 1990; Lee and Bard, 1990; Jeon and Anson, 1992; Borgwarth et al., 1996).

A unique advantage of SECM is the ability to design experiments in which the mediator interacts with the substrate surface to provide chemical and electrochemical activity maps at micrometer and submicrometer resolution. In reaction rate feedback (Fig. 1C), the imaging conditions are manipulated so that the overall mediator turnover rate at the substrate is limited by the rate of substrate-mediator electron transfer (Wipf and Bard, 1991a,b; Casillas et al., 1993). Experiments have demonstrated that the microscope can be used to map variations in electron transfer (ET) rate at metallic electrode surfaces (Wipf and Bard, 1991b) and enzymes in biological materials (Pierce and Bard, 1993).

In the generation/collection (GC) mode, the tip signal arises from a species generated at the surface of the imaged material (Fig. 1D). Ideally, the tip acts only as a passive sensor with the ability to produce concentration maps of a particular chemical species near the substrate surface. In amperometric GC, a UME tip detects species by electrolysis and the first reported use of the method

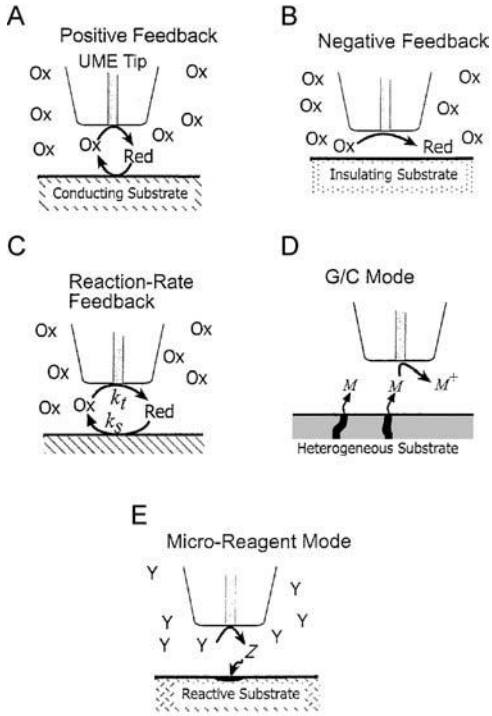


Figure 1. Modes of operation of the SECM. Ox and Red refer to oxidized and reduced forms of intentionally added mediator.

was to map electrochemically active areas on electrodes (Engstrom et al., 1986). Generation/collection SECM has been used to make high-resolution chemical concentration maps of corroding metal surfaces, biological materials, and polymeric materials. In addition, spatially resolved quantitative measurement of ion flux through porous material such as mouse skin and dental material can be made. The potentiometric GC mode uses an ISE tip, which has the advantages of minimizing perturbation of the diffusion layer, sensitivity to nonelectroactive ions, and selectivity for the imaged ion.

In addition to imaging, the SECM tip can be a microscopic source or sink of electrons and chemical reagents. With the tip positioned close to surface, these reagents perform modifications to surfaces on a microscopic scale, turning the scanning electrochemical microscope into a versatile microfabrication device (see Specimen Modification). In the "direct" mode, the tip is a counter electrode to the substrate electrode, permitting substrate electrochemical reactions at a small region of approximately tip size. Deposition of metal and polyaniline lines as thin as 0.3 μm has been reported using the direct mode (Craston et al., 1988; Hüsser et al., 1988, 1989; Wuu et al., 1989). In the microreagent mode, a reaction at the SECM tip changes the local solution composition to provide a driving force for chemical reactions at the surface (Fig. 1E). Electrolytic generation of an oxidizing agent at the SECM tip can precisely etch metal and semiconductor surfaces, and local pH changes caused by electrolysis at the SECM tip have been used to deposit metal oxides and polymers (Mandler and Bard, 1989; Zhou and Wipf, 1997; Kranz et al., 1995a).

This unit provides an overview of SECM and basic experimental apparatus. Applications of SECM will focus on experiments in corrosion, biological and polymeric materials, and the dissolution and deposition of materials. The use of SECM to measure ET rates at inert electrodes and rapid homogeneous reactions accompanying ET is not specifically covered here, but these topics are discussed in other review articles (Bard et al., 1991, 1994, 1995). The discussion of recent experiments using SECM to examine material transfer rates across phase boundaries is also available in the literature (Selzer and Mandler, 1996; Slevin et al., 1996; Tsionsky et al., 1996; Shao et al., 1997b).

Competitive and Related Techniques

Scanning electrochemical microscopy is a scanned probe method operating in an electrolyte solution. In this regard, SECM is comparable to two of the most popular SPM techniques, scanning tunneling microscopy (STM) (see SCANNING TUNNELING MICROSCOPY) and atomic force microscopy (AFM). Both techniques have much higher vertical and lateral resolution than the SECM and, even with the tip immersed in electrolyte, can achieve resolution on the nanometer scale. Atomic force microscopy is a very versatile method and can image both insulating and conducting surfaces with equal facility. The technique can achieve some chemical specificity by use of chemically modified tips to provide molecular recognition (Dammer et al., 1995) or by correlating lateral frictional forces and details of the force-distance curve with surface composition. However, other than the indirect effect that solution species have on surface forces, monitoring solution phase chemistry is not within the scope of AFM. Operating the scanning tunneling microscope in electrolyte solution presents a significant handicap. Faradaic currents must be minimized to allow the tunneling current to be observed. The STM tips must be well insulated up to the very end of the electrode and the tip-substrate bias must be chosen carefully to avoid unwanted Faradaic reactions. These problems are not insurmountable, and STM imaging of ionic or electronically conducting interfaces in electrolyte solution is common. As in AFM, direct chemical information about solution phase species is not generally available, but the electrically conducting STM tip can provide voltammetric or amperometric signals. In this case, tunneling current is no longer monitored, and the scanning tunneling microscope is essentially acting as a scanning electrochemical microscope. The STM tips are generally made from tungsten or Pt-Ir alloys, which permits formation of atomically sharp tips but restricts somewhat the scope of available electrochemical reactions. Whether one considers operation of the scanning tunneling micrscope in electrolyte solution to be an SECM experiment depends more on philosophy. In general, one can make the distinction by considering that the primary interaction in SECM between the tip and sample is mediated by diffusion of solution species between the tip and sample. The number of applications of SPM is growing at an exponential rate; comparison of SECM with other SPM methods should start with a good review of the field (Bottomley et al., 1996).

Scanning electrochemical microscopy also bears a resemblance to older methods in which a scanning microscopic reference electrode or a vibrating probe is used to detect localized current flow near corroding metal surfaces (Isaacs and Kissel, 1972; Isaacs et al., 1996) and living cells (Jaffe and Nuccitelli, 1974). These methods generally have lower spatial resolution and are not chemically specific.

PRINCIPLES OF THE METHOD

In SECM an electrochemically active tip is used to detect and produce species in solution. This is important, since a quantitative description of the tip response under most measurement conditions can be developed by using well-understood theories relating the electrochemical response, the equations for mass transport, and interfacial kinetics.

Feedback Mode

A brief description of the voltammetric response at a UME is required to understand the feedback mode. The cyclic voltammetry (CV) experiment in Figure 2 illustrates the general behavior observed at UMEs (see CYCLIC VOLTAMMETRY). The CV wave shown is for the reduction of the commonly used mediator ion, $Ru(NH_3)_6^{3+}$, at a 10-μm-diameter tip in an unstirred buffer solution. Note that the wave has a sigmoidal shape, and at potentials negative of the reduction potential, the current flow reaches a steady-state limiting current value. As the tip potential is swept back to the starting potential, the forward curve is retraced. In this experiment, the current flow is limited by the rate of diffusion of the mediator species to the electrode surface. At a UME, the electrode is smaller than the region of solution perturbed by diffusion to the electrode surface, resulting in a particularly efficient form of mass transport in which the diffusion field assumes a nearly hemispherical zone around the electrode (Wightman and Wipf, 1989). The zone is able to provide a constant supply of fresh mediator as it extends away from the electrode surface. In contrast, at short time scales or with larger electrodes, the diffusion field is smaller than the electrode dimension, and so diffusion is only possible in a direction perpendicular to the electrode surface (i.e., planar diffusion). In planar diffusion, the current decays with a square-root dependence on time and does not reach a constant value.

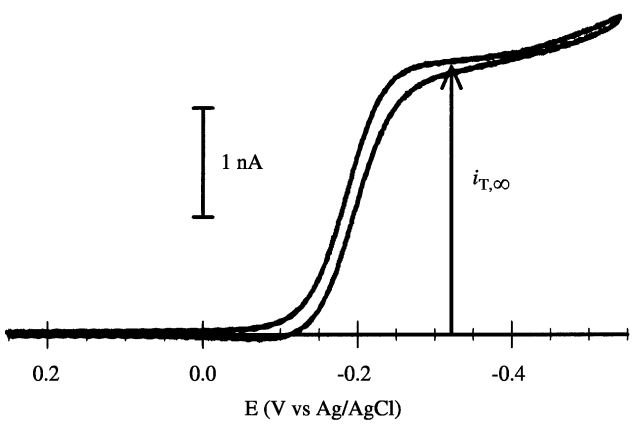


Figure 2. Cyclic voltammogram of 2.1 mM $Ru(NH_3)_6^{3+}$ in pH 7.1 phosphate-citrate buffer at a 10-μm-diameter Pt disk electrode. The voltammogram is for a scan rate of 100 mV/s.

The value of the steady-state limiting current, $i_{T,\infty}$, for a microdisk electrode embedded in an insulator is given by the equation (Wightman and Wipf, 1989)

$$i_{T,\infty} = 4nFDCa \tag{1}$$

where F is the Faraday constant, C and D are the mediator concentration and diffusion coefficient, n is the number of electrons transferred in the tip reaction, and a is the disk radius.

The imaging signal in the SECM feedback mode is the perturbation in i_T as the tip approaches the substrate surface. The notation $i_{T,\infty}$ refers to the tip signal when the tip is an infinite distance from the substrate surface. Two limiting cases in the SECM feedback mode arise when either the substrate supports ET to the mediator or when ET is blocked. As the tip approaches the blocking surface, the diffusion field normally present at the UME tip is restricted, decreasing the overall rate of diffusion of the mediator to the electrode surface and decreasing the tip current, a behavior called "negative feedback" (cf. Fig. 1B). However, if the mediator can undergo ET at the substrate, regeneration of the mediator occurs. For example, an oxidized mediator, "Ox," is reduced to "Red" at the tip (or vice versa). Red diffuses to the substrate surface and undergoes ET to restore the mediator to its original oxidation state (cf. Fig. 1A). As the tip-substrate separation decreases, this regeneration cycle becomes more rapid and the tip current increases consequently. If the ET reaction at both the tip and substrate are rapid, the tip current response is called "positive feedback."

Feedback imaging is an "active" process in the sense that the tip is used to generate a local region of a mediator species and a perturbation of that region is sensed. During imaging, a small region of the substrate is subjected to a chemical environment that is different from the bulk. Although a noncontact process, feedback imaging is strongly chemically interactive. The experimentalist has a choice of mediator species, which, in principle, permits a tuning of the imaging conditions based on chemical or electrochemical interactions between the tip and substrate. Experimental freedom is tempered by the fact that the mediator must often be added as an extra component to solution and may produce undesirable interactions; for example, an oxidizing mediator may cause etching of the substrate surface.

Numerical calculations of the steady-state tip current as a function of the tip-substrate distance (*i-d* curves) have been made (Kwak and Bard, 1989), but approximate analytical equations derived by fitting curves to the numerical results are simpler to use (Mirkin et al., 1992). Equation 2 describes the positive-feedback calculation:

$$I_T(L) = \frac{i_T}{i_{T,\infty}} = \frac{0.78377}{L} + 0.3315\ \exp\left(\frac{-1.0672}{L}\right) + 0.68 \tag{2}$$

where $L = d/a$ is the normalized distance d between the tip and substrate. This equation fits the numerical results to within 0.7% over L from 0.05 to 20. Equation 3 describes the negative-feedback *i-d* curve and is accurate to within 1.2% over the same L range:

$$I_{T,ins}(L) = \frac{1}{0.292 + 1.515/L + 0.6553\ \exp(-2.4035/L)} \quad (3)$$

The *I-L* curves for Equations 2 and 3 are shown in Figure 3. Note the increased distance sensitivity in the positive feedback compared to negative feedback.

Feedback mode topographic images are made by mapping the tip current as a function of tip position while the tip is rastered across the substrate surface. Both conducting and insulating areas can be imaged during a feedback mode scan of the substrate. Conducting areas have a positive-feedback response and can be identified by currents larger than $i_{T,\infty}$. In contrast, insulating areas have a negative-feedback response with tip currents less than $i_{T,\infty}$. Note that images made in the feedback mode have a distorted relationship to actual topography due to the nonlinear *i-d* relationship. In addition, negative-feedback images are inverted. However, if $i_{T,\infty}$ and the tip electrode radius are known, Equations 2 and 3 can be employed to convert the current image to a true topographic image (Bard et al., 1991; Macpherson and Unwin, 1995).

The lateral and vertical resolution in the feedback mode is dependent on the tip size and tip-substrate separation. One estimate of lateral resolution is the ability to discriminate between insulating and conducting regions. If both regions are located on the same plane, the ideal response would be an instantaneous change in tip current reflecting a change from positive to negative feedback, or vice versa, as the tip is scanned over the boundary. In reality, the change will occur over a region of finite width, Δx, that will be a measure of the lateral resolution. At the limit of $L = 0$, one might expect that Δx would be approximately equal to the tip electrode diameter. For $L > 0$, the lateral resolution will degrade as diffusion broadens the mediator concentration profile. An experimental study suggests that the resolution function is of the form (Wittstock et al., 1994a)

$$\Delta x/a = k_1 + k_2 L^2 \quad (4)$$

where $k_1 = 4.6$ and $k_2 = 0.111$. The best resolution predicted by this equation, 4.6 times the disk radius, runs counter to intuition and other reports (see Wipf and Bard, 1992) for superior resolution than predicted by Equation 4 and may be overly conservative, but the decrease in resolution with distance is reasonable. It should be noted that features smaller than the tip size can be recognized, but not resolved, under certain conditions. Small active regions can be distinguished from an insulating background if the regions are separated by at least a tip diameter. In this case, the image of the region will reflect the convolution of the tip and region size.

Vertical resolution in SECM depends on the tip-substrate separation and the dynamic range of the tip current measurement. One can estimate the minimum resolvable vertical change in distance by use of Equations 2 and 3 for insulating and conducting samples. For example, a change from $L = 1$ to $L = 1.0137$ gives a 1% change in I_T. Assuming this current difference is perceptible, this would correspond to the ability to distinguish a change of 0.068 μm at a 5-μm-radius tip with a 5-μm tip-substrate separation.

In principle, the resolution of SECM feedback images can be increased by application of digital image processing techniques. A degraded image can be described as the convolution of a perfect, nondegraded image with the impulse response (IR) or point spread function (PSF) and a noise function. If the PSF for feedback SECM can be found, an improved image can be made by deconvolving the degraded image from the PSF. Unfortunately, the nonlinear SECM feedback response has so far prevented a general formulation of a PSF. However, an approximate "deblurring" function based on the Laplacian of the Gaussian (LOG) filter has been shown to improve image resolution (Lee et al., 1991). Other work, so far theoretical, based on linear shift invariant processes also may be useful under restricted circumstances (Ellis et al., 1995).

There is a strong dependence on tip geometry in the feedback mode experiment. For disk-shaped electrodes, an important aspect is the radius of the insulator material in which the electrode is embedded. Tip electrodes are frequently made so that the tip assumes a truncated cone shape with the disk of the electrode centered at the truncated end. Since there is inevitably some angular offset from perpendicular as the tip approaches the surface, a smaller radius of insulating material allows closer tip-substrate approach than a larger insulating radius. Numerical simulations of the negative-feedback mode

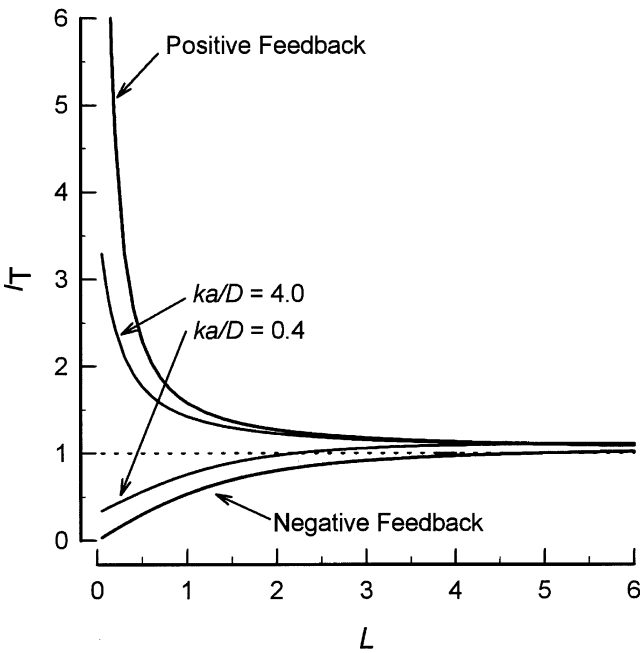


Figure 3. Theoretical current-distance curves for positive- and negative-feedback modes. Also shown are curves for a kinetically limited substrate response characterized by the parameter ka/D, where k refers to the ET coefficient and a and D are the tip radius and the diffusion coefficient, respectively. The dotted line is the baseline tip response for infinite tip-substrate separation.

demonstrate that the *i-d* relationship is very sensitive to the RG ratio, which is the ratio of the insulator radius to the disk electrode radius (Kwak and Bard, 1989). Equation 3 is only correct for RG values of 10. A small RG ratio decreases the sensitivity of the *i-d* curves in negative-feedback experiments. Conversely, the RG ratio has little effect in the positive-feedback mode. Qualitatively, the RG effect can be explained in the negative-feedback case by considering the insulator as an additional barrier for the mediator to reach the surface. Smaller RG ratios minimize the blocking effect and thus moderate the decrease in tip current. For example, at $L = 1$, $I_{T,\,ins}$ has a value of 0.534, 0.302, and 0.213 for RG ratios of 10, 100, and 1000, respectively (Kwak and Bard, 1989). Positive feedback is insensitive to the RG ratio since the tip current is mainly due to regeneration of the mediator in the region of the substrate directly below the tip.

The actual shape of the tip is another important variable in the feedback mode. Tips with diameters smaller than 1 μm are easier to make with a cone or spherical segment (e.g., hemispherical) geometry. Unfortunately, the *i-d* sensitivity of these tips in the feedback mode is significantly less than that of the disk-shaped electrodes (Bard et al., 1989; Davis et al., 1987; Mirkin et al., 1992). For cone-shaped tips, a characteristic parameter is k, the ratio of the cone height over the base radius. Calculated positive-feedback *i-d* graphs for cones with k values of 0 (disk) to 3 are shown in Figure 4A. Note the flat response at close tip-substrate distances for $k > 1$ and the significant deviation of the response even for small k values. Spherical segment tips are characterized by the parameter α_0, which is the angle between two lines from the center of the sphere, one line being a perpendicular line through the segment base and the other passing through the edge of the segment base. The positive-feedback *i-d* curves for tips with α_0 from $\pi/2$ (hemispherical) to a disk shaped are compared in Figure 4B. Although superior to conical tips, the sensitivity of hemispherical tips remains inferior to disk-shaped tips. A further difficulty is that the methods of construction (Nagahara, 1989; Penner et al., 1989; Lee et al., 1990b; Mirkin et al., 1992) of very small tips (e.g., electrochemically etched tips) are not sufficiently reproducible to produce a well-defined tip shape (Shao et al., 1997a). Fitting the experimental to simulated *i-d* curves can produce good estimates of the tip shape (Mirkin et al., 1992), but it is unlikely that fitting will be used for general imaging purposes. Thus, at present, non-disk-shaped tips should only be used to provide qualitative SECM feedback images.

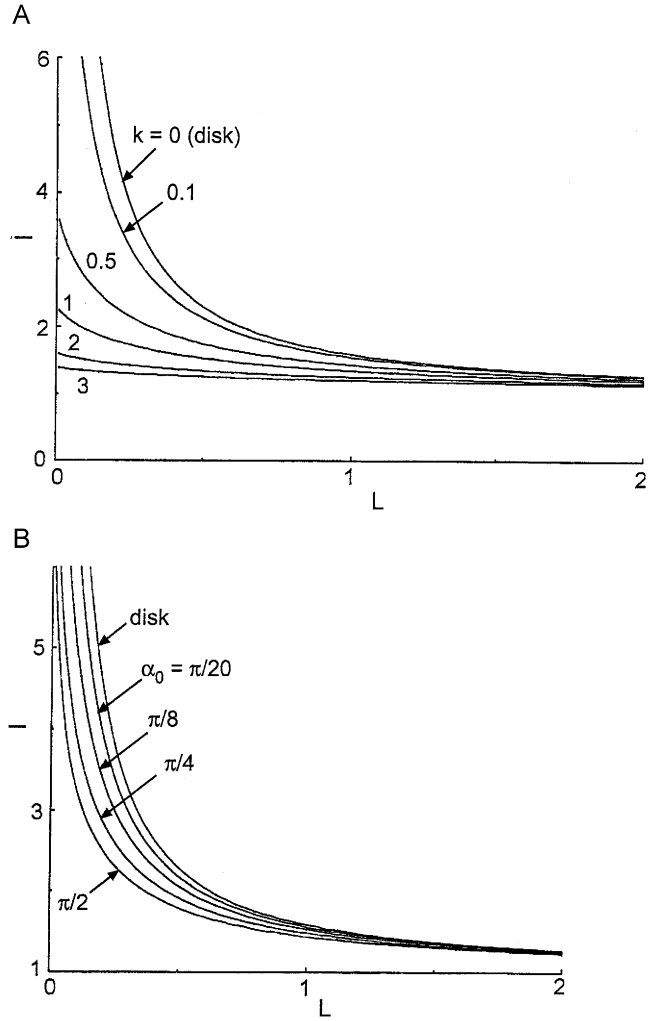


Figure 4. Theoretical current-distance curves for positive feedback at (**A**) cone-shaped and (**B**) spherical-segment type tips, where k is the ratio of cone height to radius and α_0 is the angle between a perpendicular through the base of the spherical segment and the radius of a sphere passing through the edge of the segment base. (Reprinted with permission from Mirkin et al., 1992, pp. 52, 55. Copyright © 1992 by Elsevier Science, UK.)

Reaction Rate Mode

Within the limits of positive and negative feedback, a more general case exists when the reaction rate at the tip or substrate is limited by a slow ET rate or sample conductivity rather than by diffusion rate. As the rate of reaction changes from the diffusion limit to zero, the *i-d* curves will shift from curves similar to positive feedback to what appears to be negative feedback. Note that if the rate of ET is sufficiently slow, the result is practically indistinguishable from that of negative feedback.

Limitations due to ET rate may arise even when mediators with very rapid ET kinetics are used, since very high rates of mass transport are produced as the tip-substrate distance becomes very small. In this case, the response is approximately that of a twin-electrode thin-layer cell (Anderson and Reilley, 1965; Anderson et al., 1966). Equation 5 predicts the tip current

$$i_{T,\infty} = 4nFDCA/d \qquad (5)$$

as L approaches zero, where A is the tip area. Considering that very small separations can be achieved in SECM, impressive values of the mass transfer coefficient D/d (Bard and Faulkner, 1980, p. 28) can be achieved. As an example, for $d = 0.2$ μm and $D = 1 \times 10^{-5}$ cm^2/s, the mass transfer coefficient is 0.5 cm/s. To achieve a similar rate of mass transfer, a rotating disk electrode would have to rotate at greater than 6.5×10^6 rpm! Thus, SECM can

measure rapid heterogeneous rates either for an ET reaction at the tip or the substrate as long as the opposing electrode is not allowed to limit the overall reaction. This is not generally difficult since the opposing substrate or tip can be poised at potential sufficiently past the oxidation or reduction potential to drive the reaction to the diffusion-controlled limits. Another consequence of the rapid diffusional transport in the positive-feedback mode is that the tip response becomes insensitive to convective transport, which minimizes the effect of stirring that may be produced by the tip during SECM imaging. Because of the high mass transport rate and the ease of making measurements at steady state, SECM is very promising for fundamental investigation of rapid interfacial ET rate constants. Calculations for tip-limited kinetics suggest that it is possible to measure rate constants as high as 10 to 20 cm/s under steady-state conditions (Mirkin et al., 1993).

Allowing a kinetic limitation of the feedback process at the substrate allows numerous interesting experiments, since a number of heterogeneous ET processes are accessible with SECM. At an electronically conducting surface, such as an electrode, kinetic information can be extracted at the tip-sized regions, allowing direct imaging of active or passive sites on surfaces. For example, images of a composite carbon-Au electrode clearly show the greater ET activity of the Au regions (Wipf and Bard, 1991b). Theoretical calculations of the *i-d* curves for quasireversible and irreversible ET kinetics as functions of the substrate potential have been published (Bard et al., 1992; Mirkin and Bard, 1992), permitting quantitative measurements of ET rates at micrometer-sized regions of surfaces. In addition, the feedback current can supply potential independent kinetic information for chemical reactions involving ET at the interface. Figure 3 shows *i-d* curves for the parameter ka/D, where k is the heterogeneous ET coefficient (Bard et al., 1995). As an example, these types of calculations have been used to calculate rate constants for the reaction of immobilized glucose oxidase enzyme with a tip-generated mediator (Pierce et al., 1992). Also, feedback current due to reduction of the tip-generated oxidant mediator by a Cu metal surface provides direct measurement of the interfacial Cu dissolution process (Macpherson et al., 1996).

Generation/Collection Mode

In contrast to the feedback mode, where the tip-generated mediator plays an active role in the imaging process, the GC mode is an entirely passive method. The GC mode of interest here is more properly referred to as substrate generation/tip collection (SG/TC). The GC mode is appropriate in situations where a reaction at the substrate surface produces an electrochemically detectable material. Typical examples include corroding metal surfaces (Gilbert et al., 1993; Wipf, 1994), ion movement through porous material (Scott et al., 1993b; 1995), and oxygen generation at plant leaves (Lee et al., 1990a; Tsionsky et al., 1997). Advantages to using this mode include that (1) no external mediator is required, thus eliminating the possibility of undesirable interactions of the mediator and substrate; (2) identification of the substrate-generated specie(s) is possible by using voltammetry when the tip is near the substrate surface; and (3) potentiometric tips can be used in addition to voltammetric or amperometric tips.

Two major disadvantages to this imaging mode are poor vertical sensitivity and lateral resolution. Both arise because an ion or molecule produced at the surface must diffuse to the tip to be detected. Vertical sensitivity is greatly restricted by the lack of a sharp diffusion gradient of the substrate-generated species. The concentration gradient may extend out to several hundred micrometers. In addition, the gradient is ill-defined since natural convection, vibration, and even tip movement can perturb the diffusion-based concentration profile. The lack of a sharp concentration gradient also causes a problem with positioning the tip at the substrate, since the tip signal changes little during approach. Optical microscopic examination of the tip approach is often necessary to position the tip close to the substrate surface. Lateral resolution is also compromised by diffusional broadening. A species released from a point source into an infinite sink will diffuse, on average, a distance x with time t, given by (Bard and Faulkner, 1980, p. 129)

$$x = \sqrt{2Dt} \tag{6}$$

This equation suggests that features <10 μm will be significantly broadened by diffusion at times >0.1 s. Diffusion will cause images of closely spaced features to overlap and a diffuse background signal will be present in GC images.

Figure 5 is a comparison of the difference in resolution for the feedback mode (Fig. 5A) and the amperometric GC

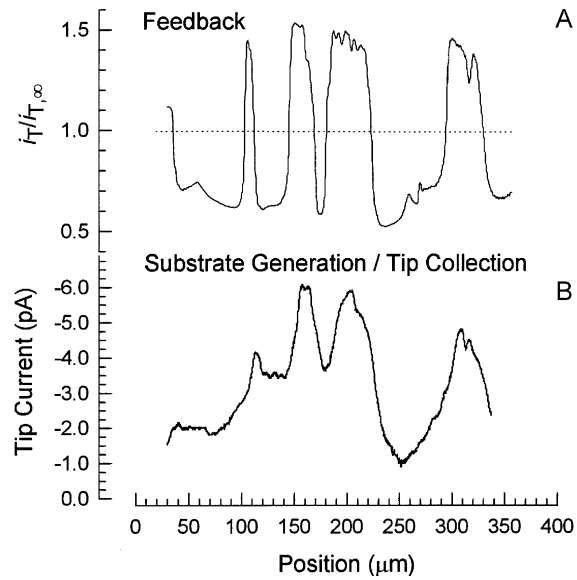


Figure 5. Single SECM line scans over an identical section of composite electrode consisting of insulating Kel-F polymer and conducting gold metal regions. Scans were acquired at a 2-μm-diameter Pt disk electrode tip in a solution of 2 mM $Fe(CN)_6^{3-}$ ion in 0.1 M KCl. Tip and substrate potentials are versus a Ag-AgCl reference electrode. (**A**) Feedback scan with the tip potential set to +300 mV and the substrate potential set to +600 mV. (**B**) Generation/collection scan with the tip potential set to +600 mV and the substrate potential set to +420 mV.

mode (Fig. 5B). The two graphs are each a single line scan using a 2-μm-diameter Pt disk electrode tip over an identical section of a composite electrode consisting of insulating and conducting regions. The mediator is the ferricyanide ion, $Fe(CN)_6^{3-}$, in an aqueous 0.1 M KCl solution. In Figure 5A, the tip is set to a sufficiently negative potential to reduce $Fe(CN)_6^{3-}$ to $Fe(CN)_6^{4-}$ (+300 mV vs. Ag/AgCl reference) while the conducting substrate is maintained at a potential sufficient to oxidize the reduced mediator. Under these conditions, positive feedback occurs while the tip is passing over conducting regions, as is apparent by current $>i_{T,\infty}$. Negative feedback occurs at the insulating regions, as is apparent by current lower than $i_{T,\infty}$. Figure 5B, showing GC imaging, was acquired with the substrate set at a potential just negative enough (+420 mV) to generate a small concentration (about 20 μM) of $Fe(CN)_6^{4-}$ while the tip was set at a positive potential (+600 mV) to detect, by oxidation, $Fe(CN)_6^{4-}$. In this case, the GC ability to detect small concentrations at a surface is illustrated, but at the cost of a lower resolution in comparison to the feedback mode. Good results in GC imaging rely on placing the tip as close as possible to the surface to avoid diffusional broadening of the detected species. Stirring may improve resolution by minimizing diffusional broadening.

In GC imaging, the tip current provides an image of the distribution of the electroactive species near a surface. At a voltammetric tip, the tip current map can be converted to a concentration map by use of the equation for the limiting current at a disk electrode (Equation 1). To do this quantitatively requires knowledge of the species diffusion coefficient and number of electrons involved in the overall reaction. In addition, the tip potential must be located on the limiting current plateau. The singular advantage of the disk electrode in feedback imaging is not present in GC imaging, and so electrodes of conical or hemispherical geometry can be used. In this case, a version of the limiting current equation appropriate for the tip geometry must be employed (Wightman and Wipf, 1989; Mirkin et al., 1992).

The potentiometric GC mode relies on an ISE as the tip. The main advantage of the potentiometric GC mode is that the tip is a passive sensor of the target ion activity and does not cause the depletion or feedback effects noted above. In addition, the tip can be used to detect ions that would be difficult or inconvenient to detect voltammetrically, such as Ca^{2+} or NH_+^4. The selectivity of the tip for the ion of interest is an advantage, but this requires separate tips for each species. The response of an ISE is, in general, a voltage proportional to the logarithm of the ion's activity, but special consideration must be made for potentiometric GC to provide a rapid response with microscopically small tips (Morf and Derooij, 1995).

PRACTICAL ASPECTS OF THE METHOD

SECM Instrumentation

Figure 6 shows a block diagram of an SECM instrument. Minimally, the instrument consists of two components: one to position the tip in three dimensions and one to measure the tip signal. A commercial instrument designed specifically for SECM has recently become available; however, most SECM investigators build their own. The one in use in the author's laboratory will serve as a model for a discussion of the instrument (cf. Fig. 6). The tip is attached to a three-axis translator stage with motion provided by linear piezoelectric motors (Inchworm motors, Burleigh Instruments). The inchworm motors provide scan rates greater than 100 μm/s over a 2.5-cm range of motion with a minimum resolution of less than 100 nm as well as backlash-free operation. An interface box (Model 6000, Burleigh) controls the speed, direction, and axis selection of the motors by accepting TTL (transistor-transistor logic) level signals from a personal computer. The motors operate in an open-loop configuration and the speed and total movement of the axis is controlled by a TTL clock signal. Use of the open-loop configuration requires that each axis be calibrated to relate the actual movement per clock pulse (i.e., nanometers per pulse). A closed-loop version of the inchworm motor is also available.

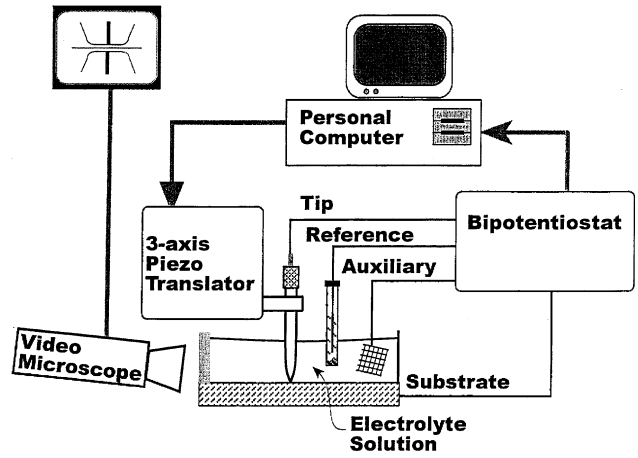


Figure 6. A scanning electrochemical microscope.

A bipotentiostat (EI-400, Cypress Systems) is used to control the tip potential and amplify the tip current. Use of a bipotentiostat, which allows simultaneous control of two working electrodes versus common reference and auxiliary electrodes, is convenient when control of the substrate potential is also required. The bipotentiostat should be sufficiently sensitive to allow measurements of the low current flow at microelectrodes, which may be in the picoampere range. Commercial bipotentiostats designed for use with rotating ring-disk electrodes are not suitable without some user modification to decrease noise and increase sensitivity. Use of a Faraday cage around the SECM cell and tip to reduce environmental noise is advisable. For potentiometric SECM operation, an electrometer is required to buffer the high impedance of the microscopic tips.

Custom software is used to program the tip movement and to display the tip and substrate signals in real time. For positioning the tip during GC imaging, a video microscope system is especially useful, since the tip can be observed approaching the substrate on the video monitor while the SECM operator is moving the tip. Descriptions of other SECM instruments are available in the literature (Bard et al., 1994; Wittstock et al., 1994b).

The SECM tip is perhaps the most important part of the SECM instrument and, at this time, disk-shaped electrodes must be constructed by the investigator. A construction method for disk electrodes with a radius of $\geq$0.6 μm is described below (see Protocol: SECM Tip Preparation).

Constant-Current Imaging

Most SECM images have been acquired in the constant-height mode where the tip is rastered in a reference plane above the sample surface. The feedback mode requires tip-substrate spacing of about one tip diameter during imaging. With submicrometer tips, it is difficult to avoid tip crashes due to vibration, sample topography, and sample tilt. Constant-current imaging can be used to avoid tip crashing and improve resolution by using a servomechanical amplifier to keep the tip-substrate distance constant while imaging. With surfaces known to be completely insulating or conducting, implementation of a constant-current mode is straightforward and similar in methodology to STM studies (Binnig and Rohrer, 1982). For example, any deviation of the tip current from the reference current is corrected by using a piezoelectric element to position the tip. A current larger than the reference level causes the tip to move toward an insulating surface and away from the conducting surface. However, this methodology cannot work when a surface has both insulating and conducting regions. Unless some additional information is provided about the substrate type or the tip-substrate separation, the tip will crash as it moves from an insulating to a conducting region or vice-versa.

One method for providing constant-current imaging is to impose a small-amplitude (100 nm or less), high-frequency vertical modulation of the tip position, i.e., tip position modulation (TPM) during feedback imaging (Wipf et al., 1993). The phase of the resulting ac current shifts by 180° when the tip is moved from an insulating to a conducting surface to provide unambiguous detection of the sample conductivity. This phase shift is used to set the proper tip motion and reference current level on the servoamplifier. The normal tip feedback signal is restored by filtering the small ac signal. A second method uses the hydrodynamic force present at a tip vibrated horizontally near the substrate surface (Ludwig et al., 1995). As the tip approaches the substrate, the amplitude of a vibration resonance mode is diminished and sensed by monitoring the diffraction pattern generated by a laser beam shining on the vibrating tip. Adjustment of the tip position to maintain a constant diffraction signal at a split photodiode allows "constant-current" imaging. Since a specific electrochemical response is not required, the hydrodynamic force method can be used for either GC or feedback imaging at a substrate. However, a special tip and cell design is required, and the method does not provide a measure of the tip-substrate separation. Another hydrodynamic method uses resonance frequency changes of a small tuning fork attached to the SECM tip (James et al., 1998). This relies on the presence of a shear-force damping as the tip approaches the surface. A third constant-current method, the "picking mode," monitors the current caused by convection during a rapid tip approach from a large tip-substrate separation (Borgwarth et al., 1994, 1995a). The transient tip current is affected by tip-substrate separation but not by substrate conductivity and, thus, can be used to position the tip automatically.

Applications in Corrosion Science

A number of researchers have recognized the possibility of using SECM for investigations in localized corrosion at metallic samples. The SECM can be used to make topographic images of samples in the feedback mode, but in this case, the more interesting images will be GC images of local ion concentrations near the corroding surface. Voltammetric or amperometric GC methods can be used to detect electroactive species produced or depleted at corroding surfaces. Examples in the literature include Fe ions (Wipf, 1994); O_2, Co, and Cr ions (Gilbert et al., 1993); and Cu and Ni ions (Küpper and Schultze, 1997b). Absolute determination of the concentration of the species is made using Equation 1 to calculate the limiting current of the SECM tip. However, the presence of multiple electroactive species may interfere with the detection. Use of a potential-time program (e.g., pulse or sweep voltammetry methods) can improve specificity and provide for multiple species detection at a stationary tip (Küpper and Schultze, 1997a). Alternatively, the use of potentiometric GC SECM images of corroding surfaces can yield information about local pH values. Changes in pH of <0.1 pH unit above a cathodic $AlFe_3$ electrode were imaged using a hydrogen-ion-sensitive neutral-charge-carrier tip (Park et al., 1996).

Scanning electrochemical microscopy can be used to preidentify sites on passive metal surfaces that will experience subsequent passive-layer corrosion and breakdown. In a study of a pit formation at a stainless steel surface, images of steel surface showed picoampere-level fluctuations in the SECM tip current over a specific region of the sample. A subsequent GC image of the same region showed current due to growth of a corrosion pit (Fig. 7; Zhu and Williams, 1997). The authors suggest that picoampere-level fluctuations in current presage a passive-layer breakdown. Corrosion precursor sites on titanium

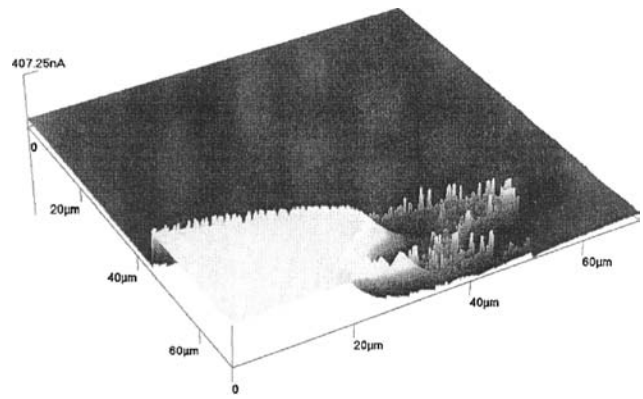


Figure 7. A GC SECM image of a corrosion pit growing on 304L stainless steel in 0.3 M $NaClO_4$/0.3 M NaCl at +230 mV vs. a saturated calomed electrode (SCE). (Reprinted with permission from Zhu and Williams, 1997, by The Electrochemical Society.)

were also identified with SECM imaging. Pitting sites were preidentified by their ability to oxidize Br^- ions to Br_2 in solution. These microscopic active sites were imaged by detecting the Br_2 at the SECM tip (Casillas et al., 1993, 1994; Garfias-Mesias et al., 1998). Pit formation was subsequently found to coincide with the Br_2-forming regions.

In the microreagent mode, the SECM can artificially stimulate corrosion at passive metal surfaces by using the SECM tip to generate solution conditions at the metal surface in which corrosion is favored. Pitting corrosion and passive-layer breakdown on Al, stainless steel (Wipf, 1994), and iron surfaces (Still and Wipf, 1997) have been stimulated by electrolytic production of Cl^- ion at the electrode tip. Since only the region of sample surface near the tip is exposed to the aggressive Cl^- ion, corrosion susceptibility of microscopic sites can be examined, and these studies have demonstrated a heterogeneous distribution of corrosion sites on iron surfaces.

SECM of Biological and Polymeric Materials

Investigation of transport of molecules or ions through skin layers is an import aspect in developing transdermal drug delivery systems. A model system is the iontophoretic transport of $Fe(CN)_6^{4-}$ through hairless mouse skin. Application of a current across a skin sample drives ions through pores, such as hair follicles and sweat glands. The GC mode of SECM can easily detect the exiting ions by making an image of the concentration of $Fe(CN)_6^{4-}$. Figure 8 is a GC image of a skin section showing two active pore regions. Using SECM allows quantitative determination of effective transport parameters such as pore diameter, pore density, and pore transport rate (Scott et al., 1993a,b; Scott et al., 1995). In a similar fashion, SECM was applied to examine fluid flow through the dentine tubules of teeth. Exposure of the dentinal material can lead to pain as pressure develops across the tubules, e.g., by osmotic pressure due to sugar. By monitoring the flux of $Fe(CN)_6^{4-}$ through a dentine sample, SECM was used to show that a calcium oxalate preparation reduces fluid flow in the tubules. In addition, the thickness and uniformity of the calcium oxalate layer on the dentine sample could be directly estimated from SECM imaging (Macpherson et al., 1995a,b).

The conductivity of electroactive polymeric materials is readily examined by feedback-mode SECM. Positive-feedback current at a fully oxidized polypyrrole film verified polymer conductivity, while a less oxidized polypyrrole film showed a decrease in feedback current as the film became less conductive (Kwak et al., 1990). A more recent result used reaction rate imaging to image regions of a polyaniline sample with significant variations in conductivity, presumably due to a variation in the levels of iodine dopant (Borgwarth et al., 1996). By examining the process of ion ejection during oxidation and reduction of conducting polymer films, processes involved with the transport of ions in these films can be studied with SECM. For example, the accepted mechanism of polyaniline oxidation is supported by use of GC SECM to monitor the ingress and egress of Cl^- and H^+ ions occurring during the polyaniline oxidation-reduction cycle (Denuault et al., 1992; Frank and Denuault, 1993, 1994). In a more quantitative study, the potential dependent kinetics of counterion transport in a polypyrrole film was investigated as a function of the nature of the solution phase ion. Ejection of electroactive ions was monitored at the SECM tip as the polypyrrole was oxidized or reduced. Figure 9 shows an example of the ejection of $Ru(NH_3)_6^{3+}$ from a polypyrrole film containing the large poly(*p*-styrenesulfonate) polyanion as a counterion. The SECM tip signal is the reduction current of $Ru(NH_3)_6^{3+}$ as it is ejected during oxidation of the film. The amount of cation ejected in the poly(*p*-styrenesulfonate)-containing film is much larger than in films containing the smaller Br^- counterion (Arca et al., 1995).

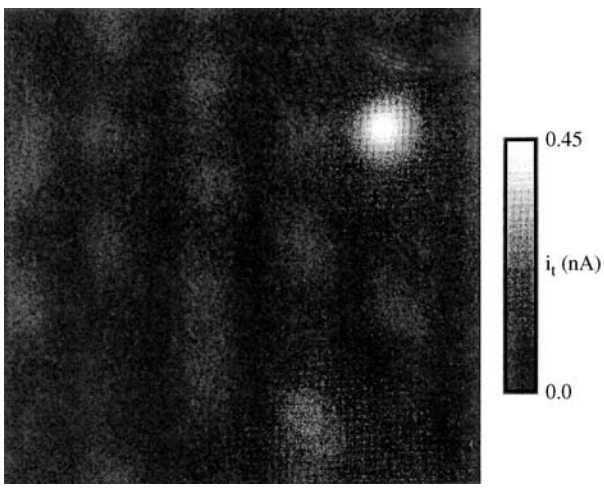


Figure 8. The SECM image of a 1 mm × 1 mm region of hairless mouse skin showing the local concentrations of the $Fe(CN)_6^{4-}$ ion at pore locations. For this image, application of an iontophoresis current of 40 $\mu A/cm^2$ drives ions from a solution of 0.1 M $Fe(CN)_6^{4-}$ through the skin into a 0.1 M NaCl solution. A 25-μm-diameter Pt electrode held at a potential of 0.4 V vs. SCE was used as the SECM tip. The scale bar is 200 μm. (Reprinted with permission from Scott et al., 1993b. Copyright © 1994 by the American Chemical Society.)

SPECIMEN MODIFICATION

Microfabrication

The SECM microscope can be used to modify surfaces in a number of different modes. The microreagent mode uses the SECM tip to produce, via electrolysis, a local chemical environment that is different from the bulk environment. With the tip positioned close to the surface, the area of the substrate exposed to the local chemical environment is only slightly larger than the tip diameter. In addition, the volume of solution between the tip and the substrate can be quite small. For example, the volume of the gap formed by a 5-μm-radius disk-shaped tip separated by 1 μm from the substrate is only 25 fL. Consequently, the local chemical environment within this small volume can be changed on a millisecond time scale. Rapid diffusion of the tip-generated reagents outside the vicinity of the tip ensures that the reaction remains limited to the tip area.

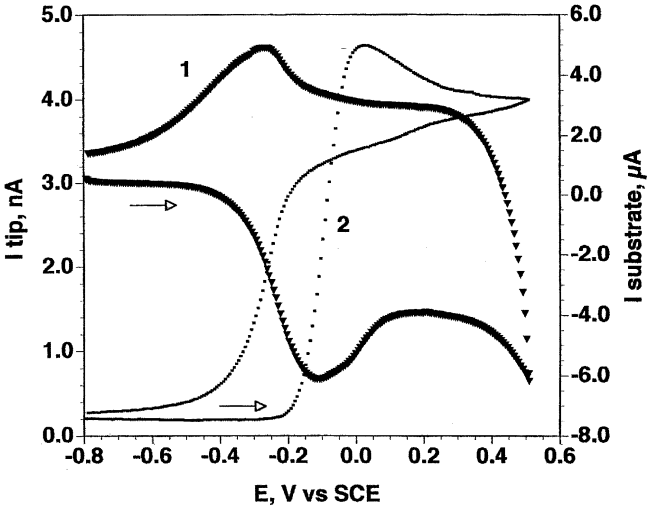


Figure 9. The SECM substrate (1) and tip (2) cyclic voltammograms of a polypyrrole film [initially containing $Ru(NH_3)_6^{2+}$ in 0.1 M K_2SO_4 solution. The polypyrrole film was prepared at +0.8 V vs SCE in a solution containing 0.2 M sodium poly(*p*-styrenesulfonate) and was then reduced at −0.8 V vs. SCE in a solution containing 0.1 M $Ru(NH_3)_6^{3+}$. The tip current is due to reduction of $Ru(NH_3)_6^{3+}$ ejected from the polypyrrole. Tip potential = −0.45 V, CV scan rate = 10 mV/s. (Reprinted with permission from Arca et al., 1995. Copyright © 1994 by the American Chemical Society.)

The number of possible species generated by electrolysis at the tip is limited principally by the investigator's imagination. Especially common are changes in pH made possible by electrolysis of water. Also, reducing and oxidizing agents are easily produced from the corresponding oxidized or reduced precursor. It should be noted that very unstable species can be produced to react with the tip surface because the diffusion time for small tip-substrate distances is very short. A species with a lifetime of >1 ms can diffuse across a 1-μm gap while a gap of 0.1 μm would allow a species with a 10-μs lifetime to react with the substrate. Further control of the reaction conditions is possible by addition of a scavenger species to solution. The "scavenger" reacts with the tip-generated species as they diffuse from the tip region. For example, a pH-buffered solution will minimize the extent of pH changes caused by OH^- generation at the tip. Likewise, an oxidizing agent can be neutralized by addition of a second reducing agent to scavenge any of the oxidant leaving the tip region.

Another substrate modification method is the "direct" mode. In this mode, the tip is used as a microscopic auxiliary electrode for the substrate reaction (Hüsser et al., 1989). With the tip positioned close to the substrate, any Faradaic reaction at the substrate is limited to the region near the tip. Any electrochemical reaction that can be produced at the bulk surface can be performed, at least in principle, at a local region of the surface using the direct mode. For example, metal etching, metal deposition, and polymer deposition are all possible at microscopic regions.

Applications in Localized Material Deposition and Dissolution

The SECM tip can be used to modify a surface on a microscopic scale by manipulating the local chemical environment or by production of a localized electric field. An example is the deposition of conducting polymer patterns from solutions of a monomer. Using a cathodic tip in the direct mode, polyaniline lines about 2 μm wide were drawn on Pt electrodes coated with a thin ionically conducting polymer (i.e., Nafion) containing the anilinium cation. The ionic polymer served to concentrate the electric field in order to oxidize anilinium cation at the metal surface (Wuu et al., 1989). Changing the local solution composition can be used to drive a reaction that leads to formation of polymer structures. Electrolytic reduction of water at the tip in a solution of aniline in H_2SO_4 increases local pH and leads to local oxidation of aniline and subsequent formation of polyaniline at a substrate electrode. Patterns with widths of <15 μm and >1 μm thickness could be formed at writing speeds of 1 μm/s (Zhou and Wipf, 1997). Polypyrrole polymer structures on electrodes can be produced by using voltage pulses in an aqueous pyrrole at the tip to produce pyrrole radical cations that subsequently polymerize on a substrate surface (Kranz et al., 1995a). Using this method, a conducting polypyrrole line was grown between two electrodes across a 100-μm insulating gap (Kranz et al., 1995b), and a 200-μm-high polypyrrole tower was grown (Fig. 10; Kranz et al., 1996). In another example, generation of the oxidizing agent Br_2 at the SECM tip formed patterns by polymerizing a thin film of a 2,5-bis(1-methylpyrrol-2yl)-thiophene monomer deposited on a substrate, and dissolution of the unreacted monomer left the polymer



Figure 10. Scanning electron micrograph of a polypyrrole tower grown by SECM from an aqueous solution of pyrrole using a 10-μm-diameter Pt microelectrode (tower height = 200 μm, width = 70 μm). (Reprinted from Kranz et al., 1996, by permission of Wiley-VCH Verlag GmbH.)

pattern (Borgwarth et al., 1995b). Due to the smaller tip size, STM deposition typically produces smaller feature sizes. The main advantage of SECM, yet to be fully realized, is the greater control of deposition conditions allowed by precise electrochemical control of the local solution composition.

Metal and metal oxide structures are readily deposited from solution using SECM. Use of the direct mode to reduce metal ions dissolved in a thin, ionically conducting polymer film deposits Au, Ag, and Cu lines as thin as 0.3 μm (Craston et al., 1988; Hüsser et al., 1988, 1989). The SECM tip induced pH changes have been used to deposit nickel hydroxide (Shohat and Mandler, 1994) and silver patterns (Borgwarth et al., 1995b). Thin Au patterns can be drawn on electronic conductors by oxidizing a gold SECM tip in Br^- solution to produce $AuBr_4^-$ ions; reduction of the ions at the sample surface produces the metal film (Meltzer and Mandler, 1995a). Three-dimensional microfabrication of a 10-μm-thick Ni column and spring was demonstrated by use of an SECM-like device in the direct mode (Madden and Hunter, 1996).

Localized etching of metals and semiconductors to form patterns of lines or spots is possible by electrolytic generation of a suitable oxidizing agent at the SECM tip. For example, Cu can be etched using a tip-generated oxidant, such as $Os(bpy)_3^{3+}$ (bpy = 2,2′-bipyridyl) generated from $Os(bpy)_3^{2+}$ (Mandler and Bard, 1989), and GaAs and Si semiconductors can be etched by Br_2 generated at the tip by oxidation of Br^- (Mandler and Bard, 1990b; Meltzer and Mandler, 1995b). Additionally, the SECM tip current observed during the etching process is a feedback current and can be used to monitor the rate of etching. In this case, the mediator is regenerated due to the redox reaction at the substrate rather than by direct ET and so the feedback current monitors the rate of oxidative dissolution directly. The behavior of the feedback current observed during etching was used to postulate a hole injection mechanism for etching of *n*-doped GaAs by tip-generated Br_2 (Mandler and Bard, 1990a). The oxidative etching kinetics of Cu by tip-generated $Ru(bpy)_3^{2+}$ was explored in the same manner (Macpherson et al., 1996).

Dissolution of nonconducting material is an area in which the ability of the SECM to produce a controlled chemical environment near the material surface provides a significant advancement. Dissolution rates of even very soluble materials can be obtained by placing the tip near (1 to 10 μm) a dissolving surface. If the solution near the surface is saturated with the dissolving material and that material is electroactive, tip electrolysis produces a local undersaturation of the dissolving material. Quantitative details about the rate and mechanism of the dissolution reaction are available from the tip current as a function of time and distance to the sample (Unwin and Macpherson, 1995). The dissolution of the (010) surface of potassium ferrocyanide trihydrate in aqueous solution is an example in which the dissolution process was found to be second order in interfacial undersaturation (Macpherson and Unwin, 1995). In a separate study, the dissolution rate constant for AgCl in aqueous potassium nitrate solution was shown to be in excess of 3 cm/s (Macpherson et al., 1995b).

The spatial resolution of the SECM was employed to examine and characterize two very different Cu^{2+} dissolution processes on a copper sulfate pentahydrate crystal in H_2SO_4. At a crystalline interface where the dislocation density is large, and thus the average distance between dislocation sites is much smaller than the tip size, dissolution is rapid and follows a first-order rate law in undersaturation (Macpherson and Unwin, 1994a). In contrast, at the (100) face of the crystal, where the average distance between dislocation sites is larger than the tip size, an initial rapid dissolution is followed by an oscillatory dissolution process (Fig. 11A; Macpherson and Unwin, 1994b). The oscillatory dissolution is modeled by assuming that nucleation sites for dissolution are only produced above a critical undersaturation value of Cu^{2+}. Thus, the production of Cu^{2+} by dissolution autoinhibits further dissolution, leading to oscillations. A micrograph (Fig. 11B) of the dissolution pit shows five ledges, in agreement with the five oscillation cycles observed in this experiment. In a related experiment, a gold SECM probe tip began oscillatory dissolution in a 1.5 M HCl solution when it was moved close to a Pt substrate electrode (Mao et al., 1995). In this case, reduction of the tip-generated $AuCl_4^-$ produced a periodic excess of Cl^- ion in the tip-substrate gap, which led to the oscillations.

PROBLEMS

A common problem in feedback SECM is a drift in the value of $i_{T,\infty}$ during an experiment due to a change in the mediator concentration by solution evaporation, a chemical reaction, or a change in electrode activity with time. For quantitative measurements, the value of $i_{T,\infty}$ should be checked several times during the experiment to verify stability. Note also that $i_{T,\infty}$ must be checked at a sufficiently large distances. An L (d/a) value of 10 will still produce an I_T value of 1.045 over a conductive surface. To get I_T values within 1% of the value of $L = \infty, L$ should be >100.

Selecting a mediator for use in feedback imaging is important to obtain good images. Ideally, mediators should be stable in both the oxidized and reduced forms and over a wide range of solution conditions and have rapid ET kinetics. A common difficulty is that a mediator may often undergo a slow chemical reaction, causing a deactivation of the tip response over the time scale of an SECM imaging session. Tip deactivation is a problem, since removal of the tip from the SECM instrument to polish it and return it to the identical position at the sample surface is difficult. Although commonly used in electrochemical experiments, ferricyanide, $Fe(CN)_6^{3-}$ ion can often cause slow deactivation of the tip signal (Pharr and Griffiths, 1997). A popular and very stable mediator that can be used in aqueous solution at pH <7.0 is the $Ru(NH_3)_6^{3+}$ ion. Lists of other mediators that might be suitable for imaging may be found in the literature (Fultz and Durst, 1982; Johnson et al., 1983; Bard et al., 1994). Before using any mediator, its stability over time should be checked. A change of 5% or less over a 1-h period is expected for $i_{T,\infty}$ in a quiet, unstirred solution. In addition, the sample

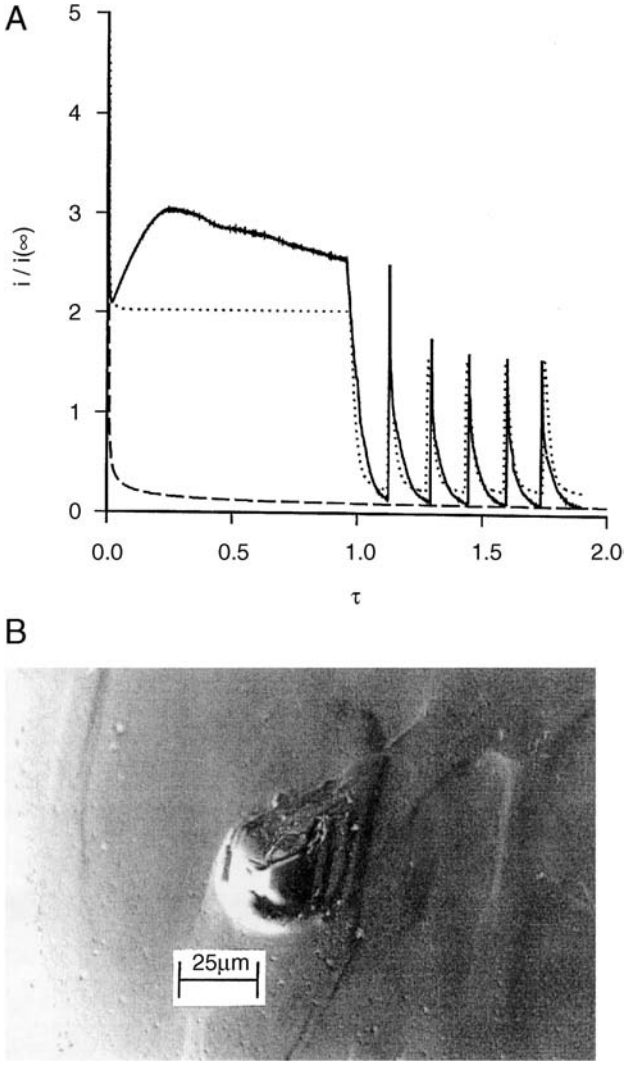


Figure 11 (**A**) Experimental current-time data (—) for the reduction of Cu^{2+} at a 25-μm-diameter Pt disk UME located 1 μm away from a copper sulfate pentahydrate (100). The curve shows the oscillations observed for SECM-induced dissolution at low dislocation density interfaces. (• • •) Theoretical model of dissolution current. (- - -) Theoretical prediction for an inert surface. (**B**) Nomarski differential interference contrast micrograph of the dissolution pit corresponding to the current-time data from (A). (Reprinted with permission from Macpherson and Unwin, 1994b. Copyright © 1994 by the American Chemical Society.)

should remain stable in the mediator solution over a similar time frame.

As discussed under Specimen Modification, SECM can be used to intentionally modify specimen surfaces. The experimenter must be careful, however, to choose imaging conditions that do not cause unintentional damage to the surface. The surface may be damaged by oxidation or reduction by the bulk mediator or by the tip-generated species. Acid-base reaction of the mediator with the surface should also be considered, particularly for tip reactions involving consumption or production of protons.

Electrode preparation is another common problem. The tip geometry must be carefully checked when quantitative feedback image characterization or kinetic measurements are desired. As discussed under Principles of the Method, disk-shaped electrodes with uniform insulation are the most theoretically characterized. Tip responses should be checked by comparing the experimental and theoretical current-distance curves at both insulating and conducting surfaces. If electrode material protrudes from or is recessed into the insulating material, the curves will not match. A comparison of theoretical and experimental curves is also a good check for proper calibration of the SECM positioning system.

Nonlevel surfaces and excessive scan speeds also cause problems during imaging. A steady-state concentration profile must be established during feedback imaging to fit the existing theoretical models. Non-steady-state conditions can arise if the scan rate is too rapid to allow equilibration. As a rule, images made on insulating surfaces are most susceptible since equilibration requires diffusion of material in or out of the tip-substrate gap. A typical ion would take about 100 ms to diffuse from underneath a 10-μm-diameter tip (cf. Equation 6). Nearly equilibrated concentration gradients should occur after 5 or 10 of these diffusional "half-lives" occur, implying an acceptable scan speed of 10 μm/s. However, good practice suggests a check of the experimental method. If possible, the image produced by tip scans in only one direction should be compared to the image produced by sweeps in the opposite direction. Under steady-state conditions, feature appearance and resolution of the two scans should be identical. In GC imaging, solution stirring by rapid tip motion and tip electrolysis will cause changes in images taken by sweeps in alternate directions.

Images made by SECM in the constant-height mode are susceptible to problems due to nonlevel samples or with samples having large changes in surface relief. Tilted samples can be noted by a change in contrast on one side or corner of the sample. Mounting the sample on a tilting mount (such as an adjustable mirror mount) permits leveling of the samples. Postacquisition correction using a mathematical "tilt correction" provides a cosmetic improvement but introduces artifacts in the processed image. A simple procedure for tilt correction is to subtract the image data from the linear regression plane passing through the data. As a rule, scans on tilted samples should proceed so that tip-substrate separation increases during the scan to avoid tipcrashes. Samples with relief of larger than a tip radius can lead to poor image contrast and tip crashes, suggesting the use of constant-current imaging.

Imaging in the GC mode at a voltammetric tip may lead to two problems. First, feedback may occur when the substrate-generated species is reduced or oxidized at the tip and this electrolyzed species diffuses back to the substrate to be restored to its original oxidation state. The regeneration of the substrate-generated species will increase the tip current and compromise the use of Equation 1 for quantitative determination of concentration values. Feedback can be avoided during GC imaging by use of short collecting or generating pulses (Engstrom et al., 1987). Since electrolysis only occurs transiently during the pulse, the

steady-state feedback process is not developed. A second possible problem with voltammetric GC is local depletion of the concentration of substrate-produced species by electrolysis at the tip. This can disrupt the equilibrium governing the release of the substrate-produced species and can confuse quantitative interpretation of the GC image. Again, the sampled GC method will minimize this problem.

PROTOCOL: SECM TIP PREPARATION

The SECM tip is perhaps the most important part of the SECM instrument and, at this time, disk-shaped electrodes must be constructed by the investigator. A construction method for disk electrodes with a radius of $\geq$0.6 μm is described below. This method is a modification of a method described previously in the literature (Wightman and Wipf, 1989). Other methods used to prepare submicrometer disk-shaped tips have also been described (Lee et al., 1990b; Shao et al., 1997a). Normally, cone-shaped tips are made by electrochemically etching Pt-Ir wire and then sealing with Apiezon wax (Nagahara, 1989; Mirkin et al., 1992). Conical tips are also available commercially from electrophysiology suppliers such as FHC (Brunswick, ME) and World Precision Instruments (Sarasota, FL).

Many types of potentiometric tips for use in GC SECM have been described. Tips based on neutral carrier membranes can be made from micropipets, which allows fabrication of submicrometer-sized tips (Morf and Derooij, 1995). The literature contains descriptions of tips and experiments used for measurement of pH (Wei et al., 1995; Park et al., 1996; Klusmann and Schultze, 1997), NH_4^+, K^+ (Wei et al., 1995), and Zn^{2+} (Toth et al., 1995; Wei et al., 1995). A pH-sensitive tip based on an antimony electrode (Horrocks et al., 1993) has the advantage that it can be converted back and forth *in situ* from a voltammetric electrode to a potentiometric electrode by oxidizing the antimony to antimony oxide. The antimony oxide surface is pH sensitive and allows operation in a potentiometric GC mode, while the antimony surface allows precision positioning using the feedback mode.

The ability to construct disk-shaped tips is vital for successful feedback imaging. The following procedure can be followed to construct robust tips of Au, Pt, Ag, Cu, and carbon fiber. Figure 12 is a schematic of a tip constructed by this method. The basic construction process involves sealing a microscopic fiber of the desired electrode material into an insulating glass sheath. Fibers are available from Goodfellow Metals (Cambridge, UK) and Johnson Matthey (Alfa Aesar, Ward Hill, MA), and although fibers as small as 0.6 μm in diameter are available commercially, initial attempts should use larger diameter wire, 50 to 100 μm, until skill is developed in working with the microwire. The smaller wires are fragile, difficult to see, costly, and easily moved by air currents. Start by preparing a well-lit and clean work surface; working on a clean sheet of white paper is helpful. Using fine tweezers or your fingers, place a 1-cm length of the microwire into the open end of a 2-mm-o.d., 1-mm-i.d., 6-cm-long glass capillary. Take care not to bend the wire during installation, since this will make insertion difficult. Fibers may be cleaned before insertion by rinsing with acetone.

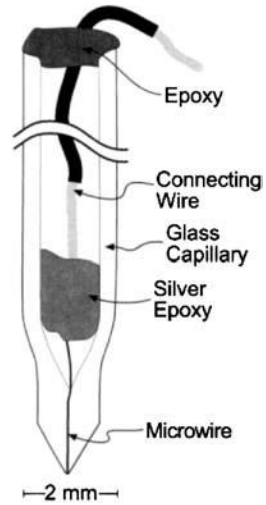


Figure 12. Tapered disk-shaped SECM electrode.

Soda-lime glass capillaries are preferable to borosilicate because of the lower melting point of the soda-lime glass. The soda-lime capillaries are not stock items from most larger supply stores but can be found from suppliers of electrophysiology equipment (e.g., FHC Co., Brunswick, ME). Do not use thin-walled capillaries, such as Pasteur pipets, since the resulting electrodes are unacceptably fragile. Sealing the electrodes in the capillary is a two-part process. Start by using a Bunsen burner flame to seal about 2 mm of the capillary end containing the microwire. A final seal is made by heating the capillary in a simple electric furnace made from a tight coil of 22-gauge chromel (or nichrome) wire. The coil should have an i.d. of 5 mm and length of 20 mm and is heated with an AC voltage from an autotransformer (Variac). To assist the sealing process, a vacuum is maintained inside the capillary. A simple method to connect the capillary to the vacuum source is to wrap the open end of the tube with several layers of Teflon tape to form a gasket. The inlet of the vacuum source is a glass tube with a slightly larger i.d. than the o.d. of the capillary tube. The Teflon gasket will form a good seal when the capillary tube is placed in the larger glass tube and a vacuum is applied.

It is important at this stage to align the capillary tube perpendicular to the ground and to center it exactly in the furnace coil. Mounting the furnace coil on a small laboratory jack allows easy placement of the capillary tube in the coil after the tube is aligned. Once vacuum is established, position the bottom 0.7 cm of the capillary tube in the coil and preheat the coil to a temperature just below a visible glow. Preheat at this temperature for 5 min to drive off any volatile material on the wire or inside the tube. Note, however, that some carbon fibers will decompose with extended preheating. After preheating, raise the coil voltage to heat the coil to a yellow-orange glow. Observe a collapse of the tube walls over a 10- to 30-s interval. A poor electrode seal will occur if melting is more or less rapid than this. Turn off the heating coil voltage and

let the electrode cool undisturbed for about 1 min. About 0.5 cm of the wire should be sealed in the glass at this point. After cooling, confirm there is a good seal under the microscope. Bubbles should not be observed along the shaft of the wire; however, an acceptable electrode may have a few widely spaced bubbles.

Use 400-grit SiC paper to grind away the end of the electrode body and to expose the wire. Ensure that the ground end is perpendicular to the electrode body. Polish with successive grades of Al_2O_3 polish on a wet polishing cloth, starting with 15 μm and continuing with successive grades down to 0.05 μm. Polish at each stage long enough to remove the scratches caused by the previous step. The final polish should leave a scratch-free surface under optical microscopic examination. Again, it should be emphasized that the electrode end must be polished perpendicular to the electrode barrel. Injecting Ag epoxy (EPO-TEK H20E, Epoxy Technology, Billerica, MA) into the tube and inserting a connection wire makes electrical contact to the microwire. A 30-gauge "wire-wrap" wire is useful as connection wire. After curing the Ag epoxy, finish the electrode by sealing the open end with a dab of "5-min" epoxy.

Platinum electrodes of 5.0- to 0.6-μm diameter can be made using a variation of the above procedure. This wire is available in the form of Wollaston wire from Goodfellow Metals. The wires are supplied with a protective 50- to 100-μm-thick Ag coating that must be removed during construction. One end of a 1-cm length of wire is bent to form a small hook, which will hold the wire securely in place and serve as a handle for manipulation. Place the wire (hook-end last) in the capillary and use a length of copper wire to force the wire down the capillary so that the end of the wire is about 2 mm from the open end of the tube. Dissolve the Ag coating with concentrated HNO_3 diluted 1:3 by distilled water. (**CAUTION**: HNO_3 is corrosive.) Dip the end of the tube so that capillary action pulls the acid solution to cover about 3 mm of the wire. Practice this technique with distilled water first. The silver will etch after a short delay with a visible gas evolution. Remove the acid by wicking it out with tissue paper. Rinse 10 times by dipping the tip into distilled water and wicking with a new tissue paper. Inspect the etched wire under a microscope after filling the tube with water again. Seal the electrode in the same way as described above. Note that the electrode will inevitably fail unless the junction between the etched and nonetched Ag is encased in glass during the sealing step.

Test the electrodes using a CV scan in a deaerated solution of 1 to 2 mM $Ru(NH_3)_6^{3+}$ ion in a pH 4.0 phosphate-citrate buffer. A 100-mV/s scan from 0.2 to −0.4 V vs. the saturated calomel electrode (SCE) should yield a sigmoid curve similar to Figure 2. The forward and back traces should be nearly superimposed, and the limiting current plateau should not have a significant slope. Check the limiting current against the theoretical value from Equation 1 (assume $D = 5.5 \times 10^{-6}$ cm^2/s for $Ru(NH_3)_6^{3+}$. Theory and experiment should agree to within 10%. Poor agreement between theory and experiment or a sloping baseline suggests a poor seal or a poor polish. In this case, first try repolishing with the 0.05-μm polish then try the whole polishing cycle again. You should expect a greater than 50% success rate for 10-μm-diameter electrodes. Electrodes that are 2 μm are much more difficult to make and an initial 1-in-3 success rate is admirable.

The final step is grinding the tip to form the truncated cone shape. This step is critical to achieve close tip-substrate separations. Start by using 400-grit SiC paper mounted on a polishing wheel. Grind the end of the electrode at a 45° angle. Rotate the electrode constantly during the grinding process to make the end circular. When the end is reduced to ≈300 μm in diameter as observed under a microscope (compare to the size of the electrode), switch to 1000-grit SiC paper and continue grinding as before to about 100 μm end diameter. This step requires practice and a good optical microscope. Use of diamond polish on the wheel will further reduce the insulator diameter, but a manual tapering process is faster and riskier. "Used" 1000-grit paper is ideal for the manual tapering step; fresh paper is too abrasive. The trick is to drag the electrode with the tapered end pointing *away* from the direction of motion. As you drag, rotate the electrode from a 45° angle to a 90° angle while simultaneously lifting the electrode away from the abrasive sheet. This removes small chunks of glass from the edge. Continue, with constant inspection under the microscope, until the desired RG diameter is reached. Endeavor to produce as circular an end as possible. With practice, one can reproducibly make electrodes with a total insulating radius of less than 10 μm.

ACKNOWLEDGMENTS

The author would like to acknowledge the National Science Foundation, the State of Mississippi, and the Mississippi EPSCOR program for support by grants CHE-94144101 and EPS-9452857.

LITERATURE CITED

Anderson, L. B., McDuffie, B., and Reilley, C. N. 1966. Diagnostic criteria for the study of chemical and physical processes by twin electrode thin-layer electrochemistry. *J. Electroanal. Chem.* 12:477–494.

Anderson, L. B. and Reilley, C. N. 1965. Thin-layer electrochemistry: steady-state methods of studying rate processes. *J. Electroanal. Chem.* 10:295–305.

Arca, M., Mirkin, M. V., and Bard, A. J. 1995. Polymer-films on electrodes: 26. Study of ion-transport and electron-transfer at polypyrrole films by scanning electrochemical microscopy. *J. Phys. Chem.* 99:5040–5050.

Bard, A. J., Fan, F.-R. F., Kwak, J., and Lev, O. 1989. Scanning electrochemical microscopy. Introduction and principles. *Anal. Chem.* 61:132–138.

Bard, A. J. and Faulkner, L. R. 1980. Electrochemical Methods: Fundamentals and Applications. John Wiley & Sons, New York.

Bard, A. J., Fan, F.-R., and Mirkin, M. V. 1994. Scanning electrochemical microscopy. *In* Electroanalytical Chemistry, Vol. 18 (A.J. Bard, ed.) pp. 244–370. Marcel Dekker, New York.

Bard, A. J., Fan, F.-R., and Mirkin, M. 1995. Scanning electrochemical microscopy. *In* Physical Electrochemistry: Principles, Methods, and Applications (I. Rubenstein, ed.) pp. 209–242. Marcel Dekker, New York.

Bard, A. J., Fan, F.-R. F., Pierce, D. T., Unwin, P. R., Wipf, D. O., and Zhou, F.M. 1991. Chemical imaging of surfaces with the scanning electrochemical microscope. *Science* 254:68–74.

Bard, A. J., Mirkin, M. V., Unwin, P. R., and Wipf, D. O. 1992. Scanning electrochemical microscopy: 12. Theory and experiment of the feedback mode with finite heterogeneous electron-transfer kinetics and arbitrary substrate size. *J. Phys. Chem.* 96:1861–1868.

Binnig, G. and Rohrer, H. 1982. Scanning tunneling microscopy. *Helv. Phys. Acta* 55:726–735.

Binnig, G., Quate, C., and Gerber, C. 1986. Atomic force microscope. *Phys. Rev. Lett.* 56:930–933.

Borgwarth, K., Ebling, D. G., and Heinze, J. 1994. Scanning electrochemical microscopy—a new scanning-mode based on convective effects. *Ber. Bunsenges. Phys. Chem.* 98:1317–1321.

Borgwarth, K., Ebling, D. G., and Heinze, J. 1995a. Applications of scanning ultra micro electrodes for studies on surface conductivity. *Electrochim. Acta* 40:1455–1460.

Borgwarth, K., Ricken, C., Ebling, D. G., and Heinze, J. 1995b. Surface characterization and modification by the Scanning Electrochemical Microscope (SECM). *Ber. Bunsenges. Phys. Chem.* 99:1421–1426.

Borgwarth, K., Ricken, C., Ebling, D. G., and Heinze, J. 1996. Surface-analysis by scanning electrochemical microscopy—resolution studies and applications to polymer samples. *Frensenius. J. Anal. Chem.* 356:288–294.

Bottomley, L. A., Courey, J. E., and First, P. N. 1996. Scanning probe microscopy. *Anal. Chem.* 68:185R–230R.

Casillas, N., Charlebois, S. J., Smyrl, W. H., and White, H. S. 1993. Scanning electrochemical microscopy of precursor sites for pitting corrosion on titanium. *J. Electrochem. Soc.* 140:L142–L145.

Casillas, N., Charlebois, S., Smyrl, W. H., and White, H. S. 1994. Pitting corrosion of titanium. *J. Electrochem. Soc.* 141:636–642.

Craston, D. H., Lin, C. W., and Bard, A. J. 1988. High resolution deposition of silver in nafion films with the scanning tunneling microscope. *J. Electrochem. Soc.* 135:785–786.

Dammer, U., Popescu, O., Wagner, P., Anselmetti, D., Günther-odt, H., and Misevic, G. N. 1995. Binding strength between cell adhesion proteoglycans measured by atomic force microscopy. *Science* 267:1173–1175.

Davis, J. M., Fan, F.-R. F., and Bard, A. J. 1987. Currents in thin layer electrochemical cells with spherical and conical electrodes. *J. Electroanal. Chem.* 238:9–31.

Denuault, G., Frank, M. H. T., and Peter, L. M. 1992. Scanning electrochemical microscopy—potentiometric probing of ion fluxes. *Faraday Discuss. Chem. Soc.*, No. 94, 23–35.

Ellis, K. A., Pritzker, M. D., and Fahidy, T. Z. 1995. Modeling the degradation of scanning electrochemical microscope images due to surface-roughness. *Anal. Chem.* 67:4500–4507.

Engstrom, R. C., Meany, T., Tople, R., and Wightman, R. M. 1987. Spatiotemporal description of the diffusion layer with a microelectode probe. *Anal. Chem.* 59:2005–2010.

Engstrom, R. C., Weber, M., Wunder, D. J., Burgess, R., and Winguist, S. 1986. Measurements within the diffusion layer using a microelectrode probe. *Anal. Chem.* 58:844–848.

Frank, M. H. T. and Denuault, G. 1993. Scanning electrochemical microscopy—probing the ingress and egress of protons from a polyaniline film. *J. Electroanal. Chem.* 354:331–339.

Frank, M. H. T. and Denuault, G. 1994. Scanning Electrochemical Microscope (SECM) study of the relationship between proton concentration and electronic charge as a function of ionic-strength during the oxidation of polyaniline. *J. Electroanal. Chem.* 379:399–406.

Fultz, M. L. and Durst, R. A. 1982. Mediator compounds for the electrochemical study of biological redox systems: A compilation. *Anal. Chim. Acta* 140:1–18.

Garfias-Mesias, L. F., Aloclan, M., James, P. I., and Smyrl, W. H. 1998. Determination of precursor sites for pitting corrosion of polycrystalline titanium by using differential techniques. *J. Electrochem. Soc.* 145:2005–2010.

Gilbert, J. L., Smith, S. M., and Lautenschlager, E. P. 1993. Scanning electrochemical microscopy of metallic biomaterials—reaction-rate and ion release imaging modes. *J. Biomed. Mater. Res.* 27:1357–1366.

Horrocks, B. R., Mirkin, M. V., Pierce, D. T., Bard, A. J., Nagy, G., and Toth, K. 1993. Scanning electrochemical microscopy: 19. Ion-selective potentiometric microscopy. *Anal. Chem.* 65:1213–1224.

Hüsser, O. E., Craston, D. H., and Bard, A. J. 1988. High-resolution deposition and etching of metals with a scanning electrochemical microscope. *J. Vac. Sci. Technol.* 6:1873–1876.

Hüsser, O. E., Craston, D. H., and Bard, A. J. 1989. Scanning electrochemical microscopy—high-resolution deposition and etching of metals. *J. Electrochem. Soc.* 136:3222–3229.

Isaacs, H. S., Aldykiewicz, A. J., Thierry, D., and Simpson, T. C. 1996. Measurements of corrosion at defects in painted zinc and zinc alloy coated steels using current-density mapping. *Corrosion* 52:163–168.

Isaacs, H. S. and Kissel, G. 1972. Surface preparation and pit propagation in stainless steel. *J. Electrochem. Soc.* 119:1626–1632.

Jaffe, L. F. and Nuccitelli, R. 1974. An ultrasensitive vibrating probe for measuring steady extracellular currents. *J. Cell Biol.* 63:614–628.

James, P. I., Garfias-Mesias, L. F., Moyer, P. J., and Smyrl, W. H. 1998. Scanning electrochemical microscopy with simultaneous independent topography. *J. Electrochem. Soc.* 145:L64–L66.

Jeon, I. C. and Anson, F. C. 1992. Application of scanning electrochemical microscopy to studies of charge propagation within polyelectrolyte coatings on electrodes. *Anal. Chem.* 64:2021–2028.

Johnson, J. M., Halsall, H. B., and Heineman, W. R. 1983. Metal complexes as mediator-titrants for electrochemical studies of biological systems. *Anal. Biochem.* 133:186–189.

Klusmann, E. and Schultze, J. W. 1997. pH-Microscopy—theoretical and experimental investigations. *Electrochim. Acta* 42: 3123–3136.

Kranz, C., Ludwig, M., Gaub, H. E., and Schuhmann, W. 1995a. Lateral deposition of polypyrrole lines by means of the scanning electrochemical microscope. *Advan. Mater.* 7:38–40.

Kranz, C., Ludwig, M., Gaub, H. E., and Schuhmann, W. 1995b. Lateral deposition of polypyrrole lines over insulating gaps—towards the development of polymer-based electronic devices. *Advan. Mater.* 7:568–571.

Kranz, C., Gaub, H. E., and Schuhmann, W. 1996. Polypyrrole towers grown with the scanning electrochemical microscope. *Advan. Mater.* 8:634–637.

Küpper, M. and Schultze, J. W. 1997a. SLCP—the scanning diffusion limited current probe: A new method for spatially resolved analysis. *Electrochim. Acta* 42:3085–3094.

Küpper, M. and Schultze, J. W. 1997b. Spatially resolved concentration measurements during cathodic alloy deposition in microstructures. *Electrochim. Acta* 42:3023–3032.

Kwak, J. Y. and Bard, A. J. 1989. Scanning electrochemical microscopy. Theory of the feedback mode. *Anal. Chem.* 61:1221–1227.

Kwak, J. Y., Lee, C. M., and Bard, A. J. 1990. Scanning electrochemical microscopy: 5. A study of the conductivity of a polypyrrole film. *J. Electrochem. Soc.* 137:1481–1484.

Lee, C. M. and Bard, A. J. 1990. Scanning electrochemical microscopy—application to polymer and thin metal oxide films. *Anal. Chem.* 62:1906–1913.

Lee, C. M., Kwak, J. Y., and Bard, A. J. 1990a. Application of scanning electrochemical microscopy to biological samples. *Proc. Natl. Acad. Sci. U.S.A.* 87:1740–1743.

Lee, C. M., Miller, C. J., and Bard, A. J. 1990b. Scanning electrochemical microscopy—preparation of submicrometer electrodes. *Anal. Chem.* 63:78–83.

Lee, C. M., Wipf, D. O., Bard, A. J., Bartels, K., and Bovik, A. C. 1991. Scanning electrochemical microscopy: 11. Improvement of image resolution by digital processing techniques. *Anal. Chem.* 63:2442–2447.

Ludwig, M., Kranz, C., Schuhmann, W., and Gaub, H. E. 1995. Topography feedback mechanism for the scanning electrochemical microscope based on hydrodynamic-forces between tip and sample. *Rev. Sci. Instrum.* 66:2857–2860.

Macpherson, J. V. and Unwin, P. R. 1994a. A novel-approach to the study of dissolution kinetics using the scanning electrochemical microscope—theory and application to copper-sulfate pentahydrate dissolution in aqueous sulfuric-acid-solutions. *J. Phys. Chem.* 98:1704–1713.

Macpherson, J. V. and Unwin, P. R. 1994b. Oscillatory dissolution of an ionic single-crystal surface observed with the scanning electrochemical microscope. *J. Phys. Chem.* 98:11764–11770.

Macpherson, J. V., Beeston, M. A., Unwin, P. R., Hughes, N. P., and Littlewood, D. 1995a. Imaging the action of fluid-flow blocking-agents on dentinal surfaces using a scanning electrochemical microscope. *Langmuir* 11:3959–3963.

Macpherson, J. V., Beeston, M. A., Unwin, P. R., Hughes, N. P., and Littlewood, D. 1995b. Scanning electrochemical microscopy as a probe of local fluid-flow through porous solids—application to the measurement of convective rates through a single dentinal tubule. *J. Chem. Soc. Faraday Trans.* 91: 1407–1410.

Macpherson, J. V. Slevin, C. J., and Unwin, P. R. 1996. Probing the oxidative etching kinetics of metals with the feedback mode of the scanning electrochemical microscope. *J. Chem. Soc. Faraday Trans.* 92:3799–3805.

Macpherson, J. V. and Unwin, P. R. 1995. Scanning electrochemical microscope induced dissolution—rate law and reaction-rate imaging for dissolution of the (010) face of potassium ferrocyanide trihydrate in nonstoichiometric aqueous-solutions of the lattice ions. *J. Phys. Chem.* 99:3338–3351.

Madden, J. D. and Hunter, I. W. 1996. Three-dimensional microfabrication by localized electrochemical deposition. *J. Microelectromechan. Syst.* 5:24–32.

Mandler, D. and Bard, A. J. 1989. Scanning electrochemical microscopy—the application of the feedback mode for high resolution copper etching. *J. Electrochem. Soc.* 136:3143–3144.

Mandler, D. and Bard, A. J. 1990a. High resolution etching of semiconductors by the feedback mode of the scanning electrochemical microscope. *J. Electrochem. Soc.* 137:2468–2472.

Mandler, D. and Bard, A. J. 1990b. Hole injection and etching studies of GaAs using the scanning electrochemical microscope. *Langmuir* 6:1489–1494.

Mao, B. W., Ren, B., Cai, X. W., and Xiong, L. H. 1995. Electrochemical oscillatory behavior under a scanning electrochemical microscopic configuration. *J. Electroanal. Chem.* 394:155–160.

Meltzer, S. and Mandler, D. 1995a. Microwriting of gold patterns with the scanning electrochemical microscope. *J. Electrochem. Soc.* 142:L82–L84.

Meltzer, S. and Mandler, D. 1995b. Study of silicon etching in HBr solutions using a scanning electrochemical microscope. *J. Chem. Soc. Faraday Trans.* 91:1019–1024.

Mirkin, M. V. and Bard, A. J. 1992. Multidimensional integral-equations—a new approach to solving microelectrode diffusion-problems: 2. Applications to microband electrodes and the scanning electrochemical microscope. *J. Electroanal. Chem.* 323:29–51.

Mirkin, M. V., Fan, F.-R. F., and Bard, A. J. 1992. Scanning electrochemical microscopy: 13. Evaluation of the tip shapes of nanometer size microelectrodes. *J. Electroanal. Chem.* 328: 47–62.

Mirkin, M. V., Richards, T. C., and Bard, A. J. 1993. Scanning electrochemical microscopy: 20. Steady-state measurements of the fast heterogeneous kinetics in the ferrocene/acetonitrile system. *J. Phys. Chem.* 97:7672–7677.

Morf, W. E. and Derooij, N. F. 1995. Micro-adaptation of chemical sensor materials. *Sensor Actuator A Phys.* 51:89–95.

Nagahara, L. A. T. T. and Lindsay, S. M. 1989. Preparation and characterization of STM tips for electrochemical studies. *Rev. Sci. Instrum.* 60:3128–3130.

Park, J. O., Paik, C. H., and Alkire, R. C. 1996. Scanning microsensors for measurement of local pH distributions at the microscale. *J. Electrochem. Soc.* 143:L174–L176.

Penner, B. D., Heben, M. J., and Lewis, N. S. 1989. Preparation and electrochemical characterization of conical and hemispherical ultramicroelectrodes. *Anal. Chem.* 61:1630.

Pharr, C. M. and Griffiths, P. R. 1997. Infrared spectroelectrochemical analysis of adsorbed hexacyanferrate species formed during potential cycling in the ferrocyanide/ferricyanide redox couple. *Anal. Chem.* 69:4673–4679.

Pierce, D. T. and Bard, A. J. 1993. Scanning electrochemical microscopy: 23. Reaction localization of artificially patterned and tissue-bound enzymes. *Anal. Chem.* 65:3598–3604.

Pierce, D. T., Unwin, P. R., and Bard, A. J. 1992. Scanning electrochemical microscopy: 17. Studies of enzyme mediator kinetics for membrane-immobilized and surface-immobilized glucose-oxidase. *Anal. Chem.* 64:1795–1804.

Scott, E. R., Laplaza, A. I., White, H. S., and Phipps, J. B. 1993a. Transport of ionic species in skin—contribution of pores to the overall skin-conductance. *Pharm. Res.* 10:1699–1709.

Scott, E. R., Phipps, J. B., and White, H. S. 1995. Direct imaging of molecular-transport through skin. *J. Investig. Dermatol.* 104: 142–145.

Scott, E. R., White, H. S., and Phipps, J. B. 1993b. Iontophoretic transport through porous membranes using scanning electrochemical microscopy—application to in vitro studies of ion fluxes through skin. *Anal. Chem.* 65:1537–1545.

Selzer, Y. and Mandler, D. 1996. A novel-approach for studying charge-transfer across an interface of 2 immiscible solutions using the scanning electrochemical microscope (SECM). *J. Electroanal. Chem.* 409:15–17.

Shao, Y. H., Mirkin, M. V., Fish, G., Kokotov, S., Palanker, D., and Lewis, A. 1997a. Nanometer-sized electrochemical sensors. *Anal. Chem.* 69:1627–1634.

Shao, Y. H., Mirkin, M. V., and Rusling, J. F. 1997b. Liquid/liquid interface as a model system for studying electrochemical catalysis in microemulsions—reduction of trans-1,2-dibromocyclohexane with vitamin-B-12. *J. Phys. Chem. B* 101:3202–3208.

Shohat, I. and Mandler, D. 1994. Deposition of nickel-hydroxide structures using the scanning electrochemical microscope. *J. Electrochem. Soc.* 141:995–999.

Slevin, C. J., Umbers, J. A., Atherton, J. H., and Unwin, P. R. 1996. A new approach to the measurement of transfer rates across immiscible liquid/liquid interfaces. *J. Chem. Soc. Faraday Trans.* 92:5177–5180.

Still, J. W. and Wipf, D. O. 1997. Breakdown of the iron passive layer by use of the scanning electrochemical microscope. *J. Electrochem. Soc.* 144:2657–2665.

Toth, K., Nagy, G., Wei, C., and Bard, A. J. 1995. Novel application of potentiometric microelectrodes—scanning potentiometric microscopy. *Electroanalysis* 7:801–810.

Tsionsky, M., Bard, A. J., and Mirkin, M. V. 1996. Scanning electrochemical microscopy: 34. Potential dependence of the electron-transfer rate and film formation at the liquid/liquid interface. *J. Phys. Chem.* 100:17881–17888.

Tsionsky, M., Cardon, Z. G., Bard, A. J., and Jackson, R. B. 1997. Photosynthetic electron-transport in single guard-cells as measured by scanning electrochemical microscopy. *Plant Physiol.* 113:895–901.

Unwin, P. R. and Macpherson, J. V. 1995. New strategies for probing crystal dissolution kinetics at the microscopic level. *Chem. Soc. Rev.* 24:109–119.

Wei, C., Bard, A. J., Nagy, G., and Toth, K. 1995. Scanning electrochemical microscopy: 28. Ion-selective neutral carrier-based microelectrode potentiometry. *Anal. Chem.* 67:1346–1356.

Wightman, R. M. 1988. Voltammetry with microscopic electrodes in new domains. *Science* 240:415–420.

Wightman, R. M. and Wipf, D. O. 1989. Voltammetry at ultramicroelectrodes. *In* Electroanalytical Chemistry, Vol. 15 (A. J. Bard, ed.) pp. 267–353. Marcel Dekker, New York.

Wipf, D. O. 1994. Initiation and study of localized corrosion by scanning electrochemical microscopy. *Colloid Surf. A* 93:251–261.

Wipf, D. O. and Bard, A. J. 1991a. Scanning electrochemical microscopy: 7. Effect of heterogeneous electron-transfer rate at the substrate on the tip feedback current. *J. Electrochem. Soc.* 138:469–474.

Wipf, D. O. and Bard, A. J. 1991b. Scanning electrochemical microscopy: 10. High resolution imaging of active sites on an electrode surface. *J. Electrochem. Soc.* 138:L4–L6.

Wipf, D. O. and Bard, A. J. 1992. Scanning electrochemical microscopy: 15. Improvements in imaging via tip-position modulation and lock-in detection. *Anal. Chem.* 64:1362–1367.

Wipf, D. O., Bard, A. J., and Tallman, D. E. 1993. Scanning electrochemical microscopy: 21. Constant-current imaging with an autoswitching controller. *Anal. Chem.* 65:1373–1377.

Wittstock, G., Emons, H., Kummer, M., Kirchhoff, J. R., and Heineman, W. R. 1994a. Application of scanning electrochemical microscopy and scanning electron-microscopy for the characterization of carbon-spray modified electrodes. *Fresenius' J. Anal. Chem.* 348:712–718.

Wittstock, G., Emons, H., Ridgway, T. H., Blubaugh, E. A., and Heineman, W. R. 1994b. Development and experimental evaluation of a simple system for scanning electrochemical microscopy. *Anal. Chim. Acta* 298:285–302.

Wuu, Y.-M., Fan, F.-R. F., and Bard, A. J. 1989. High resolution deposition of polyaniline on Pt with the scanning electochemical microscope. *J. Electrochem. Soc.* 136:885–886.

Zhou, J. F. and Wipf, D. O. 1997. Deposition of conducting polyaniline patterns with the scanning electrochemical microscope. *J. Electrochem. Soc.* 144:1202–1207.

Zhu, Y. Y. and Williams, D. E. 1997. Scanning electrochemical microscopic observation of a precursor state to pitting corrosion of stainless-steel. *J. Electrochem. Soc.* 144:L43–L45.

KEY REFERENCES

Bard et al., 1994. See above.

Provides a complete review of SECM theory and applications by one of the founders of the technique.

Bard et al., 1995. See above.

Reviews and updates the state-of-the-art in SECM.

INTERNET RESOURCES

http://www.msstate.edu/dept/chemistry/dow1/secm/secm.html

A bibliographic listing of SECM and other closely related articles.

http://www.warwick.ac.uk/electrochemistry/secan.html

An on-line overview of the SECM technique

http://chinstruments.com

CHI sells a SECM instrument. An overview of SECM principles and operations.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

a	Radius of tip electrode
A	Tip area
AFM	Atomic force microscopy
C	Concentration of redox species
CV	Cyclic voltammetry
d	Distance between tip and substrate
D	Diffusion coefficient of redox species
ET	Electron transfer
F	Faraday constant
GC	Generation/collection
d	Distance between tip and substrate
IR	Impulse response
ISE	Ion-selective electrode
$i_{T,\infty}$	Steady-state limiting current of tip electrode at infinite tip-substrate separation
I_T	Normalized tip current, $i_T/i_{T,\infty}$
$I_T(L)$	Normalized tip current at normalized distance L
$I_{T,ins}(L)$	Normalized tip current over an insulator at normalized distance L
k	Heterogeneous rate constant
L	Normalized distance from tip to substrate, $=d/a$
LOG	Laplacian of Gaussian (filter)
n	Number of electrons transferred in a redox event
Ox	A mediator in an oxidized state
PSF	point spread function
Red	A mediator in a reduced state
RG	Ratio of the insulator radius to the disk electrode radius
SCE	Saturated calomel electrode
SECM	Scanning electrochemical microscopy

SG/TC	Substrate generation/tip collection mode of SECM
SPM	Scanning probe microscopy
STM	Scanning tunneling microscopy
Substrate	Stationary electrode in SECM setup
t	time
Tip	Typically the electrode that is rastered on the scanning electrochemical microscope
TPM	Tip position modulation
TTL	Transistor-transistor logic, digital logic signals
UME	Ultramicroelectrode
x	A distance
α_0	Angle between two lines coming from the center of a sphere

DAVID O. WIPF
Mississippi State University
Mississippi State, Mississppi

THE QUARTZ CRYSTAL MICROBALANCE IN ELECTROCHEMISTRY

INTRODUCTION

The electrochemical quartz crystal microbalance (EQCM) has found wide acceptance as an analytical tool. Nanogram mass changes that occur on the surface of a quartz crystal can be correlated with the amount of charge passed during an electrochemical process. This technique has been widely reviewed (Buttry, 1991; Buttry and Ward, 1992; Ward, 1995).

That a mass change can be determined from the resonant frequency change of a quartz crystal has been extensively used in vapor deposition techniques for many years (Lu and Czanderna, 1984). However, the use of the quartz crystal microbalance (QCM) in contact with a liquid during electrochemical studies is more recent (Kaufman et al., 1984; Schumacher, 1990). Initial investigations were applied to the underpotential deposition of metals (Deakin and Melroy, 1988; Hepel et al., 1990). The EQCM has since been applied to a wide variety of electroactive materials. Examples include conducting polymers (Orata and Buttry, 1987, 1988; Daifuku et al., 1989; Servagent and Vieil, 1990), conducting polymer bilayers (Hillman and Glidle, 1994; Demoustier-Champagne et al., 1995), redox polymers (Bruckenstein et al., 1989), charge-transfer salts (Ward, 1989; Freund et al., 1990; Evans and Chambers, 1994), hydrogen uptake by a metal (Cheek and O'Grady, 1990), electrochromic behavior of metal oxide electrodes (Cordoba-Torresi et al., 1991), polymerization of disulfides (Naoi et al., 1995), self-assembled ferrocene-based monolayers (De Long et al., 1991), Prussian blue (Feldman and Melroy, 1987; Deakin and Byrd, 1989), and nickel ferricyanide (Lasky and Buttry, 1988).

The fact that the mass change of any electroactive material during any electrochemical change can be evaluated with the EQCM has led to its widespread application. The ability to simultaneously determine the mass and the charge passed during an electrochemical process in an electroactive material is critical for characterization in many instances. For example, the polymerization mechanism of an electrochemically synthesized material can be studied as well as quantifying the amount of ion and/or solvent transport during oxidation/reduction. In addition, recent applications of the QCM include simultaneous QCM and ellipsometry (Rishpon et al., 1990), investigating solvent dynamics in polymer films (Katz and Ward, 1996), and determining contact angles and surface tensions (Lin and Ward, 1996).

This unit provides a general overview of the EQCM technique and provides details about particular interpretative precautions. The QCM has also been called a thickness-shear-mode (TSM) oscillator. Related techniques involve the surface-acoustic wave and flexural-plate wave devices (Ward and Buttry, 1990).

PRINCIPLES OF THE METHOD

The EQCM can be used to study interfacial processes at the electrode surface. The piezoelectric property of quartz is utilized to record a frequency change that may be related to a mass change. Caution must be used in the interpretation of the results (see Data Analysis and Initial Interpretation) because the frequency change is not always exclusively due to a mass change. Measurements can be made to determine whether the frequency-to-mass relationship holds for a particular material deposited onto the quartz crystal.

Since 1880, it has been known that quartz exhibits piezoelectric behavior. Jacques and Pierre Curie discovered that by applying a mechanical stress to crystals such as quartz and rochelle salt, a resulting voltage is produced (Curie and Curie, 1880). The Greek word piezin means "to press." From this the "pressure-voltage" effect was named. The converse electrostrictive effect was also discovered: the application of a potential across a quartz crystal results in a corresponding mechanical strain. Therefore, when electrodes are placed on a quartz crystal and a periodic voltage is applied, the crystal can be made to vibrate at its resonant frequency. This is the basis of the use of a quartz crystal as a microbalance.

Quartz crystals are cut so that a particular mode of vibration dominates. The AT-cut is most commonly used for QCM applications. The mode of vibration for this cut is in the thickness-shear mode (see Fig. 1A) and is most responsive to changes in mass. AT-cut indicates that the quartz is sliced along a plane rotated 35°15′ about the x axis, as shown in Figure 1B. Another reason for selecting the AT-cut is that it is somewhat insensitive to temperature changes near room temperature. Variations in temperature must be minimized for the QCM. Further details about piezoelectric quartz crystals are available and well known from high-vacuum deposition applications (Lu and Czanderna, 1984).

The shear-mode deformation shown in Figure 1A indicates that the motion is parallel to the quartz surface. The resonant frequency is governed by the thickness of the quartz crystal. This frequency oscillation is induced by

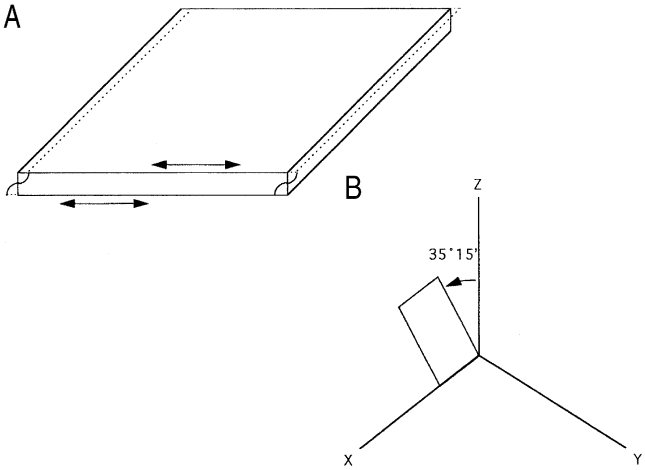


Figure 1. (**A**) A quartz crystal indicating thickness-shear mode of oscillation. (**B**) Orientation of an AT-cut quartz crystal.

the application of an electrical potential across electrodes deposited onto both sides of the quartz crystal, as shown in Figure 2A. The applied potential results in a mechanical shear deformation. An alternating potential causes the crystal to oscillate. This oscillation is extremely stable when it is near the crystal's mechanical resonance frequency due to the inherent piezoelectric properties of quartz.

An acoustic wave is produced by the vibrational motion of the quartz crystal. This wave extends through the thickness of the crystal, as shown in Figure 2B. The thickness of the quartz crystal is equal to one-half of the wavelength at resonance:

$$t_q = \frac{1}{2}\lambda_q \tag{1}$$

where t_q is the thickness of the quartz and λ_q is the wavelength of the acoustic wave in quartz. The effects of electrodes on both sides of the quartz are assumed to be negligible. This relationship can be expressed in terms of the resonant frequency of quartz (f_0) and the velocity of the acoustic wave in quartz (v_q),

$$f_0 t_q = \frac{1}{2} v_q \tag{2}$$

since

$$\lambda_q f_q = v_q \tag{3}$$

A change in the thickness therefore results in a change in the resonant frequency. An increase in the thickness causes a decrease in the resonant frequency of the quartz crystal. This leads to the principal postulate of this technique: the Sauerbrey relationship. This key equation relates a change in mass to a frequency change (Sauerbrey, 1959):

$$\Delta f = -\frac{2f_0^2 \Delta m}{A(\mu_q \rho_q)^{1/2}} \tag{4}$$

where Δf is the frequency change, f_0 the resonant frequency of the quartz crystal, Δm the mass change, A the piezoelectrically active area, ρ_q the density of quartz (2.648 g/cm^3), and μ_q the shear modulus of AT-cut quartz (2.947×10^{11} dyne/cm^2). An inherent assumption is made that for small mass changes due to foreign deposits onto the crystal, the addition of mass can be treated as an equivalent mass change of the quartz crystal. It is also assumed that the acoustic velocity and density of the foreign layer are identical to those in quartz. Even though these assumptions are used, this relationship holds true for most QCM measurements. The thickness of the deposited film should be less than 2% of the quartz crystal thickness.

Since f_0, μ_q, and ρ_q are constants of the quartz, Equation 4 can be expressed as

$$\Delta f = -C_f \Delta \mathrm{m} \tag{5}$$

where C_f is the mass sensitivity of the particular quartz crystal. This relationship emphasizes that a mass increase results in a frequency decrease and conversely a mass decrease results in a frequency increase.

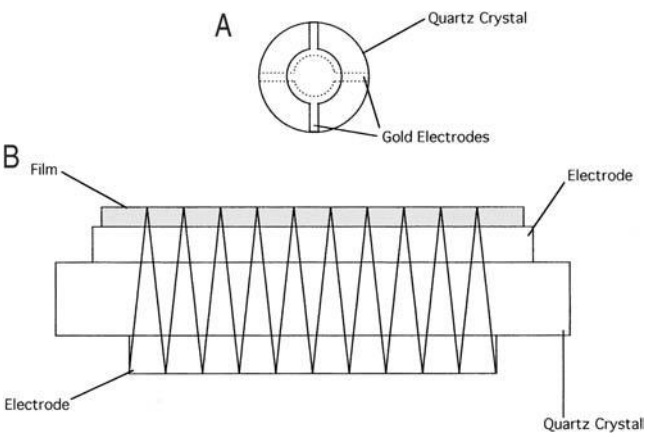


Figure 2. (**A**) Electrode pattern on both sides of a quartz crystal; dotted electrode is on bottom side. (**B**) Schematic of the transfer shear wave propagating through a quartz crystal, electrodes and film deposit.

Equivalent Circuit

W. G. Cady at Wesleyan University (Bottom, 1982) was the first (ca. 1920) to use a piezoelectric quartz crystal to control the frequency of an oscillator circuit. However, a better understanding of the piezoelectric resonator was realized in 1925 when K. S. Van Dyke, a student and colleague of Cady, found that the quartz resonator could be described by an equivalent electrical circuit. A piezoelectric quartz crystal resonator can be represented by a series resonant circuit, consisting of a resistor, capacitor, and inductor, in parallel with a capacitor, as shown in Figure 3. The series branch of the circuit is called the motional branch because it reflects the vibrational behavior of the crystal. The parallel capacitance C_0 represents the static capacitance of the quartz crystal with its electrodes and any stray parasitic

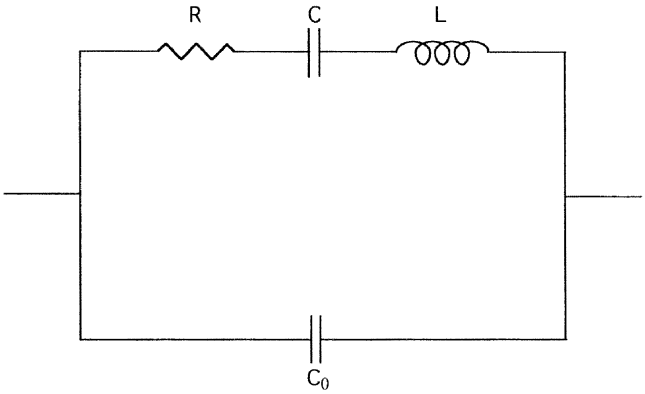


Figure 3. Equivalent electrical circuit describing the properties of a quartz resonator.

capacitances due to the connections to the crystal and its environment. The resistance R represents the energy dissipation due to internal friction, mechanical losses in the mounting system, and acoustical losses to the surrounding environment. The capacitance C represents the compliance of the quartz or the energy stored during oscillation. The inductance is the only component related to the mass; L is the inertial component representing the mass displaced during oscillation.

The electrical parameters can be expressed in terms of crystal properties (Cady, 1964):

$$C_0 = \frac{D_q \varepsilon_0 A}{t_q} \approx 10^{-12}\,\mathrm{F} \tag{6}$$

$$C = \frac{8A \in^2}{\pi^2 t_q c} \approx 10^{-14}\,\mathrm{F} \tag{7}$$

$$R = \frac{t_q^3 r}{8A \in^2} \approx 100\,\Omega \tag{8}$$

$$L = \frac{t_q^3 \rho}{8A \in^2} \approx 0.075\,\mathrm{H} \tag{9}$$

where D_q is the dielectric constant of quartz, ε_0 is the permittivity of free space, A is the piezoelectric active area, t_q is the thickness of quartz, c is the elastic constant of quartz, r is a dissipation coefficient corresponding to energy losses during oscillation, ρ is the density of quartz, and $\in$ is the piezoelectric stress constant. Typical values for a 5-MHz quartz resonator are given (Buttry and Ward, 1992).

Understanding what factors affect these terms will be useful in analyzing the results obtained from the QCM (see Data Analysis and Initial Interpretation). The parameters C, R, and L depend on $\in$, the piezoelectric stress constant. Note, however, that C_0 does not participate directly in piezoelectricity as it does not depend on $\in$. Capacitance C depends on the elastic constant of quartz; it is the component that responds to energy compliance. Resistance R depends upon the dissipation energy r, which results from thermal dissipation in the quartz resonator and coupling of the acoustic wave to the environment. An increase in the dissipation energy will therefore result in an increase in R. Inductance L depends on the density of the quartz. The quantity $t_q^3\rho/A$ represents mass per unit area and is related to the $\Delta m/A$ term in the Sauerbrey equation. An increase in mass will therefore result in an increase in L. Also note the relationship of these terms to the piezoelectric active area A. Capacitances C_0 and C increase with area, while R and L decrease with area.

Impedance Analysis

Frequency changes alone are not sufficient to ensure that all analyses made using the QCM are accurate. This is because frequency changes can occur due to other factors. Impedance analysis (also called network analysis) is needed to better characterize the deposit on the quartz crystal (Kipling and Thompson, 1990; Noel and Topart, 1994). Factors that can cause a frequency change (and not necessarily be directly related to a mass change) include (1) viscoelastic changes in the film deposited on the crystal (these properties can change with thickness as well as due to the experiment performed, such as doping of a polymer), (2) interactions with the contacting liquid that can cause swelling of a polymer, (3) morphological changes of the polymer, and (4) nonuniform mass deposition.

Impedance analysis (Honda, 1989), therefore, is useful in analyzing the properties of the quartz crystal, properties of films deposited onto the surface of the crystal, and the interaction of the crystal with a contacting liquid (Martin et al., 1991). This analysis is performed to confirm that the frequency changes observed can be related to mass changes.

The impedance of a quartz crystal is measured by applying a known voltage over a specified range of frequencies and measuring the current. By analogy with Ohm's law, $V = IZ$, where Z is the impedance, V is voltage, and I is current. The impedance is a sum of both the resistance R and the reactance X:

$$Z = R + jX \tag{10}$$

For direct current, this reduces to $Z = R$. Impedance can be represented as a complex quantity that is graphically shown in the complex plane in Figure 4, where the imaginary unit $j(\equiv \sqrt{-1})$ is introduced. The absolute value of the impedance is given by $|Z| = (R^2 + X^2)^{1/2}$.

Since the equivalent circuit for a quartz resonator consists of a static capacitance in parallel with the RLC components, it is better to use the inverse of impedance, Z, which is admittance, Y:

$$Y = G + jB \tag{11}$$

where G is the conductance (reciprocal of R) and B is the susceptance (reciprocal of X). The absolute value of the admittance is given by $|Y| = (G^2 + B^2)^{1/2}$. The unit of impedance is the ohm and the unit of admittance is therefore the inverse ohm or siemen (S). This measurement is most conveniently made with an impedance analyzer that is capable of recording these parameters of interest: impedance (Z), phase angle (θ, where $\theta = \tan^{-1}(X/R)$), admittance (Y), conductance (G), and susceptance (B).

A piezoelectric quartz resonator produces a susceptance B that rapidly changes with frequency. The frequency at

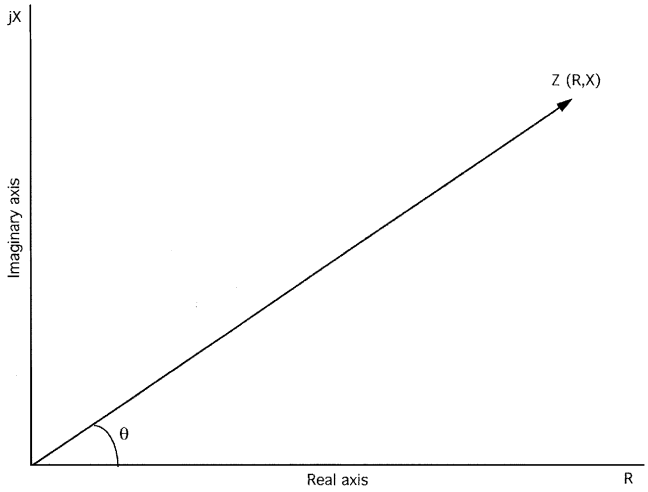


Figure 4. Impedance (Z) as a component of the real resistance (R) and the imaginary reactance (X). The polar form consists of the magnitude and the phase angle: $|Z|\angle\theta$.

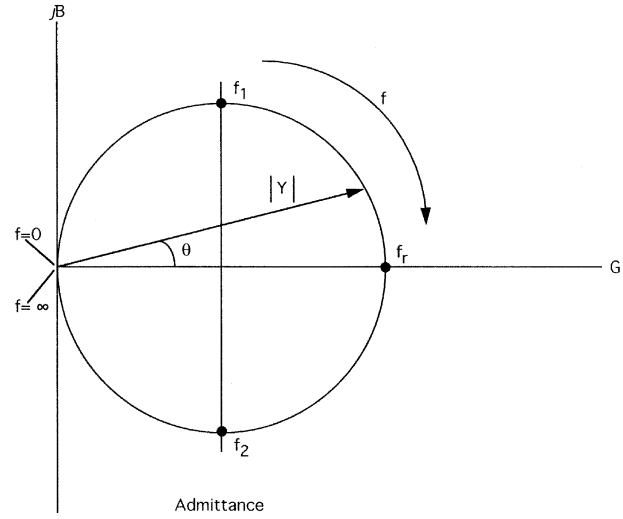


Figure 6. Admittance, Y, locus for RCL circuit: real component conductance G vs. imaginary component susceptance B. Resonance frequency f_r occurs when $Y = G$ and $\theta = 0$.

which capacitive susceptance (capacitive reactance) equals the inductive susceptance (inductive reactance) is the resonance frequency of the quartz oscillator; admittance is a maximum and impedance a minimum at this frequency ($Y = G$ or $Z = R$).This is the resonance frequency of a quartz oscillator.

This is illustrated in Figure 5 with a plot of $|X|$ vs. frequency (ω) for an ideal capacitor and an ideal inductor. The reactance of an ideal capacitor is inversely proportional to frequency. The reactance of an ideal inductor increases linearly with frequency. The point at which they are equal is the resonance frequency. The reactances cancel and therefore the impedance is at a minimum ($Z = R$).

Since admittance (and impedance) are complex quantities, it is useful to express the frequency relationship as an admittance locus as shown in Figure 6 for a series RLC circuit. Admittance is composed of both the real component G, conductance, on the abscissa and the imaginary component jB, susceptance, on the ordinate. As frequency increases, the imaginary component jB reaches a maximum at f_1. The imaginary component reaches a minimum at f_2. When the frequency crosses the real axis, G is a maximum. This is the resonance frequency (f_r) because $Y = G$ at this point. Note that the phase angle θ is also equal to zero at the resonance frequency. This is necessary to fulfill the standing-wave condition.

The RLC circuit shown in Figure 6 must be modified for a quartz resonator (Tiean et al. 1990; Buttry and Ward, 1992). The additional contribution of the static capacitance C_0 raises the admittance locus along the imaginary axis by ωC_0, as shown in Figure 7. This causes several important changes to the diagram. Resonance

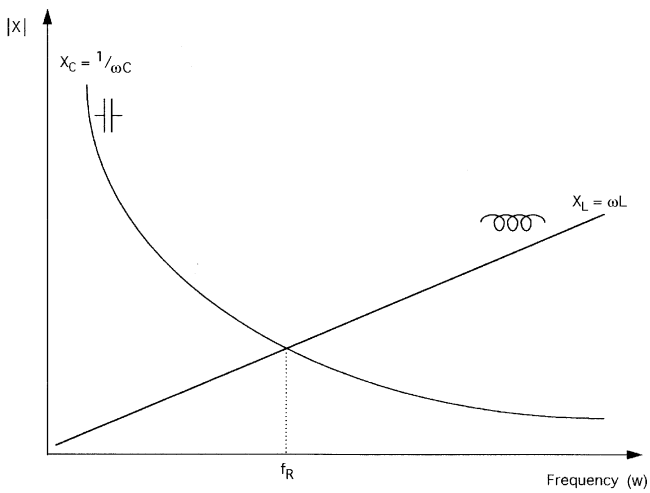


Figure 5. Absolute reactance vs. frequency plot for an ideal capacitor and an ideal inductor indicating resonance frequency.

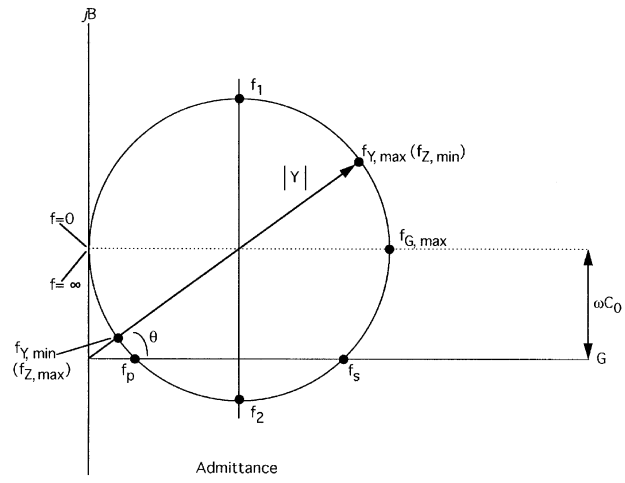


Figure 7. Admittance locus for quartz resonator: real component conductance G vs. imaginary component susceptance B. Static capacitance C_0 raises admittance locus resulting in two resonance frequencies, f_s (series frequency) and f_p (parallel frequency), when $\theta = 0$.

is now satisfied at two frequencies: f_s (series frequency) and f_p (parallel frequency), since the phase angle θ must equal zero at both f_s and f_p. However, the frequency of maximum susceptance, $f_{G,max}$, now occurs at a slightly lower frequency than f_s. The real part of the admittance at f_s is slightly less than G_{max}. The real part of the admittance of the second resonant frequency, f_p, is slightly greater than zero. The frequency of f_p is slightly less than the frequency of minimum admittance, $f_{Y,min}$. The series frequency f_s is the frequency generally measured with the QCM. When a frequency counter alone is used in QCM, the frequency most often monitored is f_s. However, the parallel frequency f_p is also useful in providing information about the surface deposit (Buttry and Ward, 1992).

The frequencies f_s and f_p can converge if R or C_0 increases. As C_0 increases, the admittance locus is raised, causing f_s and f_p to approach similar values. If R increases, the diameter of the locus decreases, thereby causing f_s and f_p to converge. It should be noted that very large values of R or C_0 will result in admittance loci that do not cross the real axis, G. The phase angle no longer equals zero and resonance cannot occur.

The impedance analyzer can provide plots of Z and θ vs. frequency, as shown in Figure 8 for a blank 5-MHz quartz crystal. The frequency of Z_{min} is approximately the frequency of $\theta = 0$, which is f_s. The frequency of Z_{max} is close in frequency to the value of the second resonant frequency, f_p, where θ also equals zero. Plots of B and G can also be displayed as shown in Figure 9. The frequency of G_{max} is close to the resonant frequency, f_s. As shown in Figure 9, the frequencies of B_{max} and B_{min} coincide with the peak width at half-height of the plot of G (noted as f_1 and f_2 Fig. 9). Current flows most easily at frequencies near $f_{G,max}$. Quartz crystals behave as bandpass filters. The bandwidth is $f_2 - f_1$. Quartz resonators have very small bandwidths; therefore, they are ideal for frequency control. The sharp conductance peak ensures that the feedback loop of the oscillator circuit is able to lock in on the narrow frequency range.

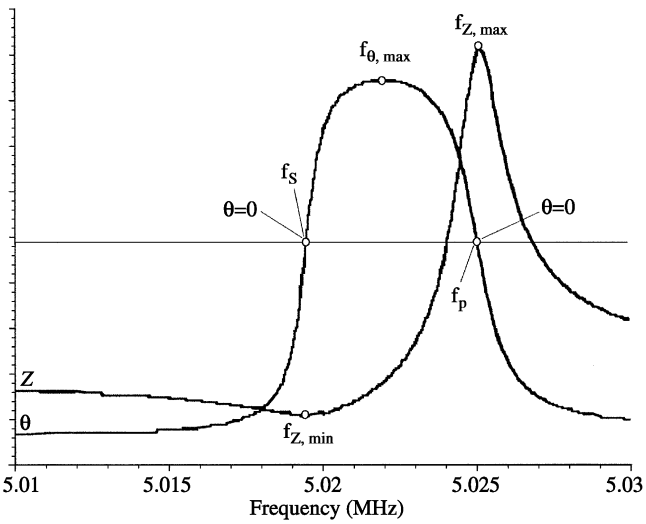


Figure 8. Impedance Z (sharp change at $f_{Z,max}$) and phase angle θ (broad peak at $f_{\theta,max}$) vs. frequency of a blank 5-MHz quartz crystal in the resonance region.

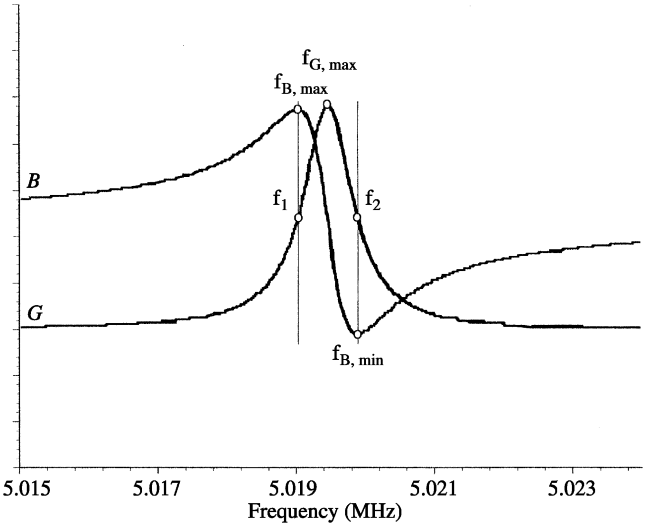


Figure 9. Plot of conductance (G) and susceptance (B) vs. frequency.

Quartz crystals have very high quality factors, Q. The bandwidth and the resonant frequency determine Q. The high values of Q for quartz crystals (~100,000) are due to their high inductance and low resistance. The quality factor can be determined from the information in Figure 9. It is the ratio of the energy stored to the energy lost during oscillation. This is approximated by $f_{G,max}$ divided by $f_2 - f_1$:

$$Q = \frac{f_s}{f_2 - f_1} = \frac{f_{G,max}}{f_2 - f_1} \tag{12}$$

The quality factor Q should be measured before, during, and after a QCM experiment. The frequency of G_{max} is influenced by changes in mass. Large changes in $f_2 - f_1$, however, indicate a large increase in the energy dissipated. Under these conditions, the Sauerbrey equation is not applicable.

A conductance plot should be made to evaluate the quality factor. A plot of G vs. frequency for a blank 5-MHz quartz crystal is shown in Figure 10A. The plot in Figure 10B illustrates how the frequency of G_{max} has decreased and the full width at half height ($f_2 - f_1$) changes slightly due to the added mass of a polymer film. However, the factor Q has only changed from 1.6×10^4 to 9.8×10^3, indicating that the Sauerbrey relationship can still be applied.

Film and Solution Effects

In actual QCM experiments, additional factors are involved that modify the equivalent circuit. The film deposited onto the crystal and the liquid in contact with the crystal make contributions. The added liquid produces a mechanical impedance. Additional components are the inductance L_f, viscosity η_f, density ρ_f, energy dissipation R_d, and elasticity C_f of the film as well as the viscosity η_L and the density ρ_L of the solution. These additional

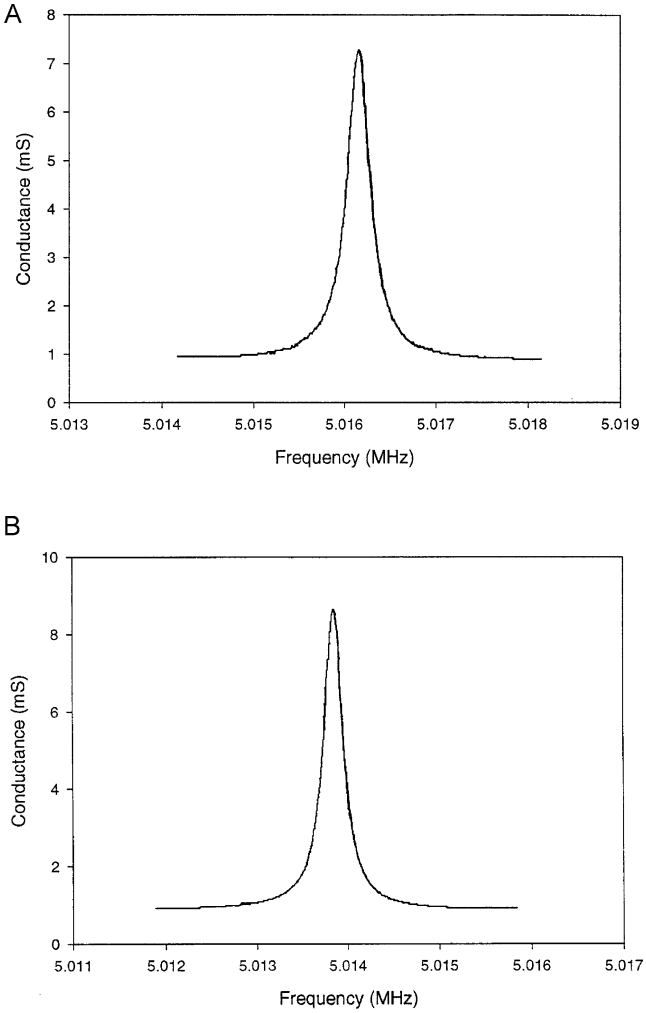


Figure 10. (**A**) Conductance G vs. frequency of a blank 5-MHz crystal. (**B**) Conductance G vs. frequency of the same 5-MHz crystal after deposition of a polymer film indicating little change in $f_2 - f_1$ and a decrease in $f_{G,\max}$. Note change in frequency range.

components are illustrated in Figure 11 and discussed in detail elsewhere (Buttry and Ward, 1992).

Recognizing that these other factors can influence the measured frequency change is important for correct data analysis. Since most EQCM studies are done with a constant amount of solution above the crystal, η_L and ρ_L can be assumed to remain constant during the measurement.

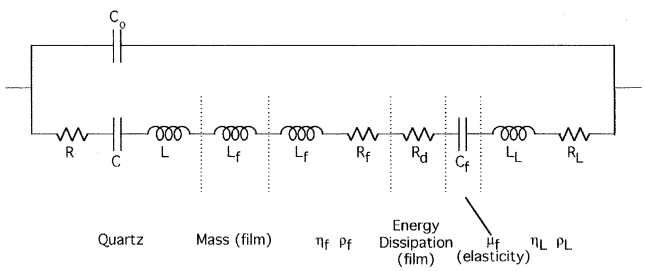


Figure 11. Equivalent electrical circuit for an AT-cut quartz resonator including film and solution components.

This also illustrates why the QCM cannot be used to measure an absolute mass value. Specifically, changes in mass are reported by the QCM.

Especially when polymer films are under investigation, the components in Figure 11 should be considered. This indicates that the measurement of the change in f_s is not enough. Other contributions from the elasticity, density, and viscosity of the film may contribute to the observed Δf_s. One method to determine if these viscoelastic effects influence a particular system is to carry out the QCM measurements over a range of film thicknesses. If the observed mass change is linear with film thickness, the viscoelastic effects are not significant. Impedance analysis can be used to evaluate the other contributions when they lead to nonideal behavior.

PRACTICAL ASPECTS OF THE METHOD

Instrumentation

Several commercial EQCM systems are available. The advantages are ease of initial set-up, including experimentation and computer analysis. Limitations are (1) the need to purchase quartz crystals with electrodes from the manufacturer (this limits experimental design changes), (2) the lack of direct access to the components (i.e., accepting values without full understanding of how they were obtained), and (3) the lack of impedance analysis to confirm the Sauerbrey relationship for a particular system.

Since a large majority of EQCMs have been home-built systems, this aspect will be addressed. The EQCM consists of a quartz crystal, one side of which acts as the working electrode, an appropriate cell holder, counter and reference electrodes, and instrumentation. A schematic of a typical system is shown in Figure 12. The top of the crystal (working electrode) is connected to ground. This limits the

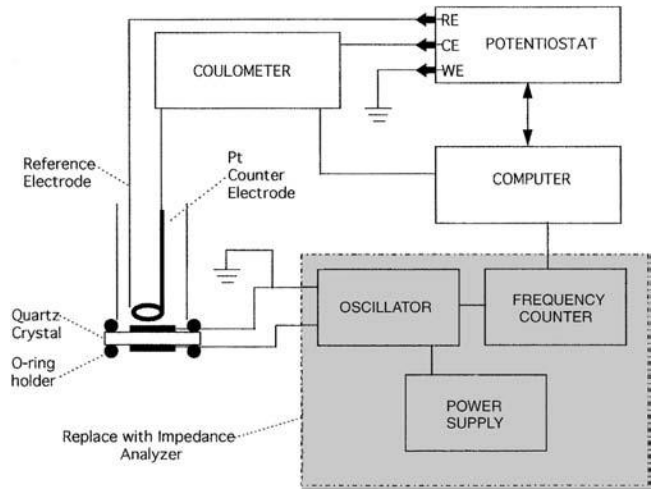


Figure 12. Schematic of a home-built electrochemical quartz crystal microbalance. The oscillator, frequency counter, and power supply can be replaced with an impedance analyzer. The electrochemical cell consists of a Pt counter electrode and reference electrode. The working electrode is the top of the quartz crystal.

use of a commercial potentiostat to measure the charge flow. A coulometer can be placed in the circuit between the potentiostat and the QCM, as shown in Figure 12. Another option is to measure the voltage drop across a known resistor, converting to current and integrating to obtain charge.

The QCM instrumentation consists of a voltage source (power supply) for the oscillator circuit and a frequency counter to record frequency changes. Several oscillator circuit diagrams are available in the literature and the oscillator circuit can be readily constructed (Bruckenstein and Shay, 1985; Benje et al., 1986; Deakin and Melroy, 1989; Buttry, 1991). Another option is for the oscillator circuit, power supply, and frequency counter to be replaced by an impedance analyzer, as shown in Figure 12. The advantages are very stable frequency measurements and the additional ability to record impedance and/or admittance, as discussed under Principles of the Method.

Recently, reports of dual QCM oscillator circuits have been published (Bruckenstein et al., 1994; Dunham et al., 1995). The spectroelectrochemical quartz crystal microbalance (SEQCM) was recently introduced (Arbuckle and Travis, 1994). It integrates spectroelectrochemistry and QCM. A configuration such as shown in Figure 13 can be utilized. The electrochemistry is performed directly in the sample compartment of the spectrometer. Modifications must be performed to allow transmission through the electrodes and quartz crystal. Thin metal electrodes are vapor deposited onto the quartz crystal so that ultraviolet (UV)/visible/near-infrared (NIR) radiation can be transmitted. The electrochemical cell is shown in Figure 13. The counter and reference electrodes are positioned so as to not interfere with the transmitted radiation.

Advantages of the SEQCM include the simultaneous measurement of spectral change and mass change during electrochemical synthesis of a material or observations of the spectral changes and mass changes during oxidation-reduction of an electroactive polymer.

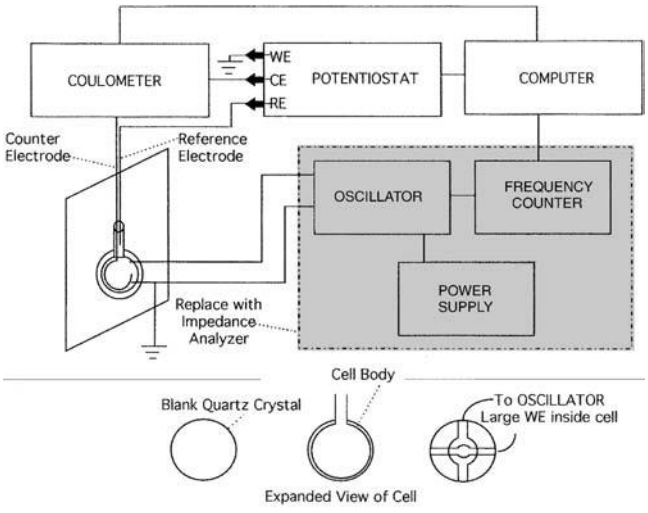


Figure 13. A spectroelectrochemical/quartz crystal microbalance. The electrochemical cell is placed directly into the sample compartment of a spectrometer. The working electrode is the quartz crystal with the active side (larger electrode) facing toward the solution. A blank (no electrodes) quartz crystal acts as the other side of the cell. The Pt counter and reference electrodes are configured to not interfere with the transmitted radiation.

Quartz Crystals

Quartz crystals are commercially available with a range of thicknesses (and hence resonant frequencies) from a variety of sources (see Internet Resources). The most commonly used quartz crystals are 5 and 10 MHz. The 5-MHz crystal is usually 1 in. in diameter and 0.013 in. thick. This is a convenient size to handle. Higher frequency crystals are thinner and more prone to breakage.

The crystals can be purchased as either plano-plano or plano-convex quartz crystals. Both surfaces are parallel in plano-plano crystals. The plano-convex crystals have a radius of curvature on the convex side that can range from 10 to 50 cm for a 5-MHz crystal. The use of both types of crystals have been reported (Buttry, 1991). Plano-convex crystals confine the region of displacement to the area defined by the smaller of the two electrodes. The displacement extends somewhat beyond the electrode for plano-plano crystals. A study of the mass sensitivity distribution of plano-plano and plano-convex AT-cut quartz crystals has been performed (Ward and Delawski, 1991; Hillier and Ward, 1992). It was found that the plano-plano crystals have a Gaussian-like sensitivity distribution, with the greatest sensitivity at the center of the electrode. Sensitivity to mass changes on the electrode tabs was found as well. The plano-convex quartz crystals display a considerably different sensitivity. The maximum at the center is much larger. The Gaussian describing the behavior has a greater amplitude and drops to a low value at the electrode edges.

In spite of these observations, bulk calibration results indicate that the overall sensitivity as defined by C_f is the same for both types of crystals. Calibration values can be obtained by electrodeposition of copper. By noting the amount of charge passed and the area of the electrode, the mass sensitivity of the particular QCM can be determined. The calibrated values are usually somewhat smaller than the value predicted by the Sauerbrey equation (0.057 Hz cm^2/ng). This may be attributed to sensitivity beyond the electrodes.

The viscosity of the liquid above the crystal has been found to influence the sensitivity constant C_f (Hillier and Ward, 1992). Therefore, care should be taken to calibrate the QCM under the conditions that will be used during an investigation. Other studies of the calibration of the EQCM have been reported (Gabrielli et al., 1991; Oltra and Efimov, 1994).

METHOD AUTOMATION

Commercial Instrumentation

Commercial EQCMs record frequency and charge changes with applied voltage. Using commercial EQCM instruments, these parameters are usually recorded by interfacing the equipment to a computer. A board may be supplied by the EQCM manufacturer or the user may need to take

the output (RS-232 or GP-IB) and interface it to a computer. Ideally, real-time data can be displayed and the course of the reaction can be monitored. Commercial software packages may be designed to provide ease of data manipulation, such as conversion from frequency to mass change and overlays of mass and charge relationships vs. time.

Home-Built EQCMs

For home-built systems, a computer interface must also be used to record the frequency (from a frequency counter or an impedance analyzer), charge (from a coulometer or the voltage drop across a resistor), and voltage (applied potential). A variety of commercial interfacing software packages are available. For example, LabView by National Instruments is a well-reviewed commercial package. Much of the instrumentation used is IEEE (GP-IB) compliant. The software directs the instrumentation to collect the measurement of frequency and/or charge as a function of time and the result is displayed on the computer screen. A real-time display is useful in monitoring the results and modifying the experiment when necessary. The amount of programming required varies from one software package to another. Advantages of home-built systems are the ability to modify the QCM, for example to simultaneously record mass change and resistance change vs. time during the doping of a conducting polymer (Hall and Arbuckle, 1996).

DATA ANALYSIS AND INITIAL INTERPRETATION

Electrochemical Reactions

Data analysis generally consists of evaluating the change in frequency and other variables such as charge and applied potential vs. time. The frequency results can readily be converted to a change in mass using the appropriate calibration factor, as discussed under Practical Aspects of the Method. The amount of mass change observed should be compared with the amount of charge passed. Correlations can be made with the theoretical mass change expected from the coulombs passed. Various hypotheses may need to be tested to explain the observed mass change. For example, in the electrochemical oxidation-reduction of a conducting polymer, the amount of mass change may be expected to correspond to the mass of the anion in the electrolyte. However, the anion will be solvated. In addition, a simple "one anion in, one anion out" scenario may not be correct. Other considerations are movement of both cation AND anion with solvent in and out of the polymer film.

Series and Parallel Frequency

Both the series and parallel frequencies f_s and f_p can be measured using an impedance analyzer. A plot of f_s and f_p vs. time for the doping of a conducting polymer is shown in Figure 14. Both frequencies decrease with doping, indicating a mass increase due to the dopant. As expected from the plot of impedance vs. frequency in Figure 8, f_p is higher. The variation in parallel frequency can be utilized to study conformational changes of deposits (Yang et al., 1993). The

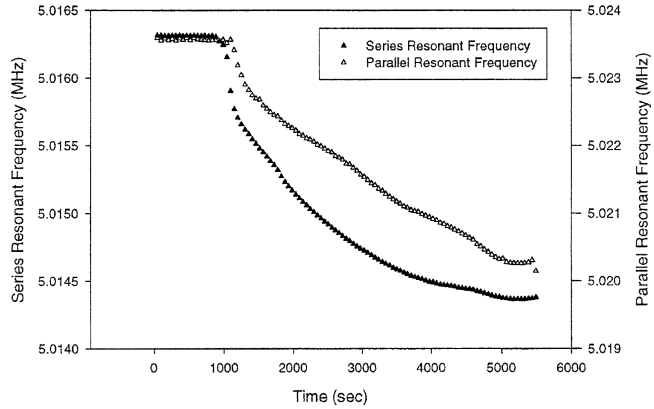


Figure 14. Series and parallel frequencies (in megahertz) vs. time for the doping of a conducting polymer, poly(*p*-phenylene vinylene). The series frequency provides the mas increase due to the dopant. The parallel frequency indicates changes in conformation.

change in parallel resonant frequency can also be correlated with changes in conductivity of the medium (Rodahl et al., 1996), which is in agreement with the result shown in Figure 14.

SAMPLE PREPARATION

Sample preparation begins with the preparation of the quartz crystals. If commercial crystals with deposited electrodes are purchased, they should still be carefully cleaned with ethanol and dried before use.

If blank quartz crystals are purchased from a supplier, electrodes must be vapor deposited onto the quartz crystal. Before vapor deposition the crystals must be thoroughly cleaned. Sonification in a dilute soap solution for 60 min has been found to be sufficient. Gold is typically used as the electrode. However, for the gold to adhere to the quartz, an underlayer of chromium is applied. Typically 100 Å of chromium is deposited followed by 1000 Å of gold. This is done on both sides of the quartz crystal in a keyhole pattern, as shown in Figure 2A.

Initially a mask to hold the quartz crystals and to provide deposition of a particular electrode pattern must be machined for vapor deposition.

After vapor deposition of the electrodes, the quartz crystals should be kept clean and gently recleaned immediately prior to an experiment. It has been found that better results are obtained with freshly deposited electrodes since the gold slightly oxidizes with time.

Depending on the experiment, either the films are electrochemically prepared during the QCM analysis or a polymer film is cast onto the quartz crystal prior to performing QCM studies. For example, a conducting polymer can be spin cast onto the quartz crystal with electrodes. The casting solvent is used to carefully remove the polymer from the edges of the quartz crystal where the electrical connections need to be attached. The electrical connections can be made (1) with flat alligator clips, (2) with spring-loaded clips, or (3) by attaching small wires with conducting solder and connecting to the wire rather than the crystal.

Since the crystals are fragile, care must be taken when attaching connections.

For an electrochemical experiment, a freshly cleaned crystal is mounted, as shown in Figure 12, and the potential is applied immediately after the measured resonance frequency stabilizes. In the case of some chemical systems, a thin film may begin to form (chemically) if the electrode is left in contact with the solution for an extended time. Therefore, the potential is applied to initiate the desired electrochemical reaction as soon as a stable resonance frequency is obtained.

SPECIMEN MODIFICATION

One of the advantages of EQCM is that the electrochemical synthesis of a material can be monitored *in situ*. For example, after a stable frequency is obtained for the quartz crystal in contact with the solution, a constant or varying potential can be applied to the working electrode (top of the crystal in Fig. 12), which will initiate an electrochemical reaction. The charge that flows and the frequency decrease (mass increase) is recorded.

For example, the process of the electrochemical synthesis of polyazulene monitored by EQCM is shown in Figure 15. The applied potential is varied from −0.05 to 1.0 V at a rate of 50 mV/sec. The overall frequency decreases with time, indicating mass deposition onto the electrode on the quartz crystal (Fig. 15A). The corresponding charge-vs.-time plot in Figure 15B parallels this mass increase. A sawtoothlike wave is superimposed onto the frequency and charge plots. This is due to the variation in potential that causes oxidation and reduction of the growing polymer. The size of the sawtooth increases as more material is deposited onto the electrode and can therefore be oxidized/reduced by each subsequent cycle.

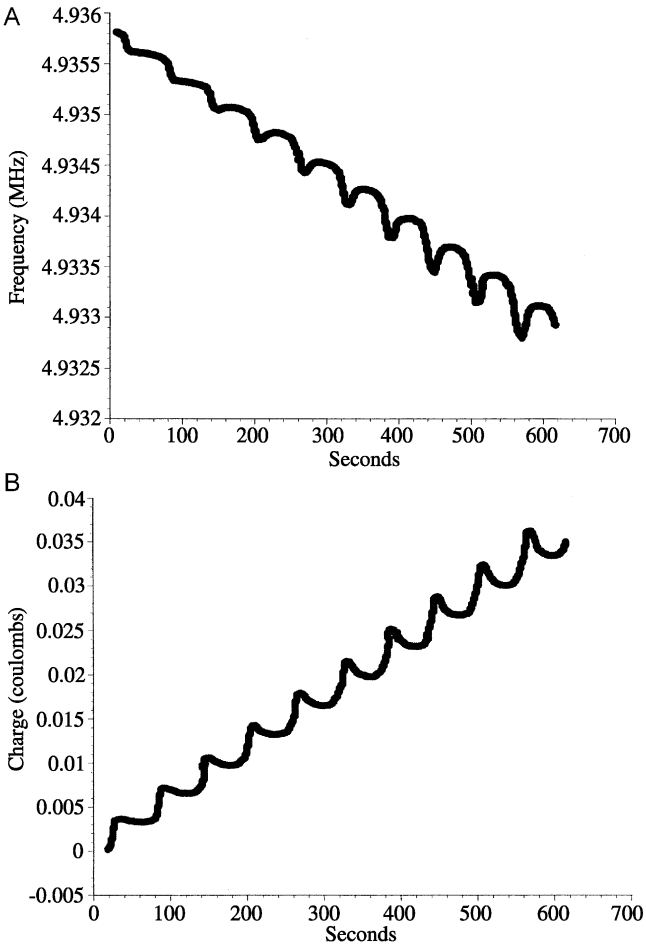


Figure 15. (**A**) Frequency (in megahertz) plot vs. time for the electrochemical synthesis of polyazulene from a solution of 0.005 M azulene in 0.1 M tetraethylammonium tetrafluoroborate in acetonitrile. The potential is cycled between −0.05 and 1.0 V vs. SCE at a rate of 50 mV/sec. (**B**) Charge (in coulombs) plot vs. time for the electrochemical synthesis of polyazulene.

PROBLEMS

Care must be taken to ensure that mass changes observed are not misinterpreted. A calibration of the particular EQCM should always be made. If any changes in the system (e.g., type of crystal, thickness of electrode, and shape of electrode) are made, a new calibration must be performed. Ideally the deposit used for the calibration and the viscosity of the medium should approximate that of the experiment as much as possible.

In addition, as mentioned under Principles of the Method, phenomena other than an actual mass change can give rise to an observed frequency change. This is especially a concern for polymer materials due to the added viscoelastic effects possible in these materials. Therefore, the polymer reaction should be studied at a variety of thicknesses (all <2% of that of the crystal) and the linearity of the results noted. Impedance analysis is especially useful as well. The quality factor Q of the blank crystal, polymer deposit, and final product should be compared. Significant changes in Q during an experiment indicate that the Sauerbrey relationship no longer applies.

When using the impedance analyzer, f_s is defined as the frequency at which $\theta = 0$. Occasionally the impedance plot does not intersect phase equal to zero. In some cases it may be necessary to increase the oscillator level applied by the impedance analyzer to the quartz resonator in order to increase the amplitude of the impedance plot so that the series frequency f_s can be determined.

ACKNOWLEDGMENTS

The author would like to thank Michael T. Kelly and Hiren Shah for assistance with figures and John Gagliardi for careful reading of the manuscript. Acknowledgement is made for an Alfred P. Sloan Research Fellowship.

LITERATURE CITED

Arbuckle, G. A. and Travis, D. A. 1994. Electrochemical polyazulene polymerization studied with the spectroelectrochemical/quartz crystal microbalance. *Polym. Mater. Sci. Eng.* 71:224–225.

Benje, M., Eierman, M., Pittermann, U., and Weil, K.G. 1986. An improved quartz crystal microbalance. Applications to the electrocrystallization and dissolution of nickel. *Ber. Bunsen-Ges. Phys. Chem.* 90:435–439.

Bottom, V. E. 1982. Introduction to Quartz Crystal Unit Design. Van Nostrand Reinhold, New York.

Bruckenstein, S. and Shay, M. 1985. Experimental aspects of use of the quartz crystal microbalance in solution. *Electrochim. Acta* 30(10):1295–1300.

Bruckenstein, S., Wilde, P. C., Shay, M., Hillman, R. A., and Loveday, D.C. 1989. Observation of kinetic effects during interfacial transfer at redox polymer films using the quartz crystal microbalance. *J. Electroanal. Chem. Interfacial Electrochem.* 258: 457–462.

Bruckenstein, S., Michalski, M., Fensore, A., and Li, Z. 1994. Dual quartz crystal microbalance oscillator circuit. Minimizing effects due to liquid viscosity, density, and temperature. *Anal. Chem.* 66:1847–1852.

Buttry, D. A. 1991. Applications of the quartz crystal microbalance to electrochemistry. *In* Electroanalytical Chemistry (A. J. Bard, ed.) pp. 1–86. Marcel Dekker, New York.

Buttry, D. A. and Ward, M. D. 1992. Measurement of interfacial processes at electrode surfaces with the electrochemical quartz crystal microbalance. *Chem. Rev.* 92:1355–1379.

Cady, W. G. 1964. Piezoelectricity. Dover, New York.

Cheek, G. T. and O'Grady, W. E. 1990. Measurement of hydrogen uptake by palladium using a quartz crystal microbalance. *J. Electroanal. Chem. Interfacial Electrochem.* 277:341–346.

Cordoba-Torresi, S. I., Gabrielli, C., Hugot-Le Goff, A., and Torresi, R. 1991. Electrochromic behavior of nickel oxide electrodes I. Identification of the colored state using quartz crystal microbalance. *J. Electrochem. Soc.* 138(6):1548–1553.

Curie, P. and Curie, J. 1880. Crystal physics: Development of polar electricity in hemihedral crystals with inclined faces. *C. R. Acad. Sci. Ser. 2* 91:294.

Daifuku, H., Kawagoe, T., Yamamoto, N., Ohsaka, T., and Oyama, N. 1989. A study of the redox reaction mechanisms of polyaniline using a quartz crystal microbalance. *J. Electroanal. Chem. Interfacial Electrochem.* 274:313–318.

Deakin, M. R. and Byrd, H. 1989. Prussian blue coated quartz crystal microbalance as a detector for electroinactive cations in aqueous solution. *Anal. Chem.* 61:290–295.

Deakin, M. R. and Melroy, O. 1988. Underpotential metal deposition on gold, monitored *in situ* with a quartz crystal microbalance. *J. Electroanal. Chem. Interfacial Electrochem.* 239: 321–331.

Deakin, M. R. and Melroy, O. R. 1989. Monitoring the growth of an oxide film on aluminum *in situ* with the quartz crystal microbalance. *J. Electrochem. Soc.* 136(2):349–351.

De Long, H. C., Donohue, J. J., and Buttry, D. A. 1991. Ionic interactions in electroactive self-assembled monolayers of ferrocene species. *Langmuir* 7:2196–2202.

Demoustier-Champagne, S., Reynolds, J. R., and Pomerantz, M. 1995. Charge- and ion-transport properties of polypyrrole/poly(styrene-sulfonate): Poly(3-octylthiophene) bilayers. *Chem. Mater.* 7:277–283.

Dunham, G. C., Benson, N. H., Petelenz, D., and Janata, J. 1995. Dual quartz crystal microbalance. *Anal. Chem.* 67:267–272.

Evans, C. D. and Chambers, J. Q. 1994. Electrochemical quartz crystal microbalance study of tetracyanoquinodimethane conducting salt electrodes. *Chem. Mater.* 6:454–460.

Feldman, B. J. and Melroy, O. R. 1987. Ion flux during electrochemical charging of Prussian blue films. *J. Electroanal. Chem. Interfacial Electrochem.* 234:213–227.

Freund, M. S., Brajter-Toth, A., and Ward, M. D. 1990. Electrochemical and quartz crystal microbalance evidence for mediation and direct electrochemical reactions of small molecules at tetrathiafulvalene-tetracyanoquinodimethane (TTF-TCNQ) electrodes. *J. Electroanal. Chem. Interfacial Electrochem.* 289: 127–141.

Gabrielli, C., Keddam, M., and Torresi, R. 1991. Calibration of the electrochemical quartz crystal microbalance. *J. Electrochem. Soc.* 138:2657–2660.

Hall, J. W. and Arbuckle, G. A. 1996. Comparing and contrasting the effects of iodine doping on different types of polyacetylene films. *Macromolecules* 29(2):546–552.

Hepel, M., Kanige, K., and Bruckenstein, S. 1990. Expulsion of borate ions from the silver/solution interfacial region during underpotenial deposition discharge of Pb(II) in borate buffers. *Langmuir* 6:1063–1067.

Hillier, A. C. and Ward, M. D. 1992. Scanning electrochemical mass sensitivity mapping of the quartz crystal microbalance in liquid media. *Anal. Chem.* 64(21):2539–2554.

Hillman, A. R. and Glidle, A. 1994. Electroactive bilayers employing conducting polymers: Part 5, Electrochemical quartz crystal microbalance studies of the overall switching process. *J. Electroanal. Chem. Interfacial Electrochem.* 379:365–372.

Honda, M. 1989. The Impedance Measurement Handbook. Yokogawa-Hewlett-Packard LTD.

Katz, A. and Ward, M. D. 1996. Probing solvent dynamics in concentrated polymer films with a high-frequency shear mode quartz resonator. *J. Appl. Phys.* 80(7):4153–4163.

Kaufman, J. H., Kanazawa, K. K., and Street, G. B. 1984. Gravimetric electrochemical voltage spectroscopy: *In situ* mass measurements during electrochemical doping of the conducting polymer polypyrrole. *Phys. Rev. Lett.* 53(26):2461–2464.

Kipling, A. L. and Thompson, M. 1990. Network analysis method applied to liquid-phase acoustic wave sensors. *Anal. Chem.* 62(14):1514–1519.

Lasky, S. J. and Buttry, D. A. 1988. Mass measurements using isotopically labeled solvents reveal the extent of solvent transport during redox in thin films on electrodes. *J. Am. Chem. Soc.* 110:6258–6260.

Lin, Z. and Ward, M. D. 1996. Determination of contact angles and surface tensions with the quartz crystal microbalance. *Anal. Chem.* 68(8):1285–1291.

Lu, C. and Czanderna, A. W. 1984. Applications of Piezoelectric Quartz Crystal Microbalances. Elsevier Science Publishing, New York.

Martin, S. J., Granstaff, V. E., and Frye, G. C. 1991. Characterization of a quartz crystal microbalance with simultaneous mass and liquid loading. *Anal. Chem.* 63(20):2272–2278.

Naoi, K., Oura, Y., Maeda, M., and Nakamura, S. 1995. Electrochemistry of surfactant-doped polypyrrole film: Formation of columnar structure by electropolymerization. *J. Electrochem. Soc.* 142(2):417–422.

Noel, M. A. M. and Topart, P. A. 1994. High-frequency impedance analysis of quartz crystal microbalance 1. General considerations. *Anal. Chem.* 66(4):484–491.

Oltra, R. and Efimov, I. O. 1994. Calibration of an electrochemical quartz crystal microbalance during localized corrosion. *J. Electrochem. Soc.* 141(7):1838–1842.

Orata, D. and Buttry, D. A. 1987. Determination of ion populations and solvent content as functions of redox state and pH in polyaniline. *J. Am. Chem. Soc.* 109:3574–3581.

Orata, D. and Buttry, D. A. 1988. Virtues of composite structures in electrode modification. Preparation and properties of poly (aniline)/Nafion composite films. *J. Electroanal. Chem. Interfacial Electrochem.* 257:71–82.

Rishpon, J., Redondo, A., Derouin, C., and Gotttesfeld, S. 1990. Simultaneous ellipsometric and microgravimetric measurements during the electrochemical growth of polyaniline. *J. Electroanal. Chem. Interfacial Electrochem.* 294:73–85.

Rodahl, M., Hook, F., and Kasemo, B. 1996. QCM operation in liquids: An explanation of measured variations in frequency and *Q* factor with liquid conductivity. *Anal. Chem.* 68:2219–2227.

Sauerbrey, G. 1959. The use of a quartz crystal oscillator for weighing thin layers and for microweighing applications. *Z. Phys.* 155:206.

Schumacher, R. 1990. The quartz crystal microbalance: A novel approach to the *in situ* investigation of interfacial phenomena at the solid/liquid junction. *Angew. Chem.* 29: 329–343.

Servagent, S. and Vieil, E. 1990. *In situ* quartz microbalance study of the electrosynthesis of poly(3-methylthiophene). *J. Electroanal. Chem. Interfacial Electrochem.* 280:227–232.

Tiean, Z., Liehua, N., and Shouzhuo, Y. 1990. On equivalent circuits of piezoelectric quartz crystals in a liquid and liquid properties Part 1. Theoretical derivation of the equivalent circuit and effects of density and viscosity of liquids. *J. Electroanal. Chem. Interfacial Electrochem.* 293:1–18.

Ward, M. D. 1989. Probing electrocrystallization of charge transfer salts with the quartz crystal microbalance. *J. Electroanal. Chem. Interfacial Electrochem.* 273:79–105.

Ward, M. D. 1995. Principles and applications of the electrochemical quartz crystal microbalance. *In* Physical Electrochemistry. Principles, Methods, and Applications (I. Rubenstein, ed.) pp. 293–338. Marcel Dekker, New York.

Ward, M. D. and Buttry, D. A. 1990. *In situ* interfacial mass detection with piezoelectric transducers. *Science* (Washington, D.C.) 249:1000–1007.

Ward, M. D. and Delawski, E. J. 1991. Radial mass sensitivity of the quartz crystal microbalance in liquid media. *Anal. Chem.* 63(9):886–890.

Yang, M., Chung, F. L., and Thompson, M. 1993. Acoustic network analysis as a novel technique for studying protein adsorption and denaturation at surfaces. *Anal. Chem.* 65(24):3713–3716.

KEY REFERENCES

Buttry, 1991. See above.

Book chapter summarizing applications of the QCM in electrochemistry.

Buttry and Ward, 1992. See above.

Provides an overall analysis of EQCM with emphasis on impedance analysis and data interpretation.

Ward, 1995. See above.

Book chapter detailing various recent applications of the EQCM.

INTERNET RESOURCES

http://www.princeton.edu/~nliu/piezo.html

A very good general site listing companies in frequency control, timing, and piezoelectric devices. This site also contains links to major equipment manufactuers.

Source of Crystals (Valpey-Fisher, Oak Frequency Control, International Crystal Manufacturing, Maxtek Inc.)

APPENDIX: INSTRUMENTS

See Internet site for equipment manufacturers.

Quartz crystal microbalances were reviewed recently (*Anal. Chem.* 1996, 625A)

Manufacturers: EG&G Princeton Applied Research, Elchema, Maxtek, QCM Research, Universal Sensors

Home-built systems require:Potentiostat (EG&G Princeton Applied Research, Pine), impedance analyzer (Agilent Technologies, Schlumberger Technologies, United Kingdom), and coulometer (Electrosynthesis, EG&G Princeton Applied Research).

In place of the impedance analyzer a frequency counter (Agilent or Fluke) can be used, the oscillator circuit must be powered by a power supply (Agilent or Keithley).

GEORGIA A. ARBUCKLE-KEIL
Rutgers University
Camden, New Jersey

This page intentionally left blank

OPTICAL IMAGING AND SPECTROSCOPY

INTRODUCTION

All materials absorb and emit electromagnetic radiation. The characteristics of a material frequently manifest themselves in the way it interacts with radiation. The fundamental basis for the interaction varies with the wavelength of the radiation, and the information gleaned from the material under study will likewise vary. Materials interact with radiation in some fashion across the entire electromagnetic spectrum, which can be thought of as spanning the frequency realm from near DC oscillations found in circuits (<1 Hz) through the extremely high frequency oscillations typical of gamma radiation ($>10^{19}$ Hz). Shown in the Table 1 is a representation of the electromagnetic spectrum, giving the correspondence of wavelength, frequency, and common terminology applied to the sub-bands.

Because human vision depends on the interaction of light with the rods and cones of the retina in the range of photon wavelengths from about 0.7 to 0.4 micrometers, this "visible" spectrum is particularly important to us. Materials are directly observed using instrumentation that magnifies and records visible light; this is the subject of the optical microscopy units in this chapter. As an adjunct to the value of direct observation of the structure of materials, often, a desirable material property arises from its interaction with visible light. The visible and ultraviolet regions are the basis for the units in this chapter, and, as seen in Table 1, they occupy the midpoint on the energy scale. Yet, it is valuable to note the interactions of electromagnetic radiation from the full-spectrum perspective, both for understanding the role that optical methods play and for realizing their relationship to those presented in other chapters throughout *Characterization of Materials*. Thus, we briefly consider here how all radiation—from DC to x ray—interacts with materials.

In the most general sense, radiation can be considered to interact with materials either by absorption, reflection, or scattering; this is the basis for Kirchoff's law, which is discussed in COMMON CONCEPTS IN MATERIALS CHARACTERIZATION.

As indicated in Table 1, the nature of the interaction between photons and matter governs the observed effects. At very low frequencies, we think of the photons more as a time-dependent electric field than as a packet of energy. Interactions with matter in this regime are exploited in the electrical and electronic characterization of transients and charge carriers in materials. This is the topic of many of the units in ELECTRICAL AND ELECTRONIC MEASUREMENTS. At higher frequencies, the nuclear (RF) and electronic (microwave) Zeeman effects become observable by virtue of transitions between states in the presence of the internal quadrupole field of the nuclei (nuclear quadrupole resonance) and/or in the presence of an externally applied field (nuclear magnetic resonance and electron spin resonance, see RESONANCE METHODS).

The microwave region is also an important regime for materials characterization, as the measurement of dielectric properties such as permittivity and dielectric loss tangent, surface resistance, and skin depth become feasible using radiation in this region. Some of these measurements are discussed in detail in ELECTRICAL AND ELECTRONIC MEASUREMENTS. As we continue up in frequency, we enter the domain of the lowest-frequency molecular vibrations in materials, lattice vibrations, or phonons, which couple to the oscillating electric field of the probe radiation. Phonons can also be probed using neutron scattering, a complementary method described in detail in NEUTRON TECHNIQUES, and by inelastic scattering of radiation (Brillouin spectroscopy).

The next jump in frequency enters the infrared radiation region, in which the electromagnetic field is coupled with the intermolecular motions of chemical bonds in the material structure. Also known as vibrational spectroscopy, the study of the spectra of materials in this regime provides information on chemical composition, as several characteristic frequency bands can be directly related to the atom and bond types involved in the vibrational motion. An important detail deserves mention here, as a shift occurs at this point in convention and nomenclature. Researchers in the lower frequency fields (DC through far infrared) typically report their measurements in frequency units. In infrared work the most common unit employed is the wavenumber, or reciprocal centimeter, although there is an increasing use of wavelength (reported in microns) in some disciplines. Analogous to Brillouin spectroscopy is Raman spectroscopy, which measures vibrational transitions by observing inelastically scattered radiation. Raman spectroscopy is complementary to infrared spectroscopy by virtue of the selection rules governing transitions between vibrational states—whereas a change in overall dipole moment in the molecule gives rise to infrared absorption, a change in overall polarizability of the molecule gives rise to the vibrational structure observed in Raman spectroscopy.

At still higher frequencies, we access transitions between electronic states within molecules and atoms, and again the conventions change. Here, convention favors units of wavelength. The underlying phenomena studied range from overtones of the molecular vibrations (near-infrared) to vibrational-electronic (vibronic) transitions, to charge transfer transitions and pure electronic spectra (visible through ultraviolet spectroscopy). Electronic spectroscopy, like vibrational spectroscopy, also reflects the chemical makeup of the material. The intensity of electronic transitions, a function of the type of molecular orbitals

Table 1. The Electromagnetic Spectrum: A Materials Perspective[a]

Spectral Region	Frequency Range (Hz)	Wavelength Range (m)	Unit Conventions	Interactions with Matter	Representative Materials Properties Accessible
Extremely low frequency	$<3 \times 10^2$	$>1 \times 10^6$	Hz	Electron transport	Conductivity, free carrier concentration
Radiofrequency					
Very low frequency (VLF)	3×10^2 to 3×10^4	1×10^6 to 1×10^4	Hz	Nuclear magnetic resonance; nuclear quadrupole resonance	Microscopic structure; chemical composition
Low frequency (LF)	3×10^4 to 3×10^5	1×10^4 to 1×10^3	kHz		
Medium frequency (MF)	3×10^5 to 3×10^6	1×10^3 to 1×10^2	kHz		
High frequency (HF)	3×10^6 to 3×10^7	1×10^2 to 1×10^1	kHz		
Very high frequency (VHF)	3×10^7 to 3×10^8	1×10^1 to 1×10^0	MHz		
Ultra high frequency	3×10^8 to 3×10^9	1×10^0 to 1×10^{-1}	MHz		
Microwave					
Super high frequency (SHF)	3×10^9 to 3×10^{10}	1×10^{-1} to 1×10^{-2}	GHz	Electron paramagnetic resonance; molecular rotations	Microscopic structure; surface conductivity
Extremely high frequency (EHF)	3×10^{10} to 3×10^{11}	1×10^{-2} to 1×10^{-3}	GHz		
Submillimeter	3×10^{11} to 3×10^{12}	1×10^{-3} to 1×10^{-4}	THz		
Infrared					
Far infrared (FIR)	$\sim 3 \times 10^{12}$ to 1.2×10^{13}	$\sim 1 \times 10^{-4}$ to 2.5×10^{-5}	THz, cm^{-1}, kiloKayser (kK, = 1000 cm^{-1}), micron (μm)	Librations, molecular vibrations, vibrational overtones, vibronic transitions	Microscopic and macroscopic structure, phases, chemical composition
Intermediate infrared	1.2×10^{13} to 3×10^{14}	2.5×10^{-5} to 1×10^{-6}			
Near infrared (NIR)	3×10^{14} to 3.8×10^{14}	1×10^{-6} to 7.9×10^{-7}			
Visible	3.8×10^{14} to 7.5×10^{14}	7.9×10^{-7} to 4×10^{-7}	kK, nm, μm	Valence shell, π bonding electronic transitions	Chemical composition and concentration
Ultraviolet					
Near ultraviolet	7.5×10^{14} to 1.5×10^{15}	4×10^{-7} to 2×10^{-7}	nm	Inner shell, σ bonding electronic transitions	Chemical composition, bond types, and bond strength
Vacuum ultraviolet	1.5×10^{14} to 3×10^{16}	2×10^{-7} to 1×10^{-8}	eV		
X ray	3×10^{16} to 3×10^{19}	1×10^{-8} to 1×10^{-11}	keV	Core shell electronic transitions, nuclear reactions	Elemental analysis, nature of chemical bonding
γray	$>\sim 3 \times 10^{19}$	$<\sim 1 \times 810^{-11}$	MeV	Nuclear transitions	Elemental analysis, profiling

[a] Note that values preceded by represent "soft definitions," i.e., frequencies that bridge disciplines and may be defined differently depending upon the field of study. In fact, many frequency definitions, especially at the boundaries between spectral regions, are "soft."

in the material, govern the appearance of the observed spectra. Absorption of radiation in this regime is often accompanied by electronic relaxation processes, and, depending on the nature of the electronic states involved, can cause fluorescence and/or phosphorescence, providing another window on the chemical makeup and structure of the material.

Visible and ultraviolet spectroscopies are frequently coupled with other measurement techniques to provide additional insight into the makeup and nature of materials. Changes in materials induced by electrochemical reactions, for instance, can change the observed spectra, or even lead directly to chemiluminescence (ELECTROCHEMICAL TECHNIQUES). Of particular interest to semiconductor studies is the luminescence (photoluminescence, or PL) that follows the promotion of electrons into the valence band of the material by incident photons (ELECTRICAL AND ELECTRONIC MEASUREMENTS). Electrons may even be ejected completely from the material, providing a measure of the work function (the photoelectric effect).

At the high end of the ultraviolet range (vacuum UV) and beyond, yet another shift in convention occurs. This regime traditionally employs the unit of electron volts (eV). Phenomenology of interactions ranges from inner-shell electronic transitions, as observed in ultraviolet and x-ray photoelectron spectroscopy, to nuclear rearrangement, as observed in gamma ray emission in nuclear reactions (ION-BEAM TECHNIQUES). Emitted electrons themselves carry spectroscopic information as kinetic energy; this is the basis for ESCA (electron spectroscopy for chemical analysis), energy dispersive spectroscopy, and Auger electron spectroscopy (ELECTRON TECHNIQUES).

The Mössbauer effect relies on the recoil-free emission of nuclear transition gamma radiation from the nuclei in a solid (the source) and its resonance absorption by identical nuclei in another solid (the absorber). Because recoilless resonance absorption yields an extremely sharp resonance, Mössbauer spectroscopy is a very sensitive probe of the slightest energy level shifts and therefore of the local environment of the emitting and absorbing atoms.

The units collected in this chapter present methods that are useful in the optical characterization of materials. The preceding discussion illustrates how spectroscopy permeates the entire *Characterization of Materials* volume—called upon where appropriate to specific materials studies. Spectroscopy, as a general category of techniques, will certainly remain central to the characterization of materials as materials science continues its advance.

ALAN C. SAMUELS

OPTICAL MICROSCOPY

INTRODUCTION

Among the numerous investigative techniques used to study materials, optical microscopy, with its several diverse variations, is important to the researcher and/or materials engineer for obtaining information concerning the structural state of a material. In the field of metallurgy, light optical metallography is the most widely used investigative tool, and to a lesser degree its methods may be extended to the investigation of nonmetallic materials. The state of microstructure of a metal or alloy, or other engineering material, whether ceramic, polymer or composite, is related directly to its physical, chemical, and mechanical properties as they are influenced by processing and/or service environment. To obtain qualitative, and in many cases quantitative information about the microstructural state of a material, it is often necessary to perform a sequence of specimen-preparation operations, which will reveal the microstructure for observation. Selecting a specimen, cutting it from the bulk material, grinding and polishing an artifact-free surface, then etching the surface to reveal the microstructure, is the usual sequence of specimen-preparation operations (see SAMPLE PREPARATION FOR METALLOGRAPHY). This scheme is appropriate for metals and alloys as well as some ceramic, polymer, and composite materials. Specimens thus prepared are then viewed with a reflected-light microscope (see REFLECTED-LIGHT OPTICAL MICROSCOPY). The microstructure of transparent or translucent ceramic and polymer specimens may be observed using a transmitted-light microscope when appropriately thinned sections are prepared. In all cases, observation of the microstructure is limited by the relatively shallow depth of focus of optical microscopes, the depth of focus being inversely related to the magnification.

Optical microscopy as related to the study of engineering materials can be divided into the categories of reflected-light microscopy for opaque specimens and transmitted-light microscopy for transparent specimens. On the whole, reflected light microscopy is of more utility to materials scientists and engineers because the vast bulk of engineering materials are opaque. Within each of the categories, there are a number of illumination variations (bright field, dark field, polarized light, sensitive tint, phase contrast, and differential interference contrast), each providing a different or enhanced observation of microstructural features. It is fortunate that a majority of illumination variations of reflected-light microscopy parallel those of transmitted-light microscopy.

In general, one does not set out to employ optical microscopy to measure a material property or set of properties. Information gained using optical microscopy is more often than not of a qualitative nature, which, to a skilled observer, translates to a set of correspondences among a material's physical, mechanical, chemical properties as influenced by chemical composition, processing and service history. One can use the measurement methods of quantitative microscopy to follow the progress of microstructural evolution during processing or during the course of exposure to a service environment. Such techniques may be used to follow grain growth, the coarsening of precipitates during a process, or the evolution of microstructure that would lead to catastrophic failure under service conditions. Such a case might be the evolution of the microstructure of steam pipes carrying superheated steam at high pressure that will ultimately fail by creep rupture.

In comparison with other instruments used to produce images of microstructure—transmission electron microscope (TRANSMISSION ELECTRON MICROSCOPY and SCANNING TRANSMISSION ELECTRON MICROSCOPY), scanning electron microscope (SCANNING ELECTRON MICROSCOPY), field ion microscope, scanning tunneling microscope (SCANNING TUNNELING MICROSCOPY), atomic force microscope, and electron beam probe—the light-optical microscope is the instrument of choice for a majority of studies. The reasons are as follows: (1) instrument cost is relatively low, (2) less technician training is required, (3) specimen preparation is relatively straightforward, and (4) the amount of information that can be obtained by an experienced microscopist is large. Because the maximum resolution is limited by the wavelength of light, useful magnification is limited to ~1600 times. At these upper limits of resolution and magnification, microstructural features as small as ~325 nm may be observed. Most of the other instruments are better suited to much higher magnifications, but the area of specimen surface is severely limited, so sampling of the microstructure state is better using optical microscopy.

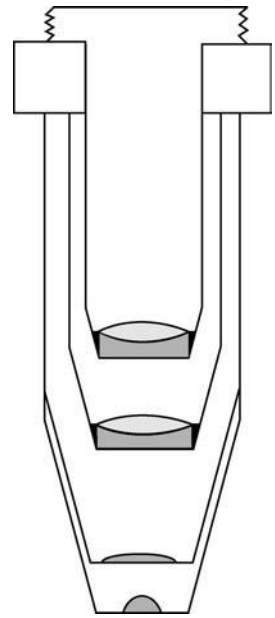


Figure 1. Schematic diagram of a typical objective lens. Note that the plano-convex lens is the "front" lens element of the objective and faces the specimen.

PRACTICAL ASPECTS OF THE METHOD

The Basic Optical Microscope

Early microscopes, though rudimentary compared to today's instruments, provided suitable magnifications, image resolution, and contrast to give early researchers and engineers suitably detailed visual information necessary for the advancement of science and materials application. Modern instruments are indeed sophisticated and incorporate a number of devices designed to enhance image quality, to provide special imaging techniques, and to record images photographically or digitally. Regardless of the sophistication of an instrument, the fundamental components that are of particular importance are a light source, an objective lens, and an ocular or eyepiece.

Significant development of the microscope was not forthcoming until the wave nature of light became known and Huygens proposed a principle that accounted for the propagation of light based on an advancing front of wavelets (Jenkins and White, 1957). Rayleigh was the first to develop a criterion for resolution, which was later refined by Abbe, who is often credited as the "father of optical microscopes" (Gifkins, 1970). Abbe's theory of the microscope was founded upon the concept that every subject consisted of a collection of diffracting points like pinholes, each pinhole behaving like a Huygens light source (Jenkins and White, 1957). It is the function of the objective lens to collect diffracted rays from an illuminated specimen to form an image by constructive and destructive interference from a collection of Huygens sources. Therefore, the function of a lens through which the specimen is viewed is to gather more diffraction information than could be gathered by the unaided eye. This means that an objective lens must be of such geometric configuration as to gather as many orders of diffracted light as possible, to give an image of suitable resolution. The more orders of diffracted rays that can be collected by the objective lens, the greater is the resolution of the lens, and generally the higher the magnification attainable.

By its most basic definition, a lens is a piece of glass or other transparent material bounded by two surfaces, each of a different and/or opposite curvature. From a geometric point of view, light rays passing through a lens either converge or diverge. Any device that can be used to concentrate or disperse light by means of refraction can be called a lens. Objective lenses usually contain a first element (nearest the specimen) that is plano-convex, having a plane front surface and a spherical rear surface (Fig. 1). Additionally, the objective lens may contain a number of elements arranged in several groups, secured in a lens barrel to keep the elements and groups in proper spatial relationship to one another. Likewise the eyepiece or ocular (Fig. 2), usually consists of two lenses, or two or more groups of lenses, each group containing two or more elements.

In a microscope the objective lens, acting like a projector lens, forms a real image of the subject near the top of the microscope tube. Because the image is formed in air, it is often called the aerial image (Restivo, 1992). The eyepiece acts to magnify this real image as a virtual image which is projected by the cornea on the retina of the eye, or it may be projected on the plane of a photographic film or on the surface of some other image recording device. The final magnification observed is the product of the magnification of the objective lens times the magnification of the ocular, except when the microscope contains a zoom magnification device or a magnification multiplier.

To a microscopist, the properties of the objective and the ocular lenses are a first consideration in setting out to use a microscope; therefore, two important properties of objective lenses, magnification and numerical aperture, are usually engraved on the barrel of objective lenses. Like-

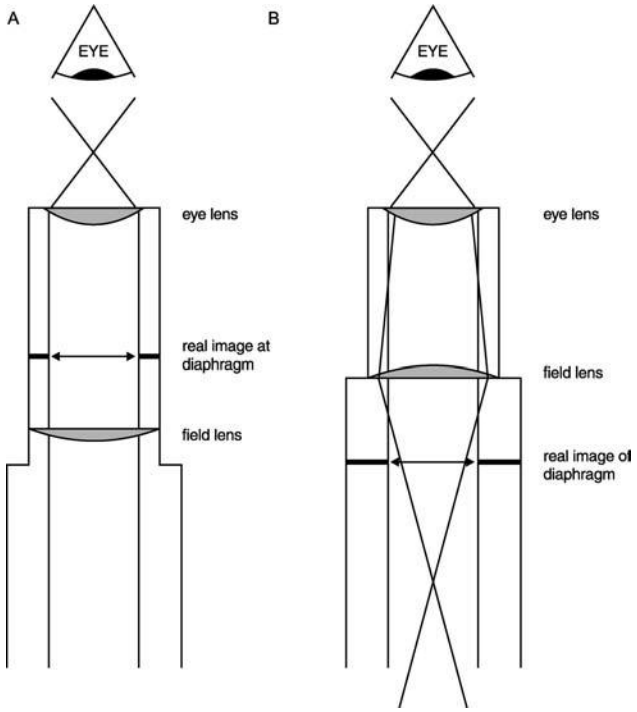


Figure 2. Schematic diagrams of (**A**) a Huygenian ocular and (**B**) a Ramsden ocular. Note the relative orientation of the field lens and the field diaphragms.

wise, magnification is engraved on the ocular. For an objective and ocular to work together properly, their focal lengths must be such that they are separated by a specific distance, so that a virtual image is produced at the eye or a real image is produced at the viewing plane. This distance is the tube length, defined as the distance between the rear focal plane of the objective and the front focal plane of the ocular. It is universally accepted that for any given microscope the objectives and oculars are designed to work together. Objectives and oculars designed for an instrument must be designed for the same tube length. As a practical rule, objective lenses and oculars should not be mixed or transferred to other microscopes. A number of recent instruments are infinity-corrected, meaning that the eyepieces must also be infinity corrected. Infinity-corrected systems allow for variations in the length of the optical path, so that other devices can be inserted for special purposes.

Characteristics of the Microscope

The characteristics of the microscope that are of the greatest importance to the user are the magnification, resolution, and the absence of lens aberrations.

While *magnification* appears to be a simple concept, without concurrent consideration of resolution it becomes relatively unimportant. Magnification alone cannot assist in revealing detail unless the system also possesses resolving power. The microscopist defines resolution as the distance between two closely spaced points in the subject that can be made observable; thus, resolution, expressed as a linear measure, is better the smaller the distance between the points. Attainable resolution is established by the properties of the objective lens, the wavelength of the illumination, and the index of refraction of the medium between the objective lens and the specimen. When the magnification is great enough for the resolution to be perceived by the eye, no amount of additional magnification will cause the image to have more resolution; no more detail will be visible. Regardless of the resolution of photographic films or charge-coupled devices, the resolution of the microscope, primarily that of the objective lens, is controlling in terms of image detail that can be detected.

Calibration of magnification is altogether straightforward. A ruled scale, usually in tenths and hundredths of a millimeter, is placed on the stage of the microscope and brought into sharp focus. It is customary to photograph such a scale at the full series of magnifications of which the instrument is capable. By measuring the photographic images of the magnified scale, one can easily determine the precise magnification.

The degree of resolution is dependent upon the amount of diffraction information that can be gathered by the objective lens. This factor is most conveniently expressed in terms of the numerical aperture (N.A.) of the objective lens. The N.A. of an objective lens is given by $\text{N.A.} = n \sin \mu$, where n is the index of refraction of the medium between the specimen and the lens (for air, $n = 1$), and μ is the half-angle of the cone of light capable of being accepted by the objective (Gifkins, 1970).

Resolution (R) is dependent upon the wavelength of light (λ) illuminating the specimen, and the N.A. of the objective lens: $R = \lambda/(2 \times \text{N.A.})$ (Gifkins, 1970). The larger the numerical aperture and/or the shorter the wavelength of light, the finer the detail that can be resolved. This expression for resolution is used expressly for reflected-light microscopy, where the objective lens carries light to the specimen, then acts again to collect light reflected from the specimen. Resolution of other reflected-light microscopes, particularly those that operate at low magnification and utilize an off-axis external light source, is not a major concern, because the low-power objectives have small N.A. values.

In transmitted-light microscopy a separate condenser lens supplies light through the specimen; thus, the expression becomes $R = \lambda/(\text{N.A.}_{\text{condenser}} + \text{N.A.}_{\text{objective}})$. By way of example, if one illuminates a specimen with green light ($\lambda = 530$ nm) and uses an objective with a 0.65 N.A., then the $R = 407.7$ nm (4.077×10^{-4} mm). R is the distance between two points on the specimen that can be resolved by the microscope. Even though these points may be resolved by the microscope, they will not appear as separate points unless the magnification of the microscope is sufficient to place them within the resolving power of the human eye. Most people can detect an angular separation of from 1 to 2 min of arc. This amounts to a distance between points of $\sim$0.11 mm at a viewing distance of 250 mm. Therefore, by dividing 0.11 mm by 4.077×10^{-4} mm, one obtains a system magnification of $\sim 270\times$. At least this amount of magnification is required in order to make the two points perceptible to the eye. Any magnification greater than this only makes the image more easily perceived by the eye, but the resolution is not further enhanced.

In practice, the largest numerical aperture is ~1.6. Using this value for numerical aperture and again using green light, one finds that the maximum resolution occurs at a magnification of the order of 650×. A magnification of 1000 or 1600 times may be justified by the increased ease of perception. Some microscope manufacturers suggest, by "rule of thumb," that the practical magnification should be 1000 times the numerical aperture. Other sources recommend lower magnifications depending on the type and magnification of the objective.

Because the resolving power of a microscope is improved as the wavelength of the illumination decreases, it is evident that blue light will give slightly better resolution than green light. Wavelengths of <400 nm have been used successfully to enhance resolution. Because conventional optical glasses absorb wavelengths below ~300 nm, the use of special optical elements of quartz, fluorite, calcite, or combinations thereof are required. Thus, all wavelengths down to ~200 nm can be transmitted through the microscope. Mercury vapor lamps produce a large amount of short-wavelength illumination and have proven suitable for ultraviolet (UV) microscopy. Because of the shortness of the wavelength, images can be damaging to the eye if viewed directly; therefore the UV part of the spectrum must be filtered out. Photomicroscopy presents a significant problem in focusing the image. If the image is focused at the film plane with the ultraviolet wavelengths filtered out, the image produced by the shorter wavelengths will not be in focus at the film plane. In 1926, Lucas developed a method for focusing the image on a fluorescent screen. Although resolution can theoretically be enhanced by as much as 60%, a more practical enhancement of 33% can be obtained using a wavelength of 365 nm rather than 540 nm (Kehl, 1949). Ultraviolet microscopy has not become a major technique in materials science because of the great amount of skill required to produce exceptionally flat and artifact-free surfaces. Also, the development of scanning electron micro-scopy (see SCANNING ELECTRON MICROSCOPY) has made it possible to get exceptional resolution, with great depth of field, without great concern for sample preparation.

A general knowledge of lens aberrations is important to the present-day microscopist. While some modern instruments are of very high quality, the kinds of objectives and oculars that are available are numerous, and it must be understood what aberrations are characteristic of each. In addition, various oculars (or projector lenses) are available for correction of aberrations introduced by objective lenses.

Lens aberrations are of two general classes: spherical and chromatic (Gifkins, 1970; Vander Voort, 1984; Kehl, 1949). Spherical aberrations are present even with monochromatic illumination. Its subclasses are spherical aberration itself, coma, curvature of field, astigmatism, and distortion. Chromatic aberration is primarily associated with objectives.

Spherical aberration occurs when a single lens is ground with truly spherical surfaces. Parallel rays come to focus at slightly different points according to their path through the lens—through the center portion of the lens, or through its outer portions. Thus, the image of a point source at the center of the field becomes split into a series of overlapping point source images resulting in an unsharp circle of confusion instead of a single sharp point source image. For spherical positive (convex or convergent) lenses, the spherical aberration is exactly the opposite of that produced by a spherical negative (concave or divergent) lens ground to the same radius of curvature. Correction for spherical aberration can take two forms. The first involves the use of a doublet (two lenses cemented together—one of positive curvature and one of negative curvature). The development of lens formulas for a doublet is rather more complex than it would seem, for should both the positive and negative lenses of the doublet be ground to the exact same radius, then the magnification of the doublet would be 1×. A second method of correcting for spherical aberration is to use a single lens with each face ground to a different radius; such a lens is said to be aspherical.

Coma affects those portions of the image away from the center. Differences in magnification resulting from ray paths that meet the lens at widely differing angles cause the image of a point to appear more like a comma. Lens formulas for the correction of coma provide for the sine of the angle made by each incident ray with the front surface of a lens to have a constant ratio to the sine of each corresponding refracted ray.

Curvature of field arises because the sharpest image of an extended object lies on a curved surface rather than a flat plane. While slight curvature of field is not a serious problem for visual work, because the eye can partially adjust for the visual image and fine focus adjustments can overcome that for which the eye cannot compensate—it becomes a serious problem for photographic work. At high magnifications, the center of the image may be sharp, with sharpness falling off radially away from the center. Attempts to focus the outer portions of the image only cause the central portion of the image to go out of focus. Correction for curvature of field is done by introducing correction in the ocular with the addition of specially designed lens groups. For this reason, it must be understood that oculars are designed to work with a specific objective or set of objectives. One should not attempt to use oculars in another instrument or with objectives for which they are not designed.

Astigmatism is a defect that causes points to be imaged as lines (or discs). Lenses designed to correct for astigmatism are called anastigmats. Correction for astigmatism is done by introducing an additional departure from spherical lens surfaces, together with the matching of refractive indices of the lens components.

Distortion is mainly an aberration found in oculars, where straight lines appear as curved lines in the outer portion of the image. The curvature may be either positive or negative and can often be observed when viewing the image of a stage micrometer. Again it is produced by unequal magnifications from various zones of the lenses in an ocular. Most modern microscopes have oculars that are well corrected for distortion.

Chromatic aberration arises because all wavelengths of light are not brought to focus at the same distance from the optical center of a lens (longitudinal chromatic aberration). If longitudinal chromatic aberration is serious and

broad wavelength spectrum illumination is used, one can observe color fringes around the image of a subject. If a black-and-white photograph is taken of the image, it will appear slightly out of focus. Lateral chromatic aberration results from a lens producing differing magnifications for differing wavelengths of light. A simple solution is to use a narrow-band-pass filter to make the illumination nearly monochromatic; however, this precludes color photography, particularly at higher magnifications. Fortunately, lens formulas exist for the correction of chromatic aberration. Such formulas employ optical glasses of different indices of refraction.

The cost of making lenses corrected to various degrees for chromatic aberrations increases dramatically as the amount of correction increases. From the amount of correction one finds that there are several classes or names for objective lenses: nonachromatic, achromatic, semi-apochromatic, and apochromatic. For work in the materials science and engineering field, non-achromatic lenses are not used.

Achromatic lenses are corrected for spherical aberration such that all rays within a limited wavelength range of the spectrum (yellow-green: 500 to 630 nm) are brought into focus essentially at the same position on the optical path. It is common practice to make chromatic correction for two specific wavelengths within this range. Correction for spherical aberration is also made for one wavelength within the range. Achromatic lenses are useful at low to medium magnifications up to 500× magnification, provided that for photography one uses black-and-white film and also a green filter to provide illumination in the yellow-green portion of the spectrum.

Improved correction for chromatic aberration can be obtained by including fluorite (LiF) lens elements in the objective (Kreidl and Rood, 1965). LiF is used because of its higher index of refraction than the normal borosilicate crown glass elements. The effect of proper optical design using fluorite elements in conjunction with borosilicate elements is to shift the focal plane of the longer wavelengths toward that of the image formed by the shorter wavelengths. Objective lenses thus corrected for chromatic aberrations are called fluorites or semiapochromatics. Such objectives provide good correction over the spectral range of 450 to 650 nm.

Apochromatic lenses are corrected for chromatic aberrations at three specific wavelengths and simultaneously for spherical aberration at two specific wavelengths. Wavelengths in the range of 420 to 720 nm are brought to the same focus. To provide optimal correction for spherical aberration, apochromatic lenses are somewhat undercorrected for color. Thus, they should be used with compensating oculars. Apochromatic lenses are of special importance to high magnifications (>800×) and for color photography.

No discussion of objective lenses is complete without comments on working distance and on depth of focus. Normal working distances (the distance between the front element of an objective lens and the specimen) are rather small even for relatively low-magnification objectives. The working distance may range from ~1 cm for a 5× objective to a fraction of a millimeter for a 100× objective. Objective lenses designed to have larger working distances for special applications are available, but they are resolution-limited due to their smaller numerical apertures. Consequently, long-working-distance objectives are not employed for high-magnification work.

Depth of focus is a property of an objective lens related to its numerical aperture. The depth of focus is the distance normal to the specimen surface that is within acceptable focus when the microscope is precisely focused on the specimen surface. Depth of focus is of concern when the specimen surface is not perfectly flat and perpendicular to the optic axis. In practice, the preparation of specimens to obtain flatness, or the depth of etching, become important factors in the utility of a given objective lens. Based on a criterion devised by Rayleigh, the depth of focus is given by $d \approx \pm n\lambda/(\text{N.A.})^2$, where d is the depth of focus, n is the index of refraction for the medium occupying the space between the objective lens and the specimen, λ is the wavelength of illumination, and N.A. is the numerical aperture (Gifkins, 1970).

Thus, for a 0.65 N.A. objective operating in air with illumination of 530-nm wavelength, the depth of focus is ~2.35 times the wavelength of light. This means that the surface of the specimen must be flat to within ~4.7 times the wavelength of light or, in this case, 2500 nm.

Oculars (eyepieces) used in modern microscope are of four main types: Huygenian, Ramsden (orthoscopic), compensating, and wide-angle (Restivo, 1992).

Huygenian oculars (Fig. 2A) consist of two simple lenses fitted with a fixed-field diaphragm located between the field lens and the eye lens. A real image is focused at the plane of the field diaphragm. Such oculars are basic low-magnification eyepieces with little correction for objective lens aberrations.

Ramsden oculars (Fig. 2B) contain the same components; however the field lens is in an inverted position and the field diaphragm is located below the field lens. As in the Huygenian ocular, a real image is located at the plane of the field diaphragm. Ramsden oculars are normally used for intermediate magnifications.

Compensating oculars have the greatest amount of correction for chromatic variations in magnification, as well as curvature of field. These eyepieces are recommended for use at high magnification and for color photomicrography with apochromatic objectives.

Wide-angle oculars in modern instruments are also designed as high-eyepoint lenses, giving a much larger relief distance between the eye lens and the eye. They contain many lens elements and their design may vary considerably from manufacturer to manufacturer. Being highly corrected for chromatic aberration and for distortion, they are often the only oculars supplied with a microscope, particularly when the instrument is fitted with flat-field objectives.

Practical Adjustment of the Microscope

Simplified schematic diagrams of transmitted- and reflected-light microscopes are given in Figure 3. Experienced microscopists have learned that the most successful use of any microscope begins by following the light path

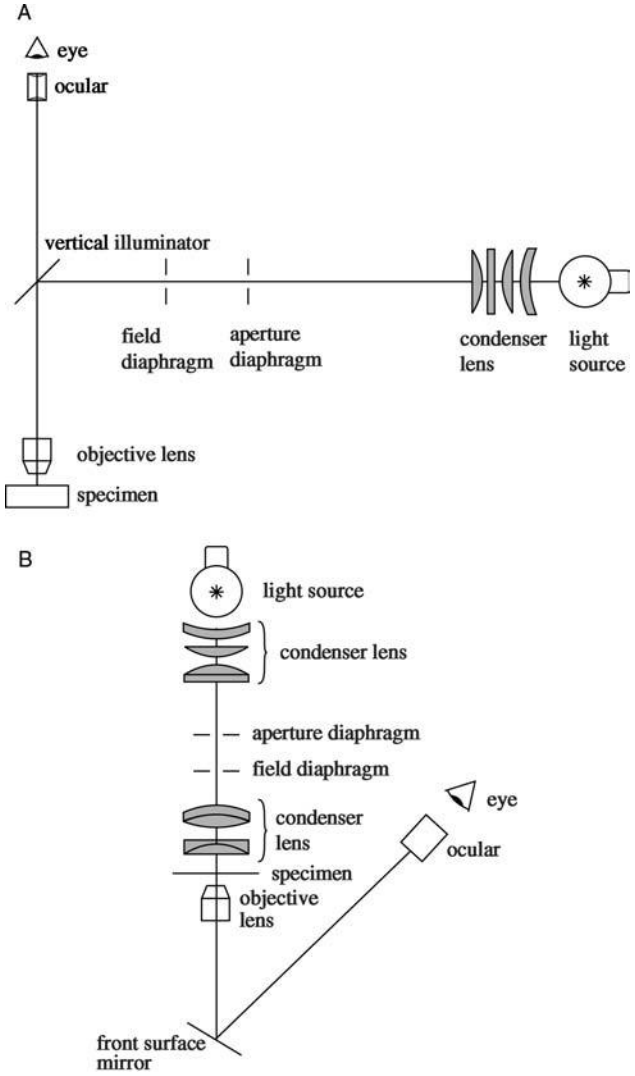


Figure 3. Schematic diagrams of (**A**) inverted reflected-light and (**B**) transmitted-light microscopes. The important components are labeled. Note that the transmitted-light microscope (panel B) does not have a vertical illuminator.

through the instrument and making necessary adjustments, starting with the source of illumination. By consistently following the sequence, the chance of failure in the form of a poor photomicrograph or poor visual image is greatly reduced.

Illumination of the specimen is accomplished by two or more components including the lamp and a condenser or collector lens. Sometimes, the aperture diaphragm, field diaphragm, heat-absorbing filter, narrow band-pass filter, and neutral-density filter are also included as components of the illumination system. This author chooses not to include the additional components as parts of the illumination system, although some microscope makers mount them in a single modular unit of the instrument.

Proper centering and focusing of the light from the illuminator is essential to obtaining satisfactory results with any optical microscope. The image of the lamp filament or arc must be focused at the plane of the aperture diaphragm. In some instruments, the light path is sufficiently open so that one can visually inspect the focus of the lamp on a closed-aperture diaphragm. In closed systems, specific instructions for obtaining proper focus are usually provided with the instrument. Centering of the light source is accomplished by adjustments that laterally translate the lamp in two directions. A focusing adjustment moves the lamp closer to or further away from the condenser to focus the lamp at the plane of the aperture diaphragm.

The aperture diaphragm must be properly adjusted to obtain the optimum resolution commensurate with image contrast, for either a transmission- or reflected-light microscope. The full numerical aperture of the objective lens must be utilized to obtain optimum resolution. From a practical standpoint, the aperture diaphragm is interposed in the optical path to control the diameter of the pencil of light so that it just fills the full circular area of the rear lens group of the objective lens. In a transmitted-light microscope the aperture diaphragm is located in the condenser lens barrel, while in a reflected-light microscope the aperture diaphragm is located in the optical path just beyond the condenser (collector) lens of the light source. In both instruments, the image of the diaphragm is focused in a plane coincident with the rear element of the objective lens. The aperture diaphragm is always partially closed to reduce internal light scattering within the microscope, a circumstance that degrades contrast in the image. In practice, the method for obtaining the optimum setting of the aperture diaphragm is to open the diaphragm fully, bring the specimen into focus, and then close the diaphragm slowly until the image begins to dim perceptibly. At this point the aperture is opened slightly to bring the image back to full brightness.

The image of the field diaphragm is focused on a plane coincident with the plane of the specimen. Its function is to reduce extraneous light scatter between the specimen surface and the front element of the objective in order to enhance contrast. To adjust the field diaphragm, one starts with it in the full-open position and gradually closes it while viewing the focused specimen either through the eyepieces or on a photographic viewing screen. When the image of the diaphragm is observed to just intrude into the field of view or into the photographic field, it should be opened slightly.

Vertical illuminators are required for reflected-light microscopes, because the specimens are opaque and the objective lens must act both to supply light to, and to collect reflected light from, the specimen. The light path must be passed to the specimen through the objective lens, then allow the reflected image-forming light to pass back through the objective to the ocular. From Figure 3A, note that the light path is diverted 90° from the horizontal to the specimen, which rests inverted on the stage (inverted microscope). This arrangement is of particular utility in using specimens that must have prepared surfaces, because it does not require that the specimen have a back surface exactly parallel to the prepared surface. Several schemes have been used for vertical illuminators, including various prism configurations and thin, optically flat glass plates. Of these, the thin, optically flat glass plate

(or disc) is almost universally used in recent instruments. Generally, vertical illuminators have no adjustment controls except for a provision for dark field illumination, a topic that is discussed in REFLECTED-LIGHT OPTICAL MICROSCOPY.

The objective lens requires only two adjustments: centering the optic axis of the lens with the optic axis of the instrument and focusing the lens by moving it closer to or further away from the specimen. Focusing the objective is as simple as moving the focusing adjustments while viewing the specimen through the eyepieces and observing when the specimen appears to be sharp. A number of instruments have their objective lens mounted in a turret to facilitate changing from one objective to another. Many instruments are said to have "parafocaled" objectives. If the objectives are parafocaled, one may readily switch objectives without having to refocus the instrument. In practice, however, it is wise to use the fine-focus adjustment each time an objective is changed in order to assure that the optimum focus is obtained.

Centering of the objective lens on an instrument with an objective turret is accomplished by centering on the objective through the centering of the axis of rotation of the turret. This involves moving the axis of rotation of the turret by shifting it in one or two orthogonal directions of a plane normal to the axis of rotation. Such an operation is not recommended except when performed by competent service personnel. Fortunately, once a turret is centered and locked into the instrument, frequent centering is not required. Other instruments, in which objectives are placed in the instrument individually, require that each lens be centered individually by adjustment set screws on the mounting plate of each lens. Often the instrument contains one "standard" objective that is centered at the factory in the final fitting stages of instrument assembly. The mounting plate of this objective has no centering adjustments. Other objectives may be centered individually using the "standard" lens as a reference, provided the instrument is fitted with a rotating stage that is also centered with the optic axis of the instrument. A simple method involves starting with the "standard" objective, aligning a reflective cross-hair (or reflective point target) on the stage with the center of the field as viewed through the ocular, and making certain that the stage is itself centered. The stage is then rotated to ascertain that the cross-hair stays in the center of the field as the stage is rotated. Without moving the stage or the cross-hair, a centerable objective is substituted for the "standard" objective. If the objective is centered, the cross-hair will again fall in the center of the field and will remain there when the stage is rotated. To center the objective, it will be necessary to use the adjustment set screws on the mounting plate to move the optic axis of the objective in x and y directions in a plane normal to the optic axis. A number of instruments include special accessories for centering the objectives.

Ocular adjustment amounts only to focusing the ocular. Instruments with a single eyepiece generally do not have an adjustment for focus, because the eyepiece is set at its proper focal distance from the objective. Since most microscopes now have binocular eyepieces, some care must be taken to assure that the eyepieces are correctly focused to match the eyes of the user. One eyepiece has a focusing ring, while the other is fixed. Depending on the instrument, focusing methods may vary; however, they are all variations on the method outlined in the following. A stage micrometer, or other similar device, is placed on the stage and brought into focus with the fixed eyepiece using the objective focusing adjustment. Once this is done, the eyepiece focusing ring should be used to focus the image for the other eye. This method accounts for differences between the two eyes, and is necessary to reduce eye strain from long sessions with the microscope.

It is important to emphasize that if the microscopist normally wears eyeglasses, they should be worn while using a microscope. In recent years wide-field objective lenses and matching wide-field, high-eyepoint oculars have been introduced. Matching optics of this type provide a wider field of view of the specimen and are also particularly suited to users who wear glasses. Multifocal eyeglasses (bifocal or trifocal) present special problems that can be overcome by the adaptability of the microscopist.

Autofocusing has yet to be applied to microscopes except in the case of instruments designed to make precise measurements. Such devices are only of use when one of several turret-mounted objective lenses is brought into sharp focus. The autofocusing feature will then cause each of the other lenses in the turret to be brought into sharp focus when they are rotated into the optic axis of the microscope. Critical focusing depends on the skill of the microscopist in being able to judge when sharp focus is attained. Focusing for edge sharpness could probably be developed for instruments used in conjunction with image analysis computers, wherein the focusing would be accomplished by feedback of the contrast gradient at a sharp edge. However, obtaining the best overall focus for the entire field of view might be difficult, either because of a limited range of the gray scale or because of the lack of flatness of specimens.

Devices for calibrating resolution or the depth of focus are generally not used. Observable degradation of the image is a signal to inspect the optical components for contaminants and for physical defects such as scratched or etched surfaces. Regular annual service by a competent service technician is recommended, to maintain the microscope in optimum working condition.

Damage from contamination and damage to objectives is, for the most part, up to the microscopist in maintaining a clean environment for the microscope. The use of appropriate lubricants for the mechanical parts of the instrument is of paramount importance. Physical damage of the front element of the objective lens comes primarily from two sources: (1) contact of the front element with the specimen and (2) etching of the front element by chemical reagents used in specimen preparation. Careful operation by the microscopist is the most effective way to keep the objective lens from contacting the specimen; however, some microscopes, usually upright instruments, employ a stage lock to prevent the objective from contacting the specimen. To prevent contamination of the objective lenses with specimen-preparation reagents, it is essential that specimens be thoroughly cleaned and dried

prior to placing them on the microscope. This is particularly important when HF or other corrosive reagents have been used in specimen preparation.

In the past, special stages for providing special specimen environments have been available as regular production accessories for some instruments. However, with the advent of the scanning electron microscope (SEM), most studies that require hot or cold environments are now undertaken in the SEM. Studies in liquid environments are especially difficult, particularly at high magnifications, because the index of refraction of the liquid layer interposed between the specimen and the objective lenses is different from that of air. Therefore, some corrections can be designed into objectives for certain biological microscopes. Objectives designed for use with glass slides and coverslips are also corrected for the presence of the thin glass layer of the coverslip, which is of the order of 0.1 mm thickness. A number of specialized environmental stages have been designed and constructed by microscopists.

Photomicrography, on film and with charge-coupled devices, as well as specimen preparation, are of paramount importance to successful optical microscopy for the materials scientist or engineer. These topics are to be covered in subsequent units of this chapter.

LITERATURE CITED

Gifkins, R. C. 1970. Optical Microscopy of Metals. Elsevier Science Publishing, New York.

Jenkins, F. A. and White H. E. 1957. Fundamentals of Optics, 3rd ed. McGraw-Hill, New York.

Kehl, G. L. 1949. The Principles of Metallography Practice. McGraw Hill, New York.

Kreidl, N. and Rood, J. 1965. Optical Materials. Vol. I. Applied Optics and Optical Engineering. Academic Press, New York and London.

Restivo, F. 1992. A Simplified Approach to the Use of Reflected Light Microscopes. Form No. 200-853, LECO Corporation, St. Joseph, MI.

Vander Voort, G. 1984. Metallography Principles and Practice. McGraw-Hill, New York.

KEY REFERENCES

Gifkins, 1970. See above.

An excellent reference for microscope principles and details of techniques, using copious illustrations.

Jenkins and White, 1957. See above.

A classic text on optics that has been a mainstay reference for students of physics and engineering.

Kehl, 1949. See above.

A standard text on metallography used extensively for many years. It contains much useful information concerning microscopy of opaque specimens.

Kreidl and Rood, 1965. See above.

Contains a compilation of the optical properties of a large variety of materials.

Restivo, 1992. See above.

While this is a brief treatment of the use of reflected light microscopes, primarily from the standpoint of service technicians, it contains excellent schematic drawings with simplified text.

Vander Voort, 1984. See above.

This is the current standard text for students of metallography and a valuable reference for practicing metallographers. Chapter 4, dealing with light microscopy, is a superior overview containing 100 references.

INTERNET RESOURCES

http://www.carlzeiss.com

Carl Zeiss Web site.

http://www.olympus.com

Olympus Web site.

http://www.nikon.com

Nikon Web site.

http://www.LECO.com

LECO Corporation Web site listing special Olympus reflected-light microscopes.

http://microscopy.fsu.edu

This site provides a historical review of the development of the microscope.

RICHARD G. CONNELL, JR.
University of Florida
Gainesville, Florida

REFLECTED-LIGHT OPTICAL MICROSCOPY

INTRODUCTION

Reflected-light microscopes were developed for imaging the surfaces of opaque specimens. For illumination of the specimen, early microscopes used sunlight played on the specimen surface by mirrors and/or focused by simple condenser lenses. Optical microscopes began to evolve toward modern instruments with the understanding that engineering properties of materials were dependent upon their microstructures. Current higher-magnification instruments (50× to 2000×), for obtaining images of polished and etched specimens, employ vertical illumination devices to supply light to the specimen along the optical axis of the microscope. Vertical illumination overcomes the shadowing of surface features, which are enhanced by off-axis illumination, and improves the attainable resolution. On the other hand, lower-power (5× to 150×) stereo microscopes, most commonly used for imaging bulk specimens such as microelectronic circuits or fracture surfaces, utilize off-axis lighting supplied by focused lamps or fiber-optic light systems.

Microscopy, regardless of the instrumental technique, has as its sole objective the production of visual images that can be further analyzed by counting measurements to quantify any of a number of microstructural properties including grain size, the volume fractions of phases, count (or identification) of microconstituents or microstructural features such as twins, number of inclusions, surface area

of grains and second phases, and more esoteric microstructural measurements such as the total mean curvature of the grain boundary in a given specimen. Quantification of microstructural features may be accomplished by manual counting while viewing the image through the microscope, taking measurements from photographic or digital images, or using a computer for digital image analysis. Obtaining quantitative microstructural information constitutes the specialized field of quantitative microscopy or stereology, based upon geometric and topological considerations and statistics. Stereology is a branch of quantitative microscopy that provides for the inference of three-dimensional structure information from two-dimensional measurements. The reader is referred to DeHoff and Rhines (1968), Gifkins (1970), and Vander Voort (1984) as primary references on quantitative microscopy.

While reflected-light microscopy alone is not intended for the measurement of material properties, coupled with the techniques of quantitative microscopy it does provide the connecting links among properties, chemical composition, and processing parameters. However, an experienced microscopist can obtain much useful qualitative information related to processing and properties. For instance, one may assess the degree of annealing of a metal or alloy by the amount of the microstructure that appears to consist of recrystallized grains. In a production setting, a microscopist may quickly determine if a material meets specified properties merely by observing the microstructure and judging it on the basis of experience.

It is the intent of this unit to describe the more common techniques of reflected-light microscopy, but not to delve into microstructural property measurement. Emphasis will be placed on the use of the higher-power instruments (metallographs), because of the variety of useful illumination modes and image-forming techniques that are available on these instruments.

PRACTICAL ASPECTS OF THE METHOD

Reflected-light Microscopes

The reflected-light microscope, shown by the light-path diagram in Figure 1, contains a vertical illuminator, in this case employing a half-mirror to direct light to the specimen and to allow light from the objective lens to pass to the ocular (eyepiece). In arrangement of the specimen stage, with respect to the axis of the horizontal incoming light path, there are two types of microscopes—upright and inverted. Generally, the inverted instrument is preferred because it ensures that a specimen surface is always normal to the optic axis of the objective lens. Further, it allows for easier observation of large unmounted specimens and specimens of irregular shape that have a prepared surface (Vander Voort, 1984).

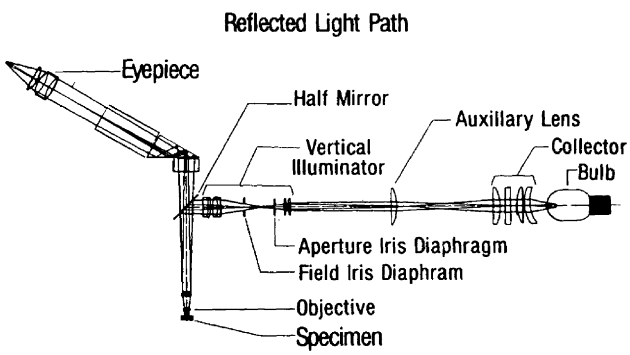


Figure 1. Light path diagram for a reflected-light microscope (courtesy LECO Corporation).

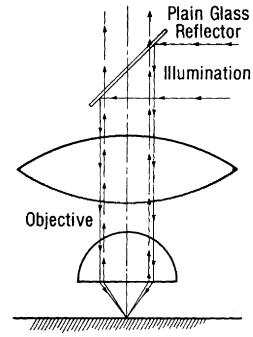


Figure 2. Light-path diagram for a plain-glass reflector for a reflected-light microscope (courtesy LECO Corporation).

Vertical illuminators have taken several forms during their evolution. For the most part, two types of vertical illuminators are in use today; the plain-glass reflector and the half-mirror reflector. The plain-glass reflector consists of a very thin, optically flat glass plate inclined at 45° with respect to the path of the incoming illumination (Fig. 2). Part of the illumination is reflected at 90° through the objective to illuminate the specimen. A portion of the light that returns from the specimen through the objective lens passes through the glass plate on to the ocular. Note that the objective first acts as an illuminating (condenser) lens and second as the image-forming lens. When the plain-glass reflector is properly centered, plain axial illumination is the result and no shadows of specimen surface relief are formed. While the loss of light intensity is significant, this type of reflector can be used with any objective. The half-mirror reflector is a totally reflecting surface, usually silvered glass, with a thin clear coated glass area in its center. The clear central area is elliptical in shape, so that when the reflecting plate is placed at 45° in the optical path a round beam of light can be reflected (Restivo, 1992). An advantage of this type of reflector is that it is particularly well suited for dark-field illumination, and it also works well with bright-field illumination and polarized light (see the discussion below under "Illumination Modes and Image-Enhancement Techniques" for a description of these illumination modes). Light rays reflected back through the objective lens from the specimen surface pass through the central clear region of the reflector to form an aerial image that is further magnified by the ocular. Because the reflecting surface reflects all of the incident light, there is no loss of light intensity as with the plane glass reflector (Restivo, 1992). Figure 3 shows the light path for the half-mirror reflector as it is used in dark-field illumination mode. Other vertical illuminator/

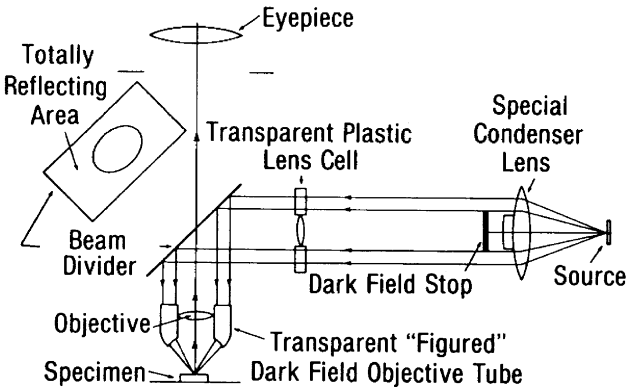


Figure 3 Light path diagram for a half-mirror reflector for a reflected-light microscope. (courtesy LECO Corporation).

reflector schemes have been used with some success; however, they have fallen out of favor because of limitations or expense in producing suitable reflectors.

Illumination Modes and Image-Enhancement Techniques

Provisions for a number of variations in illumination modes and image enhancement techniques may be included as accessories on many of the current production microscopes. Among the more common of these are: bright-field illumination, oblique-light illumination, dark-field illumination, polarized-light, sensitive-tint, phase-contrast, and differential-interference-contrast. Each of these will be discussed in the following paragraphs.

Bright-field illumination is the most widely used mode of illumination for metal specimens that have been prepared by suitable polishing and etching. The incident illumination to the specimen surface is on axis with the optic axis through the objective lens, and is normal to the specimen surface. As expected, those surfaces of the specimen that are normal or nearly so to the optic axis reflect light back through the objective lens, giving rise to bright portions of the image. Those features of the microstructure that have surfaces that are not normal to the optic axis appear dark and thus give the image contrast (Fig. 4). Image contrast also occurs in bright-field illumination because of variations in reflectivity within a specimen. For example, in properly prepared specimens of irons and steels that contain MnS (manganese sulfide)

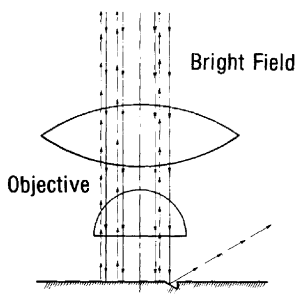


Figure 4. Light path diagram for bright-field illumination (courtesy LECO Corporation).

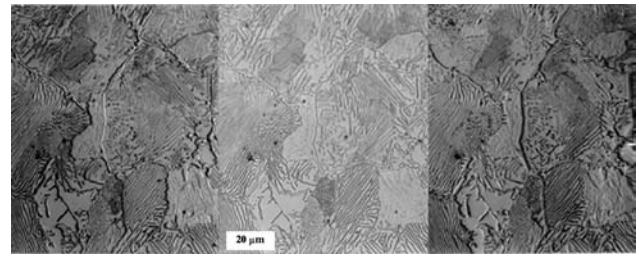


Figure 5. Bright-field photomicrograph (center) of pearlite and proeutectoid cementite in an annealed high-carbon steel. The left and right panels show oblique-light photomicrographs of the same field, with the right panel illuminated with light from the left and the left panel illuminated with light from the right.

as a microconstituent, the MnS constituent appears somewhat darker than the surrounding α-ferrite iron, and possibly lighter than Fe_3C (cementite). The image contrast in this case is due to the optical properties of the constituents that render their polished surfaces more or less reflective. In fact, if the specimen preparation is very well done, the MnS appears to have a smooth dark-slate color. Certain specimen-preparation techniques and spectral image enhancements provide a variety of colors to establish image contrast. The vast majority of metallographic work utilizes bright-field illumination. A typical bright-field photomicrograph appears in Figure 5 (center panel).

Oblique-light illumination is a variation of bright-field illumination wherein the condenser system is shifted off axis. The result is that relief in the specimen surface is shadowed. While the image appears to possess enhanced contrast, and detailed features become more easily observed, the optical resolution is decreased; fine details become broadened. In addition, by shifting the incoming light beam off axis, the effective numerical aperture of the objective lens is decreased, thereby reducing resolution (see OPTICAL MICROSCOPY). To some extent, compensation for this deficiency is made by opening the aperture diaphragm, but at some sacrifice to contrast. On some microscopes oblique illumination is obtained by shifting only the aperture diaphragm off axis, an adjustment that requires that the aperture diaphragm be opened, or else the specimen will be unevenly illuminated. Oblique illumination is illustrated in the photomicrographs of Figures 5, left and right. These photomicrographs are of the same specimen and of the same area as that of the bright-field photomicrograph of Figure 5 (center). Both the left and right panels of Figure 5 are taken under conditions of oblique light, but the off-axis shifts are in opposite directions. Note how certain microstructural features appear raised in one photomicrograph, while the same features appear depressed in the other.

Dark-field illumination is a useful technique to enhance image contrast by making the normally bright areas of an image dark while making dark features light. Dark-field illumination is effective because of the perception characteristics of human vision. It has been demonstrated that a small white disc on a black background is more easily perceived than a small black disc on a white background. The size limit of the white disc on the black background

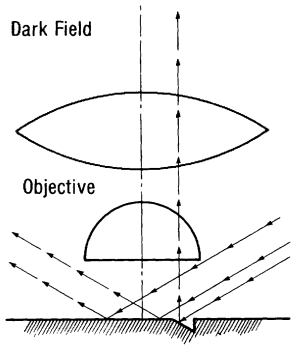


Figure 6 Light path diagram for dark-field illumination (courtesy LECO Corporation).

depends entirely on its reflected luminous flux, or the total light reflected per unit area (Gifkins, 1970).

Dark-field illumination occurs when the light incident on the specimen arrives at a low angle. If the surface is featureless, reflected light does not reenter the objective lens and the image is dark. However, any feature on the surface that is inclined a relatively small angle can reflect light back through the objective lens to form an image. To effect dark-field illumination, a small circular stop is inserted in the light path ahead of the vertical illuminator to produce an annulus of light larger in inside diameter than the back element of the objective lens. The objective lens is surrounded by a "shaped" or "figured" transparent tube that directs the incident light to the specimen at a relatively low angle to the specimen surface. Reflected light from tilted features of the specimen reflects light back into the objective lens (Fig. 6). Figures 7 A and B are bright-field and dark-field photomicrographs, respectively, of a specimen of low-carbon iron. Note that the grain boundaries and inclusions are enhanced in dark-field.

Polarized-light microscopy has significant use in imaging microstructural details in some materials, notably, alloys of uranium, zirconium, beryllium and, from the author's experience, high-purity aluminum (Connell, 1973). Specimens of these materials are difficult to etch to reveal microstructural details. Before the invention of the electron-beam probe, polarized light was used as an aid in identifying inclusions in alloys. Also, polarized light has been of paramount importance in mineralogical work for the identification of various minerals.

Polarized-light microscopy depends upon a characteristic of light that cannot be detected by the human eye. When considering the electromagnetic wave character of light, it is easy to understand that, the greater the amplitude of the wave, the brighter the light. Likewise, variations in wavelength produce variations in color; the shorter the wavelength, the more blue the light, while the longer the wavelength, the more red the light. Two other characteristics of light—phase and polarization—cannot be directly detected by the eye. From an elementary point of view, one may consider that light waves travel in a straight line, and the light waves oscillate (vibrate) along the line in random directions around the line. Under certain conditions of reflection defined by Brewster's law, and when light passes through certain transparent crystalline

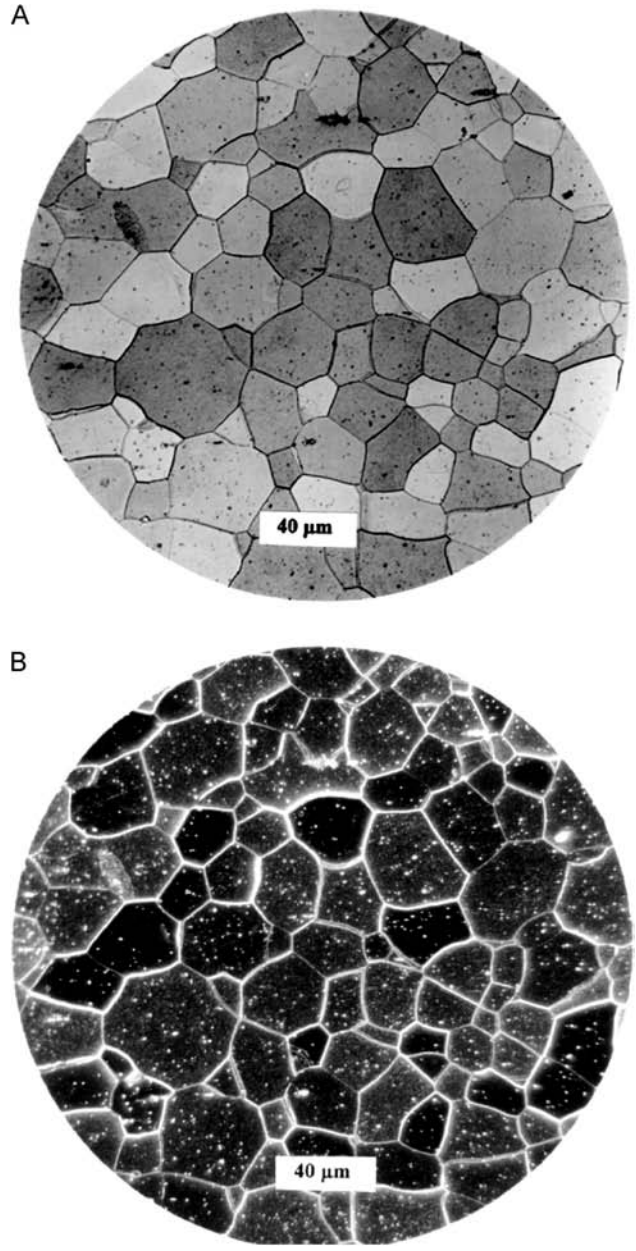


Figure 7. Photomicrographs of ferrite in low-carbon iron: (**A**) bright field; (**B**) dark field.

materials, some of the vibration directions around the light axis become attenuated, and there may be rotation of other directions, so that certain directions of vibration are predominant (Jenkins and White, 1957; Gifkins, 1970). From a practical standpoint, when such interaction of light waves with matter occurs, vibrations of the light are confined to one plane, and the light is said to be plane-polarized. It is to be further stated that, light being electromagnetic waves, it is the electrical or magnetic interaction of the light waves with matter that result in polarization (Gifkins, 1970).

In microscopy, the usual way to produce polarized light from a normal source is to pass the light through a transparent plate of some specific optically active material,

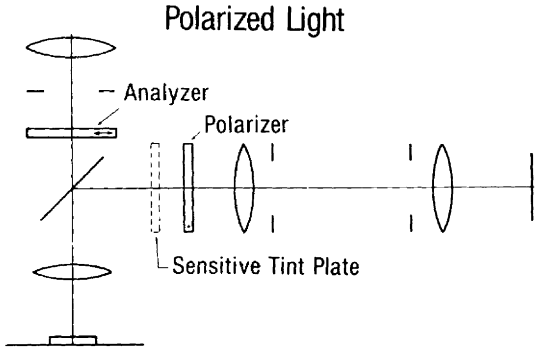


Figure 8. Light path diagram for a reflected-light microscope fitted with polarizer and analyzer. Note also the position of the sensitive-tint plate (courtesy LECO Corporation).

called a polarizer. Most polarizers in use in microscopes are made of synthetic material in sheet form, called by its proprietary name, Polaroid. If two sheets of polarizing material are placed parallel to one another separated by some distance (4 or 5 cm), and light is passed through them, one can observe that light passes through the first although it is polarized. The light may also pass through the second, but by rotation of the second sheet, one will observe that the light no longer passes through it. This is the case of extinction, and one can say that the polarizers are crossed. The second sheet, called the analyzer, will not allow the polarized light through, because its plane of polarization is 90° to the first sheet, the polarizer.

Both a polarizer and an analyzer are employed in a polarizing microscope. The polarizer is placed in the light path ahead of the vertical illuminator and the analyzer is placed in the light path between the vertical illuminator and the ocular (Fig. 8). The polarizer usually is fixed so that it will not rotate. When polarized light is incident on the reflective surface of an isotropic specimen at a right angle, the polarization of the reflected light will not be altered (there is no polarization dependence of reflectivity in isotropic specimens). If one rotates an analyzer located in the light path at the back end of the objective lens, extinction will occur when the polarizer and analyzer are crossed. The image turns dark and indicates that the originally polarized light has not been altered by the reflecting surface. If the specimen is replaced with an anisotropic material such as zirconium, whose reflectivity depends on orientation of polarization with respect to crystal axes, then it can be observed that some grains appear dark while others appear in various shades of gray all the way to white, indicating that each grain has a different crystallographic orientation. When the specimen stage is rotated, the orientation of individual grains will change with respect to the polarized light and go through several extinctions as the stage is rotated through 360°. In this way, the microscopist can observe the grain structure in a material for which etching to reveal the grain structure is difficult. An example of a photomicrograph of martensite taken under polarized light is given in Figure 9. Martensite, being body-centered tetragonal, is anisotropic and therefore responds to the polarization of the light according to its crystal orientation.



Figure 9 Polarized-light photomicrograph of martensite in a hardened steel specimen.

Isotropic metals can be examined by polarized light if they can be etched to develop regular patterns of closely spaced etch pits or facets on each grain. Anodizing the polished specimen surface of isotropic metals can also make the material respond to polarized light. In the case of aluminum, a relatively thick transparent film is formed by anodizing. The observed polarizing effect results from double reflection from the surface irregularities in the film (Vander Voort, 1984). The author has used this technique to observe the subgrain structure in creep-deformed high-purity aluminum. Angular misorientations as small as 2.5° between adjacent subgrains have been observed (Connell, 1973).

Sensitive tint is an image-enhancement technique used with polarized light. Because the human eye is more sensitive to minute differences in color than to minute differences in gray scale, the sensitive-tint technique is a valuable tool in viewing and photographing images that lack contrast. Therefore, preferred orientation may be qualitatively assessed, counts to determine grain size are more easily made, and the presence of twins is more easily detected.

A sensitive-tint plate is placed in the light path of the microscope between the polarizer and the analyzer (Fig. 8). The material of the sensitive plate is a birefrigent. Birefringence occurs in certain anisotropic crystals when these are cut along specific crystallographic planes to produce double refraction (Jenkins and White, 1957). One sees a double image when viewing through such a plate. The sensitive-tint plate (also called magenta-tint or whole-wave plate) is made of quartz sliced parallel to its optic axis. The thickness of the plate is such that it is equivalent to a whole number of wavelengths of light, where the wavelength is in the middle of the visual spectrum (green, 5400 Å). Rays of plane-polarized light pass through the sensitive-tint plate and become divided

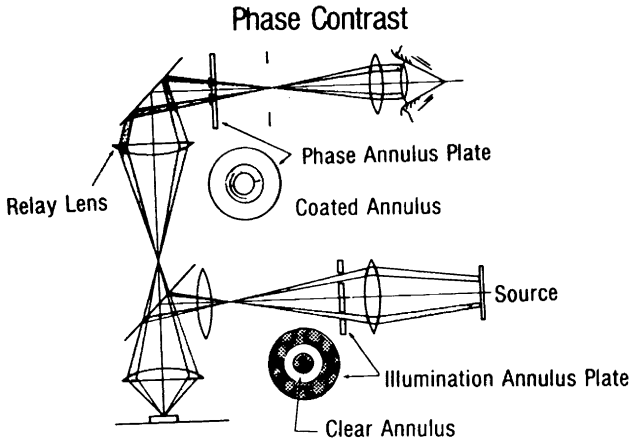


Figure 10. Light path diagram for phase-contrast microscopy. Note the positions of the illumination- and phase annulus plates in a reflected-light microscope (courtesy LECO Corporation).

into ordinary rays and extraordinary rays (Gifkins, 1970), the latter being refracted at some angle to the former. When the ordinary and extraordinary rays emerge from the plate, they are out of phase by one wavelength. The two rays recombine to form a single plane-polarized ray that is extinguished by the analyzer, while all other wavelengths pass through the analyzer. The result is that white light minus the green light excluded by the analyzer produces light that is magenta in color. Typically, the hue imparted to sensitive-tint images have a magenta hue, with differences in grain orientation displayed as different colors. Rotation of the specimen causes each grain to change color. Images are striking and usually aesthetically pleasing. Color control can be obtained by rotation of the sensitive-tint plate through a range of $\sim 15^\circ$ to 20°.

Phase-contrast illumination is an image enhancement that is dependent upon another property of light that cannot be observed by the eye. Subtle variations in polished and/or etched metallographic specimens sometimes are so minute that phase differences from the difference in optical path produce no image contrast. Differences in height as little as 5 nm can be observed using phase contrast (Vander Voort, 1984; Restino, 1992).

The physical arrangement of the microscope is seen in Figure 10, where an illumination annulus and a phase annulus are inserted in the optical path as shown. The illumination annulus is opaque, with a clear central annulus, while the phase annulus is a clear glass onto which two separate materials have been applied by vacuum coating. One material is a thin, semitransparent metallic film (antimony) designed to reduce transmission of the annulus, while the other (magnesium fluoride) is for the purpose of introducing a 1/4 wavelength difference between the light that passes through the annulus versus the light that passes through the other part of the phase plate. An alternate construction for the phase plate is to grind a shallow annular groove in the glass (Gifkins, 1970).

In using phase-contrast illumination, proper and precise adjustment of the instrument is critical for obtaining satisfactory images. Some instruments have sets of matching illumination and phase annuli for use with different objective lenses. It is critical that the phase annulus and illumination annulus match. It is generally not sufficient merely to focus on the specimen and insert the annuli to obtain enhanced contrast, because the specimen surface is not generally exactly normal to the optic axis. To make the appropriate adjustment, the specimen must be sharply focused and the image of the illuminating annulus must be centered with the phase annulus. This requires replacing the ocular with an auxiliary lens (Bertrand lens) that is focused on the phase annulus. Once the ocular is reinstalled, phase contrast is realized. The technique is at best fussy, because specimen translation often throws the annuli out of coincidence and the alignment must be done all over again.

Phase-contrast illumination gained some popularity in the early 1950s for metallurgical work, but today it has lost favor because of the critical and frequent instrumental adjustments required. The technique remains viable for image enhancement in transmission microscopy, particularly for biological specimens.

Differential interference contrast (DIC; also known as Nomarski optics), because of its versatility, has become an image-enhancement technique highly favored by microscopists. A number of other interference techniques have been developed over the years, some requiring special instruments or elaborate accessories attached to rather standard instruments, for example two-beam interference microscopy. Of all these, the device that produces contrast by differential interference (Padawar, 1968) appears to be the most readily adaptable to standard instruments, because it requires no internal optically flat reference surfaces. Images produced by DIC display some of the attributes of other illumination modes including dark-field, polarized light, and sensitive-tint. Nonplanar surfaces in specimens are contrast-enhanced, grains in anisotropic materials display extinction behavior, and the color can be of benefit in enhancing perceived contrast.

Differential interference contrast can be classified as an application of polarizing interference, wherein the underlying principle is image duplication in some birefringent material. The image consists of two parts, one formed by the ordinary ray and a second formed by the extraordinary ray.

A Wollaston prism (Jenkins and White, 1957; Gifkins, 1970) is the birefringent unit used in DIC. This device consists of two quartz plates, w_1 and w_2, cut so that their optic axes are at right angles to each other and so that both make a right angle with the optic axis of the microscope (Fig. 11). The Wollaston prism is located at the focal plane of the objective and can be moved laterally across the optic axis. In operation, the incident light is plane polarized after passing through the polarizer, and the critical setting of the illuminating system is such that the beam is focused in w_1 at point x. The reflected light from the specimen is focused in w_2 at point z. Points x and z are symmetrical in the Wollaston prism; thus the path-length difference, d, of the ordinary and extraordinary rays in their passage through the prism to the specimen is the same as that produced on their return passage through the prism, but it is of opposite sign. So the path-length difference after passing through the Wollaston prism the second time, D,

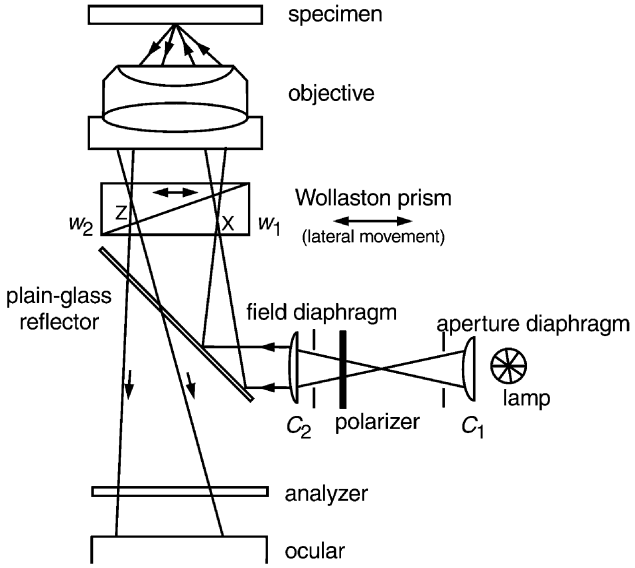


Figure 11. Light path diagram for differential interference contrast (DIC) microscopy. Note that the two sections (w_1 and w_2) of the Wollaston prism have optic axes that are normal to each other, and both optic axes are normal to the light path of the microscope. The axis in w_1 is normal to the plane of the figure, while that in w_2 is in the plane as shown. Points x and z are symmetrically located about the interface between w_1 and w_2.

becomes zero. With a crossed polarizing analyzer in place, and the equalization of path-length differences between the ordinary and extraordinary rays ($D = 0$), a dark field results. The analyzer suppresses both the ordinary rays and the extraordinary rays. Should a specimen contain tilted surfaces such as those that exist at grain boundaries, twins, or slip bands, the reflected rays will pass through the Wollaston prism at nonsymmetrical positions, and the path-length differences between the ordinary and extraordinary rays will not be equalized ($D \neq 0$); therefore, the ordinary and extraordinary rays will not be suppressed by the analyzer. Any path-length difference between the ordinary and extraordinary rays is proportional to the amount of tilt, so that the intensity of the light passed to the ocular is also proportional the amount of tilt.

An important feature of DIC is the ability to introduce color to the image. By laterally shifting the Wollaston prism across the optical path, the path-length difference, D, can be set to values other than zero, and the background takes on a hue dependent upon the path-length difference. For example, if the position of the prism is set so that $D = 0.565$ μm (a wavelength in the yellow region of the spectrum), the background takes on a purple color—i.e., white light minus yellow. Portions of the specimen surface that are tilted will impose other path-length differences, which will result in other colors.

Figure 12A and B are photomicrographs that illustrate a comparison between bright-field illumination and DIC, respectively.

Sample Preparation

Successful use of reflected light microscopes depends to a large measure on specimen preparation. Without a thorough discussion of specimen-preparation techniques, it is sufficient to make the reader aware that opaque specimens must have a properly prepared surface from which light may be reflected. Such a surface is usually prepared by mechanical grinding and polishing using appropriate abrasives until the surface is scratch-free and as specular as a mirror (see SAMPLE PREPARATION FOR METALLOGRAPHY). In most cases it is necessary to chemically (or electrochemically) etch the surface to reveal the details of the microstructure.

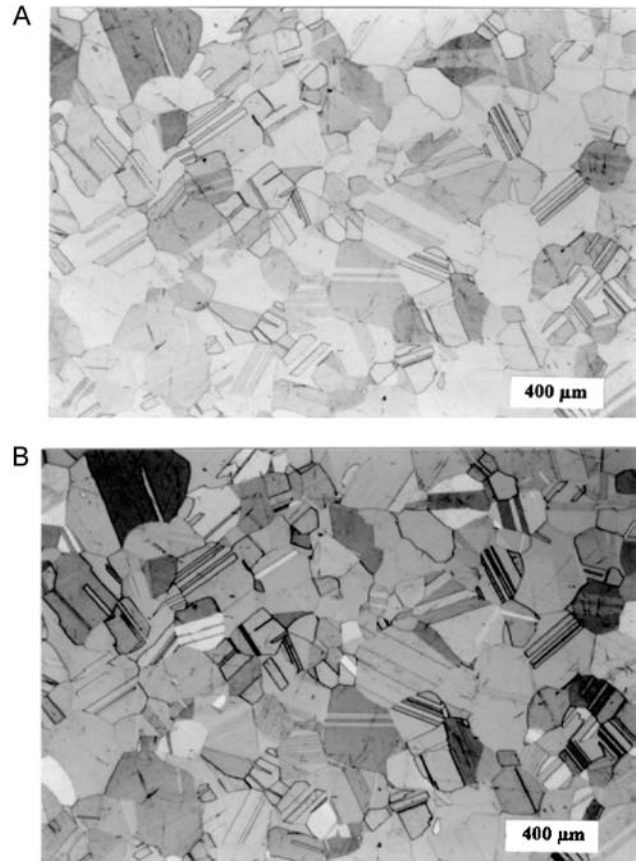


Figure 12. Photomicrographs of α-brass (**A**) bright-field, (**B**) DIC. Grain and twin boundaries are easily seen in the DIC image making it possible to determine grain size easily.

Metallographic preparation techniques have been developed for a wide variety of materials, but with some the metallographer must use his intuition and experience. An excellent source related to specimen preparation is Vander Voort's (1984) book.

PROBLEMS

When inadequate results are obtained with the use of reflected-light microscopy, they are most often associated either with specimen preparation or the inappropriate setting of the controls of the instrument. Improper or poor specimen preparation may introduce artifacts such as surface pitting or inclusion of abrasive particles in the surface. Such artifacts can be confused with second-phase

precipitates. When some portion of the image appears to be in focus, while others appear to be poorly focused, the problem is usually associated with lack of flatness of the specimen surface. In addition, one should pay attention to the action of polishing abrasives on the specimen surface. Some combinations of abrasive materials result in undesired chemical attack of the specimen surface, and the true structure of the specimen may not be revealed. Poor etching (over- or underetching) is a major concern to the metallographer. Overetching often causes what appears to be lack of resolution; fine features can be difficult or impossible to resolve. Underetching may leave some fine details not revealed at all. A better understanding of some aspects of metallographic specimen preparation is given in SAMPLE PREPARATION FOR METALLOGRAPHY.

A practical guide to the setting of the controls is given in OPTICAL MICROSCOPY. Some of the common errors in making settings that degrade resolution of the image are: (1) too small an aperture diaphragm setting, (2) incorrect selection of objective lens, and (3) too high a magnification. Closing the aperture diaphragm beyond the point where the rear element of the objective lens is just covered with light will not allow the objective lens to collect all the diffraction information needed to realize the full resolution capabilities of the lens. The numerical aperture (N.A.) of the objective lens must be consistent with the intended magnification of the final image. As a practical upper limit, the final magnification should be no greater than 1000 times the numerical aperture of the objective lens. For a 0.65 N.A. objective lens, the final magnification should be 650× or less.

Dirty or damaged optical components seriously degrade image quality. Dirt and dust on lens, mirror, or prism surfaces, as well as degradation of lens coatings, are of primary concern. It is wise not to attempt optical surface cleaning other than that which can be done with a soft brush or light blast of air. It is well worth the expense to have a skilled service person thoroughly clean the microscope on an annual basis.

The forgoing discussions are not intended to be a comprehensive treatment of reflected-light microscopy. A number of interesting special techniques have been excluded because they require specialized instrumentation. Almost all of the illumination modes and image-enhancement techniques discussed can be done with current research-grade metallographs. One feature that most modern instruments do not include is a precision rotating stage that allows for rotation of the specimen without disturbing the focus of the instrument. Such a stage was incorporated on instruments manufactured through the mid-1970s. Stage rotation, wherein the specimen remains normal to the optic axis of the instrument, greatly enhances the ability to use polarized light, sensitive tint, and differential interference contrast.

LITERATURE CITED

Connell, R. G. 1973. Microstructural Evolution of High-Purity Aluminum During High-Temperature Creep. Ph.D. thesis, University of Florida, Gainesville, Fla.

DeHoff, R. T. and Rhines, F. N. 1968. Quantitative Microscopy. McGraw-Hill, New York.

Gifkins, R. C. 1970. Optical Microscopy of Metals. Elsevier Science Publishing, New York.

Jenkins, F. A. and White, H. E. 1957. Fundamentals of Optics. McGraw-Hill, New York.

Padawar, J. 1968. The Nomarski interference-contrast microscope, an experimental basis for image interpretation. *J. R. Microsc. Soc.* 99:305-3–49.

Restivo, F. A. 1992. Simplified Approach to the Use of Reflected-Light Microscopes. Form No. 200–853. LECO Corporation, St. Joseph, Mich.

Vander Voort, G. F. 1984. Metallography, Principles and Practice. McGraw-Hill, New York.

KEY REFERENCES

DeHoff and Rhines, 1968. See above.

A textbook of the well-established fundamentals underlying the field of quantitative microscopy, which contains the details of applied measurement techniques.

Gifkins, 1970. See above.

Chapter 6 of this book provides an excellent treatment of interference techniques.

Kehl, G.L. 1949. The Principles of Metallographic Laboratory Practice. McGraw-Hill, New York.

Excellent reference on microscopes and photomicrography for use by technicians, although some of the instrumentation is dated.

Vander Voort, 1984. See above.

Chapter 4 of this book is a particularly good overview reference for reflected-light microscopy.

INTERNET RESOURCES

http://microscopy.fsu.edu/primer/resources/tutorials.html

List of educational resources for microscopy.

http://www.people.virginia.edu/jaw/mse310L/w4/mse4-1.html

Optical micrography reference.

RICHARD G. CONNELL, JR.
University of Florida
Gainesville, Florida

PHOTOLUMINESCENCE SPECTROSCOPY

INTRODUCTION

Photoluminescence (PL) spectroscopy is a powerful technique for investigating the electronic structure, both intrinsic and extrinsic, of semiconducting and semi-insulating materials. When collected at liquid helium temperatures, a PL spectrum gives an excellent picture of overall crystal quality and purity. It can also be helpful in determining impurity concentrations, identifying defect complexes, and measuring the band gap of semiconductors. When

performed at room temperature with a tightly focused laser beam, PL mapping can be used to measure micrometer-scale variations in crystal quality or, in the case of alloys and superlattices, chemical composition.

The information obtainable from PL overlaps with that obtained by absorption spectroscopy, the latter technique being somewhat more difficult to perform for several reasons. First, absorption spectroscopy requires a broadband excitation source, such as a tungsten-halogen lamp, while PL can be performed with any above-band-gap laser source. Second, for thin-film samples absorption spectroscopy requires that the substrate be etched away to permit transmission of light through the sample. Third, for bulk samples the transmitted light may be quite weak, creating signal-to-noise ratio difficulties. On the other hand, absorption spectroscopy has the advantage of probing the entire sample volume, while PL is limited to the penetration depth of the above-band-gap excitation source, typically on the order of a micrometer.

The areas of application of PL also overlap somewhat with Raman spectroscopy (RAMAN SPECTROSCOPY OF SOLIDS), but the latter is much harder to perform because of the inherent weakness of nonlinear scattering processes and because of its sensitivity to crystallographic orientation. On the other hand, this sensitivity to orientation gives Raman spectroscopy a capability to detect variations in crystallinity that PL lacks.

There are also more elaborate techniques, such as photoluminescence excitation (PLE), time-resolved PL, and resonant Raman scattering, but these techniques are more useful for basic scientific investigations than for routine crystal characterization.

PRINCIPLES OF THE METHOD

When a semiconductor is illuminated by a light source having photon energy greater than the band-gap energy (E_g) of the material, electrons are promoted to the conduction band, leaving holes behind in the valence band. When an electron-hole pair recombines, it emits a photon that has a wavelength characteristic of the material and the particular radiative process. Photoluminescence spectroscopy is a technique for extracting information about the electronic structure of the material from the spectrum of light emitted.

In the basic process of PL at low temperature (liquid nitrogen temperature and below), an incoming photon is absorbed to create a free electron-hole pair. This transition must be essentially vertical to conserve crystal momentum, since the photon momentum is quite small. At cryogenic temperatures the electrons (holes) rapidly thermalize to the bottom (top) of the conduction (valence) band by emitting phonons. For this reason there is typically no observable luminescence above the band-gap energy. Once they have relaxed to the band edges, an electron and hole can bind together to form an exciton, which then recombines radiatively as described below. Alternatively, the electron or hole (or both) can drop into a defect state by nonradiative recombination and then undergo radiative recombination.

Excitons and Exciton-Polaritons

It is impossible to analyze luminescence spectra without considering the role of excitons. Formally an exciton must be thought of as an additional bound state of the crystal that appears when the Coulomb interaction between electrons and holes is incorporated into the crystal Hamiltonian, but there is also a simple, mechanistic picture that leads to the same result. The exciton can be thought of as a hydrogenlike atom that results when the electron and hole bind together. The binding energy of an exciton is then given by

$$E_n = -\frac{E_0}{n^2} \qquad E_0 > 0 \qquad n = 1, 2, 3 \tag{1}$$

where E_0 is the free exciton binding energy. The total energy of the exciton is $E_g + E_n < E_g$, so when the exciton recombines, the energy of the emitted photon is below E_g.

This simple, hydrogenlike picture of the exciton is actually an oversimplification. In reality, the free exciton should be thought of as a polarization wave that can propagate through the crystal. This polarization wave can interact with an electromagnetic wave to produce *exciton-polariton* states. The dispersion curve for these states has two branches, as illustrated in many texts in solid-state and semiconductor physics (e.g., Yu and Cardona, 1996, p. 273). Given this continuum of states, the polariton emission spectrum can be quite complex, the peak positions being determined by the relative lifetimes of states at various positions on the dispersion curve. In particular, polaritons near the transverse free exciton energy have particularly long lifetimes, creating a "bottleneck" that leads to accumulation of exciton population. Hence there is typically a peak at the free exciton energy, which can be thought of as a "pure" free exciton peak and is often labeled as such in the literature and in this unit.

Bound Excitons

At low temperatures excitons can bind to donor or acceptor sites via van der Waals forces to form states of lower energy than the polariton states. Since bound exciton states are localized, they have no dispersion. The dissociation energy for a bound exciton is dependent on the species of donor or acceptor; this fact is sometimes helpful in identifying impurities. Figure 1 shows donor- and acceptor-bound exciton luminescence for CdTe.

Alloy Broadening

In an alloy, random microscopic composition variations in the crystal cause a broadening in the energy spectrum of electronic states. Figure 2 shows the exciton spectrum for $Cd_{0.96}Zn_{0.04}Te$. Much of the fine structure that was apparent for pure CdTe has been obscured by alloy broadening.

Phonon Replicas

There may also be phonon replicas of any of the aforementioned peaks. That is, the system may emit part of its energy as one or more phonons—typically longitudinal-

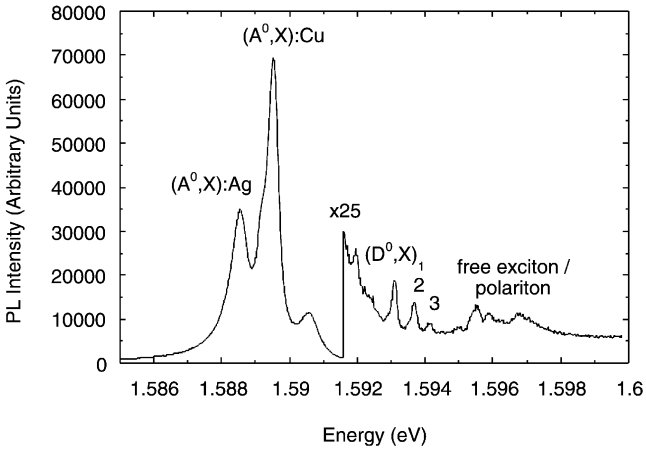


Figure 1. Free and bound exciton luminescence of CdTe at 4.2 K showing bound exciton luminescence from two acceptor species.

optical (LO) phonons—and the remainder as a photon. The emitted photon is thus of longer wavelength than that for the no-phonon process. In some cases the emitted phonon corresponds to a local vibrational mode of the defect to which the exciton is bound; this provides another means for identifying defects. In Figure 2, the first phonon replica of the free exciton luminescence is visible. The phonon replicas are less well-resolved than the no-phonon lines, because of the slight dispersion in the LO phonon energy. The LO phonon energy can be read from the spacing of the peaks as ~22 meV.

Donor-Acceptor and Free-to-Bound Transitions

Some electrons and holes drop by nonradiative processes from the conduction and valence bands into donor and acceptor states, respectively. When these charge carriers recombine, they emit photons of lower energy than do excitonic or direct band-to-band transitions. Possible transitions include those in which a conduction band electron

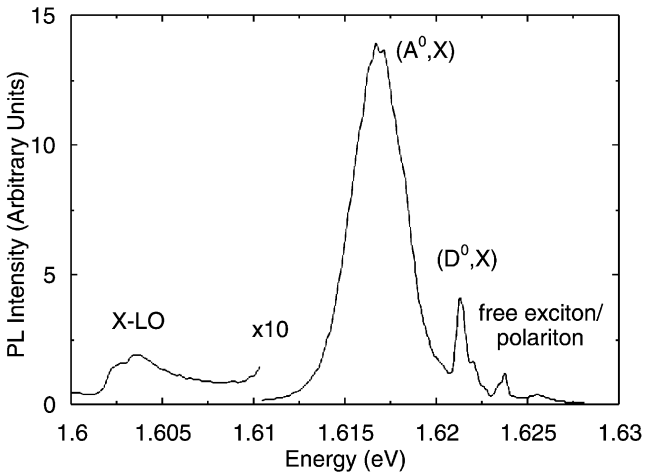


Figure 2. Exciton luminescence of $Cd_{0.96}Zn_{0.04}Te$ at 4.2 K. Many of the peaks that are present in the CdTe spectrum are not resolvable because of alloy broadening.

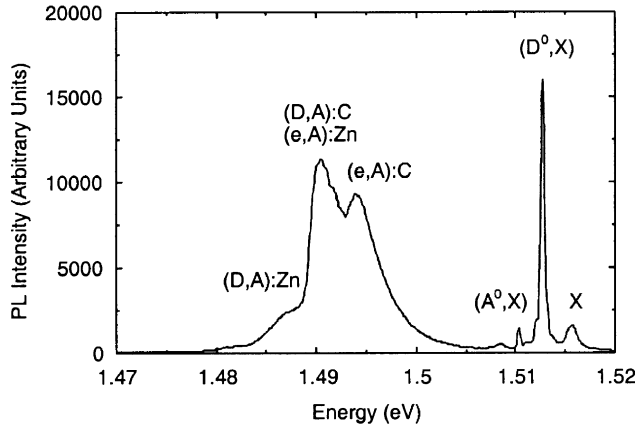


Figure 3. PL spectrum of semi-insulating GaAs at 4.2 K showing separate donor-acceptor and band-to-acceptor peaks for C and Zn.

recombines with a hole in an acceptor level (e,A), an electron in a donor level combines with a hole in the valence band (D,h), and an electron in a donor level recombines with a hole in an acceptor level (D,A). The (D,A) luminescence is below the band gap by the sum of the donor and acceptor binding energies minus the Coulomb energy of attraction between the ionized donor and acceptor in the final state:

$$h\nu = E_g - (E_A + E_D) + \frac{e^2}{\varepsilon r_n} \quad \text{(Gaussian units)} \qquad (2)$$

where r_n is the distance between the nth nearest neighbor donor and acceptor sites and can take on only those values consistent with the crystal lattice. The discrete values of r imply that there are a large number of separate (D,A) peaks, but for relatively large values of n, they merge into a single broad band. With high-resolution measurements on high-quality crystals at low temperature, it is sometimes possible to detect separate peaks for low values of n (Yu and Cardona, 1996, p. 346).

There can also be multiple donor or acceptor levels in a given material, each exhibiting its own set of peaks in the spectrum. A well-known example is C and Zn in GaAs, which give rise to two sets of overlapping (e,A) and (D,A) peaks (Stillman et al., 1991). A representative 4.2 -K GaAs spectrum is shown in Figure 3.

Transitions Involving Defect Levels

Many kinds of defects produce energy states closer to the middle of the band gap than the shallow donor and acceptor states mentioned above. Recombination processes involving these defect levels result in emission of longer-wavelength photons. Because these defect states are localized, they have broad energy spectra. Also, they tend to be strongly coupled to the lattice and hence have a large number of phonon replicas that often smear together into a single broad band. An example CdZnTe spectrum is shown in Figure 4. The band at ~1.4 eV is believed to be related to the A-center (Cd vacancy plus donor), and the band at 1.1 eV is believed to be related to Te vacancies.

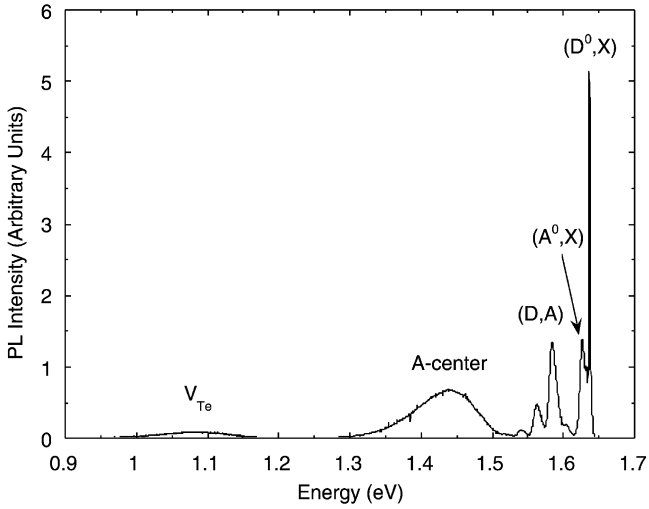


Figure 4. PL spectrum of $Cd_{0.9}Zn_{0.1}Te$ at 4.2 K showing excitonic, donor-acceptor, and deep-level luminescence.

Band-to-Band Recombination

At room temperature there is a nonzero occupancy of electron states near the bottom of the conduction band and hole states near the top of the valence band. Hence, electron-hole pairs can recombine to emit photons over a range of energies, producing a broad band rather than a sharp peak. Donor, acceptor, and defect states are generally fully ionized at room temperature and do not contribute significantly to the observed spectrum.

Band-to-band recombination produces a luminescence spectrum that is the product of the joint density of states and the Fermi-Dirac distribution (see Bhattacharya, 1994):

$$I(E) \propto \frac{\sqrt{E - E_g}}{1 + \exp[(E - E_g)/kT]} \tag{3}$$

Observed room temperature luminescence typically does not follow Equation 3 except at the upper end of the energy range. Equation 3 has a sharp lower cutoff at the band-gap energy, while the measured spectrum has a long tail below the peak. This is an indication that the spectrum is only partly due to band-to-band recombination and also includes excitonic and phonon-assisted transitions (Lee et al., 1994b).

PRACTICAL ASPECTS OF THE METHOD

Experimental Setup

The apparatus for low-temperature PL can be set up easily on an optical table. The sample is enclosed in an optical cryostat and illuminated with an above-band-gap light source, in most cases a laser. An optical band-pass filter is typically used to remove unwanted plasma lines from the laser emission, and a low-pass filter is often placed at the entrance of the monochromator to remove stray light at shorter wavelengths than the range of interest. Both of these sources of extraneous light can show up in the spectrum in second order. The light emitted by the sample is collected with a collimating lens and then focused onto the entrance slits of the monochromator with a second lens.

A photon detector, typically either a photomultiplier tube (PMT) or a positive-intrinsic-negative (PIN) diode detector, is placed at the exit of the monochromator. The output signal of the detector is routed either to a photon counter or to a lock-in amplifier. In the latter case, a light chopper is used to synchronize the signal with a reference. In the photon-counting mode, it is desirable to cool the detector with liquid nitrogen to reduce noise; N_2 gas must then be allowed to flow over the face of the detector to keep it from frosting.

The instrumentation for room temperature PL mapping is similar, with the addition of an *x-y* translation stage. Generally, the optics are kept fixed and the sample translated. The excitation source is focused onto the sample with a microscope objective, which also serves to collimate the luminescence. A beam splitter or a mirror containing a small hole is used to allow transmission of the laser beam while reflecting the luminescence into the monochromator.

An alternative scheme for PL uses an optical multichannel analyzer. In this case the monochromator disperses the light onto an array of charge-coupled devices (CCDs) so that the entire spectrum is captured at once. This configuration has the advantage of greater speed, which is especially valuable in room temperature mapping, but it generally suffers from reduced wavelength resolution.

Following is a list of instrumentation and representative manufacturers for low-temperature PL along with some of the requirements and considerations for each apparatus:

1. *Detectors* (Hamamatsu, Phillips, Burr-Brown, Electron Tubes, Oriel, EG&G, Newport)—The principal trade-off is between bandwidth and sensitivity. The need for cooling is a further consideration. Photomultiplier tubes are unmatched in sensitivity but typically have limited long-wavelength response. Most PMTs can be operated at room temperature in the current mode but may require cooling for optimal performance in the photon-counting mode. Narrow-band-gap semiconductor detectors, such as germanium PIN diodes, have better long-wavelength response but poorer sensitivity at shorter wavelengths. They generally must be cooled with liquid nitrogen.
2. *Spectrometers/monochromators* (Jobin Yvon-Spex, Oriel, Jarrell Ash, Acton Research, McPherson, Ocean Optics, CVI Laser)—For high resolution, a large, double-grating spectrometer is required. For lower resolution applications, compact, inexpensive prism spectrometers with integrated CCD detectors are available.
3. *Lasers* (Coherent, Uniphase, Kimmon, Spectra-Physics, Newport, Melles Griot, Anderson Lasers)—The principal requirement is that the photon energy be higher than the band gap of the material

at the temperature at which the measurements are made. For direct band-gap materials, output power of a few milliwatts is typically sufficient. For indirect band-gap materials, higher power may be desirable. Single-line operation is preferred, since it permits the use of a laser line filter to remove extraneous lines that can appear in the measured spectrum in second order.

4. *Lock-in amplifiers/photon counters* (EG&G, Stanford Research, Oriel)—Both analog and digital lock-in amplifiers are available, the digital instruments generally having superior signal-to-noise ratios.
5. *Complete PL spectroscopy/mapping/imaging systems* (Phillips Analytical, Perkin Elmer, Oriel, ChemIcon)—Some systems are configurable with various laser modules, gratings, and detectors to be optimized for the application.

Use of Low-Temperature PL for Crystal Characterization

The most basic use of low-temperature PL spectra is as an indicator of overall crystal quality. The principal indicator of quality is the ratio of excitonic luminescence to donor-acceptor and deep-level luminescence. Figure 5 shows PL spectra for two CdZnTe samples of different quality. In the high-quality sample, the spectrum is dominated by a donor-bound exciton peak (D^0,X), while in the poor-quality sample the excitonic luminescence is outweighed by donor-acceptor and defect bands.

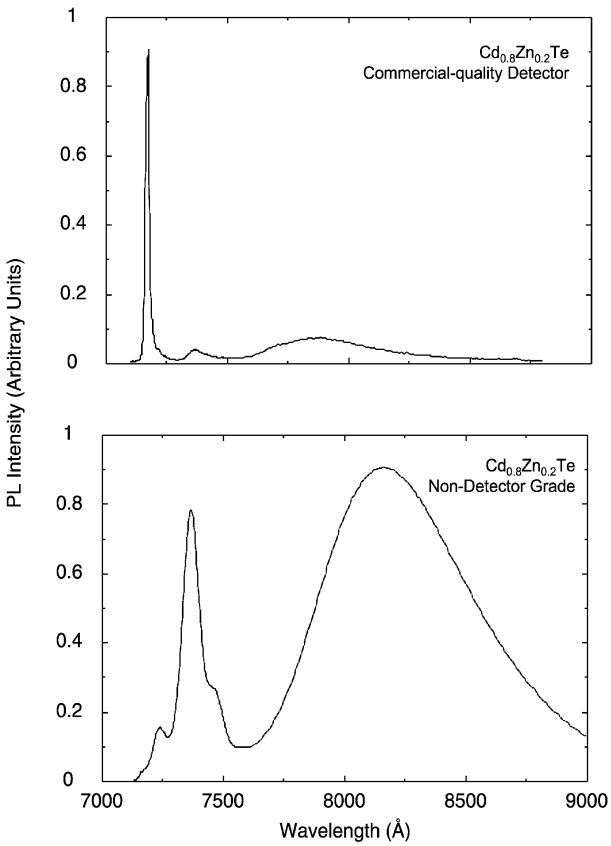


Figure 5. The 4.2-K PL spectra for $Cd_{0.8}Zn_{0.2}Te$ samples of differing quality. In the high-quality (commercial-quality detector grade) sample (top), the spectrum is dominated by excitonic luminescence. In the poor-quality (non-detector grade) sample the donor-acceptor and deep-level luminescence dominate.

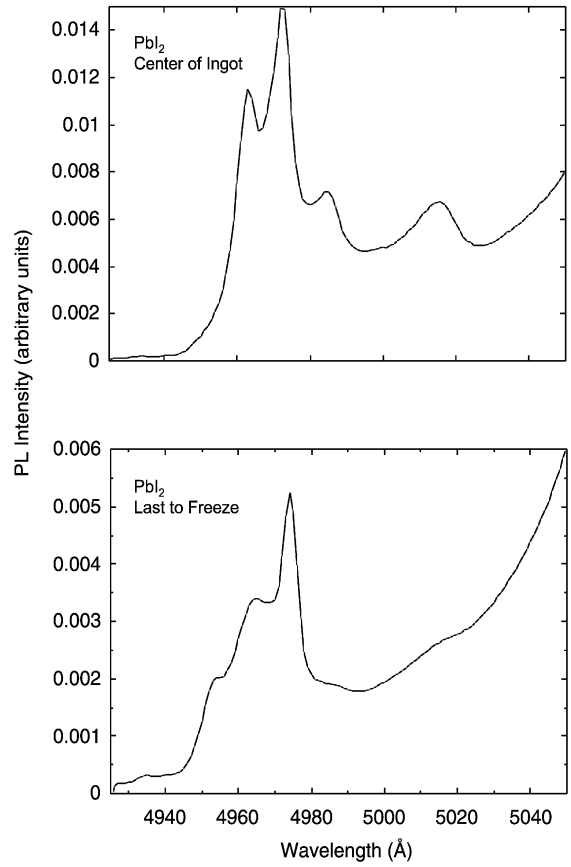


Figure 6. 4.2 K PL spectra for PbI_2 samples from center (top) and last-to-freeze portions of ingot. The spectrum for the sample from the middle of the ingot is sharper, indicating better crystal quality.

A second measure of crystal quality is the sharpness of the spectrum—that is, the extent to which adjacent lines can be resolved. Figure 6 shows near-band-edge PL spectra for two PbI_2 samples, one from the center of the ingot and the other from the last-to-freeze portion. The spectrum of the sample from the center is sharper, indicating a more uniform crystal potential on a microscopic scale.

Aside from indicating overall crystallinity, the low-temperature PL spectrum can sometimes be helpful in identifying specific impurities. The example of the C and Zn (D,A) bands in GaAs was given in Figure 4, but the bound exciton luminescence can also be used. For example, in the CdTe spectrum shown in Figure 1, the (A^0,X) peak at 1.5895 eV is due to Cu, while that at 1.5885 eV is due to Ag (Molva et al., 1984).

The ratio of bound to free exciton luminescence can also be used to measure impurity concentrations, but this application requires careful control over excitation intensity and sample temperature (Lightowlers, 1990). It also requires that calibration curves be generated based on chemical analysis.

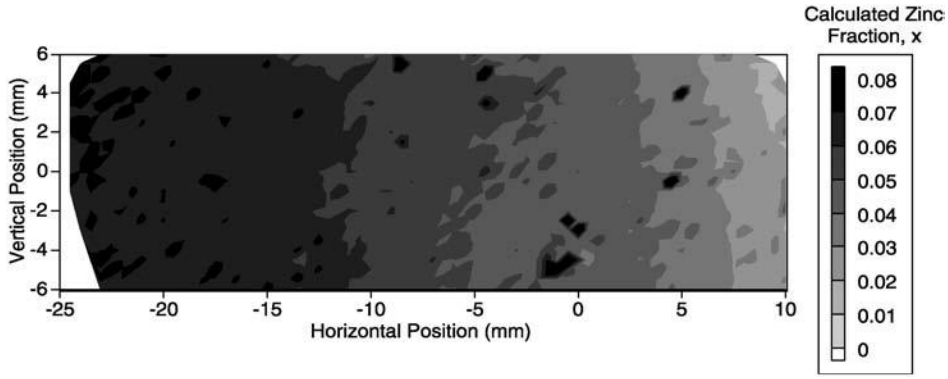


Figure 7. Room temperature PL composition map of $Cd_{1-x}Zn_xTe$ showing segregation of zinc along the growth axis.

Uses of Room Temperature PL Mapping

One application of room temperature PL is composition mapping of alloys. Figure 7 shows a PL composition map for CdZnTe grown by a high-pressure Bridgman method. Segregation of zinc along the growth axis is clearly visible.

In addition to the systematic variation in composition along the growth axis, local variations are sometimes seen. Figure 8 shows an example in which minute striations are present, perhaps due to temperature fluctuations during growth. Note that the dark bands are only ~20 μm wide, demonstrating that high spatial resolution is important.

Photoluminescence mapping is also useful for studying structural defects. PL maps of CnZnTe often contain isolated dark spots that can be attributed to inclusions, possibly of tellurium or zinc. High-spatial-resolution PL maps of these inclusions show a region of near-zero PL intensity surrounded by a region in which the band gap is shifted to higher energy—hence the dark spots in the larger-scale maps. It is unclear whether the shift is caused by gettering of zinc around the inclusion or by strain.

METHOD AUTOMATION

Complete automated systems for PL spectroscopy and mapping are now commercially available. If a system is constructed from scratch, it is a relatively simple matter to integrate GPIB-controllable equipment using general-purpose laboratory software packages.

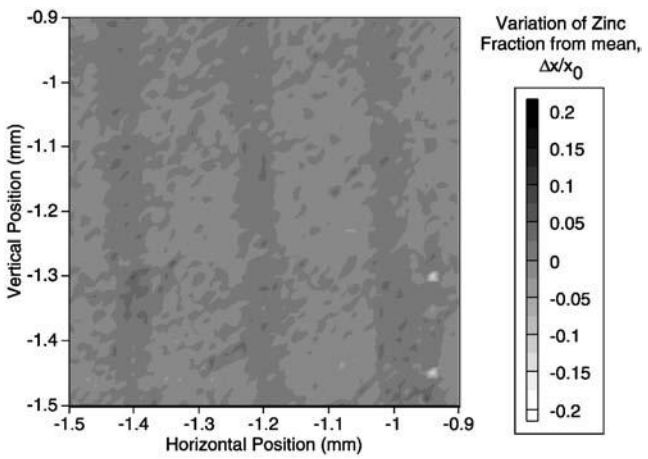


Figure 8. Room temperature PL composition map of $Cd_{1-x}Zn_xTe$ showing growth striations.

DATA ANALYSIS AND INITIAL INTERPRETATION

Identification of Bands in a Low-Temperature PL Spectrum

Qualitative interpretation of a low-temperature PL spectrum is quite simple if there are reliable sources in the literature to identify the origins of the various bands. If there are bands of uncertain origin, the intensity and temperature dependence of the spectrum can give clues to their nature.

As the sample temperature is raised, the luminescence is quenched as the centers involved in the transition are ionized by thermal excitation. The activation energy or energies of the level can then be determined by a curve fit to the following equation, which follows from Boltzmann statistics (Bimberg et al., 1971):

$$I(T) = I(0)\left[1 + C_1\exp\left(-\frac{-E_1}{kT}\right) + \cdots + C_n\exp\left(-\frac{-E_n}{kT}\right)\right]^{-1} \quad (4)$$

where $I(T)$ is the integrated intensity of the band at absolute temperature T, k is Boltzmann's constant, E_i is the activation energy of the ith process involved in the dissociation, and C_i is formally the degeneracy of the level but in practice is left as a fitting parameter. Most processes can be modeled as one- or two-step dissociations. An example of a one-step dissociation is that of a free exciton into a free electron and hole, whereas the prototypical two-step dissociation is bound exciton quenching, in which the exciton first dissociates from the donor or acceptor to form a free exciton, which then dissociates in a second step. This difference in temperature dependence is one means of distinguishing between free and bound excitons in an unfamiliar spectrum.

Temperature dependence can also be used to distinguish between (D,A) and (e,A) transitions, since the former, which involve a shallow donor, will be quenched quickly as the donors are ionized.

The dependence of the spectrum on the excitation intensity can be used to distinguish between excitonic and

donor-acceptor bands. In both cases the luminescence intensity varies with the excitation intensity as a power law, but for excitonic lines the exponent is >1.0, whereas for donor-acceptor and free-to-bound transitions it is <1.0 (Schmidt et al., 1992).

One additional technique that can help to identify peaks is infrared quenching. As the sample is subjected to below-band-gap infrared radiation in addition to laser excitation, donors and acceptors are ionized, causing quenching of (D^0,X) and (A^0,X) transitions and an increase in (D^+,X) luminescence. It has been shown that (A^-,X) complexes cannot exist in most semiconductors (Hopfield, 1964), so that this provides another way of distinguishing between donor- and acceptor-bound excitons. Infrared quenching has also been used quantitatively to measure impurity concentrations (Lee et al., 1994a).

Curve Fitting in Room Temperature PL Mapping

Extracting material parameters from a room temperature PL spectrum in a systematic way requires that the calculated line shapes for various types of transitions (Bebb and Williams, 1972) be combined. This approach has been used with some success in cadmium telluride (Lee et al., 1994b), and in cadmium zinc telluride (Brunnett et al., 1999). In the alloy, variation in the band gap and exciton and phonon energies introduces additional uncertainty in the extracted composition. The composition determined from this procedure cannot be regarded as highly accurate, but variations can be detected with precision.

SAMPLE PREPARATION

Photoluminescence generally requires no special sample preparation, other than to provide a reasonably smooth surface so that a large fraction of the light is not scattered away. Thin-film samples generally can be analyzed as grown, whereas bulk samples that have been cut typically require etching to remove the damaged layer produced by the cutting.

SPECIMEN MODIFICATION

Low-temperature PL requires relatively low optical power density, so that the risk of damaging the material is minimal. Room temperature mapping, however, does use rather high power densities that cause significant sample heating. Most group II–VI and III–V compounds produce adequate luminescence at low enough intensity to avoid damaging the sample, but with less robust materials, such as PbI_2 and HgI_2, room temperature mapping may not be practical.

POTENTIAL PROBLEMS

A difficulty with extracting accurate composition information from room temperature PL spectra is that the peak position depends on the excitation wavelength and intensity. The spectrum shifts to lower energy as the intensity is raised, and the peak position is at lower energy for the shorter-wavelength light. The explanation for this effect probably lies in local heating of the sample by the rather high intensities used. The heating is greater for higher-energy (blue) photons because a greater fraction of each photon's energy is converted to phonons.

LITERATURE CITED

Bebb, H. B. and Williams, E. W. 1972. Photoluminescence I: Theory. *In* Semiconductors and Semimetals, Vol. 8 (R. K. Willardson and A. C. Beer, eds.) pp. 183–217. Academic Press, New York.

Bhattacharya, P. 1994. Semiconductor Optoelectronic Devices. Prentice Hall, Englewood Cliffs, N.J.

Bimberg, D., Sondergeld, M., and Grobe, E. 1971. Thermal dissociation of excitons bound to neutral acceptors in high-purity GaAs. *Phys. Rev. B* 4:3451–3455.

Brunnett, B. A., Schlesinger, T. E., Toney, J. E., and James, A. B. 1999. Room-temperature photoluminescence mapping of CdZnTe detectors. *Proc. SPIE 3768*. In press.

Hopfield, J. J. 1964. The quantum chemistry of bound exciton complexes. *In* Proceedings of 7th International Conference on the Physics of Semiconductors (M. Hulin, ed.) pp. 725–735. Dunod, Paris.

Lee, J., Giles, N. C., Rajavel, D., and Summers, C. J. 1994b. Room-temperature band-edge photoluminescence from cadmium telluride. *Phys. Rev. B* 49:1668–1676.

Lee, J., Myers, T. H., Giles, N. C., Dean, B. E., and Johnson, C. J. 1994a. Optical quenching of bound excitons in CdTe and $Cd_{1-x}Zn_xTe$ alloys: A technique to measure copper concentration. *J. Appl. Phys.* 76:537–541.

Lightowlers, E. C. 1990. Photoluminescence characterization. *In* Growth and Characterization of Semiconductors (R. A. Stradling and P. C. Klipstein, eds.) pp. 135–163. Adam Hilger, Bristol, U.K.

Molva, E., Pautrat, J. L., Saminadayar, K., Milchberg, G., and Magnea, N. 1984. Acceptor states in CdTe and comparison with ZnTe. General trends. *Phys. Rev. B* 30:3344–3354.

Schmidt, T., Lischka, K., and Zulehner, W. 1992. Excitation-power dependence of the near-band-edge photoluminescence of semiconductors. *Phys. Rev. B* 45:8989–8994.

Stillman, G. E., Bose, S. S., and Curtis, A. P. 1991. Photoluminescence characterization of compound semiconductor optoelectronic materials. *In* Advanced Processing and Characterization Technologies, Fabrication and Characterization of Semiconductor Optoelectronic Devices and Integrated Circuits (P. H. Holloway, ed.) pp. 34–37. American Institute of Physics, Woodbury, N.Y.

Yu, P. Y. and Cadona, M. 1996. Fundamentals of Semiconductors. Springer-Verlag, Berlin.

KEY REFERENCES

Bebb and Williams, 1972. See above.

A thorough, mathematical treatment of the theory of PL.

Demtroder, W. 1996. Laser Spectroscopy, 2nd ed. Springer-Verlag, Berlin.

Contains a detailed discussion of instrumentation used in optical spectroscopy.

Perkowitz, E. 1993. Optical Characterization of Semiconductors. Academic Press, London.

Contains a good summary of the general principles of PL and some helpful case studies dealing with stress and impurity emission.

Peyghambarian, N., Loch, S. W., and Mysyrowicz, A. 1993. Introduction to Semiconductor Optics. Prentice-Hall, Englewood Cliffs, N.J.

A thorough, pedagogical introduction to optics of semiconductors with emphasis on properties of excitons.

Swaminathan, V. and Macrander, A. T. 1991. Materials Aspects of GaAs and InP Based Structures. Prentice-Hall, Englewood Cliffs, N.J.

Gives a fine, conceptual introduction to PL.

JAMES E. TONEY
Spire Corporation
Bedford, Massachusetts

ULTRAVIOLET AND VISIBLE ABSORPTION SPECTROSCOPY

INTRODUCTION

Ultraviolet and visible (UV-Vis) absorption spectroscopy is the measurement of the attenuation of a beam of light after it passes through a sample or after reflection from a sample surface. This unit will use the term UV-Vis spectroscopy to include a variety of absorption, transmittance, and reflectance measurements in the ultraviolet (UV), visible, and near-infrared (NIR) spectral regions. These measurements can be at a single wavelength or over an extended spectral range. As presented here, UV-Vis spectroscopy will be found under other names, including UV-Vis spectrometry, UV-Vis spectrophotometry, and UV-Vis reflectance spectroscopy. This unit provides an overview of the technique and does not attempt to provide a comprehensive review of the many applications of UV-Vis spectroscopy in materials research. In this regard, many of the references were chosen to illustrate the diversity of applications rather than to comprehensively survey the uses of UV-Vis spectroscopy. A unifying theme is that most of the measurements discussed in this unit can be performed with simple benchtop spectrometers and commercially available sampling accessories.

The UV-Vis spectral range is approximately 190 to 900 nm. This definition originates from the working range of a typical commercial UV-Vis spectrometer. The short-wavelength limit for simple UV-Vis spectrometers is the absorption of UV wavelengths <180 nm by atmospheric gases. Purging a spectrometer with nitrogen gas extends this limit to 175 nm. Working beyond 175 nm requires a vacuum spectrometer and a suitable UV light source and is not discussed in this unit. The long-wavelength limit is usually determined by the wavelength response of the detector in the spectrometer. Higher end commercial UV-Vis spectrometers extend the measurable spectral range into the NIR region as far as 3300 nm. This unit includes the use of UV-Vis-NIR instruments because of the importance of characterizing the NIR properties of materials for lasers, amplifiers, and low-loss optical fibers for fiber-optic communications and other applications. Table 1 summarizes the wavelength and energy ranges of the UV-Vis and related spectral regions.

Ultraviolet-visible spectroscopy is one of the more ubiquitous analytical and characterization techniques in science. There is a linear relationship between absorbance and absorber concentration, which makes UV-Vis spectroscopy especially attractive for making quantitative measurements. Ultraviolet and visible photons are energetic enough to promote electrons to higher energy states in molecules and materials. Figures 1 and 2 illustrate typical absorption spectra for the absorption processes for molecules and semiconductor materials, respectively. Therefore, UV-Vis spectroscopy is useful to the exploration of the electronic properties of materials and materials precursors in basic research and in the development of applied materials. Materials that can be characterized by UV-Vis spectroscopy include semiconductors for electronics, lasers, and detectors; transparent or partially transparent optical components; solid-state laser hosts; optical fibers, waveguides, and amplifiers for communication; and materials for solar energy conversion. The UV-Vis range also spans the range of human visual acuity of approximately 400 to 750 nm, making UV-Vis spectroscopy useful in characterizing the absorption, transmission, and reflectivity of a variety of technologically important materials, such as pigments, coatings, windows, and filters.

The use of UV-Vis spectroscopy in materials research can be divided into two main categories: (1) quantitative measurements of an analyte in the gas, liquid, or solid phase and (2) characterization of the optical and electronic properties of a material. The first category is most useful as a diagnostic tool for the preparation of materials, either to quantitate constituents of materials or their precursors or as a process method to monitor the concentrations of reactants or products during a reaction (Baucom et al., 1995; Degueldre et al., 1996). In quantitative applications it is often only necessary to measure the absorbance or reflectivity at a single wavelength. The second more

Table 1. Approximate Wavelengths, Energies, and Type of Excitation for Selected Spectral Regions

Spectral Region	Wavelength Range (nm)	Energy Range (cm^{-1})	Energy Range (eV)	Types of Excitation
Vacuum-UV	10–180	1×10^6–55,600	124–6.89	Electronic
UV	200–400	50,000–25,000	6.89–3.10	Electronic
Visible	400–750	25,000–13,300	3.10–1.65	Electronic
Near IR	750–2500	13,300–4000	1.65–0.496	Electronic, vibrational overtones
IR	2500–25,000	4000–400	0.496–0.0496	Vibrations, phonons

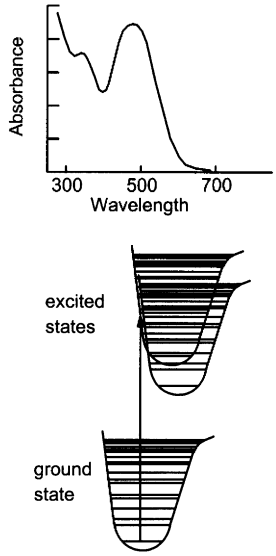


Figure 1. Typical molecular absorption spectrum and a schematic of the absorption process in a molecule.

qualitative application usually requires recording at least a portion of the UV-Vis spectrum for characterization of the optical or electronic properties of materials (Kovacs et al., 1997). Much of the discussion of UV-Vis spectroscopy will follow these two main applications of quantitative measurements and materials characterization.

Materials Properties Measurable by UV-Vis Spectroscopy

Electronic excitations in molecules usually occur in the UV and visible regions. Charge transfer bands can occur from the UV to the NIR. The band gap of a semiconductor depends on the specific material and physical dimensions and can range from the UV to NIR. A reduction in semiconductor particle size or dimensions (<10 nm) will shift the band edge to shorter wavelength (higher energy) due to quantum confinement effects (Bassani et al., 1997). The NIR light excites vibrational overtones and combination bands in molecules. Although these transitions are very weak, they become important in certain cases such as optical fibers in which a very long path length is of interest.

In organic molecules and polymers, the UV-Vis spectrum can help to identify chromophores and the extent of electronic delocalization (Yoshino et al., 1996). Similarly, absorption measurements can be correlated with bulk physical properties (Graebner, 1995). For inorganic complexes, the UV-Vis spectrum can provide information about oxidation states, electronic structure, and metal-ligand interactions. For solid materials, the UV-Vis spectrum can measure the band gap and identify any localized excitations or impurities (Banerjee et al., 1996).

Optical materials that rely on interference can be utilized throughout the UV-Vis-NIR spectral range. Collective excitations that determine the absorption or reflectivity of a material likewise occur throughout this range. For example, the peak absorption of the plasmon resonance absorption of thin films of aluminum, gold, and semiconductors occurs in the UV, visible, and NIR regions, respectively, For small metal particles, the plasmon resonance absorption also varies due to particle size (Ali et al., 1997).

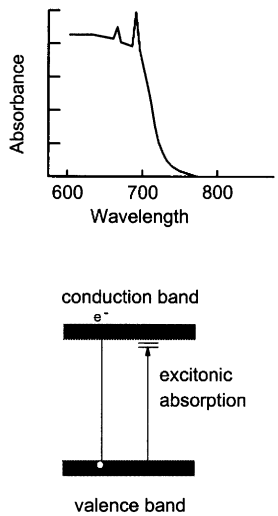


Figure 2. Typical absorption spectrum for a semiconductor and a schematic of the absorption process in a direct-bandgap semiconductor.

Competitive and Related Techniques for Qualitative and Quantitative Analysis

The spectrum of a molecule depends on its energy-level structure, and the UV-Vis absorption spectrum can sometimes be useful for identifying molecular compounds. However, UV-Vis spectra have broad features that are of limited use for sample identification. The broad lines do facilitate making accurate quantitative measurements. Lanthanide ions and some transition metals in molecular complexes have sharp UV-Vis bands and are easily identified by their absorption spectrum. Qualitative identification and structural analysis of molecules usually require multiple analytical methods, including C, H, and O elemental analysis; infrared (IR) absorption; Raman spectroscopy (see RAMAN SPECTROSCOPY OF SOLIDS); nuclear magnetic resonance (NMR) spectroscopy (see NUCLEAR MAGNETIC RESONANCE IMAGING); mass spectrometry; size exclusion chromatography; and x-ray diffraction (XRD, see X-RAY TECHNIQUES). The elemental composition of solids can be determined by x-ray fluorescence, microprobe techniques (see X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS), and surface spectroscopies such as x-ray photoelectron spectroscopy (XPS), Auger electron spectroscopy (AES, see AUGER ELECTRON SPECTROSCOPY), or secondary-ion mass spectrometry (SIMS). For crystalline or polycrystalline solids, x-ray (X-RAY TECHNIQUES), electron (ELECTRON TECHNIQUES), and neutron diffraction (NEUTRON TECHNIQUES) are the most reliable means of structure identification.

Quantitative measurements with benchtop UV-Vis spectrometers are most commonly performed on molecules and inorganic complexes in solution. Photoluminescence for fluorescent analytes is usually more sensitive than absorption measurements, because the fluorescence signal

is measured on a nominally zero-noise background. Absorbance measurements require that the change in light power due to absorption be greater than the noise of the light source. Quantitating gas phase species or analytes on surfaces or in thin films using UV-Vis spectrometers is possible for high concentrations or strongly absorbing species (Gregory, 1995). In practice, polymer or Langmuir-Blodgett films are more often measured using IR absorption or some type of laser spectroscopy (Al-Rawashdeh and Foss, 1997). In addition to UV-Vis reflectance spectroscopy, inorganic surfaces often require analysis with multiple spectroscopies such as Raman spectroscopy, extended x-ray absorption fine structure (EXAFS, see XAFS SPECTROSCOPY), XPS, AES, and SIMS.

Laser spectroscopic absorption techniques can provide a variety of advantages compared to conventional UV-Vis spectroscopy. Lasers have much higher intensities and can provide much shorter time resolution than modulated lamp sources for measuring transient phenomena. The recent commercialization of optical parametric oscillators (OPOs) alleviates the limited scan range of dye lasers. Cavity-ringdown laser absorption spectroscopy (Paul and Saykally, 1997; Scherer et al., 1997) and intracavity laser absorption (Stoeckel and Atkinson, 1985; Kachanov et al., 1997) are much more sensitive than measurements using UV-Vis spectrometers. Similarly, photoacoustic and photothermal spectroscopy can provide very high sensitivity absorption measurements of opaque samples and analytes on surfaces (Bialkowski, 1995; Childers et al., 1986; Palmer, 1993; Welsch et al., 1997). Nonlinear laser spectroscopies, such as second-harmonic generation (SHG), can be surface selective and provide enhanced selectivity and sensitivity for analytes on surfaces compared to absorption or fluorescent measurements (Di Bartolo, 1994).

Competitive and Related Techniques for Material Characterization

Because of the diverse applications to which UV-Vis spectroscopy can be applied, there are a multitude of related techniques for characterizing the electronic and optical properties of materials. Characterizing complex materials or processes often requires the results of multiple analytical techniques (see, e.g., Lassaletta et al., 1995; Takenaka et al., 1997; Weckhuysen et al., 1996). Electronic properties are often measured by a variety of methods, and UV-Vis spectral results are correlated with electrical measurements, surface spectroscopies, and electrochemical measurements (Ozer et al., 1996; Yoshino et al., 1996). Characterization of the electronic and optical properties of semiconductors often requires time-resolved laser spectroscopic methods, photoluminescence spectroscopy (see PHOTOLUMINESCENCE SPECTROSCOPY), and sophisticated modulation techniques such as piezospectroscopy (Ramdas and Rodriguez, 1992) and modulated photoreflectance (Glembocki and Shanabrook, 1992).

Measuring the transmittance or reflectance of most optical materials in the UV-Vis region rarely requires more sophisticated instrumentation than a commercial spectrometer. Measuring an extremely small absorbance can require some type of laser-based method. Measuring polarization properties such as optical rotation and circular dichroism of materials also requires more sophisticated techniques and is usually performed on dedicated polarimetry, circular dichroism, or ellipsometry instruments. Many technologically important materials, such as light-emitting diodes (LEDs), optical amplifiers, laser materials, and lamp and display phosphors, require characterization of the emission properties as well as the absorption properties. For these types of materials, measurements of the photoluminescence or electroluminescence spectrum and the fluorescence quantum efficiency are as important as characterizing the absorption spectrum and absorbance.

PRINCIPLES OF THE METHOD

Ultraviolet-visible spectroscopy measures the attenuation of light when the light passes through a sample or is reflected from a sample surface. The attenuation can result from absorption, scattering, reflection, or interference. Accurate quantitation requires that the measurement record the attenuation due only to absorption by the analyte, so the spectroscopic procedure must compensate for the loss of light from other mechanisms. The cause of the attenuation is often not important for many optical materials, and the total resulting transmittance or reflectance is sufficient to determine the suitability of a material for a certain application.

Experimental measurements are made in terms of transmittance T:

$$T = P/P_0 \tag{1}$$

where P is the radiant power (radiant energy on unit area in unit time) after it passes through the sample and P_0 is the initial radiant power. This relationship will also be found in terms of light intensities:

$$T = I/I_0 \tag{2}$$

Percent transmittance is simply $T \times 100\%$ ($\%T$). The parameters P and P_0 are not always well-defined and can depend on the UV-Vis application. Figures 3 and 4 illustrate the differences that are encountered in defining P and P_0 in quantitative and qualitative characterization measurements.

Quantitative Analysis

Within certain limits, the absorbance of light by a sample is directly proportional to the distance the light travels through the sample and to the concentration of the absorbing species. This linear relationship is known as the Beer-Lambert law (also called the Beer-Lambert-Bouguer law or simply Beer's law) and allows accurate concentration measurements of absorbing species in a sample. The general Beer-Lambert law is usually written as

$$A = a \times b \times c \tag{3}$$

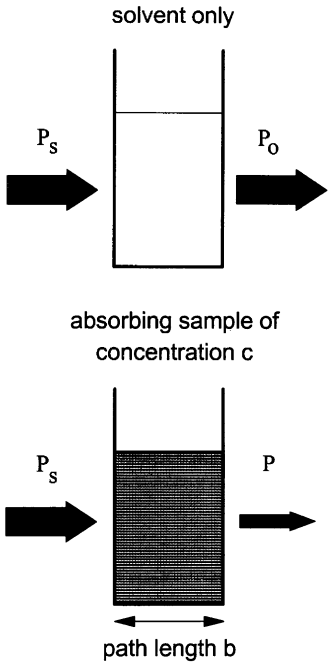


Figure 3. Schematic of a transmittance measurement for the quantitative determination of an analyte in solution.

where A is the measured absorbance, a is a wavelength-dependent absorptivity coefficient, b is the path length, and c is the analyte concentration. When working in concentration units of molarity, the Beer-Lambert law is written as

$$A = \varepsilon \times b \times c \tag{4}$$

where ε is the wavelength-dependent molar absorptivity coefficient with units of reciprocal molarity per centimeter and b has units of centimeters. If multiple species that absorb light at a given wavelength are present in a sample, the total absorbance at that wavelength is the sum due to all absorbers:

$$A = (\varepsilon_1 \times b \times c_1) + (\varepsilon_2 \times b \times c_2) + \cdots \tag{5}$$

where the subscripts refer to the molar absorptivity and concentration of the different absorbing species that are present.

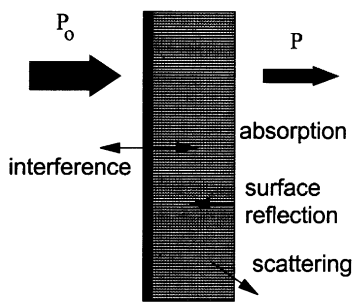


Figure 4. Schematic of a transmittance measurement of an optical component such as a dielectric-coated filter.

The linear relationship of the Beer-Lambert law depends on several conditions. The incident radiation must be monochromatic and collimated through a sample. Samples, including calibration standards, must be homogeneous without irreproducible losses due to reflection or scattering. For analytes in a matrix, such as a solution or a glass, the absorber concentration must be low enough so that the absorbing species do not interact. Interactions between absorbing species can lead to deviations in the absorptivity coefficient (at a given wavelength) as a function of concentration. More details on the most common limitations are described below (see Problems).

The relationship between absorbance A and the experimentally measured transmittance T is

$$A = -\log T = -\log(P/P_0) \tag{6}$$

where T and A are both unitless. Absorption data and spectra will often be presented using A, T, $\%T$, or $1 - T$. An absorption spectrum that uses $1 - T$ versus wavelength will appear similar to a plot of absorbance A versus wavelength, but the quantity $1 - T$ does not directly correspond to absorber concentration as does A.

Figure 3 shows the quantitative measurement of P and P_0 for an analyte in solution. In this example, P_s is the source light power that is incident on a sample, P is the measured light power after passing through the analyte, solvent, and sample holder, and P_0 is the measured light power after passing through only the solvent and sample holder. The measured transmittance in this case is attributed to only the analyte.

If the absorptivity coefficient of an absorbing species is not known, an unknown concentration can be determined using a working curve of absorbance versus concentration derived from a set of standards of known concentration. Calibration with standards is almost always necessary for absorbance measurements made in a reflectance geometry or with a fiber-optic light delivery system.

Material Characterization

Figure 4 illustrates the measurement of P and P_0 to determine the transmittance of a dielectric coating on a substrate. This example is a typical measurement to characterize the optical properties of a material such as a window, filter, or other optical component. In this case P is the light power after the sample and P_0 is the light power before the sample. Figure 5 shows two examples of transmittance spectra for optical components. The transmittance of the colored glass filter (top spectrum) does not reach 1.0 at wavelengths >600 nm due to surface reflection losses. The oscillations in the bottom spectrum arise from the interference origin of the transmittance properties. In these examples, the measured transmittance is due to all sources of light attenuation, which for practical applications is often unimportant compared to knowing the resulting transmittance at any given wavelength. These example spectra show the transmittance spectrum for an incidence angle of 0° through an optical component. Depending on the application, the transmittance or reflectance spectrum as a

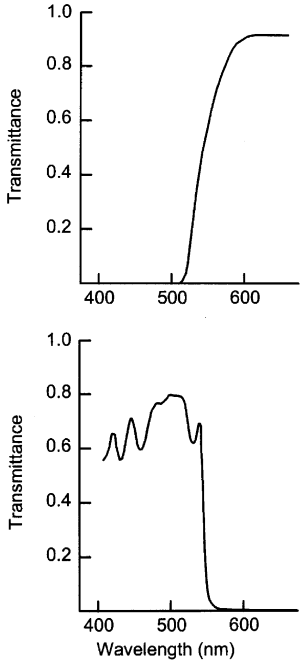


Figure 5. Transmittance spectra of a colored glass filter (top) and a dielectric-coated interference filter (bottom).

function of the angle the light beam makes with the optical component might be of more interest.

PRACTICAL ASPECTS OF THE METHOD

General Considerations

Most commercial UV-Vis absorption spectrometers use one of three overall optical designs: a scanning spectrometer with a single light beam and sample holder, a scanning spectrometer with dual light beams and dual sample holders for simultaneous measurement of P and P_0, or a nonscanning spectrometer with an array detector for simultaneous measurement of multiple wavelengths. Most commercial instruments use a dispersive design, but Fourier transform UV-Vis spectrometers find applications in high-speed, low-resolution and other specialized applications (Thorne, 1991). In single- and dual-beam spectrometers, the light from a lamp is dispersed before reaching the sample cell. In an array detector instrument, all wavelengths pass through the sample and the dispersing element is between the sample and the array detector. The specific designs and optical components vary widely depending on the required level of spectrometer performance. The following paragraph describes common features of the UV-Vis instruments followed by descriptions of the three different designs and sampling options. Specific details about UV-Vis instruments can be found in the manufacturers' literature.

Common Spectrometer Components

The light source is usually a deuterium discharge lamp for UV measurements and a tungsten-halogen lamp for visible and NIR measurements. The instruments automatically swap lamps when scanning between the UV and visible regions. The wavelengths of these continuous light sources are typically dispersed by a holographic grating in a single or double monochromator or spectrograph. The spectral bandpass is then determined by the monochromator slit width or by the array element width in array detector spectrometers. Spectrometer designs and optical components are optimized to reject stray light, which is one of the limiting factors in quantitative absorbance measurements. The detector in single-detector instruments is a photodiode, phototube, or photomultiplier tube (PMT). The UV-Vis-NIR spectrometers utilize a combination of a PMT and a Peltier-cooled PbS IR detector. The light beam is redirected automatically to the appropriate detector when scanning between the visible and NIR regions, and the diffraction grating and instrument parameters, such as slit width, can also change automatically.

Single-Beam Spectrometer

The simplest UV-Vis spectrometer design has one light beam and uses one sample cell. These instruments are simple and relatively inexpensive and are useful for routine quantitative measurements. Making a quantitative measurement requires calibrating the instrument at the selected wavelength for 0 and 100% transmittance (or a more narrow transmittance range if accurate standards are available). As in Figure 3, the 100% calibration is done with the sample cell containing only the solvent for quantitative measurements in solution. These measurements are stored by the instrument, and the spectrometer readout can also display in absorbance units (A.U.).

Absorbance and reflectivity measurements can be made with discrete light sources without dispersing optics in specialized applications. One example of a very small and robust optical spectrometer is the Mars Oxidant Experiment, which was developed for the Mars96 mission. This spectrometer consisted of LED light sources, a series of thin-film sensors, a photodiode array detector, detection electronics, and optical fibers to get light from an LED to a sensor and from that sensor to a pixel of the array detector. A variety of sensor films provide a measure of the oxidizing power of the Martian surface and atmosphere. The reflectivity of the sensors changed depending on the chemical environment, as recorded by the attenuation of the LED power reaching the detector (Grunthaner et al., 1995).

Dual-Beam Spectrometer

Spectra can be recorded with a single-beam UV-Vis instrument by manually recording absorbance measurements at different wavelength settings. Both a reference and the sample must be measured at each wavelength, which makes recording a wide spectral range or a small step size a tedious procedure. The dual-beam design greatly simplifies this process by simultaneously measuring P and P_0 of the sample and reference cells, respectively. Most spectrometers use a mirrored rotating chopper wheel to alternately direct the light beam through the sample and reference cells. The detection electronics or software program can then manipulate the P and P_0 values as the

wavelength scans to produce the spectrum of absorbance or transmittance as a function of wavelength.

Array Detector Spectrometers

In a large number of applications absorbance spectra must be recorded very quickly, such as in process monitoring or kinetics experiments. Dispersing the light after it passes through a sample allows the use of an array detector to simultaneously record the transmitted light power at multiple wavelengths. These spectrometers use photodiode arrays (PDAs) or charge-coupled devices (CCDs) as the detector. The spectral range of these array detectors is typically 200 to 1000 nm. Besides allowing rapid spectral recording, these instruments are relatively small and robust and have been developed into PC-card portable spectrometers that use optical fibers to deliver light to and from a sample. These instruments use only a single light beam, so a reference spectrum is recorded and stored in memory to produce transmittance or absorbance spectra after recording the raw sample spectrum.

Sample Handling

Most benchtop UV-Vis spectrometers contain a 10- to 200-mm-wide sample compartment through which the collimated light beam(s) pass. Standard mounts in this sample compartment hold liquid sample cuvettes or optical components and can include thermostatted temperature regulation. Custom mounts can hold transparent solid samples of arbitrary shapes. The standard sample holder is a 10-mm-path-length quartz cuvette, which can be purchased in matched sets for use in dual-beam spectrometers. Microcuvettes and capillary cuvettes hold less than 10 μL of sample for limited amounts of liquid samples. These microcuvettes have shorter path lengths than the standard sample cells, and the smaller volume can introduce more error into a quantitative measurement. Sample cells and mounts with path lengths up to 100 mm are also available. Measuring longer path lengths requires conventional or fiber-optic accessories to deliver and re-collect the light beam outside of the sample compartment.

For samples that are opaque or strongly absorbing, the light absorption must be measured in a reflectance geometry. These measurements are done with reflectance accessories of various designs that fit internally or externally to the sample compartment. Variable-angle specular reflectance measurements can measure bulk optical properties of thin films, such as the refractive index or film thickness (Lamprecht et al., 1997; McPhedran et al., 1984; Larena et al., 1995). For quantitative or spectral characterization, diffuse reflectance spectra are recorded with an integrating sphere. A variation of a reflectance sampling method is attenuated total reflectance (ATR). In ATR the light beam travels through a waveguide or crystal by total internal reflection. The evanescent wave of the light extends beyond the surface of the waveguide and is attenuated by strongly absorbing samples on the outer surface of the waveguide. An ATR accessory can be used remotely by transferring the light to and from the ATR waveguide with optical fibers.

The use of optical fibers, or optical fiber bundles, allows the delivery of a light beam to and from a sample that is remote from the spectrometer sample compartment. This delivery mechanism is extremely useful for process control or to make remote measurements in clean rooms or reaction chambers. These accessories are usually used without a reference sample and can be used with single-beam, dual-beam, or array detector instruments. The transmission properties of optical fibers can limit the wavelength range of a single spectral scan, and complete coverage of the UV-Vis region can require measurements with multiple fibers. The delivery and return fibers or bundles can be arranged in a straight-through transmission geometry, a folded sample compartment geometry that uses a mirror in a compact probe head, or a reflectance mode to measure nontransparent samples.

METHOD AUTOMATION

Most modern dual-beam UV-Vis absorption spectrometers include an integrated microprocessor or are controlled by a software program running on a separate microcomputer. Spectrometer parameters, such as slit width (spectral bandpass), scan range, scan step size, scan speed, and integration time, are set by the program. These parameters can be stored to disk as a method and recalled later to make the same types of measurements. Because UV-Vis spectroscopy is often used for repetitive measurements, instrument manufacturers offer a variety of automated sample-handling accessories. These automated accessories are controlled by the software programs, which also automatically store the data to disk after measurement. Most of these accessories are designed for liquid-containing sample cells, but some of the accessories can be adapted for measuring solid samples. Automated cell changers and multicell transporters can cycle eight to sixteen samples through the spectrometer. These units are useful for repetitive measurements on a small number of samples or to follow spectral changes in a set of multiple samples. As an example, the automated cell transporter can cycle through a set of systematically varied samples to determine the reaction kinetics as a function of reactant concentrations or reaction conditions.

Sipper systems use a peristaltic pump to deliver liquid sample to the sample cell in the spectrometer. A sipper system can automate repetitive measurements and also reduces the error in absorbance measurements due to irreproducibility in positioning the sample cell. Autosamplers are available to automatically measure large numbers of samples. An autosampler can use a sipper system to sequentially transfer a series of samples to the sample cell with intermediate rinsing or to sequentially place a fiber-optic probe into a set of sample containers. Autosamplers are used for repetitive measurements on the order of 100 to 200 different liquid samples.

DATA ANALYSIS AND INITIAL INTERPRETATION

Quantitative Analysis

An unknown concentration of an analyte can be determined by measuring the amount of light that a sample

absorbs and applying the Beer-Lambert law as described above (see Principles of the Method). Modern UV-Vis absorption spectrometers will display the spectral data in a variety of formats, with the most common data presentation being in transmittance, percent transmittance, absorbance, reflectance, or percent reflectance (%R). For quantitative measurements, instrument control programs will store measurements of standards and automatically calculate the concentrations of unknowns from the absorbance measurement, including multicomponent mixtures. Spectra will usually be plotted versus wavelength (in angstroms, nanometers, or micrometers) or energy (in electron volts, or wavenumbers, in reciprocal centimeters). The data can be processed and displayed after a variety of processing methods, such as first to fourth derivative, the Kubelka-Munk function for turbidity measurements (Vargas and Niklasson, 1997), Savitzky-Golay smoothing (Savitzky and Golay, 1964; Barak, p. 1995), peak fitting, and others.

Standards are available from the National Institute of Standards and Technology (NIST) Standards Reference Materials Program to calibrate the transmittance, absorbance, wavelength, and reflectance of UV-Vis spectrometers. Many of the instrument manufacturers offer calibration standards and validation procedures for their UV-Vis spectrometers, as do third-party vendors (see the *Anal. Chem. Labguide* listed in Key References for more information). Table 4 in Dean (1995) gives the percent transmittance and absorbance values from 220 to 500 nm for a laboratory-prepared K_2CrO_4 standard solution, which is useful for checking the calibration of UV-Vis spectrometers. This handbook also lists other useful data such as the UV cutoff of common solvents and UV-Vis absorption methods for the analysis of metals and nonmetals (Dean, 1995).

Material Characterization

Several manufacturers offer software packages for specific applications. Many of these programs are for biological applications, but they can be adapted to materials work. For example, an enzyme kinetics program that works in conjunction with an automated multicell holder can be adapted for monitoring the kinetics of the solution-phase synthesis of materials. Software modules are also available to automatically calculate the refractive index or thickness of a thin film from the interference pattern in transmittance or reflectance measurements. Another materials application is the characterization of color, including analysis of the CIE (1986) tristimulus values in diffuse reflectance measurements (e.g., CHROMA software package from Unicam UV-Visible Spectrometry). These colorimetry standards are also applicable in transmission experiments (Prasad et al., 1996).

Two examples are presented here to illustrate the use of UV-Vis spectrometry for materials characterization. In the first example, Figure 6 shows the UV-Vis absorption spectra of two samples of CdS quantum dots (Tsuzuki and McCormick, 1997).

These materials were prepared by a mechanochemical reaction:

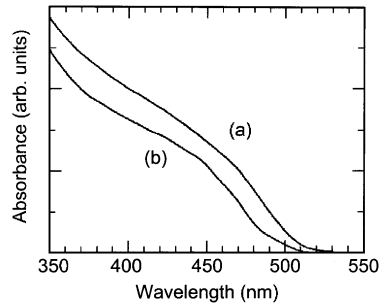


Figure 6. Absorption spectra of CdS quantum dots produced using 12.7-mm balls (curve *a,* average dot diameter 8.3 nm) and 4.8-mm balls (curve *b*, 4.3 nm average dot diameter). (Adapted with permission from Tsuzuki and McCormick, 1997. Copyright © 1997 by Springer-Verlag.)

$$CdCl_2 + Na_2S \rightarrow CdS + 2NaCl \qquad (7)$$

After reaction in a ball mill for 1 h, the NaCl is removed by washing and the CdS quantum dots are dispersed in $(NaPO_3)_6$ solution to obtain an absorption spectrum. Curve *a* in Figure 6 is the spectrum of 8.3-nm-diameter quantum dots produced using 12.7-mm balls in the ball mill, and curve *b* is the spectrum of 4.3-nm quantum dots produced using 4.8-mm balls. The smaller ball size produces lower collision energies, which results in smaller crystallite size. The absorption threshold is taken as the inflection point in the spectra. The 8.3-nm CdS has an absorption threshold of 510 nm, which is close to the threshold of 515 nm for bulk CdS. The blue shift due to quantum confinement is very evident for the 4.3-nm quantum dots, which exhibits an absorption threshold of 470 nm. Determining the absorption edge provides a convenient means of monitoring particle size for semiconductor and metal particles (Ali et al., 1997).

In the second example, Figure 7 shows reflectance spectra of GaN films under different uniaxial stresses in the *c* plane (Yamaguchi et al., 1997). The spectra show a narrow

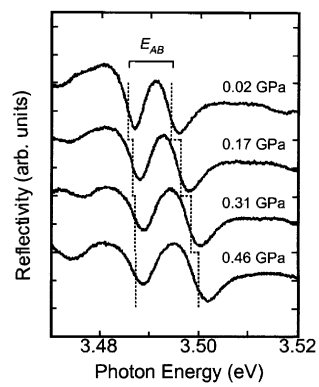


Figure 7. Reflectance spectra of GaN films under different uniaxial stresses in the *c* plane. The dashed lines indicate exciton energies as determined by theoretical fitting. The symbol E_{AB} refers to the energy separation between the A and B excitons. (Adapted with permission from Yamaguchi et al., 1997. Copyright © 1997 by the American Institute of Physics.)

spectral region (3.47 eV = 357.3 nm, 3.52 eV = 352.2 nm) that encompasses two of three dispersive-shaped exciton absorption bands that occur near the valence band maximum. The closely spaced excitons lead to a large density of states at the valence band maximum, which results in a high threshold current for blue GaN laser diodes. Theory predicts that breaking the symmetry in the c plane should increase the energy splitting of the A and B bands and decrease the density of states. The dashed lines in Figure 7 show that the energy splitting does increase with increasing stress. The dependence of E_{AB} on the uniaxial stress, with other polarization measurements, allowed the authors to determine deformation potentials to be $D_4 = -3.4$ eV and $D_5 = -3.3$ eV. These results confirm the theoretical predictions that uniaxial strain should reduce the density of states and ultimately improve the performance of GaN laser diodes by reducing the lasing threshold current.

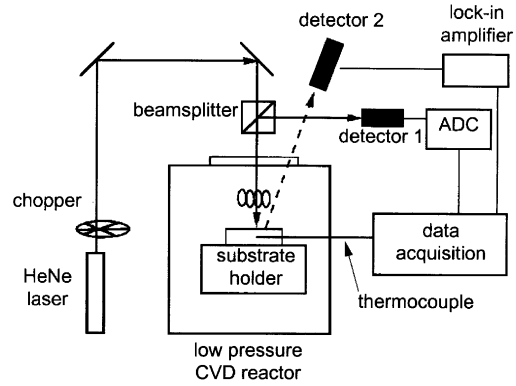


Figure 8. Experimental schematic of a CVD chamber that uses reflectivity (detector 1) and scattering (detector 2) to monitor diamond film growth on a Ni substrate. (Adapted with permission from Yang et al., 1997. Copyright © 1997 by the American Institute of Physics.)

SAMPLE PREPARATION

Sample preparation for UV-Vis spectroscopy is usually straightforward. Absorbing analytes that are in solution or can be dissolved are measured in an appropriate solvent or pH buffer. The pH and competing equilibria must be controlled so that the absorbing species are present in the same form in all samples and standards. Metals and nonmetals are quantitated by adding reagents that produce absorbing complexes in solution (Dean, 1995). To avoid saturation of the absorbance, strongly absorbing samples must be diluted, measured in a reflectance or ATR geometry, or measured in a short-path-length cell (Degiorgi et al., 1995). Because UV-Vis absorption bands are typically broad, quantitative analysis requires that a sample contain only one or a few absorbing species. Very complex mixtures usually require separation before making quantitative measurements. Similarly, particulates that scatter light should be removed by centrifugation or filtration to avoid interfering with the absorbance measurement.

Solid samples require clean and polished surfaces to avoid contaminant and scattering losses. If the solid has nonparallel surfaces, it will act as a prism and bend the light beam. For regular shaped samples, a compensating optic can redirect the light beam to its original path. Fiber-optic light delivery and collection provides a simple means of directing the light through irregular shaped samples. Preparing samples for reflectance measurements can be much more difficult than transparent samples. The reflectivity of a surface or film can depend on a variety of material parameters, including the presence of absorbing species, surface roughness, particle size, and packing density (Childers et al., 1986).

Specular reflectance is usually treated similarly to transmittance measurements, i.e., as the fraction or percentage of reflected light power compared to the incident light power. Diffuse reflectance is usually described as the relative reflectance r_∞, where $r_\infty = r_{\text{sample}}/r_{\text{reference}}$. The absorptive and scattering components can be extracted from the relative reflectance using the Kubelka-Munk equation:

$$\frac{k}{s} = \frac{(1 - r_\infty)^2}{2r_\infty} \tag{8}$$

where k is the absorption coefficient and s is the scattering coefficient (Childers et al., 1986).

The reflectivity dependence on morphology and composition can be used to monitor the growth or processing of a thin film or surface. Figure 8 shows an experimental layout to use reflectivity and scattering to monitor diamond film growth on Ni in a low-pressure chemical vapor deposition (CVD) chamber (Yang et al., 1997). The 632.8-nm beam of a HeNe laser is incident on the substrate/film surface, and detector 1 monitors reflectivity and detector 2 monitors scattered light. The optical monitoring used bandpass interference filters in front of the detectors and a chopped light source with phase-sensitive detection to discriminate against the large background light emission from the hot filament. Figure 9 shows typical changes in reflected and scattered light as the substrate temperature increases. Correlation of these curves with scanning electron micrographs indicated that the time to change the reaction conditions from pretreatment of the diamond seeds to diamond growth should occur immediately after the drop in scattered light intensity.

SPECIMEN MODIFICATION

Specimen modification rarely occurs in commercial UV-Vis spectrometers due to the fairly low intensity of UV and visible light that passes through or strikes the sample. Some samples can be very susceptible to photodegradation; obvious examples are photocurable polymers and photoresist materials. To minimize photochemical degradation, most instruments disperse the light before it passes through the sample, so that wavelengths that are not being measured are also not irradiating the sample. Photodegradation is more problematic in diode array and

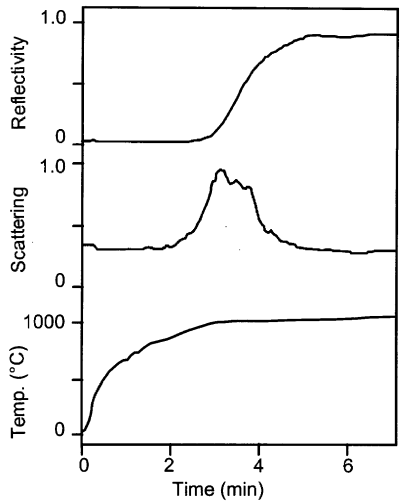


Figure 9. Typical changes in reflectivity (top curve, arbitrary units) and scattering (middle curve, arbitrary units) of a HeNe laser beam from a growing diamond film as the substrate temperature increases (bottom curve). (Adapted with permission from Yang et al., 1997. Copyright © 1997 by the American Institute of Physics.)

Fourier transform spectrometers since all wavelengths of light pass through the sample before dispersion and detection. Photosensitive samples can be protected to some degree by placing a filter between the light source and the sample to block shorter wavelengths that are not of interest in the measurement.

PROBLEMS

The problems that can arise in UV-Vis spectroscopy can be divided into sample problems and instrumental limitations. These problems mostly affect the linear dependence of absorbance versus concentration in quantitative measurements, but in severe cases they can alter the qualitative nature of a UV-Vis spectrum. Sample problems can include the presence of absorbing or scattering impurities and degradation of very photochemically active molecules or materials. Very intense fluorescence from the analyte or other sample components is usually not a problem in scanning spectrometers but can cause problems in array detector spectrometers.

Problems due to chemical effects such as analyte-solvent interaction, pH effects, and competing equilibria can be minimized by making measurements of all samples under identical conditions. Quantitative measurements are usually calibrated with external or internal standards, rather than relying only on published absorptivity coefficients. For cases where the chemical behavior varies in different samples, the measurement can be calibrated using the standard addition method (Ingle and Crouch, 1988).

The two most common instrumental factors that affect quantitative measurements are the amount of stray light reaching the detector and the non-monochromatic nature of the light beam. The amount of stray light depends on the quality of the spectrometer components and the shielding of the sample compartment or fiber-optic probe. It is usually fixed for a given instrument, although it can vary significantly between the UV-Vis and NIR spectral regions. Maintaining measurement accuracy of approximately 1% requires stray light of $<0.01\%$ (Ingle and Crouch, 1988). The effects of stray light can also be minimized by measuring analytes at low concentration so that P and P_0 remain large compared to the stray light power. Rule-of-thumb ranges for quantitative absorbance measurements are 0.1 to 2.5 A.U. for a high-quality spectrometer and 0.1 to 1.0 A.U. for a moderate-quality spectrometer (Ingle and Crouch, 1988; also see manufacturer's literature).

The unavoidable use of a fixed bandpass of light, which is non-monochromatic, results in a detector response due to multiple wavelengths around the central wavelength at which the spectrometer is set. The total amount of light reaching the detector will depend on the absorptivities of all wavelengths within the spectral bandpass and will vary nonlinearly with analyte concentration. This problem is eliminated by making measurements at a level part in the absorption spectrum with a spectrometer bandpass that is small compared to the width of the absorption line. The rule-of-thumb for accurate measurements is to use a spectrometer bandpass-to-linewidth ratio of approximately 0.1 (Ingle and Crouch, 1988). This criterion is easily met for most molecular absorption bands, which tend to be broad. Recording spectra with a bandpass that is larger than the absorption linewidths can drastically alter the qualitative appearance of a spectrum.

ACKNOWLEDGMENTS

The author thanks Professor Paul Deck for critically reading the manuscript and gratefully acknowledges support from a National Science Foundation Career Award (CHE-9502460) and a Research Corporation Cottrell Scholars Award.

LITERATURE CITED

Ali, A. H., Luther, R. J., and Foss, C. A., Jr. 1997. Optical properties of nanoscopic gold particles adsorbed at electrode surfaces: The effect of applied potential on plasmon resonance absorption. *Nanostruct. Mater.* 9:559–562.

Al-Rawashdeh, N. and Foss, C. A., Jr. 1997. UV/visible and infrared spectra of polyethylene/nanoscopic gold rod composite films: Effects of gold particle size, shape and orientation. *Nanostruct. Mater.* 9:383–386.

Banerjee, S., Chatterjee, S., and Chaudhuri, B. K. 1996. Optical reflectivity of superconducting $Bi_{3.9}Pb_{0.1}Sr_3Ca_3Cu_4O_x$ obtained by glass-ceramic route in 0.6–6.2 eV range. *Solid State Commun.* 98:665–669.

Barak, P. 1995. Smoothing and differentiation by an adaptive-degree polynomial filter. *Anal. Chem.* 67:2758–2762.

Bassani, F., Arnaud d'Avitaya, F., Mihalcescu, I., and Vial, J. C. 1997. Optical absorption evidence of quantum confinement in Si/CaF_2 multilayers grown by molecular beam epitaxy. *Appl. Surf. Sci.* 117: 670–676.

Baucom, K. C., Killeen, K. P., and Moffat, H. K. 1995. Monitoring of MOCVD reactants by UV absorption. *J. Electron. Mater.* 24: 1703–1706.

Bialkowski, S. E. 1995. Photothermal Spectroscopy Methods for Chemical Analysis. John Wiley & Sons, New York.

Childers, J. W., Rohl, R., and Palmer, R. A. 1986. Direct comparison of the capabilities of photoacoustic and diffuse reflectance spectroscopies in the ultraviolet, visible, and near-infrared regions. *Anal. Chem.* 58:2629–2636.

Commission Internationale de l'Eclairage (CIE, International Commission on Illumination). 1986. Colorimetry, 2nd ed. Publication CIE 15.2-1986. CIE, Vienna, Austria.

Dean, J. A. 1995. Analytical Chemistry Handbook. McGraw-Hill, New York.

Degiorgi, E., Postorino, P., and Nardone, M. 1995. A cell of variable thickness for optical studies of highly absorbing liquids. *Meas. Sci. Technol.* 6:929–931.

Degueldre, C., O'Prey, S., and Francioni, W. 1996. An in-line diffuse reflection spectroscopy study of the oxidation of stainless steel under boiling water reactor conditions. *Corros. Sci.* 38:1763–1782.

Di Bartolo, B. and Bowlby, B. (eds). 1994. Nonlinear spectroscopy of solids: Advances and applications. *In:* NATO ASI Series B*:* Physics, Vol. 339. Plenum, New York.

Glembocki, O. J. and Shanabrook, B. V. 1992. Photoreflectance spectroscopy of microstructures. *In* The Spectroscopy of Semiconductors (D. G. Seiler and C. L. Littler, eds.). pp. 221–292. Academic Press, San Diego.

Graebner, J. E. 1995. Simple correlation between optical absorption and thermal conductivity of CVD diamond. *Diam. Relat. Mater.* 4: 1196–1199.

Gregory, N. W. 1995. UV-vis vapor absorption spectrum of antimony(III) chloride, antimony(V) chloride, and antimony(III) bromide. The vapor pressure of antimony(III) bromide. *J. Chem. Eng. Data* 40:963–967.

Grunthaner, F. J., Ricco, A. J., Butler, M. A., Lane, A. L., McKay, C. P., Zent, A. P., Quinn, R. C., Murray, B., Klein, H. P., Levin, G. V., Terhune, R. W., Homer, M. L., Ksendzov, A., and Niedermann, P. 1995. Investigating the surface chemistry of Mars. *Anal. Chem.* 67:605A–610A. (Author's note: Unfortunately the Mars96 orbiter did not arrive at Mars.)

Ingle, J. D. and Crouch, S. R. 1988. Spectrochemical Analysis. Prentice-Hall, New York.

Kachanov, A. A., Stoeckel, F., and Charvat, A. 1997. Intracavity laser absorption measurements at ultrahigh spectral resolution. *Appl. Opt.*36:4062–4068.

Kovacs, L., Ruschhaupt, G., and Polgar, K. 1997. Composition dependence of the ultraviolet absorption edge in lithium niobate. *Appl. Phys. Lett.* 70:2801–2803.

Lamprecht, K., Papousek, W., and Leising, G. 1997. Problem of ambiguity in the determination of optical constants of thin absorbing films from spectroscopic reflectance and transmittance measurements. *Appl. Opt.* 36:6364–71.

Larena, A., Pinto, G., and Millan, F. 1995. Using the Lambert-Beer law for thickness evaluation of photoconductor coatings for recording holograms. *Appl. Surf. Sci.* 84:407–411.

Lassaletta, G., Fernandez, A., and Espinos, J. P. 1995. Spectroscopic characterization of quantum-sized TiO_2 supported on silica: Influence of size and TiO_2-SiO_2 interface composition. *J. Phys. Chem.* 99:1484–1490.

McPhedran, R. C., Botten, L. C., and McKenzie, D. R. 1984. Unambiguous determination of optical constants of absorbing films by reflectance and transmittance measurements. *Appl. Opt.* 23:1197–1205.

Ozer, N., Rubin, M. D., and Lampert, C. M. 1996. Optical and electrochemical characteristics of niobium oxide films prepared by sol-gel process and magnetron sputtering. A comparison. *Sol. Energ. Mat. Sol. Cells* 40:285–296.

Palmer, R. A. 1993. Photoacoustic and photothermal spectroscopies. In Physical Methods of Chemistry, 2nd ed., Vol. 8 (Determination of Electronic and Optical Properties) (B. W. Rossiter and R. C. Baetzold, eds.) pp. 61–108. John Wiley & Sons, New York.

Paul, J. B. and Saykally, R. J. 1997. Cavity ringdown laser absorption spectroscopy. *Anal. Chem.* 69:287A–292A.

Prasad, K. M. M. K., Raheem, S., Vijayalekshmi, P., and Kamala Sastri, C. 1996. Basic aspects and applications of tristimulus colorimetry. *Talanta* 43:1187–1206.

Ramdas, A. K. and Rodriguez, S. 1992. Piezospectroscopy of semiconductors. *In* The Spectroscopy of Semiconductors (D. G. Seiler and C. L. Littler, eds.) Academic Press, San Diego.

Savitzky A. and Golay, M. J. E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36:1627–1639.

Scherer, J. J., Paul, J. B., O'Keefe, A., and Saykally, R. J. 1997. Cavity ringdown laser absorption spectroscopy: History, development, and application to pulsed molecular beams. *Chem. Rev.* 97:25–51.

Stoeckel, F. and Atkinson, G. H. 1985. Time evolution of a broadband quasi-cw dye laser: Limitations of sensitivity in intracavity laser spectroscopy. *Appl. Opt.* 24:3591–3597.

Takenaka, S., Tanaka, T., and Yamazaki, T. 1997. Structure of active species in alkali-ion-modified silica- supported vanadium oxide. *J. Phys. Chem. B* 101:9035–9040.

Thorne, A. P. 1991. Fourier transform spectrometry in the ultraviolet. *Anal. Chem.* 63:57A–65A.

Tsuzuki, T. and McCormick, P. G. 1997. Synthesis of CdS quantum dots by mechanochemical reaction. *Appl. Phys. A* 65:607–609.

Vargas, W. E. and Niklasson, G. A. 1997. Applicability conditions of the Kubelka-Munk theory. *Appl. Opt.* 36:5580–5586.

Weckhuysen, B. M., Wachs, I. E., and Schoonheydt R. A. 1996. Surface chemistry and spectroscopy of chromium in inorganic oxides. *Chem. Rev.* 96:3327–3349.

Welsch, E., Ettrich, K., and Blaschke, H. 1997. Investigation of the absorption induced damage in ultraviolet dielectric thin films. *Opt. Eng.* 36:504–514.

Yamaguchi, A. A., Mochizuki, Y., Sasaoka, C., Kimura, A., Nido, M., and Usui, A. 1997. Reflectance spectroscopy on GaN films under uniaxial stress. *Appl. Phys. Lett.* 71:374–376.

Yang, P. C., Schlesser, R., Wolden, C. A., Liu, W., Davis, R. F., Sitar, Z., and Prater, J. T. 1997. Control of diamond heteroepitaxy on nickel by optical reflectance. *Appl. Phys. Lett.* 70:2960–2962.

Yoshino, K., Tada, K., Yoshimoto, K., Yoshida, M., Kawai, T., Zakhidov, A., Hamaguchi, M., and Araki, H. 1996. Electrical and optical properties of molecularly doped conducting polymers. *Syn. Metals* 78:301–312.

KEY REFERENCES

Analytical Chemistry. American Chemical Society.

This journal publishes several special issues that serve as useful references for spectroscopists. The June 15 issue each year is a set of reviews, and the August 15 issue is a buyer's guide

(Labguide) of instruments and supplies. The most relevant reviews to this unit are Coatings and Surface Characterization reviews in the Application Reviews issue (published odd years) and Ultraviolet and Light Absorption Spectrometry review in the Fundamental Reviews issue (published even years). An on-line version of the Labguide is currently available at http://pubs.acs.org/labguide/.

Demtröder, W. 1996. Laser Spectroscopy: Basic Concepts and Instrumentation, 2nd ed. Springer-Verlag, Berlin.

A graduate-level text on laser spectroscopy, including theory, laser-spectroscopic techniques, and instrumentation. Provides an overview of advanced absorption spectroscopies and competitive techniques such as photoacoustic spectroscopy.

Hollas, J. M. 1996. Modern Spectroscopy, 3rd ed. John Wiley & Sons, New York.

A graduate-level text that covers the principles and instrumentation of spectroscopy. Specific topics include quantum mechanics and interaction of light and matter, molecular symmetry, rotational spectroscopy, electronic spectroscopy, vibrational spectroscopy, photoelectron and surface spectroscopies, and laser spectroscopy.

Ingle and Crouch, 1988. See above.

A graduate-level text that covers the principles and instrumentation of spectroscopy. Specific techniques include atomic absorption and emission spectroscopy, molecular absorption and fluorescence spectroscopy, infrared absorption spectroscopy, and Raman scattering.

Settle, F. A. 1997. Handbook of Instrumental Techniques for Analytical Chemistry. Prentice-Hall, New York.

A broad comprehensive handbook of analytical techniques. It covers separation techniques, optical spectroscopy, mass spectrometry, electrochemistry, surface analysis, and polymer analysis.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS USED

a	Absorptivity
A	Absorbance, $= abc = \varepsilon bc$
A.U.	Absorbance unit
b	Path length
c	Analyte concentration; also velocity of light in vacuum (2.99795×10^8 m/s), $= \lambda\nu$
D_4, D_5	Deformation potentials
E	Energy of a photon, $= h\nu$
E_{AB}	Energy separation of A and B bands
h	Planck's constant (6.626×10^{-34} Js)
I	Light intensity
I_0	Initial light intensity
k	Absorption coefficient
P	Radiant power (radiant energy on unit area in unit time)
P_0	Initial radiant power; power transmitted through sample holder and solvent only
P_s	Power of source light
r_∞	Relative reflectance, $= r_{sample}/r_{reference}$
R	Reflectance (includes diffuse and specular reflectance)
$\%R$	Percent reflectance
s	Scattering coefficient
T	Transmittance
$\%T$	Percent transmittance
ε	Molar absorptivity
λ	Wavelength
ν	Frequency

BRIAN M. TISSUE
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

RAMAN SPECTROSCOPY OF SOLIDS

INTRODUCTION

Raman spectroscopy is based on the inelastic scattering of light by matter and is capable of probing the structure of gases, liquids, and solids, both amorphous and crystalline. In addition to its applicability to all states of matter, Raman spectroscopy has a number of other advantages. It can be used to analyze tiny quantities of material (e.g., particles that are $\sim$1 μm on edge), as well as samples exposed to a variety of conditions such as high temperature and high pressure and samples embedded in other phases, so long as the surrounding media are optically transparent.

Succinctly stated, Raman scattering results from incident radiation inducing transitions in the atoms/molecules that make up the scattering medium. The transition can be rotational, vibrational, electronic, or a combination (but first-order Raman scattering involves only a single incident photon). In most studies of solids by Raman spectroscopy, the transitions observed are vibrational and these will be the focus of this unit.

In a Raman experiment, the sample is irradiated with monochromatic radiation. If the sample is transparent, most of the light is transmitted, a small fraction is elastically (Rayleigh) scattered, and a very small fraction is inelastically (Raman) scattered. The inelastically scattered light is collected and dispersed, and the results are presented as a Raman spectrum, which plots the intensity of the inelastically scattered light as a function of the shift in wavenumber of the radiation. (The wavenumber of a wave is the reciprocal of its wavelength and is proportional to its momentum in units of reciprocal centimeters.) Each peak in the spectrum corresponds to one or more vibrational modes of the solid. The total number of peaks in the Raman spectrum is related to the number of symmetry-allowed, Raman active modes. Some of the modes may be degenerate and some may have Raman intensities that are too low to be measured, in spite of their symmetry-allowed nature. Consequently, the number of peaks in the Raman spectrum will be less than or equal to the number of Raman active modes. The practical usefulness of Raman spectroscopy resides largely in the fact that the Raman spectrum serves as a fingerprint of the scattering material. In fact, Raman activity is a function of the point group symmetry of a molecule and the space group symmetry of a crystalline solid; it can provide a range of information, including the strength of interatomic and intermolecular bonds, the mechanical strain present in a solid, the

composition of multicomponent matter, the degree of crystallinity of a solid, and the effects of pressure and temperature on phase transformations.

Competitive and Related Techniques

As mentioned above, Raman spectroscopy is extremely versatile and can be used to investigate the structures of solids, liquids, and gases. This unit is concerned with the analyses of solids. In general, two major types of experimental techniques are used to obtain information about the structures of solids: spectroscopy and diffraction. Spectroscopy is based on the interaction of electromagnetic radiation with matter. Diffraction occurs whenever a wave is incident on an array of regularly spaced scatterers in which the wavelength of the wave is similar to the spacing of the scatterers.

X rays, electrons, and neutrons with wavelengths similar to the interatomic spacing of a crystal will be diffracted by the crystal. The diffraction pattern will provide detailed information about, e.g., the crystal structure, crystal size, interplanar spacings, long- and short-range order, and residual lattice strain. While diffraction can provide some information about gases, liquids, and amorphous solids, the greatest amount of information is obtained for crystals.

Spectroscopic methods can probe the electronic, vibrational, rotational, and nuclear states of matter. In one family of spectroscopic techniques, the material absorbs electromagnetic radiation of energy equivalent to the difference in energy between particular states of the material. For example, transitions between adjacent electronic states generally require energies equivalent to that of ultraviolet and visible radiation (UV-vis electronic spectroscopy; see ULTRAVIOLET AND VISIBLE ABSORPTION SPECTROSCOPY). Vibrational transitions typically make use of infrared radiation (infrared absorption spectroscopy, IRAS). In the gas-phase molecules may absorb microwave radiation and undergo a change in rotational state. Finally, in nuclear magnetic resonance (NMR) the nuclei of atoms with angular momentum (spin) and magnetic moment absorb energy in the radio frequency region while subjected to magnetic fields that are alternating synchronous with the natural frequencies of the nucleus (see NUCLEAR MAGNETIC RESONANCE IMAGING).

Raman spectroscopy, unlike the above-mentioned spectroscopies, involves the scattering of electromagnetic radiation rather than absorption. In classical terms, the energy of the Raman scattered radiation is shifted from that of the incident radiation because of modulation by the vibrations of the scattering medium. Thus, Raman spectroscopy (RS), like IRAS, probes the vibrational spectra of materials. Because IRAS involves absorption of radiation and RS involves inelastic scattering of radiation, the two techniques are complementary. Since Raman spectroscopy makes use of optical radiation, it is a highly flexible technique and can probe the structures of materials in relatively inaccessible locations, such as in high-pressure chambers, high-temperature furnaces, and even aqueous solutions. All that is necessary is the ability to get laser light onto the sample and to collect the scattered light.

For materials that are both Raman active and IR active, the Raman scattering cross-section is much smaller than the IR absorption cross-section so the incident radiation needs to be more intense and the detectors need to be more sensitive for Raman spectroscopy than for IR spectroscopy. Because of the fundamental differences between absorption and scattering of electromagnetic radiation, some vibrational modes of a material may be Raman active but not IR active. In particular, for a molecule with a center of symmetry, vibrational modes that are Raman active are IR inactive and vice versa. Consequently, when deciding between IR and Raman spectroscopies, the point group or space group symmetry of the material needs to be taken into account.

A major advantage of Raman spectroscopy compared to x-ray diffraction is the ability of Raman spectroscopy to provide detailed structural information of amorphous materials. The Raman spectrum of an amorphous solid, for example, is directly related to its complete density of (vibrational) states. However, all metals are Raman inactive so Raman spectroscopy cannot be used for the structural analyses of this entire class of materials.

PRINCIPLES OF THE METHOD

Theoretical Background

This discussion of Raman spectroscopy will begin with a description of the fundamental phenomenon that underlies the scattering of light by matter, namely the generation of electromagnetic radiation by the acceleration of charged particles. As a light-scattering process, Raman scattering is rigorously described by treating both the incident and scattered radiation and the scattering medium quantum mechanically. In this unit a full quantum mechanical description of Raman scattering is not given. Instead, Raman scattering is treated less rigorously but in a manner that allows all of the practically important phenomena associated with Raman spectroscopy of solids to be identified. Specifically, a combination of classical physics and semiclassical physics (treating the scattering medium quantum mechanically while approximating light as a sinusoidally varying electric field) is used to describe Raman scattering.

Generation of Electromagnetic Radiation—Classical Physics. Classical physics dictates that accelerating charged particles emit electromagnetic radiation whose electric field ε (a vector quantity) is directly proportional to the particle's acceleration $\mathbf{a}$ (a vector quantity; see Fig. 1; Rossi, 1957):

$$\varepsilon_{\mathbf{r}}(r,\theta,\phi,t) = \varepsilon_{\phi}(r,\theta,\phi,t) = 0 \tag{1a}$$

and

$$\varepsilon_{\theta}(r,\theta,\phi,t) = \left(\frac{q\sin\theta}{4\pi c^2 r}\right)a \tag{1b}$$

where c is the velocity of light; a is the magnitude of the acceleration of the charged particle of charge q; r, θ, and ϕ are the spherical coordinates with the particle located at

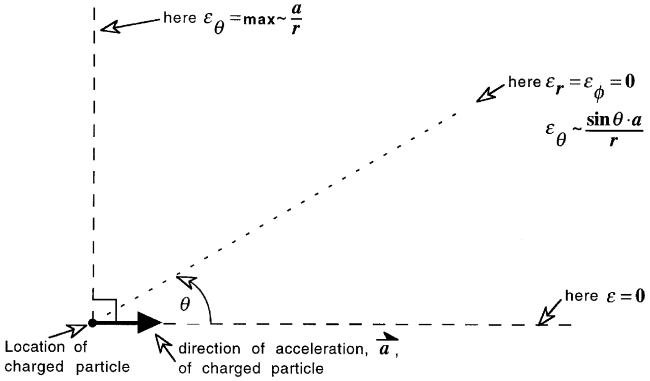


Figure 1. Electromagnetic radiation associated with an accelerating charged particle.

the origin and; t is time. The term $\varepsilon_i(r, \theta, \phi, t)$ is the ith component of the electric field at the point (r, θ, ϕ) and time t. The other terms are defined in Figure 1.

The optical behavior of matter, e.g., the emission of sharp spectral lines by elemental gases excited by an electric discharge, requires that matter, when treated by classical physics, be composed of charged particles that behave as harmonic oscillators when displaced from their equilibrium positions. When matter is illuminated by monochromatic optical radiation, the electrons are driven from their equilibrium positions by the sinusoidally varying electric field of light, creating a time-varying electric polarization $\mathbf{p}$ (a vector quantity) that is directly proportional to the acceleration of the electrons. Approximating the polarized molecule by a dipole, $\mathbf{p}$ is given by

$$\mathbf{p} = -e\mathbf{x} = -e\mathbf{x}_0 \cos \omega_{\text{inc}} t \tag{2a}$$

where $\mathbf{x}$ is the displacement of the center of negative charge of the molecule from the center of positive charge and $\mathbf{x}_0$ is the maximum value of this displacement.

$$\mathbf{p} = -\left(\frac{\mathbf{e}}{\omega_{\text{inc}}^2}\right)\mathbf{a} \tag{2b}$$

where $-e$ is the electric charge of the electron and ω_{inc} is the frequency of oscillation of the electron, which matches the frequency of the incident light. The electrons thereby emit radiation whose electric field is proportional to the polarization of the molecule:

$$\varepsilon_\theta \left(\frac{\omega_{\text{inc}}^2 \sin\theta}{4\pi\varepsilon_0 c^2 r}\right)|\mathbf{p}| \tag{3}$$

where $\mathbf{x}$ is the displacement of the center of negative charge of the molecule from the center of positive charge and $\mathbf{x}_0$ is the maximum value of this displacement.

$$S_{\text{r}} = \left(\frac{\omega_{\text{inc}}^4 \sin^2\theta}{16\pi^2\varepsilon_0^2 c^4 r^2}\right)|\mathbf{p}|^2 \tag{4}$$

where ε_0 is the dielectric strength of the medium in which the radiation is propagating and S_{r} is the radial component of the Poynting vector, which is the only nonzero component of the Poynting vector. The Poynting vector is the energy flux of the light wave. It is equal to the amount of energy flowing through a unit area in unit time.

When analyzing the effects of irradiating a molecule or a large quantity of matter with electromagnetic radiation, it is not necessary to treat each electron individually. Rather, it is possible to model the molecule(s) by atoms that are composed of heavy, positively charged ion cores and a mobile negative charge, which together form an electric dipole. If the centers of positive and negative charges are not coincident, the molecule has a permanent dipole moment. If the negative charge is displaced from its equilibrium position, e.g., by colliding with another particle, it will harmonically oscillate about its equilibrium position and emit electromagnetic radiation until all of the energy it gained in the collision is radiated away. Similarly, if the molecule is irradiated with an electromagnetic wave, the negative charge will be driven away from its equilibrium position by the applied electric field and will harmonically oscillate about its equilibrium position. The oscillating electric dipole will emit electromagnetic radiation with an electric field given by Equation 4 with $|\mathbf{p}| = p_0 \cos \omega_{\text{inc}} t$, where ω_{inc} is the frequency of the incident radiation and p_0 is the magnitude of the polarization when the centers of positive and negative charges experience their maximum separation.

The polarization $\mathbf{p}$ induced by the applied electric field of the incident electromagnetic radiation varies with time at the same frequency as the applied electric field. There is a second contribution to the time dependency of the material's polarization, namely that caused by the time-dependent nuclear displacements (thermally excited molecular vibrations), which, as illustrated in Figure 2, act to modulate the radiation generated by the electrons oscillating under the influence of the incident radiation (Long, 1977). As a result, there are three frequency components to the scattered light, which is the radiation emitted by the time-varying, molecular electric dipoles. One component (the Rayleigh line) has the same frequency, ω_{inc} as the incident radiation and the other two components have a slightly lower and a slightly higher frequency, $\omega_{\text{inc}} + \omega_{\text{v}}$ and $\omega_{\text{inc}} - \omega_{\text{v}}$, respectively, where ω_{v} is the vibrational frequency of the atoms/molecules making up the scattering medium. These last two components represent the Stokes and anti-Stokes components, respectively, of the Raman scattered radiation.

Classically, the expression for the intensity of the Raman scattered radiation is

$$I = \langle |S| \rangle \frac{dA}{d\Omega} \tag{5}$$

where A is the surface area of the sphere centered at the source of the Raman scattered radiation, Ω is the solid angle through which the radiation passes, and $\langle |S| \rangle$ is the time-averaged value of the Poynting vector of the Raman scattered radiation. Its form is identical to that given above for S_{r} except that the time-varying part of $\mathbf{p}$ (see Equation 4) is now given by

$$\mathbf{p} \sim \left(\frac{\partial\alpha}{\partial\mathbf{Q}_k}\right)_0 \mathbf{Q}_k \boldsymbol{\varepsilon} \tag{6}$$

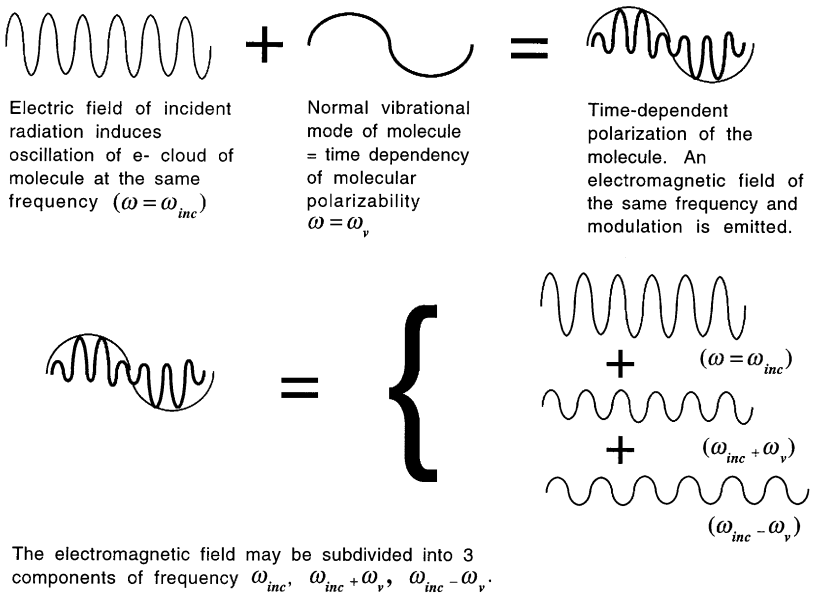


Figure 2. Polarization of vibrating molecule irradiated by light.

where $\underline{\alpha}$ (a tensor quantity) is the polarizability of the material, $\mathbf{Q}_k$ (a vector quantity) is the normal coordinate of the kth vibrational mode, and $(\partial\underline{\alpha}/\partial\mathbf{Q}_k)_0$ (a tensor quantity) is the differential polarizability, which indicates the substance's change in polarizability at its equilibrium position during the kth normal mode of vibration. Consequently, the material property that is directly probed by Raman spectroscopy is the differential polarizability $(\partial\underline{\alpha}/\partial\mathbf{Q}_k)_0$. It is a symmetric tensor quantity, and at least one component must be nonzero for Raman scattering to occur. Knowing the point group symmetry of a molecule or the space group symmetry of a crystalline solid, it is possible to use group theoretical techniques to determine if the symmetry permits nonzero components of the differential polarizability tensor. The use of molecular and crystal symmetry greatly simplifies the analyses of Raman scattering by a material, as will be demonstrated below (see Overview of Group Theoretical Analysis of Vibrational Raman Spectroscopy).

Light Scattering—Semiclassical Physics. From the perspective of semiclassical physics, Raman scattering results from the incident radiation inducing a transition (considered here to be vibrational) in the scattering entity, as schematically illustrated in Figure 3. The material is placed by the sinusoidally varying electric field into a higher energy state, termed a virtual state, which may be thought of as the distortion of the molecule by the electric field. As shown in Figure 3, when the molecule returns to a lower energy state, it emits radiation of frequency $\omega_{\text{inc}}(=>$ Rayleigh, or elastic scattering), $\omega_{\text{inc}} - \omega_v(=>$ Stokes Raman scattering), or $\omega_{\text{inc}} + \omega_v$ $(=>$ anti-Stokes Raman scattering). The probability of these transitions occurring is, according to quantum mechanical perturbation theory, given by a_{mn}^2, where

$$a_{mn} \sim \int \psi_m^* \mathbf{p} \psi_n dV \tag{7}$$

Here, ψ_i is the wave function of the ith state and ψ_i^* is its complex conjugate.

$$\mathbf{p} = \underline{\alpha}\boldsymbol{\varepsilon} \tag{8}$$

When expressed in terms of the individual components, p_i, Equation 8 becomes

$$p_i = \sum \alpha_{ij}\varepsilon_j \tag{9}$$

where α_{ij} is the ij matrix component of the polarizability, ε_j is the j component of the electric field, and the differential dV includes differential displacements of all quantum vibrational states. For Rayleigh scattering, $m = n$ is the vibrational ground state; for Stokes Raman scattering, n is the vibrational ground state and m is the first vibrational excited state; for anti-Stokes Raman scattering, n

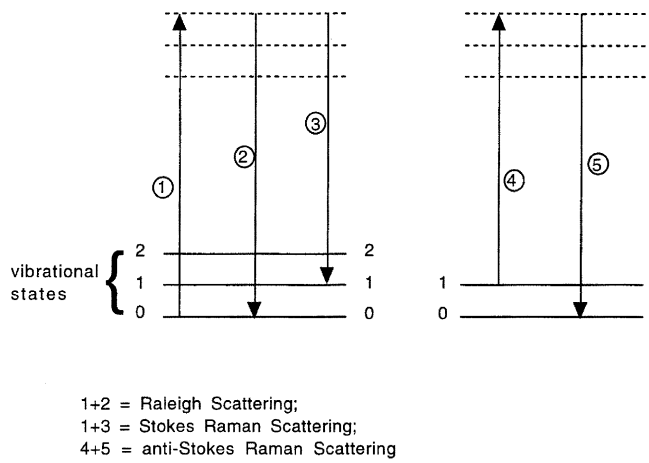


Figure 3. Rayleigh and Raman scattering. $1 + 2$, Rayleigh scattering; $1 + 3$, Stokes Raman scattering; $4 + 5$, anti-Stokes Raman scattering.

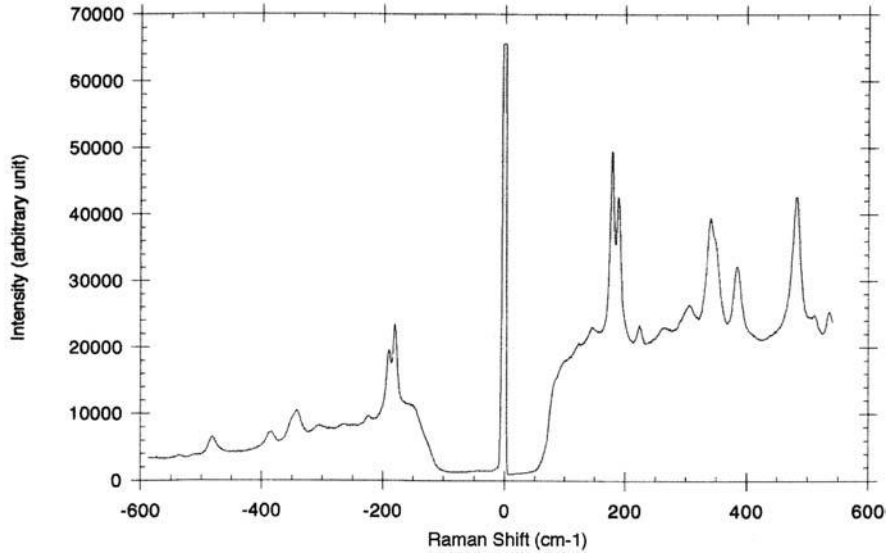


Figure 4. Raman spectrum of impure, polyphase (monoclinic+tetragonal) ZrO_2. (C. S. Kumai, unpublished results, 1998.)

is the first vibrational excited state and m is the vibrational ground state.

Since the population of energy levels in a collection of atoms/molecules is governed by the Maxwell-Boltzmann distribution function, the population of the vibrational ground state will always be greater than the population of the first vibrational excited state. As a result, the intensity of the Stokes-shifted Raman scattering will always be greater than the intensity of the anti-Stokes-shifted Raman scattered radiation. This effect, which is predicted quantum mechanically but not classically, is illustrated in Figure 4, which presents the Raman spectrum of polycrystalline ZrO_2 (mixed monoclinic and tetragonal). The Raman spectrum is depicted in terms of the shift (in wavenumbers) in the wavelength of the scattered radiation with respect to that of the incident light. Thus the Rayleigh peak is seen in the center, and the anti-Stokes and Stokes lines are to the left and right, respectively. Note that for each peak in the Stokes Raman section of the spectrum there is a corresponding peak of lower intensity in the anti-Stokes Raman section.

Overview of Group Theoretical Analysis of Vibrational Raman Spectroscopy

The use of group theory in determining whether or not a given integral can have a value of zero may be illustrated using the ground vibrational state $\psi_0(x)$ and first vibrational excited state $\psi_1(x)$ of a harmonic oscillator. The equations describing these two vibrational states are, respectively,

$$\psi_0(x) = \exp\left(-\frac{2\pi m\omega x^2}{\hbar}\right) \tag{10}$$

$$\psi_1(x) = \left(\frac{4\pi^{1/2} m\omega}{\hbar}\right)^{1/2} \exp\left(-\frac{2\pi m\omega x^2}{\hbar}\right) \tag{11}$$

where $\hbar = h/2\pi$ and h is Planck's constant. These two vibrational states are plotted in Figure 5.

By inspection,

$$\int_{-\infty}^{\infty} \psi_1(x)dx = 0 \tag{12}$$

That is, performing the integration by summing up the infinite number of terms given by

$$\sum \psi_1(x_n)\Delta x_n \tag{13}$$

(where $\Delta x_n = x_n - x_{n-1}$), it is seen that for every positive contribution (to the sum) given by $\psi_1(x_i)\Delta x_i$ there is a term equal in magnitude but opposite in sign [$\psi_1(x_{-i})\Delta x_{-i} = -\psi_1(x_i)\Delta x_i$].

Thus, the entire sum in Equation 13 is zero.

Similarly, by inspection,

$$\int_{-\infty}^{\infty} \psi_0(x)dx \neq 0 \tag{14}$$

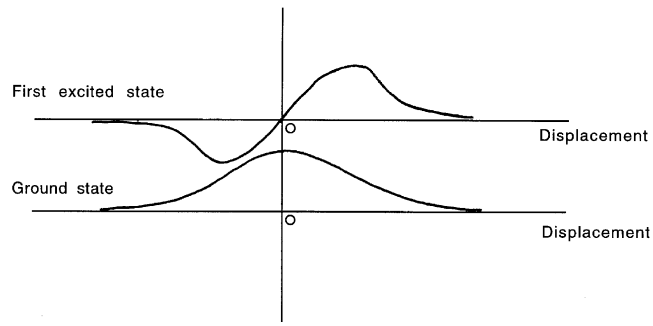


Figure 5. Ground-state vibrational wave function and first excited-state vibrational wave function of harmonic oscillator.

That is, every term $\psi_0(x_i)\Delta x_i$ in the sum $\sum \psi_0(x_n)\Delta x_n$ has the identical sign and, hence,

$$\int_{-\infty}^{\infty} \psi_0(x)dx \approx \sum \psi_0(x_n)\Delta x_n \neq 0 \qquad (15)$$

The above analysis of the integrals in Equations 12 and 13 can be repeated using group theory. To begin, there needs to be a mathematical description of what is meant by the symmetry of a function. The difference between the symmetry of $\psi_1(x)$ and $\psi_0(x)$ can be expressed as follows. Imagine that there is a mirror plane parallel to the *YZ* plane and that passes through the origin, (0, 0, 0); then the operation of this mirror plane on the function $\psi_0(x)$ is described by the phrase "a mirror plane acting on $\psi_0(x)$ yields $\psi_0(x)$."

If the effect of the mirror plane acting on $\psi_0(x)$ is represented by a matrix $[R]$, then

$$\text{Effect of mirror plane acting on } \psi_0(x) = [R][\psi_0(x)] \qquad (16)$$

In this case, $[R] = [I]$ is the identity matrix.

Similarly, for $\psi_1(x)$, the effect of a mirror plane acting on $\psi_1(x)$ equals $-\psi_1(x)$. In matrix notation, this statement is given as

$$\text{Effect of mirror plane acting on } \psi_1(x) = [R]\psi_1(x) = -[I]\psi_1(x) \qquad (17)$$

Note that the effect of the mirror operation on x is to produce $-x$ and that $\psi(-x) = -\psi(x)$. The character of the one-dimensional matrix in Equation 16 is $+1$ and that in Equation 17 is -1. Consider now a molecule that exhibits a number of symmetry elements (e.g., mirror planes, axes of rotation, center of inversion). If the operation of each of those on a function leaves the function unchanged, then the function is said to be "totally symmetric" with respect to the symmetry elements of the molecule (or, in other words, the function is totally symmetric with respect to the molecular point group, which is the collection of all the molecule symmetry elements). For the totally symmetric case, the character of each one-dimensional (1D) matrix representing the effect of the operation of each symmetry element on the function is 1 for all symmetry operations of the molecule. For such a function, the value of the integral over all space of $\psi(x)dx$ will be nonzero.

If the operation of any symmetry element of the molecule (in the molecular point group) generates a matrix that has a character other than 1, the function is not totally symmetric and the integral of $\psi(x)dx$ over all space will be zero.

Now consider the application of group theory to vibrational spectroscopy. A vibrational Raman transition generally consists of the incident light causing a transition from the vibrational ground state to a vibrational excited state. The transition may occur if

$$\int_{-x}^{x} \psi_1(x)\alpha_{ij}\psi_0(x)dx \neq 0 \qquad (18)$$

Table 1. Character Table of Point Group C_{2v}

C_{2v}	E	$C_2(z)$	$\sigma_v(xz)$	$\sigma'_v\ (yz)$	Basis Functions
A_1	$+1$	$+1$	$+1$	$+1$	$z; x^2; y^2; z^2$
A_2	$+1$	$+1$	-1	-1	xy
B_1	$+1$	-1	$+1$	-1	$x; xz$
B_2	$+1$	-1	-1	$+1$	$y; yz$

That is, the integral may have a nonzero value if the integrand product is totally symmetric over the range of integration. The symmetry of a function that is the product of two functions is totally symmetric if the symmetries of the two functions are the same. In addition, for the sake of completeness, the last statement will be expanded using group theoretical terminology, which will be defined below (see Example: Raman Active Vibrational Modes of α-Al_2O_3). The symmetry species of a function that is the product of two functions will contain the totally symmetric irreducible representation if the symmetry species of one function contains a component of the symmetry species of the other.

The ground vibrational state is totally symmetric. Hence, the integrand is totally symmetric and the vibrational mode is Raman active if the symmetry of the excited vibrational state $\psi_1(x)$ is the same as, or contains, the symmetry of the polarizability operator α_{ij}.

As shown below (see Appendix), the symmetry of the operator α_{xy} is the same as that of the product xy. Thus, the integrand $\psi_1\alpha_{xy}\psi_0$ is totally symmetric if ψ_1 has the same symmetry as xy (recall that ψ_0, the vibrational ground state, is totally symmetric).

A molecular point group is a mathematical group (see Appendix) whose members consist of all the symmetry elements of the molecule. Character tables summarize much of the symmetry information about a molecular point group. Complete sets of character tables may be found in dedicated texts on group theory (e.g., Bishop, 1973). The character table for the point group C_{2v} is presented in Table 1. In the far right-hand column are listed functions of interest in quantum mechanics and vibrational spectroscopy in particular. The top row lists the different symmetry elements contained in the point group and illustrated in Figure 6. Here, C_2 is a twofold rotational axis that is parallel to the z axis. The parameters σ_{xz} and σ_{yz} are mirror planes parallel to the xz and yz planes, respectively. In each row beneath the top row are listed a set of numbers. Each number is the character of the matrix that represents the effect of the symmetry operation acting on any of the functions listed in the right-hand end of the same row. In the point group C_{2v}, e.g., the functions z, x^2, y^2, and z^2 have the same symmetry, which in the terminology of group theory is identified as A_1. The functions y and yz have the same symmetry, B_2.

As shown in Figure 6, the water molecule, H_2O, exhibits the point group symmetry C_{2v}. There are three atoms in H_2O and hence there are $3N - 6 = 3$ normal vibrational modes, which are depicted in Figure 7. These exhibit symmetries A_1, A_1, and B_2. That is, the vibrational wave functions of the first excited state of each of these modes

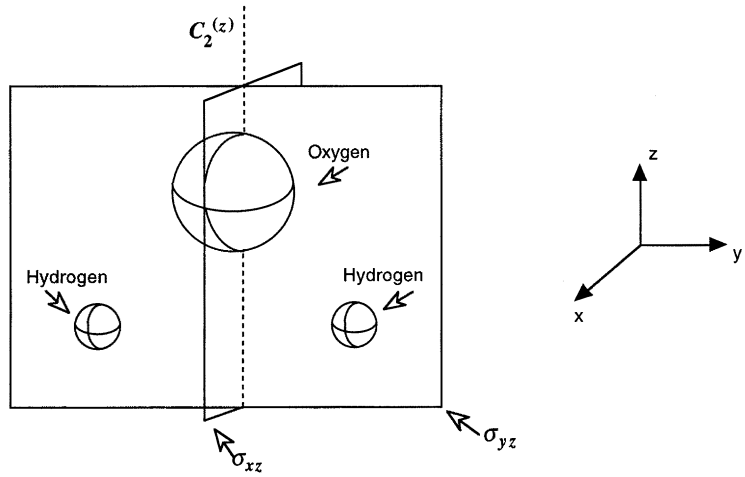


Figure 6. Symmetry operations of H_2O (point group C_{2v}).

possess the symmetries A_1, A_1, and B_2, respectively. Since x^2, y^2, and z^2 exhibit A_1 symmetry, so too would α_{xx}, α_{yy} and α_{zz}. Hence the product of α_{xx}, α_{yy} and α_{zz} with ψ_1 for the vibrational modes with A_1 symmetry would be totally symmetric, meaning that the integral over the whole space $\psi_{A1}\alpha_{xx(\text{or}yy,zz)}\psi_0 dV$ may be nonzero, meaning that this vibrational mode is Raman active (the integrand $\psi_{A1}\alpha_{xx(\text{or}yy,zz)}\psi_0 dV$ is the same as the integrand in Equation 7).

Now consider the Raman vibrational spectra of solids. One mole of a solid will have $3 \times 6.023 \times 10^{23} - 6$ normal modes of vibration. Fortunately, it is not necessary to consider all of these. As demonstrated below (see Appendix), the only Raman active modes will be those near the center of the Brillouin zone (BZ; $k = 0$). Solids with only one atom per unit cell have zero modes at the center of the BZ and hence are Raman inactive (see Appendix). Solids with more than one atom per unit cell, e.g., silicon and Al_2O_3, are Raman active.

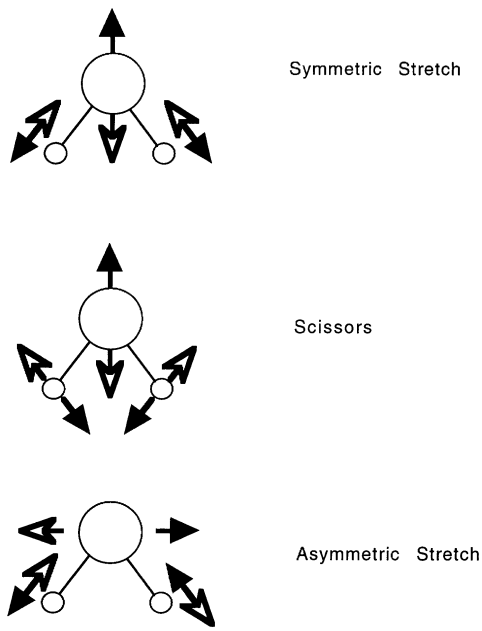


Figure 7. Normal vibrational modes of H_2O.

The symmetry of a crystal is described by its space group. However, as demonstrated below (see Appendix), for the purposes of vibrational spectroscopy, the symmetry of the crystal is embodied in its crystallographic point group. This is an extremely important and useful conclusion. As a consequence, the above recipe for deciding the Raman activity of the vibrational modes of H_2O can be applied to the vibrational modes of any solid whose unit cell is multiatomic.

PRACTICAL ASPECTS OF THE METHOD

The weakness of Raman scattering is the characteristic that most strongly influences the equipment and experimental techniques employed in Raman spectroscopy.

A conventional Raman facility consists of five major components: (1) a source of radiation, (2) optics for illuminating the sample and collecting the Raman scattered radiation, (3) a spectrometer for dispersing the Raman scattered radiation, (4) a device for measuring the intensity of the Raman scattered light, and (5) a set of components that control the polarization of the incident radiation and monitor the polarization of the Raman scattered radiation (Chase, 1991; Ferraro and Nakamoto, 1994).

Sources of Radiation

Prior to the introduction of lasers, the primary source of monochromatic radiation for Raman spectroscopy was the strongest excitation line in the visible region (blue, 435.83 nm) of the mercury arc. The power density of the monochromatic radiation incident on the sample from a mercury arc is very low. To compensate, relatively large samples and complicated collection optics are necessary to, respectively, generate and collect as many Raman scattered photons as possible. Collecting Raman photons that are scattered in widely different directions precludes investigating the direction and polarization characteristics of the

Raman scattered radiation. This is valuable information that is needed to identify the component(s) of the crystal's polarizability tensor that is (are) responsible for the Raman scattering (Long, 1977).

In general, laser radiation is unpolarized. However, in ionized gas lasers, such as argon and krypton ion lasers, a flat window with parallel faces and oriented at the Brewster angle (θ_{Brewster}) relative to the incident beam is positioned at the end of the tube that contains the gas. This window is called a Brewster window. The Brewster angle is also called the polarizing angle and is given by the arctangent of the ratio of the indices of refraction of the two media that form the interface (e.g., for air/glass, $\theta_{\text{Brewster}} = \tan^{-1}1.5 \approx 57^\circ$). If an unpolarized beam of light traveling in air is incident on a planar glass surface at the Brewster angle, the reflected beam will be linearly polarized with its electric field vector parallel to the plane of incidence. (The plane of incidence contains the direction of propagation of the light and the normal to the interface between the two media.) A beam of light that is linearly polarized with its electric field vector in the plane of incidence and is incident on a Brewster window at θ_{Brewster} will be entirely transmitted. Thus, the light that exits from a laser tube that is capped with a Brewster window is linearly polarized (Fowles, 1989; Hecht, 1993).

Lasers have greatly increased the use of Raman spectroscopy as a tool for both research and chemical identification. In particular, continuous-wave gas lasers, such as argon, krypton, and helium-neon, provide adequately powered, monochromatic, linearly polarized, collimated, small-diameter beams that are suitable for obtaining Raman spectra from all states of matter and for a wide range of sample sizes using relatively simple systems of focusing and collection optics.

Optics

As the light exits the laser, it generally first passes through a filter set to transmit a narrow band of radiation centered at the laser line (e.g., an interference filter). This reduces the intensity of extraneous radiation, such as the plasma lines that also exit from the laser tube. A set of mirrors and/or lenses will make the laser light incident on the sample at the desired angle and spot size.

The focusing lens reduces the beam's diameter to

$$d = \frac{4\lambda f}{\pi D} \tag{19}$$

where λ is the wavelength of the laser radiation, f is the focal length of the lens, and D is the diameter of the unfocused laser beam. Note that for a given focusing lens and incident light, the size of the focused spot is inversely proportional to the size of the unfocused beam. Thus, a smaller spot size can be generated by first expanding the laser beam to the size of the focusing lens. The distance l over which the beam is focused is proportional to the square of the focused beam's diameter (Hecht, 1993):

$$l = \frac{16\lambda f^2}{\pi D^2} = \frac{\pi d^2}{\lambda} \tag{20}$$

Thus, to reduce the diameter of the focal spot from an argon laser ($\lambda = 514.5$ nm) to 1 μm, the distance between the focusing lens and the sample must be accurate to within $\pm\frac{1}{2}\pi(1\ \mu\text{m})^2/(0.5145\ \mu\text{m}) \approx 3\ \mu\text{m}$.

The specularly reflected light is of no interest in Raman spectroscopy. The inelastically scattered light is collected using a fast lens (i.e., low f number, defined as the focal length per diameter of the lens). Often, the lens from a 35-mm camera is a very cost-effective collection lens. The collection lens collimates the Raman scattered light and transmits it toward a second lens that focuses the light into the entrance slit of the spectrometer. The f number of the second lens should match that of the spectrometer. Otherwise, the spectrometer's grating will either not be completely filled with light, which will decrease the resolving power (which is directly proportional to the size of the grating), or be smaller than the beam of radiation, which will result in stray light that may reflect off surfaces inside the spectrometer and reduce the signal-to-noise ratio.

There are many sources of stray light in a Raman spectrum. One of the major sources is elastically scattered light (it has the same wavelength as the incident radiation), caused by Rayleigh scattering by the molecules/atoms responsible for Raman scattering and Mie scattering from larger particles such as dust. Rayleigh scattering always accompanies Raman scattering and is several orders of magnitude more intense than Raman scattering. The Rayleigh scattered light comes off the sample at all angles and is captured by the collection optics along with the Raman scattered light and is directed into the spectrometer. This would not be a problem if the Rayleigh scattered light exited the spectrometer only at the 0-cm^{-1} position. However, because of imperfections in the gratings and mirrors within the spectrometer, a portion of the Rayleigh scattered light exits the spectrometer in the same region as Raman scattered radiation that is shifted in the range of 0 to $\pm 250\ \text{cm}^{-1}$ from the incident radiation. Even though only a fraction of the Rayleigh scattered light is involved in this misdirection, the consequence is significant because the intensity of the Rayleigh component is so much greater than that of the Raman component. Raman peaks located within $\approx 250\ \text{cm}^{-1}$ of the laser line can only be detected if the intensity of the Rayleigh scattered light is strongly reduced.

The intensity of stray light that reaches the detector can be diminished by passing the collimated beam of elastically and inelastically scattered light collected from the sample into a notch filter. After exiting the notch filter, the light enters a lens that focuses the light into the entrance of the spectrometer.

A notch filter derives its name from the narrow band of wavelengths that it filters out. In a Raman experiment, the notch filter is centered at the exciting laser line and is placed in front of the spectrometer. (Actually, as mentioned above and explained below, the notch filter should be located in front of the lens that focuses the scattered light into the spectrometer.) The notch filter then removes the Rayleigh scattered radiation (as well as Brillouin scattered light and Mie scattering from dust particles). Depending on the width of the notch, it may also remove a significant amount of the Raman spectrum.

Dielectric notch filters consist of multiple, thin layers of two materials with different indices of refraction. The materials are arranged in alternating layers so that the variation of index of refraction with distance is a square wave. Typically, dielectric filters have a very wide notch with diffuse edges and have nonuniform transmission in the region outside the notch. As a consequence, a dielectric notch filter removes too much of the Raman spectrum (the low-wavenumber portion) and distorts a significant fraction of the portion it transmits. In contrast, holographic notch filters have relatively narrow widths, with sharp edges and fairly uniform transmission in the region outside the notch (Carrabba et al., 1990; Pelletier and Reeder, 1991; Yang et al., 1991; Schoen et al., 1993).

A holographic notch filter (HNF) is basically an interference filter. In one manufacturing process (Owen, 1992), a photosensitive material consisting of a dichromate gelatin film is placed on top of a mirror substrate. Laser light that is incident on the mirror surface interferes with its reflected beam and forms a standing-wave pattern within the photosensitive layer. The angle between the incident and reflected beams determines the fringe spacing of the hologram. If the incident beam is normal to the mirror, the fringe spacing will be equal to the wavelength of the laser light.

Chemical processing generates within the hologram-exposed material an approximately sinusoidal variation of index of refraction with distance through the thickness of the layer. The wavelength of the modulation of the index of refraction determines the central wavelength of the filter. The amplitude of the modulation and the total thickness of the filter determine the bandwidth and optical density of the filter.

The precise shape and location of the "notch" of the filter is angle tunable so the filter should be located in between the lens that collects and collimates the scattered light and the lens that focuses the Raman scattered radiation into the spectrometer such that only collimated light passes through the filter.

A holographically generated notch filter can have a relatively narrow, sharp-edged band of wavelengths that will not be transmitted. The band is centered at the wavelength of the exciting laser line and may have a total width of $\approx$300 cm^{-1}. The narrow width of the notch and its high optical density permit the measurement of both the Stokes and anti-Stokes components of a spectrum, as is illustrated in the Raman spectrum of zirconia presented in Figure 4.

In addition to a notch filter, it is also possible to holographically generate a bandpass filter, which can be used to remove the light from the plasma discharge in the laser tube (Owen, 1992). In contrast to dielectric filters, holographic filters can have a bandpass that is five times more narrow and can transmit up to 90% of the laser line. Dielectric bandpass filters typically transmit only 50% of the laser line.

Optical Alignment

Alignment of the entire optical system is most easily accomplished by use of a second laser, which need only have an intensity of $\approx$1 mW. The spectrometer is set to transmit the radiation of the second laser, which is located at the exit slit of the spectrometer. For convenience, instead of having to remove the light detector, the light from the second laser can enter through a second, nearby port and made to hit a mirror that is inserted inside the spectrometer, just in front of the exit slits. The light reflected from the mirror then passes through the entire spectrometer and exits at the entrance slits, a process termed "back illumination." The back-illuminated beam travels the same path through the spectrometer (from just in front of the exit slits to the entrance slits) that the Raman scattered radiation from the sample will travel in getting to the light detector. Consequently, the system is aligned by sending the back-illuminated beam through the collection optics and into coincidence with the exciting laser beam on the sample's surface.

Spectrometer

The dispersing element is usually a grating, and the spectrometer may typically have one to three gratings. Multiple gratings are employed to reduce the intensity of the Rayleigh line. That is, light that is dispersed by the first grating is collimated and made incident on a second grating. However, the overall intensity of the dispersed light that exits from the spectrometer decreases as the number of gratings increases. The doubly dispersed light has a higher signal-to-noise ratio, but its overall intensity is significantly lower than singly dispersed light. Since the intensity of Raman scattered radiation is generally very weak (often by one or more orders of magnitude), there is an advantage to decreasing the intensity of the Rayleigh light that enters the spectrometer by using a single grating in combination with a notch filter rather than by using multiple gratings (see Optics).

Modern spectrometers generally make use of interference or holographic gratings. These are formed by exposing photosensitive material to the interference pattern produced by reflecting a laser beam at normal incidence off a mirror surface or by intersecting two coherent beams of laser light. Immersing the photosensitized material into a suitable solvent, in which the regions exposed to the maximum light intensity experience either enhanced or retarded rates of dissolution, produces the periodic profile of the grating. The surface of the grating is then typically coated with a thin, highly reflective metal coating (Hutley, 1982).

The minimum spacing of grooves that can be generated by interference patterns is directly proportional to the wavelength of the laser light. The minimum groove spacing that can be formed using the 458-nm line of an argon ion laser is 0.28 μm (3500 grooves/mm). The line spacing of a grating dictates one of the most important characteristics of the grating, its resolving power.

The resolving power of a grating is a measure of the smallest change of wavelength that the grating can resolve:

$$\frac{\lambda}{\Delta\lambda} = \frac{mW}{d} \tag{21}$$

where λ is the wavelength at which the grating is operating, m is the diffraction order, W is the width of the grating, and d is the spacing of the grating grooves (or steps). This relation indicates the importance of just filling the grating with light and the influence of the spacing of the grating grooves, which is generally expressed in terms of the number of grooves per millimeters, on its resolving power.

A second parameter of interest for characterizing the performance of a grating is its absolute efficiency, which is a measure of the fraction of the light incident on the grating that is diffracted into the required order. It is a function of the shape of the groove, the angle of incidence, the wavelength of the incident light, the polarization of the light, the reflectance of the material that forms the grating, and the particular instrument that houses the gratings (Hutley, 1982). The efficiency of a grating can be measured or calculated with the aid of a computer. The measurement of the absolute efficiency is performed by taking the ratio of the flux in the diffracted beam to the flux in the incident beam. Generally, it is not necessary for a Raman spectroscopist to measure and/or calculate the absolute efficiencies of gratings. Such information may be available from the manufacturer of the gratings.

Measurement of the Dispersed Radiation

A device for measuring the intensity of the dispersed radiation will be located just outside the exit port of the spectrometer. Since the intensity of the Raman scattered radiation is so weak, a very sensitive device is required to measure it. Three such devices are a photomultiplier tube (PMT), a photodiode array, and a charge-transfer device.

The key elements in a PMT are a photosensitive cathode, a series of dynodes, and an anode (Long, 1977; Ferraro and Nakamoto, 1994). The PMT is located at the exit of the spectrometer, which is set to transmit radiation of a particular wavelength. For Stokes-shifted Raman spectroscopy, the wavelength is longer than that of the exciting laser. As photons of this energy exit the spectrometer, they are focused into the photocathode of the PMT. For each photon that is absorbed by the photocathode, an electron is ejected and is accelerated toward the first dynode, whose potential is $\approx$100 V positive with respect to the photocathode. For each electron that hits the first dynode, several are ejected and are accelerated toward the second dynode. A single photon entering the PMT may cause a pulse of 10^6 electrons at the anode. Long (1977) describes the different procedures for relating the current pulse at the anode to the intensity of the radiation incident on the PMT.

The efficiency of the PMT is increased through the use of a photoemission element consisting of a semiconductor whose surface is coated with a thin layer ($\approx$2 nm) of material with a low work function (e.g., a mixture of cesium and oxygen; Fraser, 1990). For a heavily doped p-type semiconductor coated with such a surface layer, the electron affinity in the bulk semiconductor is equal to the difference between its band-gap energy and the surface layer work function. If the work function is smaller than the band gap, the semiconductor has a negative electron affinity. This means that the lowest energy of an electron in the conduction band in the bulk of the semiconductor is higher than the energy of an electron in vacuum and results in an increase in the number of electrons emitted per absorbed photon (i.e., increased quantum efficiency). The gain of the PMT can be increased by treating the dynodesurface in a similar fashion. A dynode with negative electron affinity emits a greater number of electrons per incident electron. The overall effect is a higher number of electrons at the anode per photon absorbed at the cathode.

It is important to note that the PMT needs to be cooled to reduce the number of thermally generated electrons at the photocathode and dynodes, which add to the PMT "dark count." Heat is generally extracted from the PMT by a thermoelectric cooler. The heat extracted by the cooler is generally conducted away by flowing water.

A PMT is well suited to measuring the weak signal that exits a spectrometer that has been set to pass Raman scattered radiation of a single energy. The entire Raman spectrum is measured by systematically changing the energy of the light that passes through the spectrometer (i.e., rotating the grating) and measuring its intensity as it exits from the spectrometer. For a double monochromator fitted with a PMT, it may take minutes to tens of minutes to generate a complete Raman spectrum of one sample.

In a PMT, current is generated as photons hit the cathode. In contrast, each active segment in a photodiode array (PDA) and a charge-coupled device (CCD) stores charge (rather than generating current) that is created by photons absorbed at that location. The Raman spectrum is generated from the spatial distribution of charge that is produced in the device. In both a PDA and CCD, photons are absorbed and electron-hole pairs are created. The junction in which the electron-hole pairs are created are different in the two devices. In a PDA, the junction is a reverse-biased p-n junction. In a CCD, the junction is the depletion zone in the semiconductor at the semiconductor-oxide interface of a metal-oxide-semiconductor.

When the reverse-biased p-n junction is irradiated by photons with an energy greater than the band gap, electron-hole pairs are generated from the absorbed photons. Minority carriers from the pairs formed close (i.e., within diffusion distance) to the charge-depleted layer of the junction are split apart from their complementary, oppositely charged particles and driven in the opposite direction by the electric field in the junction. The increase in reverse saturation current density is proportional to the light intensity hitting the photodiode. To minimize the "dark counts", the photodiode must be cooled to reduce the number of thermally generated electron-hole pairs.

If the junction is at equilibrium and is irradiated with photons of energy greater than the band gap, electron-hole pairs generated in the junction are separated by the built-in electric field. The separated electrons and holes lower the magnitude of the built-in field. If an array of diodes is distributed across the exit of the spectrometer, the distribution of charge that is created in the array provides a measure of the intensity of the radiation that is dispersed by the spectrometer across the focal plane for the exiting radiation.

The time to measure a Raman spectrum can be greatly lowered by measuring the entire spectrum at once, rather than one energy value at a time, as is the case with a PMT. Multichannel detection is accomplished by, e.g., placing a PDA at the exit of a spectrometer. There are no slits at the exit, which is filled by the PDA. Each tiny diode in the array measures the intensity of the light that is dispersed to that location. Collectively, the entire array measures the whole spectrum (or a significant fraction of the whole spectrum) all at once. Generally, an individual photodiode in a PDA is not as good a detector as is a PMT. However, for some experiments the multichannel advantage compensates for the lower quality detector. Unfortunately, a PDA do not always provide the sensitivity at low light intensities that are needed in Raman spectroscopy.

A CCD is a charge-transfer device that combines the multichannel detection advantage of a PDA with a sensitivity at low light intensities that rivals that of a PMT (Bilhorn et al., 1987a,b; Epperson, 1988). Charge-coupled devices are made of metal-oxide-semiconductor elements and, in terms of light detection, function analogously to photographic film. That is, photons that hit and are absorbed at a particular location of the CCD are stored at that site in the form of photogenerated electrons (for *p*-type semiconductor substrate) or photogenerated holes (for *n*-type semiconductor substrate).

One example of a CCD is a heavily doped, *p*-type silicon substrate whose surface is coated with a thin layer of SiO_2 on top of which are a series of discrete, closely spaced metal strips, referred to as gates. The potential of each adjacent metal strip is set at a different value so that the width of the depletion layer in the semiconductor varies periodically across its surface. When a positive potential is applied to a metal strip, the mobile holes in the *p*-type silicon are repelled from the surface region, creating a depleted layer. If a high enough potential is applied to a metal strip, significant bending of the bands in the region of the semiconductor close to the oxide layer causes the bottom of the conduction band to approach the Fermi level. Electrons occupy states in the conduction band near the surface forming an inversion layer, i.e., an *n*-type surface region in the bulk *p*-type semiconductor. Photons that are absorbed in the surface region generate electron-hole pairs in which the electrons and holes are driven apart by the field in the depletion layer. The holes are expelled from the depletion layer and the electrons are stored in the potential well at the surface. As the intensity of light increases, the number of stored electrons increases. The light exiting the spectrometer is dispersed and so hits the CCD at locations indicative of its wavelength. The dispersed light creates a distribution of charge in the depletion layer from which the sample Raman spectrum is generated. The distribution of charge corresponding to the Raman spectrum is sent to a detector by shifting the periodic variation in potential of the metal strips in the direction of the detector. In this manner the charge stored in the depletion layer of the semiconductor adjacent to a particular metal strip is shifted from one strip to the next. Charge arrives at the detector in successive packets, corresponding to the sequential distribution of the metal gates across the face of the CCD.

An illustration of the increased efficiency of measuring Raman spectra that has emerged in the past 10 years is provided by a comparison of the time required to measure a portion (200 to 1500 cm^{-1}) of the surface-enhanced Raman spectrum (see the following paragraph) of thin passive films grown on samples of iron immersed in aqueous solutions. Ten to 15 min (Gui and Devine, 1991) was required to generate the spectrum by using a double spectrometer (Jobin-Yvon U1000 with 1800 grooves/mm holographic gratings) and a PMT (RCA c31034 GaAs), while the same spectrum can now be acquired in $\approx$5 s (Oblonsky and Devine, 1995) using a single monochromator (270M Spex), in which the stray light is reduced by a notch filter (Kaiser Optics Super Holographic Notch Filter centered at 647.1 nm), and the intensity of the dispersed Raman scattered radiation is measured with a CCD (Spectrum One, 298 $\times$ 1152 pixels). The enhanced speed in generating the spectrum makes it possible to study the time evolution of samples.

In the previous paragraph, mention was made of surface-enhanced Raman scattering (SERS), which is the greatly magnified intensity of the Raman scattered radiation from species adsorbed on the surfaces of specific metals with either roughened surfaces or colloidal dimensions (Chang and Furtak, 1982). SERS is not a topic of this unit, but a great deal of interest during the past 15 years has been devoted to the use of SERS in studies of surface adsorption and surface films.

Polarization

By controlling the polarization of the incident radiation and measuring the intensity of the Raman scattered radiation as a function of its polarization, it is possible to identify the component(s) of the polarizabilty tensor of a crystal that are responsible for the different peaks in the Raman spectrum (see Example: Raman Active Vibrational Modes of α-Al_2O_3, below). This information will then identify the vibrational mode(s) responsible for each peak in the spectrum. The procedure is illustrated below for α-Al_2O_3 (see Example: Raman Active Vibrational Modes of α-Al_2O_3). The configuration of the polarizing and polarization measuring components are discussed in Scherer (1991).

Fourier Transform Raman Spectroscopy

The weak intensity of Raman scattering has already been mentioned as one of the major shortcomings of Raman spectroscopy. Sample heating and fluorescence are two other phenomena that can increase the difficulty of obtaining Raman spectra. Sample heating is caused by absorption of energy during illumination with laser beams of high power density. Fluorescence is also caused by absorption of laser energy by the sample (including impurities in the sample). When the excited electrons drop back down to lower energy levels, they emit light whose energy equals the difference between the energies of the initial (excited) and final states. The fluorescence spectrum is (practically speaking) continuous and covers a wide range of values. The intensity of the fluorescence can be much greater than that of the Raman scattered radiation.

Both sample heating and fluorescence can be minimized by switching to exciting radiation whose energy is too low to be absorbed. For many materials, optical radiation in the red or near-IR range will not be absorbed. If such radiation is to be used in Raman spectroscopy, two consequences must be recognized and addressed. First, since Raman scattering is a light-scattering process, its intensity varies as λ^{-4}. Hence, by switching to longer wavelength exciting radiation, the intensity of the Raman spectrum will be significantly lowered. Second, the sensitivity of most PMTs to red and especially near-infrared radiation is very low. Consequently, the use of long-wavelength radiation in Raman spectroscopy has been coupled to the use of Fourier transform Raman spectroscopy (FTRS) (Parker, 1994).

In FTRS, an interferometer is used in place of a monochromator. The Raman scattered radiation is fed directly into a Michelson interferometer and from there it enters the detector. The Raman spectrum is obtained by taking the cosine-Fourier transform of the intensity of the radiation that reaches the detector, which varies in magnitude as the path difference between the moving mirror and the fixed mirror within the interferometer changes. Since the Raman scattered radiation is not dispersed, a much higher Raman intensity reaches the detector in FTRS than is the case for conventional Raman spectroscopy. The much higher intensity of the Raman scattered radiation in FTRS permits the use of longer wavelength incident radiation (e.g., $\lambda = 1064$ nm of Nd-YAG), which may eliminate fluorescence.

DATA ANALYSIS AND INITIAL INTERPRETATION

Raman scattering may result from incident radiation inducing transitions in electronic, vibrational, and rotational states of the scattering medium. Peaks in the Raman spectrum of a solid obtained using visible exciting radiation are generally associated with vibrational modes of the solid. The vibrations may be subdivided into internal modes that arise from molecules or ions that make up the solid and external modes that result from collective modes ($k = 0$) of the crystal. In either case, the presence of a peak in the Raman spectrum due to the vibration requires that the Raman transition be symmetry allowed, i.e., that the integrand in Equation 7 have a nonzero value. In addition, for a measurable Raman intensity, the integral in Equation 7 must have a large enough magnitude. It is generally difficult to calculate the intensity of a Raman peak, but it is rather straightforward to use group theoretical techniques to determine whether or not the vibrational mode is symmetry allowed.

Example: Raman Active Vibrational Modes of α-Al_2O_3

The extremely useful result that the crystallographic point group provides all the information that is needed to know the symmetries of the Raman active vibrational modes of the crystal is derived below (see Appendix). The left-hand column in the character table for each point group lists a set of symbols such as A_1, B_1, and B_2 for the point group C_{2v}. The term A_1 is the list of characters of the matrices that represent the effect of each symmetry operation of the point group on any function listed in the far right-hand column of the character table. A remarkable result of group theory is that the list of characters of the matrices that represent the effect of each symmetry operation of the point group on "any" function can be completely described by a linear combination of the few lists presented in the character table. For example, for the point group C_{2v}, the symmetry of the function x^2 is described by A_1. This means that when the effect of each of the symmetry operations operating on x^2 is represented by a matrix, the characters of the four matrices are 1, 1, 1, 1. In fact, within the context of the point group C_{2v}, the symmetry of any function is given by $aA_1 + bB_1 + cB_2$, where a, b, and c are integers. In the language of group theory, A_1, B_1, and B_2 are called the irreducible representations of the point group C_{2v}. The linear combination of irreducible representations that describe the symmetry of a function is called the symmetry species of the function.

The irreducible representations of the normal vibrational modes of a crystal are most easily identified by the correlation method (Fately et al., 1971). The analysis for the lattice vibrations of α-Al_2O_3 will be summarized here. The details of the analysis are contained in Fately et al. (1971).

The space group of α-Al_2O_3 is D^6_{3d}. There are two molecules of α-Al_2O_3 per Bravais cell. Each atom in the Bravais cell has its own symmetry, termed the site symmetry, which is a subgroup of the full symmetry of the Bravais unit cell. The site symmetry of the aluminum atoms is C_3 and that of the oxygen atoms is C_2. From the character table for the C_3 point group, displacement in the z direction is a basis for the A irreducible representation. Displacements parallel to the x and y axes are bases for the E irreducible representation. The correlation tables for the species of a group and its subgroups (Wilson et al., 1955) correlate the A species of C_3 to A_{1g}, A_{2g}, A_{1u}, and A_{2u} of the crystal point group D_{3d}.

Displacement of the oxygen atom in the z direction is a basis of the A irreducible representation of the C_2 point group, which is the site symmetry of the oxygen atoms in α-Al_2O_3. Displacements of the oxygen atom that are parallel to the x and y axes are bases for the B irreducible representation of C_2. The correlation table associates the A species of C_2 to A_{1g}, E_g, A_{1u}, and E_u of D_{3d}. The B species of C_2 correlates to A_{2g}, E_g, A_{2u}, and E_u of D_{3d}.

After accounting for the number of degrees of freedom contributed by each irreducible representation of C_2 and C_3 to an irreducible representation of D_{3d} and removing the irreducible representations associated with the rigid translation of the entire crystal, the irreducible species for the optical modes of the corundum crystal are determined to be

$$2A_{1g} + 2A_{1u} + 3A_{2g} + 2A_{2u} + 5E_g + 4E_u \qquad (22)$$

Consulting the character table of the point group D_{3d} indicates that the Raman active vibrational modes would span

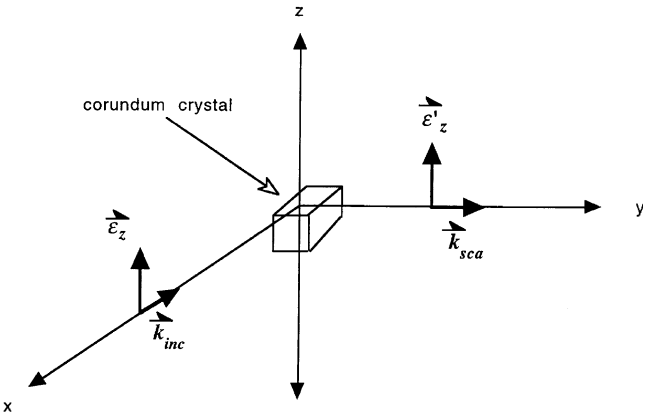


Figure 8. Directions and polarizations of incident and scattered radiation for the $x(zz)y$ configuration.

the symmetry species A_{1g} and E_g. The basis functions for these symmetry species are listed in the character table:

$$x^2 + y^2 \quad \text{and} \quad z^2 \quad \text{for } A_{1g} \tag{23}$$

and

$$(x^2 - y^2, xy) \quad \text{and} \quad (xz, yz) \quad \text{for } E_g \tag{24}$$

Thus there are a total of seven Raman active modes so there may be as many as seven peaks in the Raman spectrum. The question now is which peaks correspond to which vibrational modes.

Figure 8 depicts a Cartesian coordinate system, and a crystal of corundum is imagined to be located at the origin with the crystal axes parallel to x, y, and z. If light is incident on the crystal in the $-x$ direction and polarized in the z direction, and scattered light is collected in the y direction after having passed through a y polarizer, then the experimental conditions of incidence and collection are described by the Porto notation (Porto and Krishnan, 1967) as $x(zz)y$. The first term inside the parentheses indicates the polarization of the incident light and the second term denotes the polarization of the scattered light that is observed. In general, if z polarized light is incident on a crystal in the x direction, then the induced polarization of the molecule will be given by $[p_x, p_y, p_z] = [\alpha_{xz}E_z, \alpha_{yz}E_z, \alpha_{zz}E_z]$. Light scattered in the y direction would be the result of the induced polarization components p_x and p_z. If the light scattered in the y direction is passed through a z polarizer, then the source of the observed radiation would be the induced polarization along the z direction. Consequently, only the α_{zz} component of the polarizability contributes to scattered light in the case of $x(zz)y$. Thus, peaks present in the Raman spectrum must originate from Raman scattering by vibrational modes that span the A_{1g} irreducible species.

Figure 9 presents the Raman spectra reported by Porto and Krishnan (1967) for single-crystal corundum under a variety of incidence and collection conditions. Figure 9A is the Raman spectrum in the $x(zz)y$ condition that was discussed above so the two peaks present in the spectrum correspond to vibrational modes with A_{1g} symmetry. The peaks present in the spectrum in Figure 9B can originate from vibrational modes with either A_{1g} or E_g symmetries. Consequently, the peaks present in Figure 9B but missing in Figure 9A have E_g symmetry. The spectra in Figures 9C-E exhibit only peaks with E_g symmetry. Collectively, the spectra reveal all seven Raman active modes and distinguish between the A_{1g} modes and the E_g modes.

The magnitude of the Raman shift in wavenumbers of each peak with respect to the exciting laser line can be related to the energy of the vibration. Combined with the knowledge of the masses of the atoms involved in the vibrational mode and, e.g., the harmonic oscillator assumption, it is possible to calculate the force constant associated with the bond between the atoms participating in the vibration. If one atom in the molecule were replaced by another without changing the symmetry of the molecule, the peak would shift corresponding to the different strength of the bond and the mass of new atom compared to the original atom. This illustrates why the Raman spectrum can provide information about bond strength and alloying effects.

Other Examples of the Use of Raman Spectroscopy in Studies of Solids

Raman spectroscopy can provide information on the amorphous/crystalline nature of a solid. Many factors can contribute to the widths of peaks in the Raman spectra of a solids. The greatest peak broadening factor can be microcrystallinity. For a macroscopic-sized single crystal of silicon at room temperature, the full width at half-maximum (FWHM) of the peak at 522 cm^{-1} is $\approx$3 cm^{-1} (Pollak, 1991). The breadth of the peak increases dramatically as the size of the crystal decreases below 10 nm. The cause of peak broadening in microcrystals is the finite size of the central peak in the Fourier transform of a phonon in a finite-size crystal. In an infinitely large crystal, the Fourier transform of a phonon consists of a single sharp line at the phonon frequency. The peak width Δk (i.e., FWHM) in a crystal of dimension L is given approximately by

$$\Delta k = \frac{2\pi(v/c)}{L} \tag{25}$$

where v is phonon velocity and c the velocity of light. Assuming $v = 1 \times 10^5$ cm/s, $\Delta k = 20$ cm^{-1} for $L = 10$ nm and $\Delta k = 200$ cm^{-1} for $L = 1$ nm. In the extreme case, the silicon is amorphous, in which case the $k = 0$ selection rule breaks down and all phonons are Raman active. In that case, the Raman spectrum resembles the phonon density of states and is characterized by two broad peaks centered at 140 and 480 cm^{-1} (Pollak, 1991). Consequently, the Raman spectrum can distinguish between the crystalline and amorphous structures of a solid and can indicate the grain size of microcrystalline solids.

Raman spectroscopy can be used to nondestructively measure the elastic stress in crystalline samples. Elastic straining of the lattice will alter the spring constant of a chemical bond and, hence, will shift the frequency of the

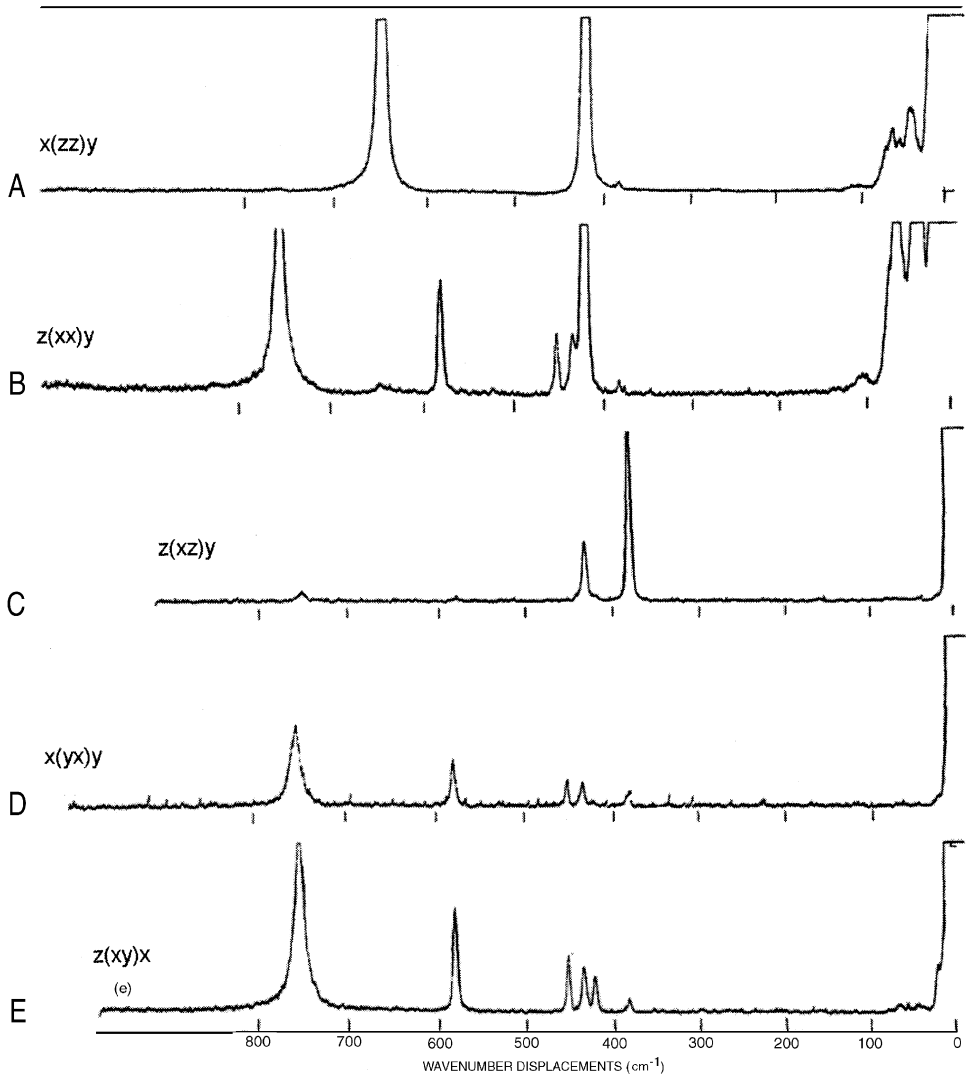


Figure 9. Raman spectra of corundum as a function of crystal and laser light polarization orientations (Porto and Krishnan, 1967).

vibrational mode associated with the strained bond. The direction and magnitude of the shift in peak location will depend on the sign (tensile or compressive) and magnitude, respectively, of the strain. Raman spectroscopy has been used to measure the residual stresses in thin films deposited on substrates that result in lattice mismatches and in thermally cycled composites where the strains result from differences in thermal expansion coefficients of the two materials (Pollak, 1991).

Another example of the use of Raman spectroscopy to nondestructively investigate the mechanical behavior of a solid is its use in studies of Al_2O_3-ZrO_2 composites (Clarke and Adar, 1982). These two-phase structures exhibit improved mechanical toughness due to transformation of the tetragonal ZrO_2 to the monoclinic form. The transformation occurs in the highly stressed region ahead of a growing crack and adds to the energy that must be expended in the propagation of the crack. Using Raman microprobe spectroscopy, it was possible to measure, with a spatial resolution of $\approx$1 μm, the widths of the regions on either side of a propagating crack in which the ZrO_2 transformed. The size of the transformed region is an important parameter in theories that predict the increased toughness expected from the transformation of the ZrO_2.

Phase transformations resulting from temperature or pressure changes in an initially homogeneous material have also been studied by Raman spectroscopy (Ferraro and Nakamoto, 1994). Raman spectra can be obtained from samples at temperatures and pressures that are markedly different from ambient conditions. All that is needed is the ability to irradiate the sample with a laser beam and to collect the scattered light. This can be accomplished by having an optically transparent window (such as silica, sapphire, or diamond) in the chamber that houses the sample and maintains the nonambient conditions. Alternatively, a fiber optic can be used to transmit the

incident light to the sample and the scattered light to the spectrometer.

The influence of composition on the Raman spectrum is in evidence in semiconductor alloys such as $Ga_{1-x}AL_xAs$ $(0 < x < 1)$, which exhibit both GaAs-like and AlAs-like longitudinal optical (LO) and transverse optical (TO) phonon modes. The LO modes have a greater dependence on composition than do the TO modes. The AlAs-like LO mode shifts by 50 cm^{-1} and the GaAs-like mode shifts by $\approx$35 cm^{-1} as x varies from 0 to 1. Thus the Raman spectrum of $Ga_{1-x}Al_xAs$ $(0 < x < 1)$ can be used to identify the composition of the alloy (Pollak, 1991).

Similarly, the Raman spectra of mixtures of HfO_2 and ZrO_2 are strong functions of the relative amounts of the two components. Depending on the particular mode, the Raman peaks shift by $\approx$7 to $\approx$60 cm^{-1} as the amount of HfO_2 increases from 0 to 100% spectroscopy (Ferraro and Nakamoto, 1994).

Raman spectroscopy has long been used to characterize the structure of polymers: identifying functional groups, end groups, crystallinity, and chain orientation. One word of caution: As might be expected of any Raman study of organic matter, fluorescence problems can arise in Raman investigations of polymers (Bower and Maddams, 1989; Bulkin, 1991; Rabolt, 1991).

Raman Spectroscopy of Carbon

Because carbon can exist in a variety of solid forms, ranging from diamond to graphite to amorphous carbon, Raman spectroscopy is the single most effective characterization tool for analyzing carbon. Given the wide spectrum of structures exhibited by carbon, it provides a compelling example of the capability of Raman spectroscopy to analyze the structures of a large variety of solids. The following discussion illustrates the use of Raman spectroscopy to distinguish between the various structural forms of carbon and describes the general procedure for quantitatively analyzing the structure of a solid by Raman spectroscopy.

Figure 10, which was originally presented in an article by Robertson (1991), who compiled the data of a number of researchers, presents the Raman spectra of diamond, large crystal graphite, microcrystalline graphite, glassy carbon, and several forms of amorphous carbon. Diamond has two atoms per unit cell and therefore has a single internal vibrational mode $(3N - 5 = 1)$. This mode is Raman active and is located at 1332 cm^{-1}. Graphite, with four atoms per unit cell, has six internal vibrational modes $(3N - 6 = 6)$, two of which are Raman active. The rigid-layer mode of graphite spans the symmetry species E_{2g} and occurs at 50 cm^{-1}, which is too close to the incident laser line to be accessible with most Raman collection optics and spectrometers. The second Raman active mode of graphite is centered at 1580 cm^{-1}, and it too spans the symmetry species E_{2g}. The relative displacements of the carbon atoms during this in-plane mode are presented next to the Raman spectrum in Figure 10. Thus, Raman spectroscopy can easily distinguish between diamond and large-crystal graphite.

The third spectrum from the top of Figure 10 was obtained from microcrystalline graphite. The next lower spectrum is similar and belongs to glassy carbon. There

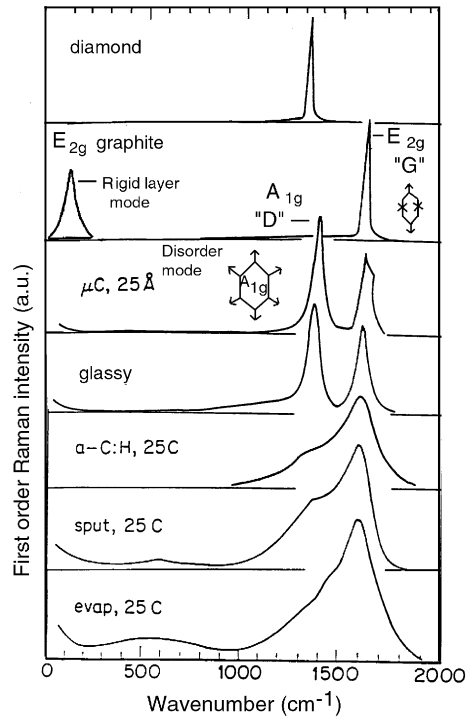


Figure 10. First-order Raman spectra of diamond, highly oriented pyrolytic graphite (hopg), microcrystalline graphite, glassy C, plasma-deposited a-C:H, sputtered a-C, and evaporated a-C. (From Robertson, 1991.)

are two peaks in these spectra, one at 1580 cm^{-1} and the other at 1350 cm^{-1}. Given these peak positions, originally there was some thought that the structures were a mixture of graphitelike (sp^2 bonding) and diamondlike (sp^3 bonding) components. However, experiments that revealed the effects of heat treatments on the relative intensities of the two peaks clearly showed that the material was polycrystalline graphite and the peak at 1350 cm^{-1} is a consequence of the small size of the graphite crystals. The new peak at 1350 cm^{-1} is labeled the disorder, or "D," peak, as it is a consequence of the breakdown of the crystal momentum conservation rule. In the Raman spectrum of an infinite crystal, the peak at 1350 cm^{-1} is absent. This peak results from the vibrational mode sketched next to the spectrum for microcrystalline graphite in Figure 10, and it spans the symmetry species A_{1g}. Its Raman activity is symmetry forbidden in a crystal of infinite size. Finite crystal size disrupts translational symmetry, and so the D mode is Raman active in microcrystalline graphite. The ratio of the integrated intensities of the peaks associated with the D and G modes is proportional to the size of the crystal:

$$\frac{I(D\,\text{mode})}{I(G\,\text{mode})} = \frac{k}{d} \tag{26}$$

where d is the crystal size. This relationship may be used, for example, to calculate the relative sizes of two microcrystals and the relative mean grain diameters of two

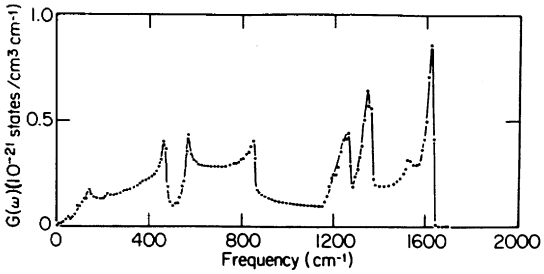


Figure 11. Phonon density of states for graphite.

polycrystalline aggregates, so long as the grain size is greater than 12 Å. The behavior of crystals smaller than 12 Å begins to resemble that of amorphous solids.

An amorphous solid of N atoms may be thought of as a giant molecule of N atoms. All $3N - 6$ of its internal vibrational modes are Raman active, and the Raman spectrum is proportional to the phonon density of states. This is demonstrated by a comparison of the bottom three spectra in Figure 10 with the calculated phonon density of states of graphite in Figure 11.

At this point several important distinctions between x-ray diffraction and Raman spectroscopy can be appreciated. Like Raman spectroscopy, x-ray diffraction is also able to distinguish between crystals, microcrystals, and amorphous forms of the same material. Raman spectroscopy not only is able to distinguish between the various forms but also readily provides fundamental information about the amorphous state as well as the crystalline state. This is due to the fact that x-ray diffraction provides information about the long-range periodicity of the arrangement of atoms in the solid, while Raman spectroscopy provides information concerning the bonds between atoms and the symmetry of the atomic arrangement. On the other hand, while x-ray diffraction is in principle applicable to every crystal, Raman spectroscopy is only applicable to those solids with unit cells containing two or more atoms.

Quantitative Analysis

The quantitative analysis of a spectrum requires fitting each peak to an expression of Raman scattered intensity vs. frequency of the scattered light. Typically, a computer is used to fit the data to some mathematical expression of the relationship between the intensity of the Raman scattered radiation as a function of its frequency with respect to that of the incident laser line. One example of a peak-fitting relationship is the expression (DiDomenico et al., 1968)

$$\frac{dI}{d\omega} = \frac{(\text{const})\Gamma\omega_0}{\{\omega_0^2 - (\Delta\omega)^2\}^2 + 4\Gamma^2\omega_0^2(\Delta\omega)^2} \qquad (27)$$

where $dI/d\omega$ is the average Raman scattered intensity per unit frequency and is proportional to the strength of the signal at the Raman shift of $\Delta\omega$; Γ is the damping constant, ω_0 is the frequency of the undamped vibrational mode, and $\Delta\omega$ is the frequency shift from the laser line. The calculated fit yields the values of Γ and ω_0 for each peak in the spectrum.

An example of the quantitative analysis of a Raman spectrum is provided in the work of Dillon et al. (1984), who used Raman spectroscopy to investigate the growth of graphite microcrystals in carbon films as a function of annealing temperature. The data for the D and G peaks of graphite were fitted to Equation 27, which was then integrated to give the integrated intensity of each peak. The ratio of the integrated intensity of the D mode to the integrated intensity of the G mode was then calculated and plotted as a function of annealing temperature. The ratio increased with annealing temperature over the range from 400 to 600°C, and the widths of the two peaks decreased over the same range. The results suggest that either the number or the size of the crystal grains was increasing over this temperature range. The integrated intensity ratio of the D-to-G mode reached a maximum and then decreased with higher temperatures, indicating growth in the grain size during annealing at higher temperatures. Dillon et al. also used the peak positions defined by Equation 27 to monitor the shift in peak locations with annealing temperature. The results indicated that the annealed films were characterized by threefold rather than fourfold coordination.

In summary, a single-phase, unknown material can be identified by the locations of peaks and their relative intensities in the Raman spectrum. The location of each peak and its breadth can be defined by fitting the peaks to a quantitative expression of peak intensity vs. Raman shift. The same expression will provide the integrated intensity of each peak. Ratios of integrated intensities of peaks belonging to different phases in a multiphase sample will be proportional to the relative amounts of the two phases.

PROBLEMS

Many problems can arise in the generation of a Raman spectrum. Peaks not part of the spectrum of the sample can sometimes appear. Among the factors that can produce spurious peaks are plasma lines from the laser and cosmic rays hitting the light detector. The occurrence of peaks caused by cosmic rays increases with the time required for acquisition of the spectrum. The peaks from cosmic rays can be identified by their sharpness and their nonreproducibility. Peaks attributed to plasma lines are also sharp, but not nearly as sharp as those caused by cosmic rays. A peak associated with a plasma line will vanish when the spectrum is generated using a different laser.

An all too common problem encountered in Raman spectroscopy is the occurrence of fluorescence, which completely swamps the much weaker Raman signal. The fluorescence can be caused by the component of interest or by impurities in the sample. Fluorescence can be minimized by decreasing the concentration of the guilty impurity or by long-time exposure of the impurity to the exciting laser line, which "burns out" the fluorescence. Absorption of the exciting laser line by the impurity results in its thermal decomposition.

Fluorescence is generally red shifted from the exciting laser line. Consequently, the anti-Stokes Raman spectrum may be less affected by fluorescence than the Stokes-Raman spectrum. The intensity of the anti-Stokes Raman spectrum is much weaker than that of the Stokes Raman so this is not always a viable approach to avoiding the fluorescence signal.

Increasing the wavelength of the exciting laser line may also reduce fluorescence, as was mentioned above (see Fourier Transform Raman Spectroscopy). One potential problem with the latter approach is tied to the wavelength dependence of the intensity of the scattered light, which varies inversely as the fourth power of the wavelength. Shifting to exciting radiation with a longer wavelength may reduce fluorescence but it will also decrease the overall intensity of the scattered radiation.

Fluctuations in the intensity of the incident laser power during the generation of a Raman spectrum are particularly problematic for spectra measured using a single-channel detector. Errors in the relative intensities of different peaks can result from fluctuations in the laser power. Similarly, fluctuations in laser power will cause variations in the intensities of successive measurements of the same spectrum using a multiple-channel detector. These problems will be largely averted if the laser can be operated in a constant-light-intensity mode rather than in a controlled (electrical) power input mode.

If the sample absorbs the incident radiation, problems may occur even if the absorption does not result in fluorescence. For example, if an argon laser ($\lambda = 514.5$ nm) were used to generate the Raman spectrum of an adsorbate on the surface of a copper or gold substrate, significant reduction in the Raman intensity would result because of the strong absorption by copper and gold of green light. In this case, better results would be obtained by switching to a krypton laser ($\lambda = 647.1$ nm) or helium-neon laser ($\lambda = 632.8$ nm).

Absorption of the incident radiation can also lead to increases in the temperature of the sample. This may have a major deleterious effect on the experiment if the higher temperature causes a change in the concentration of an adsorbate or in the rate of a chemical or electrochemical reaction. If optically induced thermal effects are suspected, the spectra should be generated using several different incident wavelengths.

Temperature changes can also cause significant changes in the Raman spectrum. There are a number of sources of the temperature dependency of Raman scattering. First, the intensity of the Stokes component relative to the anti-Stokes component decreases as temperature increases. For studies that make use of only the Stokes component, an increase in temperature will cause a decrease in intensity. A second cause of the temperature dependency of a Raman spectrum is the broadening of peaks as the temperature increases. This cause of peak broadening is dependent on the decay mechanism of the phonons. Broadening of the Raman peak can also result from hot bands. Here the higher temperature results in a higher population of the excited state. As a result, the incident photon can induce the transition from the first excited state to the second excited state. Because of anharmonic effects, the energy required for this transition is different from the energy required of the transition from the ground state to the first excited state. As a result, the peak is broadened. Hence, Raman peaks are much sharper at lower temperatures. Consequently, it may be necessary to obtain spectra at temperatures well below room temperature in order to minimize peak overlap and to identify distinct peaks.

Temperature changes can also affect the Raman spectrum due to the temperature dependencies of the phonons themselves. For example, the phonon itself will have a strong temperature dependency as the temperature is raised close to that of a structural phase transition if at least one component of the oscillation coincides with the displacement during the phase transition. Temperature changes can also affect the phonon frequency. As the temperature increases, the anharmonic character of the vibration causes the well in the plot of potential energy vs. atomic displacement to be more narrow than is the case for a purely harmonic oscillation. The atomic displacements are therefore smaller, which results in a higher frequency since the amplitude of displacement is inversely proportional to frequency. In addition, thermal expansion increases the average distance between the atoms as temperature increases, leading to a decrease in the strength of the interatomic interactions. This results in a decrease in the phonon frequency with increasing temperature. In summary, superior Raman spectra are generally obtained at lower temperatures, and it may be necessary in some cases to obtain spectra at temperatures that are considerably lower than room temperature.

Changes in the temperature of the room in which the Raman spectra are measured can cause changes in the positions of optical components (lenses, filters, mirrors, gratings) both inside and outside the spectrometer. Such effects are important when attempting to accurately measure the locations of peaks in a spectrum, or, e.g., stress-induced shifts in peak locations.

One of the more expensive mistakes that can be made in systems using single-channel detectors is the inadvertent exposure of the PMT to the Rayleigh line. The high intensity of the Rayleigh scattered radiation can "burn out" the PMT, resulting in a significant increase in PMT dark counts. Either closing the shutter in front of the PMT or shutting off the high voltage to the PMT will prevent damage to the PMT from exposure to high light intensity. Typically, accidents of this type occur when the frequency of the incident radiation is changed and the operator neglects to note this change at the appropriate point in the software that runs the experiment and controls the operation of the PMT shutter.

ACKNOWLEDGMENTS

It is a pleasure to thank J. Larry Nelson and Wylie Childs of the Electric Power Research Institute for their long-term support and interest in the use of Raman spectroscopy in corrosion investigations. In addition, Gary Chesnut and David Blumer of ARCO Production and Technology have encouraged the use of surface-enhanced Raman spectroscopy in studies of corrosion inhibition.

Both organizations have partially supported the writing of this unit.

The efforts and skills of former and current graduate students, in particular Jing Gui, Lucy J. Oblonsky, Christopher Kumai, Valeska Schroeder, and Peter Chou, have greatly contributed to my continually improved understanding and appreciation of Raman spectroscopy.

LITERATURE CITED

Altmann, S. L. 1991. Band Theory of Solids: An Introduction from the Point of View of Symmetry (see p. 190). Oxford University Press, New York.

Atkins, P. W. 1984. Molecular Quantum Mechanics. Oxford University Press, New York.

Bilhorn, R. B., Epperson, P. M., Sweedler, J. V., and Denton, M. B. 1987b. Spectrochemical measurements with multichannel integrating detectors. *Appl. Spectrosc.* 41:1125–1135.

Bilhorn, R. B., Sweedler, J. V., Epperson, P. M., and Denton, M. B. 1987a. Charge transfer device detectors for analytical optical spectroscopy—operation and characteristics. *Appl. Spectrosc.* 41:1114–1125.

Bishop, D. 1973. Group Theory and Chemistry. Clarendon Press, Oxford.

Bower, D. I. and Maddams, W. F. 1989. The Vibrational Spectroscopy of Polymers. Cambridge University Press, Cambridge.

Bulkin, B. J. 1991. Polymer applications. *In* Analytical Raman Spectroscopy (J. G. Grasselli and B. J. Bulkin, eds.) pp. 45–57. John Wiley & Sons, New York.

Carrabba, M. M., Spencer, K. M., Rich, C., and Rauh, D. 1990. The utilization of a holographic Bragg diffraction filter for Rayleigh line rejection in Raman spectroscopy. *Appl. Spectrosc.* 44: 1558–1561.

Chang, R. K. and Furtak, T. E. (eds.). 1982. Surface Enhanced Raman Scattering. Plenum Press, New York.

Chase, D. B. 1991. Modern Raman instrumentation and techniques. *In* Analytical Raman Spectroscopy (J. G. Grasselli and B. J. Bulkin, eds.) pp. 45–57. John Wiley & Sons, New York.

Clarke, D. R. and Adar, F. 1982. Measurement of the crystallographically transformed zone produced by fracture in ceramics containing tetragonal zirconia. *J. Am. Ceramic Soc.* 65:284–288.

DiDomenico, M., Wemple, S. H., Perto, S. P. S., and Bauman, R. P. 1968. Raman spectrum of single-domain $BaTiO_3$. *Phys. Rev.*

Dillon, R. O., Woollam, J. A., and Katkanant, V. 1984. Use of Raman scattering to investigate disorder and crystallite formation in As-deposited and annealed carbon films. *Phys. Rev. B.*

Epperson, P. M., Sweedler, J. V., Bilhorn, R. B., Sims, G. R., and Denton, M. B. 1988. Applications of charge transfer devices in spectroscopy. *Anal. Chem.* 60:327–335.

Fateley, W. G., McDevitt, N. T., and Bailey, F. F. 1971. Infrared and Raman selection rules for lattice vibrations: The correlation method. *Appl. Spectrosc.* 25:155–173.

Ferraro, J. R. and Nakamoto, K. 1994. Introduction to Raman Spectroscopy. Academic Press, San Diego, Ca.

Fowles, G. R. 1989. Introduction to Modern Optics. Dover Publications, Mineola, N.Y.

Fraser, D. A. 1990. The Physics of Semiconductor Devices. Oxford University Press, New York.

Gui, J. and Devine,T. M. 1991. *In-situ* vibrational spectra from the passive film in iron in buffered borate solution. *Corr. Sci.* 32:1105–1124.

Hamermesh, M. 1962. Group Theory and Its Application to Physical Problems. Dover Publications, New York.

Hecht, J. 1993. Understanding Lasers: An Entry-Level Guide. IEEE Press, Piscataway, N.J.

Hutley, M. C. 1982. Diffraction Gratings. Academic Press, New York.

Kettle, S. F. A. 1987. Symmetry and Structure. John Wiley & Sons, New York.

Koster, G. F. 1957. Space groups and their representations. *In* Solid State Physics, Advances in Research and Applications, Vol. 5 (F. Seitz and D. Turnbull, eds.) 173–256. Academic Press, New York.

Lax, M. 1974. Symmetry Principles in Solid State and Molecular Physics. John Wiley & Sons, New York.

Leech, J. W. and Newman, D. J. 1969. How To Use Groups. Methuen, London.

Long, D. A. 1977. Raman Spectroscopy. McGraw-Hill, New York.

Mariot, L. 1962. Group Theory and Solid State Physics. Prentice-Hall, Englewood Cliffs, N.J.

Meijer, P. H. E. and Bauer, E. 1962. Group Theory—The Application to Quantum Mechanics. North-Holland Publishing, Amsterdam, The Netherlands.

Oblonsky, L. J. and Devine, T. M. 1995. Surface enhanced Raman spectroscopic study of the passive films formed in borate buffer on iron, nickel, chromium and stainless steel. *Corr. Sci.* 37:17–41.

Owen, H. 1992. Holographic optical components for laser spectroscopy applications. *SPIE* 1732:324–332.

Parker, S. F. 1994. A review of the theory of Fourier-transform Raman spectroscopy. *Spectrochim. Acta* 50A:1841–1856.

Pelletier, M. J. and Reeder, R. C. 1991. Characterization of holographic band-reject filters designed for Raman spectroscopy. *Appl. Spectrosc.* 45:765–770.

Pollak, F. H. 1991. Characterization of semiconductors by Raman spectroscopy. *In* Analytical Raman Spectroscopy (J. G. Grasselli and B. J. Bulkin, eds.) pp. 137–221. John Wiley & Sons, New York.

Porto, S. P. S. and Krishnan, R. S. 1967. Raman effect of corundum. *J. Chem. Phys.* 47:1009–10012.

Rabolt, J. F. 1991. Anisotropic scattering properties of uniaxially oriented polymers: Raman studies. *In* Analytical Raman Spectroscopy (J. G. Grasselli and B. J. Bulkin, eds.) pp. 45–57. John Wiley & Sons, New York.

Robertson, J. 1991. Hard amorphous (diamond-like) carbons. *Prog. Solid State Chem.* 21:199–333.

Rossi, B. 1957. Optics. Addison-Wesley, Reading, Mass.

Scherer, J. R. 1991. Experimental considerations for accurate polarization measurements. *In* Analytical Raman Spectroscopy (J. G. Grasselli and B. J. Bulkin, eds.) pp. 45–57. John Wiley & Sons, New York.

Schoen, C. L., Sharma, S. K., Henlsley, C. E., and Owen, H. 1993. Performance of a holographic supernotch filter. *Appl. Spectrosc.* 47:305–308.

Weyl, H. 1931. The Theory of Groups and Quantum Mechanics. Dover Publications, New York.

Wilson, E. B., Decius, J. C., and Cross, P. C. 1955. Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra. McGraw-Hill, New York.

Yang, B., Morris, M. D., and Owen, H. 1991. Holographic notch filter for low-wavenumber stokes and anti-stokes Raman spectroscopy. *Appl. Spectrosc.* 45:1533–1536.

KEY REFERENCES

Altmann, 1991. See above.

In Chapter 11, Bloch sums are used for the eigenvectors of a crystal's vibrational Hamiltonian.

Atkins, 1984. See above.

Provides a highly readable introduction to quantum mechanics and group theory.

Fateley et al., 1971. See above.

Recipe for determining the symmetry species of vibrational modes of crystals.

Ferraro and Nakamoto, 1994. See above.

Provides a more current description of equipment and more extensive and up-to-date discussion of applications of Raman spectroscopy than are provided in the text by Long (see below).

Long, 1977. See above.

Provides a thorough introduction to Raman spectroscopy.

Pollak, 1991. See above.

A comprehensive review of the use of Raman spectroscopy to investigate the structure and composition of semiconductors.

APPENDIX: GROUP THEORY AND VIBRATIONAL SPECTROSCOPY

This appendix was written, in part, to provide a road map that someone interested in vibrational spectroscopy might want to follow in migrating through the huge number of topics that are covered by the vast number of textbooks on group theory (see Literature Cited). Although necessarily brief in a number of areas, it does cover the following topics in some depth, selected on the basis of their importance and/or the rareness with which they are fully discussed elsewhere: (1) How character tables are used to identify the Raman activity of vibrational modes is demonstrated. (2) That the ij component, α_{ij}, of the polarizability tensor has the same symmetry as the product function $x_i x_j$ is demonstrated. Character tables list quadratic functions such as x^2, yz and identify the symmetry species (defined below) spanned by each. The same symmetry species will be spanned by the Raman active vibrational modes. (3) The proper rotations found in crystallographic point groups are identified. (4) A general approach, developed by Lax (1974) and using multiplier groups, is presented that elegantly establishes the link between all space groups, both symmorphic and nonsymmorphic, and the crystallographic point groups.

Vibrational Selection Rules

Quantum mechanically a vibrational Raman transition may occur if the integral in Equation 7 has a nonzero value. This can quickly be determined by the use of group theoretical techniques, which are summarized in the form of a Raman selection rule. The Raman selection rule is a statement of the symmetry that a vibrational mode must possess in order that it may be Raman active.

Before discussing the symmetry-based selection rules for Raman scattering, there is also a restriction based on the principles of energy and momentum conservation

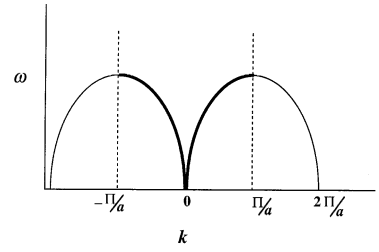


Figure 12. Dispersion of longitudinal wave in a linear monoatomic lattice (nearest-neighbor interactions only).

that is especially helpful in considering the Raman activity of solids. The requirements of energy and momentum conservation are expressed as follows:

$$\hbar\omega_i = \hbar\omega_s \pm \hbar\omega_v \tag{28}$$

$$\hbar k_i = \hbar k_s \pm \hbar k_v \tag{29}$$

where $\hbar k_i, \hbar k_s$, and $\hbar k_v$ are the momentum vectors of the incident and scattered radiation and the crystal phonon, respectively. For optical radiation $|k_i|$ is on the order of 10^4 to 10^5 cm^{-1}. For example, for $\lambda = 647.1$ nm, which corresponds to the high-intensity red line of a krypton ion laser,

$$k = \frac{2\pi}{\lambda} = 9.71 \times 10^3 \text{ cm}^{-1} \tag{30}$$

For a crystal with $a_0 = 0.25$ nm, $k_{v,\max} \approx 2.51 \times 10^8$ cm^{-1}. Hence, for Raman scattering, $k_v \ll k_{v,\max}$. In other words, k_v must be near the center of the Brillouin Zone (BZ).

The phonon dispersion curves for a simple, one-dimensional, monoatomic lattice and a 1D diatomic lattice are presented in Figures 12 and 13. Since there are no available states at the center of the BZ in a monatomic lattice [for a three-dimensional (3D) lattice as well as a one-dimensional lattice], such structures are not Raman active. Consequently, all metals are Raman inactive. Diatomic lattices, whether homonuclear or heteronuclear, do possess phonon modes at the center of the BZ (see Fig. 13) and are Raman active. Thus, crystals such as diamond, silicon, gallium nitride, and aluminum oxide are all Raman active.

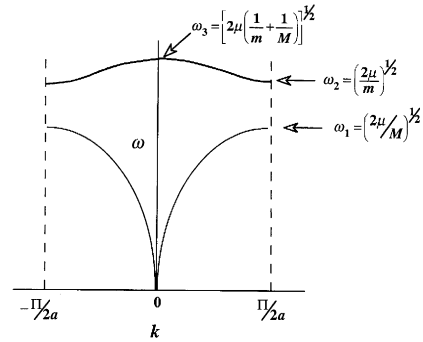


Figure 13. Dispersion of longitudinal wave in a linear diatomic lattice.

A Raman spectrum may consist of several peaks in the plot of the intensity of the scattered radiation versus its shift in wavenumber with respect to the incident radiation, as illustrated in Figure 4. If the immediate task is to identify the material responsible for the Raman spectrum, then it is only necessary to compare the measured spectrum with the reference spectra of candidate materials. Since the Raman spectrum acts as a fingerprint of the scattering material, once a match is found, the unknown substance can be identified.

Often the identity of the sample is known, and Raman spectroscopy is being used to learn more about the material, e.g., the strength of its atomic bonds or the presence of elastic strains in a crystalline matrix. Such information is present in the Raman spectrum but it must now be analyzed carefully in order to characterize the structure of the material.

Once a vibrational Raman spectrum is obtained, the first task is to identify the vibrational modes that contribute to each peak in the spectrum. Group theoretical techniques enormously simplify the analyses of vibrational spectra. At this point, the rules of group theory that are used in vibrational spectroscopy will be cited and examples of their use will then be given. In doing so, it is worth recalling the warning issued by David Bishop in the preface to his textbook on Group Theory (Bishop, 1973, p. vii): "The mathematics involved in actually applying, as opposed to deriving group theoretical formulae is quite trivial. It involves little more than adding and multiplying. It is in fact possible to make the applications, by filling in the necessary formulae in a routine way, without even understanding where the formulae have come from. I do not, however, advocate this practice."

Although the approach taken in the body of this unit ignores Bishop's advice, the intent is to demonstrate the ease with which vibrational spectra can be analyzed by using the tools of group theory. One researcher used the automobile analogy to illustrate the usefulness of this approach to group theory. Driving an automobile can be a very useful practice. Doing so can be accomplished without any knowledge of the internal combustion engine. Hopefully, the elegance of the methodology, which should be apparent from reading this appendix, will encourage the reader to acquire a firm understanding of the principles of group theory by consulting any number of texts that provide an excellent introduction to the subject (see Literature Cited, e.g., Hammermesh, 1962; Koster, 1957; Leech and Newman, 1969; Mariot, 1962; Meijer and Bauer, 1962; Weyl, 1931).

Group Theoretical Tools for Analyzing Vibrational Spectra

Point Groups and Matrix Representations of Symmetry Operations. The collection of symmetry operations of a molecule make up its point group. A point group is a collection of symmetry operations that are linked to one another by the following rules: (1) The identity operation is a member of the set. (2) The operations multiply associatively. (3) If R and S are elements, then so is RS. (4) The inverse of each element is also a member of the set. In satisfying these requirements, the point groups meet the criteria for a mathematical group so that molecular symmetry can be investigated with the aid of group theory.

At this juncture, point groups, which describe the symmetry of molecules, and the use of group theoretical techniques to identify the selection rules for Raman active, molecular vibrations will be described. Although the symmetry of crystals is described by space groups, it will be demonstrated below that the Raman active vibrational modes of a crystal can be identified with the aid of the point group associated with the crystal space group. That is, the Raman spectrum of a crystal is equated to the Raman spectrum of a single unit cell, which is treated like a molecule. Specifically, the symmetry of the unit cell is described by one of 32 point groups. All that is presented about point groups and molecular symmetry will be directly applicable to the use of symmetry and group theory for analyses of vibrational spectra of solids. After completing the discussion of point groups and molecular symmetry, the link between point groups and vibrational spectroscopy of solids will be established.

There are a total of fourteen different types of point groups, seven with one principal axis of rotation; groups within these seven types form a series in which one group can be formed from one other group by the addition of a symmetry operation. The seven other types of point groups do not constitute a series and involve multiple axes of higher order. The symmetries of the vast number of different molecules can be completely represented by a small number of point groups.

The water molecule belongs to the point group C_{2v}, which consists of the symmetry elements $\sigma_{v(xz)}, \sigma_v(yz)$, and C_2, as illustrated in Figure 6, plus the identity operation E. Each symmetry operation can be represented by a matrix that describes how the symmetry element acts on the molecule. First, a basis function that represents the molecule is identified. Examples of basis functions for the water molecule are: (1) the 1s orbitals of the two hydrogen atoms and the 2s orbital of the oxygen atom; (2) the displacement coordinates of the molecule during normal-mode vibrations; and (3) the 1s orbitals of the hydrogen atoms and the $2p_{x,y,z}$ orbitals of oxygen. In fact, there are any number of possible basis functions, although some are more efficient and effective than others in representing the molecule. The number of functions in the basis defines the dimensions of the matrix representations of the symmetry operations. In the case of a basis consisting of three components, e.g., the two 1s orbitals of hydrogen atoms and the 2s orbital of oxygen, each matrix would be 3×3. The action of the symmetry operation on the molecule would then be provided by an equation of the form

$$[f_1 f_2 f_3] = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{pmatrix} \begin{bmatrix} 2s \\ 2s_{\mathrm{a}} \\ 1s_{\mathrm{b}} \end{bmatrix} \qquad (31)$$

where $[f_1, f_2, f_3]$ represents the basis function following the action of the symmetry operation and $1s_{\mathrm{a}}$ represents the 1s orbital on the "a" hydrogen atom. If the symmetry operation under consideration is the twofold axis of rotation,

$$D(R) = \begin{bmatrix} a & b & o \\ c & d & o \\ o & o & e \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} + [e]$$

Figure 14. Matrix representation of the symmetry operation R is given as $D(R)$. In the form depicted, $D(R)$ can be visualized as the "direct sum" of a 2×2 and a 1×1 matrix.

then $[f_1, f_2, f_3]$ is $[2s, 1s_b, 1s_a]$ and the matrix representation of the twofold axis of rotation for this basis function is

$$D(C_2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \tag{32}$$

In instances in which one component of the basis function is unchanged by all of the symmetry operations, each matrix representation has the form depicted in Figure 14. The matrix can be thought of as consisting of the sum of two matrices, a 1×1 matrix and a 2×2 matrix. This type of sum is not an arithmetic operation and is referred to as a "direct sum." The process of converting all of the matrices that represent each of the symmetry operations for a given basis to the direct sum of matrices of smaller dimensions is referred to as "reducing the representation." In some cases each 2×2 matrix representation of the symmetry operations is also diagonalized, and the representation can be reduced to the direct sum of two 1×1 matrices. When the matrix representations have been converted to the direct sum of matrices with the smallest possible dimensions, the process of reducing the representation is complete.

Obviously, there are an infinite number of matrix representations that could be developed for each point group, each arising from a different basis. Fortunately, and perhaps somewhat surprisingly, most of the information about the symmetry of a molecule is contained in the characters of the matrix representations. Knowledge of the full matrix is not required in order to use the symmetry of the molecule to help solve complex problems in quantum mechanics. Often, only the characters of the matrices representing the group symmetry operations are needed.

The many different matrix representations of the symmetry operations can be reduced via similarity transformations to block diagonalized form. In block diagonalized form, each matrix is seen to consist of the direct sum of a number of irreducible matrices. An irreducible matrix is one for which there is no similarity transformation that is capable of putting it into block diagonalized form. A reduced representation consists of a set of matrices, each of which represents one symmetry operation of the point group and each of which is in the identical block diagonalized form.

Now it turns out that the character of a matrix representation is the most important piece of information concerning the symmetry of the molecule, as mentioned above. The list composed of the characters of the matrix representations of the symmetry operations of the group is referred to as the symmetry species of the representation. For each point group there is a small number of unique symmetry species that are composed of the characters of irreducible matrices representing each of the symmetry operations. These special symmetry species are called irreducible representations of the point group. For each point group the number of irreducible representations is equal to the number of symmetry classes.

Remarkably, of the unlimited number of representations of a point group, corresponding to the unlimited number of bases that can be employed, the symmetry species of each representation can be expressed by a linear combination of symmetry species of irreducible representations. The list of the symmetry species of each irreducible representation therefore concisely indicates what is unique about the symmetry of the particular point group. This information is provided in character tables of the point groups.

Character Tables. The character table for the point group C_{2v} is presented in Table 1. The top row lists the symmetry elements in the point group. The second through the fourth rows list the point group irreducible representations. The first column is the symbol used to denote the irreducible representation, e.g., A_1. The numbers in the character tables are the characters of the irreducible matrix representations of the symmetry operations for the various irreducible symmetry species. The right-hand column lists basis functions for the irreducible symmetry species. For example, if x is used as a basis function for the point group C_{2v}, the matrix representations of the various symmetry operations will all be 1×1 matrices with character of 1. The resultant symmetry species, therefore, is 1, 1, 1, 1 and is labeled A_1. On the other hand, if xy is the basis function, the matrix representations of all of the symmetry operations will consist of 1×1 matrices. The resultant symmetry species is 1, 1, -1, -1 and is labeled A_2. If a function is said to have the same symmetry as x, then that function will also serve as a basis for the irreducible representation B_1 in the point group C_{2v}. There are many possible basis functions for each irreducible representation. Those listed in the character tables have special significance for the analyses of rotational, vibrational, and electronic spectra. This will be made clear below. When a function serves as a basis of a representation, it is said to span that representation.

Vibrational Selection Rules. Where group theory is especially helpful is in deciding whether or not a particular integral, taken over a symmetric range, has a nonzero value. It turns out that if the integrand can serve as a basis for the A_1, totally symmetric irreducible representation, then the integral may have a nonzero value. If the integrand does not serve as a basis for the A_1 irreducible representation, then the integral will necessarily be zero.

The symmetry species that is spanned by the products of two functions is obtained by forming the "direct product" of the symmetry species spanned by each function. Tables are available that list the direct products of all possible combinations of irreducible representations for each point group (see, e.g., Wilson et al., 1955; Bishop, 1973; Atkins, 1984; Kettle, 1987).

Thus, if we want to know whether or not a Raman transition from the ground-state vibrational mode n to the first vibrationally excited state m can occur, we need to examine the six integrals of the form

$$\int \psi_m^* \alpha_{ij} \psi_n dV \tag{33}$$

one for each of the six independent values of α_{ij} (or fewer depending on the symmetry of the molecule or crystal). If any one is nonzero, the Raman transition may occur. If all are zero, the Raman transition will be symmetry forbidden.

The Raman selection rule states that a Raman transition between two vibrational states n and m is allowed if the product $\psi_m^* \psi_n$ of the two wave functions describing the m and n vibrational states has the same symmetry species as at least one of the six components of α_{ij}. This rule reflects the fact that for the integral $\int f_1 f_2 f_3 \, dV$ to have a nonzero value when taken over the symmetric range of variables V, it is necessary that the integrand, which is the product of the three functions f_1, f_2, and f_3, must span a symmetry species that is or contains the totally symmetric irreducible representation of the point group.

The symmetry species spanned by the product of three functions is obtained from the direct product of the symmetry species of each function (see, e.g., Wilson et al., 1955; Bishop, 1973; Atkins, 1984; Kettle, 1987). For the symmetry species of a direct product to contain the totally symmetric irreducible representation, it is necessary that the two functions span the same symmetry species or have a common irreducible representation in their symmetry species. The function x^2 acts as a basis for the A_1 irreducible representation of the point group C_{2v}. A basis for the irreducible representation B_1 is given as x. The function x^2x would not span the irreducible representation A_1, and so its integral over all space would be zero. The function xz is also a basis for the irreducible representation B_1 in the point group C_{2v}. Hence, the function $xxz(=x^2z)$ spans the irreducible representation A_1 and its integral over all space could be other than zero. Thus, for the integrand to span the totally symmetric irreducible representation of the molecular point group, it is necessary that the representation spanned by the product $\psi_m^* \psi_n$ be the same as the representation spanned by α_{ij}. This last statement is extremely useful for analyzing the probability of a Stokes Raman transition from the vibrational ground state n to the first vibrational excited state m or for an anti-Stokes Raman transition from the first vibrational excited state m to the vibrational ground state n. The vibrational ground state spans the totally symmetric species. Consequently, the direct product $\psi_m^* \psi_n$ spans the symmetry species of the first vibrational excited state m. Hence, the Raman selection rule means that the symmetry species of α_{ij} must be identical to the symmetry species of the first vibrational excited state.

The symmetry of α_{ij} can be obtained by showing that it transforms under the action of a symmetry operation of the point group in the same way as a function of known symmetry. Here, α_{ij} relates p_i to α_{ij}. If a symmetry operation R is applied to the molecule/crystal, then the applied electric field is transformed to $R = \mathbf{D}(R)$, where $\mathbf{D}(R)$ is the matrix representation of R. The polarizability, which expresses the proportionality between the applied electric field and the induced polarization of the molecule/crystal, must transform in a manner related to but in general different from that of the applied electric field because the direction of $\mathbf{p}$ will generally be different from the direction of ε_j. The polarization of the crystal induced by the applied electric field is transformed to $Rp = D(R)p$ as a result of the action of the symmetry operation R. Consequently, the polarizability transforms as $RR\alpha = D(R)D(R)\alpha$. The transformation of α can be derived as follows (taken from Bishop, 1973).

If the symmetry operation R transforms the coordinate system from x to x', then (summations are over all repeated indices)

$$\begin{aligned} p_i(x') &= \sum \alpha_{ij}(x')\varepsilon_j(x') \\ &= \sum \alpha_{ij}(x') \sum D_{jk}(R)\varepsilon_k(x) \end{aligned} \tag{34}$$

$$p_l(x) = \sum D_{ml}(R) p_m(x') \tag{35}$$

$$p_l(x) = \sum D_{ml}(R) \sum \alpha_{im}(x') \sum D_{jk}(R) \; {}_k(x) \tag{36}$$

Substituting into the last equation the expression

$$p_l(x) = \sum \alpha_{lk}(x) \; {}_k(x) \tag{37}$$

gives

$$\alpha_{lk}(x) = \sum \sum D_{ml}(R) D_{jk}(R) \alpha_{mj}(x') \tag{38}$$

This describes the action of the symmetry operation R on the polarizability α.

Now,

$$x_l = \sum D_{ml}(R) x_m \tag{39}$$

and

$$x_k = \sum D_{jk}(R) x_j \tag{40}$$

so

$$x_l x_k = \sum \sum D_{ml}(R) D_{jk}(R) x_m x_j \tag{41}$$

Comparison of Equations 19 and 22 indicates that α_{lk} transforms as the function $x_l x_k$. Any symmetry species spanned by the function $x_m x_j$ will also be spanned by α_{mj}. Thus, any vibrational mode that spans the same symmetry species as $x_l x_k$ will be Raman active. This information is included in the character tables of the point groups. In the column that lists the basis functions for the various symmetry species of the point group are listed quadratic functions such as $x^2, yz, x^2 + y^2, x^2 + y^2 - 2z^2, \ldots$.

By way of illustration, the character table indicates that if a molecule with point group symmetry C_{2v} has a vibrational mode that spans the symmetry species A_1, it will be Raman active since this symmetry species is also spanned

by the quadratic function x^2. Any vibrational mode that spans a symmetry species that has at least one quadratic function (e.g., x^2, yz) as a basis will be Raman active.

Raman Active Vibrational Modes of Solids

It is now possible to illustrate how Raman spectroscopy can identify the symmetry species of the vibrational modes of a single crystal. To begin, it is important to recognize that the symmetry of a molecule may be lowered when it is part of a crystal. This can be appreciated by considering a simpler case of symmetry reduction: the change in symmetry of a sulfate anion as a consequence of its adsorption on a solid surface. As illustrated in Figure 15, the sulfate anion consists of oxygen nuclei located on the four corners of a tetrahedron and the sulfur nucleus positioned at the geometric center of the tetrahedron. The anion exhibits T_d point group symmetry. If the sulfate is adsorbed on the surface of a solid through a bond formed between the solid and one of the oxygen anions of the sulfate, then the symmetry of the sulfate is lowered to C_{3v}. If two oxygen-surface bonds are formed, in either a bridging or nonbridging configuration, the symmetry of the sulfate is further lowered to C_{2v}. When present in a crystal, a molecule or anion will exhibit either the same or lower symmetry than the free molecule or anion.

The distinctive feature of the symmetry of a crystal is the presence of translational symmetry elements. In fact, it is possible to generate the entire lattice of a crystal by the repeated operation of the three basic primitive translation vectors that define the unit cell. When all possible arrays of lattice points in three dimensions are considered, it turns out that there are only fourteen distinctive types of lattices, called Bravais lattices. The Bravais lattice may not completely describe the symmetry of a crystal. The unit cell may have an internal structure that is not completely specified by the symmetry elements of the bare lattice. The symmetry of the crystal is completely specified by its space group, which is a mathematical group that contains translational operations and that may contain as many as three additional types of symmetry elements: rotations (proper and improper), which are the point group elements, screw axes, and glide planes. The latter two elements may be conceptually subdivided into a rotation (proper = screw axis; improper = glide plane) plus a rational fraction of a primitive lattice translation.

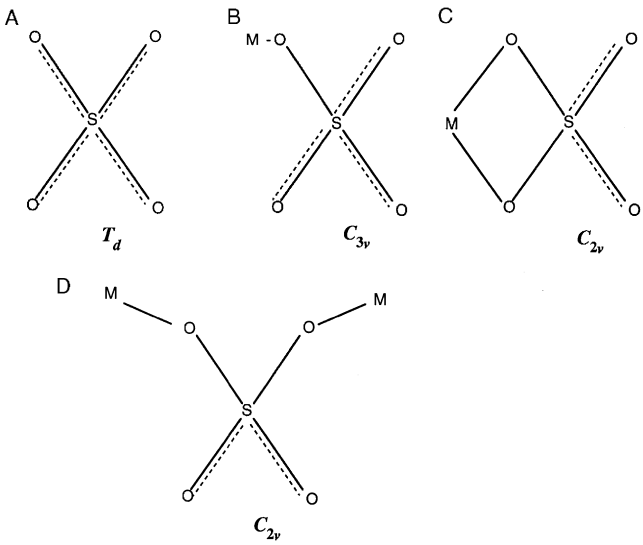


Figure 15. Influence of bonding on the symmetry of sulfate: (**A**) free sulfate; (**B**) unidentate sulfate; (**C**) bidentate sulfate; (**D**) bidentate sulfate in the bridging configuration.

Lattices can be brought into themselves by the operations of certain point groups. Each Bravais lattice type is compatible with only a particular set of point groups, which are 32 in number and are referred to as crystallographic point groups. The appropriate combinations of the 14 Bravais lattices and the 32 crystallographic point groups result in the 230 three-dimensional space groups.

The restrictions that a lattice imposes on the rotational symmetry elements of the space group can be readily illustrated. Each lattice point is located at the tip of a primitive lattice vector given by

$$\mathbf{T} = n_1\mathbf{t}_1 + n_2\mathbf{t}_2 + n_3\mathbf{t}_3 \tag{42}$$

where the $\mathbf{t}_i$ are the basic primitive translation vectors and the n_i are positive and negative integers including zero. The rotation of the lattice translates the points to new locations given by

$$\mathbf{T}' = \boldsymbol{R}T = R_{ij}n_j\mathbf{t}_j \tag{43}$$

where $R_{ij}n_j$ must be integers for all values of n_j, which are integers. Hence, the R_{ij} must be integers as well. For a proper or improper rotation of θ about the z axis,

$$\boldsymbol{R} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & \pm 1 \end{bmatrix} \tag{44}$$

The requirement of integral values for each element R_{ij} means, in particular, that

$$\mathrm{Tr}\ \boldsymbol{R} = 2\cos\theta \pm 1 = \text{integer} \tag{45}$$

The plus sign corresponds to a proper rotation, while the minus sign corresponds to an improper rotation (e.g., reflection or inversion). Thus, the requirement of integral values of R_{ij} limits the possible rotation angles θ to

$$\cos\theta = (n - \pm 1)/2 \Rightarrow \theta = 0, \pi/3, \pi/2, 2\pi/3, 2\pi \tag{46}$$

Crystallographic point groups may therefore contain the following proper rotations: C_1, C_2, C_3, C_4, and C_6, where C_n indicates a rotation angle of $180^\circ/n$ about an axis of the crystal.

Within each space group is an invariant subgroup consisting of the primitive translational operations. An invariant group is one in which a conjugate transformation

produces another member of the group; i.e., if for all elements X and t of the group G, $X^{-1}tX = t'$, where t and t' are elements of the same subgroup of G, then this subgroup is an invariant subgroup of G.

A translational operation is represented by the symbol $\underline{\varepsilon}|\mathbf{t}$, which is a particular case of the more general symbol for a rotation of α followed by a translation of $\mathbf{a}$, $\underline{\alpha}|\mathbf{a}$. The notation $\underline{\varepsilon}$ represents a rotation of 0°. The invariance of the translational subgroup is demonstrated by

$$
\begin{aligned}
(\underline{\alpha}|\mathbf{a})\mathbf{x} &= \underline{\alpha}\mathbf{x} + \mathbf{a} \\
(\underline{\beta}|\mathbf{b})(\underline{\alpha}|\mathbf{a})\mathbf{x} &= \underline{\beta\alpha}\mathbf{x} + \underline{\beta}\mathbf{a} + \mathbf{b} \\
&= (\underline{\beta\alpha}|\underline{\beta}\mathbf{a} + \mathbf{b})\mathbf{x}
\end{aligned}
$$

Since $(\underline{\alpha}|\mathbf{a})^{-1}(\underline{\alpha}|\mathbf{a}) = (\underline{\varepsilon}|0)$

$(\underline{\alpha}|\mathbf{a})^{-1}$ must be given by $(\underline{\alpha}^{-1}| - \underline{\alpha}^{-1}\mathbf{a})$

$$
\begin{aligned}
(\underline{\alpha}|\mathbf{a})^{-1}(\underline{\varepsilon}|\mathbf{t})(\underline{\alpha}|\mathbf{a})\mathbf{x} &= (\underline{\alpha}|\mathbf{a})^{-1}(\underline{\varepsilon}|\mathbf{t})(\underline{\alpha}\mathbf{x} + \mathbf{a}) \\
(\underline{\alpha}|\mathbf{a})^{-1}(\underline{\alpha}\mathbf{x} + \mathbf{t} + \mathbf{a}) &= \underline{\alpha}^{-1}\underline{\alpha}\mathbf{x} + \underline{\alpha}^{-1}\mathbf{t} + \underline{\alpha}^{-1}\mathbf{a} - \underline{\alpha}^{-1}\mathbf{a} \\
&= \mathbf{x} + \underline{\alpha}^{-1}\mathbf{t} \\
&= (\underline{\varepsilon}|\underline{\alpha}^{-1}\mathrm{t})\mathbf{x} \\
&= (\underline{\varepsilon}|\mathbf{t}')\mathbf{x} \qquad (47)
\end{aligned}
$$

Thus, a conjugate transformation of a translation operator $\underline{\varepsilon}|\mathbf{t}$ using a general rotation plus translation operator $\underline{\alpha}|\mathbf{a}$ that is also a member of the group G produces another simple translation operator $\underline{\varepsilon}|\mathbf{t}'$, demonstrating that the translational subgroup is invariant to the conjugacy operation.

The invariant property of the translational subgroup is exploited quite heavily in developing expressions for the irreducible representations of space groups. The difficulty in dealing with the symmetry of a crystal and in analyzing the vibrational modes of a crystal stems from the fact that the translational subgroup is infinite in size. Stated slightly differently, a crystal consists of a large number of atoms so that the number of normal vibrational modes is huge, i.e., $\approx 10^{23}$. Fortunately, group theory provides a method for reducing this problem to a manageable size.

Most textbooks dealing with applications of group theory to crystals develop cofactor expressions for space groups. The cofactor group is isomorphic with the crystallographic point group providing the space group is symmorphic. A symmorphic space group consists of rotational and translational symmetry elements. A nonsymmorphic space group contains at least one symmetry element $(\underline{\alpha}|\mathbf{a})$ in which the rotation α, all by itself, and/or the translation $\mathbf{a}$, all by itself, are not symmetry elements of the space group. Of the 230 three-dimensional space groups, only 73 are symmorphic. Consequently, the use of factor groups is of no benefit for analyzing the vibrational modes of crystals in the 157 nonsymmorphic space groups.

Since the use of cofactor groups is so common, albeit of limited value, the approach will be summarized here and then a more general approach that addresses both symmorphic and nonsymmorphic space groups will be developed.

The translational factor group $G|T$ of the symmorphic space group G consists of

$$G|T = T + \sum s_i T \qquad (i = 2, 3, \ldots, n) \qquad (48)$$

where the s_i are the nontranslational symmetry elements except the identity element $(s_1 = E)$ of the symmorphic space group G whose crystallographic point group is of order n; T is the translational symmetry elements of G. The translational factor group is distinct from the group G because T is treated as a single entity, rather than as the infinite number of lattice translational operations that it is in the group G. The factor group has a different multiplication rule than the space group and T acts as the identity element. As long as G is symmorphic, the factor group $G|T$ is obviously isomorphic with the crystallographic point group of G. Because of the isomorphic relationship of the two groups, the irreducible representations of the crystallographic point group will also serve as the irreducible representations of the factor group $G|T$. The irreducible representations of the factor group provide all the symmetry information about the crystal that is needed to analyze its vibrational motion. On the other hand, if G is nonsymmorphic, then it contains at least one element s_j of the form $[\underline{\alpha}|\mathbf{v}(\alpha)]$, where α and/or $\mathbf{v}(\alpha)$ are not group elements. As a consequence, $G|T$ is no longer symmorphic with any crystallographic point group.

The value of this approach, which makes use of cofactor groups, is that the total number of group elements have been reduced from $g \times 10^{23}$, where g is the number of point group operations, to g. Furthermore, the factor group $G|T$ is isomorphic with the crystallographic point group. Consequently, the irreducible representations of the point group serve as irreducible representations of the factor group. How does this help in the analyses of vibrational modes? Recall that the only Raman active modes are characterized by $k = 0$ (i.e., approximately infinite wavelength) and nonzero frequency. The modes of $k = 0$ consist of corresponding atoms in each unit cell moving in phase. The number of normal modes of this type of motion is given by $3N$, where N is the number of atoms in the unit cell. The symmetry of the unit cell is just that of the factor group. Thus, the irreducible representations of the crystallographic point group provide the irreducible representations of the Raman active vibrational modes of the crystal.

As mentioned above, a more general approach to the group theoretical analysis of crystal vibrational modes will be developed in place of the use of factor groups. This approach is enunciated by Lax (1974).

The group multiplication of symmetry elements of space groups is denoted by

$$
\begin{aligned}
[\underline{\alpha}|\mathbf{v}(\alpha)][\underline{\beta}|\mathbf{v}(\beta)]\mathbf{r} &= [\underline{\alpha}|\mathbf{v}(\alpha)][\underline{\beta}\mathbf{r} + \mathbf{v}(\beta)] \\
&= \underline{\alpha\beta}\mathbf{r} + \underline{\alpha}\mathbf{v}(\beta) + \mathbf{v}(\alpha) \\
&= \underline{\alpha\beta}\mathbf{r} + \underline{\alpha}\mathbf{v}(\beta) + \mathbf{v}(\alpha) + \mathbf{v}(\alpha\beta) - \mathbf{v}(\alpha\beta) \\
&\quad \underline{\alpha\beta}\mathbf{r} + \mathbf{v}(\alpha\beta) + \underline{\alpha}\mathbf{v}(\beta) + \mathbf{v}(\alpha) - \mathbf{v}(\alpha\beta) \\
&= \alpha\beta\mathbf{r} + \mathbf{v}(\alpha\beta) + \mathbf{t} \\
\text{where} \quad \mathbf{t} &= \underline{\alpha}\mathbf{v}(\beta) + \mathbf{v}(\alpha) - \mathbf{v}(\alpha\beta) \\
&= (\underline{\varepsilon}|\mathbf{t})(\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)) \qquad (49)
\end{aligned}
$$

Next, consider the operation of $(\underline{\varepsilon}|\mathbf{t})[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]$ on a Bloch function $\exp(i\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r})$. Functions of this form serve as eigenvectors of a crystal's vibrational Hamiltonian that is expressed in normal-mode coordinates with $u_k(\mathbf{r})$ representing the displacement in the kth normal vibrational mode of the harmonic oscillator at $\mathbf{r}$ (Altmann, 1991). Therefore, the operation of two successive symmetry elements $[\underline{\alpha}|\mathbf{v}(\alpha)]$ and $[\underline{\beta}|\mathbf{v}(\beta)]$ on the Bloch function is given by

$$(\underline{\varepsilon}|\mathbf{t})[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]\exp(i\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r}) \tag{50}$$

Applying $[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]$ to both functions that constitute the Bloch function converts the above expression to

$$(\underline{\varepsilon}|\mathbf{t})\{\exp[i\mathbf{k}\cdot(\underline{\alpha\beta}|\mathbf{v}(\alpha\beta))^{-1}\cdot\mathbf{r}](\underline{\alpha\beta}|\mathbf{v}(\alpha\beta))u_k(\mathbf{r})\} \tag{51}$$

Now focus attention on the argument of the exponential in the previous expression. It may be rewritten as

$$\begin{aligned}&[i\mathbf{k}\cdot((\underline{\alpha\beta})^{-1}|-(\underline{\alpha\beta})^{-1}\mathbf{v}(\alpha\beta))\cdot\mathbf{r}]\\&\quad=[i\mathbf{k}\cdot(\underline{\alpha\beta})^{-1}\cdot[\mathbf{r}-\mathbf{v}(\alpha\beta)]]\\&\quad=[i(\underline{\alpha\beta})\mathbf{k}\cdot(\mathbf{r}-\mathbf{v}(\alpha\beta))]\end{aligned} \tag{52}$$

(1) Substituting this expression for the argument of the exponential back into Equation 51, (2) operating $\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)$ on $\mathbf{r}$ in $u_k(\mathbf{r})$, and (3) including $\underline{\varepsilon}|\mathbf{t}$ in the arguments of both functions of the Bloch function convert Equation 51 to

$$\exp\{i(\alpha\beta)\mathbf{k}\cdot(\underline{\varepsilon}|\mathbf{t})^{-1}\cdot[\mathbf{r}-\mathbf{v}(\alpha\beta)]\}u_k\{(\underline{\varepsilon}|\mathbf{t})^{-1}[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]^{-1}\mathbf{r}\} \tag{53}$$

Now,

$$u_k\{(\underline{\varepsilon}|\mathbf{t})^{-1}[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]^{-1}\mathbf{r}\}=u_k\{[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]^{-1}\mathbf{r}\} \tag{54}$$

because

$$u_k(\mathbf{r}-\mathbf{t})=u_k(\mathbf{r}) \tag{55}$$

where $\mathbf{t}$ is the lattice vector as $u_k(\mathbf{r})$ has the periodicity of the lattice. Equation 53 now becomes

$$\begin{aligned}&\exp\{i(\underline{\alpha\beta})\mathbf{k}\cdot[\mathbf{r}-\mathbf{v}(\alpha\beta)-\mathbf{t}]\}u_k\{[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]^{-1}\mathbf{r}\}\\&\exp[-i(\underline{\alpha\beta})\mathbf{k}\cdot\mathbf{t}]\exp\{i(\underline{\alpha\beta})\mathbf{k}\cdot[\mathbf{r}-\mathbf{v}(\alpha\beta)]\}u_k\{[\underline{\alpha\beta}|\mathbf{v}(\alpha\beta)]^{-1}\mathbf{r}\}\\&\exp[-i(\underline{\alpha\beta})\mathbf{k}\cdot\mathbf{t}](\underline{\alpha\beta}\mathbf{v}(\alpha\beta))\psi(\mathbf{r},\mathbf{k})\end{aligned} \tag{56}$$

where $\psi(\mathbf{r},\mathbf{k})$ is the Bloch function. At this point it is no longer possible to continue in a completely general fashion. To include in Equation 56 all possible values of $\mathbf{k}$, it would be necessary to develop two separate expressions of Equation 56: one for symmorphic space groups and the other for nonsymmorphic space groups. Alternatively, a single expression for Equation 56 can be developed for both symmorphic and nonsymmorphic space groups if we consider only one value of $\mathbf{k}$: the long-wavelength limit, $\mathbf{k}=0$. For $\mathbf{k}=0$, the above group multiplication becomes

$$(\underline{\alpha}|\mathbf{v}(\underline{\alpha}))(\underline{\beta}|\mathbf{v}(\underline{\beta}))=(\underline{\alpha\beta}|\mathbf{v}(\underline{\alpha\beta})) \tag{57}$$

which is the same group multiplication rule obeyed by symmetry elements of point groups (i.e., groups consisting of elements of the form $[\underline{\alpha}|0]$). Thus, the representations of space group elements $[\underline{\alpha}|\mathbf{v}(\alpha)]$ are identical to the representations of the point group elements $[\underline{\alpha}|0]$.

THOMAS M. DEVINE
University of California
Berkeley, California

ULTRAVIOLET PHOTOELECTRON SPECTROSCOPY

INTRODUCTION

Photoemission and Inverse Photoemission

Ultraviolet photoelectron spectroscopy (UPS) probes electronic states in solids and at surfaces. It relies on the process of photoemission, in which an incident photon provides enough energy to bound valence electrons to release them into vacuum. Their energy E, momentum $\hbar\mathbf{k}$, and spin σ provide the full information about the quantum numbers of the original valence electron using conservation laws. Figure 1 depicts the process in an energy diagram. Essentially, the photon provides energy but negligible momentum (due to its long wavelength $\lambda=2\pi/|\mathbf{k}|$), thus shifting all valence states up by a fixed energy ("vertical" or "direct" transitions). In addition, secondary

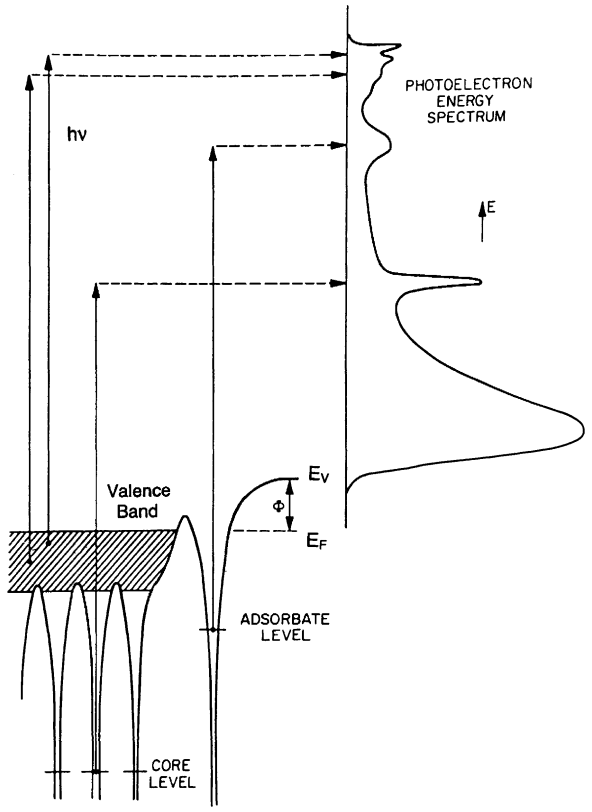


Figure 1. Photoemission process (Smith and Himpsel, 1983).

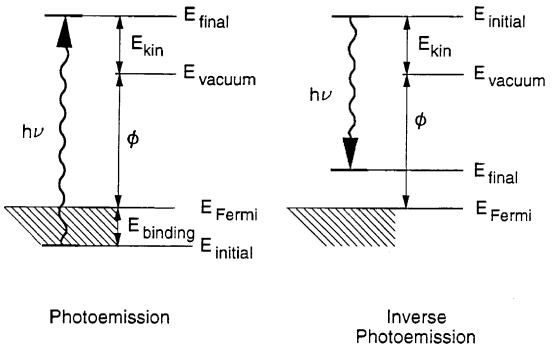


Figure 2. Photoemission and inverse photoemission as probes of occupied and unoccupied valence states: ϕ = work function (Himpsel and Lindau, 1995).

processes, such as energy loss of a photoelectron by creating plasmons or electron-hole pairs, produce a background of secondary electrons that increases toward lower kinetic energy. It is cut off at the vacuum level E_V, where the kinetic energy goes to zero. The other important energy level is the Fermi level E_F, which becomes the upper cutoff of the photoelectron spectrum when translated up by the photon energy. The difference $\phi = E_V - E_F$ is the work function. It can be obtained by subtracting the energy width of the photoelectron spectrum from the photon energy. For reviews on photoemission, see Cardona and Ley (1978), Smith and Himpsel (1983), and Himpsel and Lindau (1995). Information specific to angle-resolved photoemission is given by Plummer and Eberhardt (1982), Himpsel (1983), and Kevan (1992).

Photoemission is complemented by a sister technique that maps out unoccupied valence states, called inverse photoemission or bremsstrahlung isochromat spectroscopy (BIS). This technique is reviewed by Dose (1985), Himpsel (1986), and Smith (1988). As shown in Figure 2, inverse photoemission represents the reverse of the photoemission process, with an incoming electron and an outgoing photon. The electron drops into an unoccupied state and the energy is released by the photon emission.

Both photoemission and inverse photoemission operate at photon energies in the ultraviolet (UV), starting with the work function threshold at ~4 eV and reaching up to 50- to 100-eV photon energy, where the cross-section of valence states has fallen off by an order of magnitude and the momentum information begins to get blurred. At kinetic energies of 1 to 100 eV, the electron mean free path is only a few atomic layers, making it possible to detect surface states as well as bulk states.

Characterization of Valence Electrons

In atoms and molecules, the important parameters characterizing a valence electron are its binding energy (or ionization potential) plus its angular momentum (or symmetry label in molecules). This information is augmented by the vibrational and rotational fine structures, which shed light onto the interatomic potential curves (Turner et al., 1970). In a solid, it becomes necessary to consider not only energy but also momentum. Electrons in a crystalline solid are completely characterized by a set of quantum numbers. These are energy E, momentum $\hbar\mathbf{k}$, point group symmetry (i.e., angular symmetry), and spin. This information can be summarized by plotting $E(\mathbf{k})$ band dispersions with the appropriate labels for point group symmetry and spin. Disordered solids such as random alloys can be characterized by average values of these quantities, with disorder introducing a broadening of the band dispersions. Localized electronic states (e.g., the 4*f* levels of rare earths) exhibit flat $E(\mathbf{k})$ band dispersions. Angle-resolved photoemission, combined with a tunable and polarized light source, such as synchrotron radiation, is able to provide the full complement of quantum numbers. In this respect, photoemission and inverse photoemission are unique among other methods of determining the electronic structure of solids.

Before getting into the details of the technique, its capabilities are illustrated in Figure 3, which shows how much information can be extracted by various techniques about the band structure of Ge. Optical spectroscopy integrates over the momentum and energy of the photoelectron and leaves only the photon energy $h\nu$ as variable. The resulting spectral features represent regions of momentum and energy space near critical points at which the density of transitions is high (Fig. 3A). By adjusting an empirical band structure to the data (Chelikowski and Cohen, 1976; Smith et al., 1982), it is possible to extract rather accurate information about the strongest critical points. Angle-integrated photoemission goes one step further by detecting the electron energy in addition to the photon energy (Fig. 3B). Now it becomes possible to sort out whether spectral features are due to the lower state or the upper state of the optical transition. To extract all the information on band dispersion, it is necessary to resolve the components of the electron momentum parallel and perpendicular to the surface, $\mathbf{k}^{\parallel}$ and $k^{\perp}$, by angle-resolved photoemission with variable photon energy (Fig. 3C).

Photoemission and inverse photoemission data can, in principle, be used to derive a variety of electronic properties of solids, such as the optical constants [from the $E(\mathbf{k})$ relations; see Chelikowski and Cohen (1976) and Smith et al. (1982)], conductivity (from the optical constants), the electron lifetime [from the time decay constant τ or from the line width δE; see Olson et al. (1989) and Haight (1995)], the electron mean free path [from the attenuation length λ or from the line width δk; see Petrovykh et al. (1998)], the group velocity of the electrons [via the slope of the $E(\mathbf{k})$ relation; see Petrovykh et al. (1998)], the magnetic moment (via the band filling), and the superconducting gap (see Olson et al., 1989; Shen et al., 1993; Ding et al., 1996).

Energy Band Dispersions

How are energy band dispersions determined in practice? A first look at the task reveals that photoemission (and inverse photoemission) provides just the right number of independent measurable variables to establish a unique correspondence to the quantum numbers of an electron in a solid. The energy E is obtained from the kinetic energy of the electron. The two momentum components parallel to

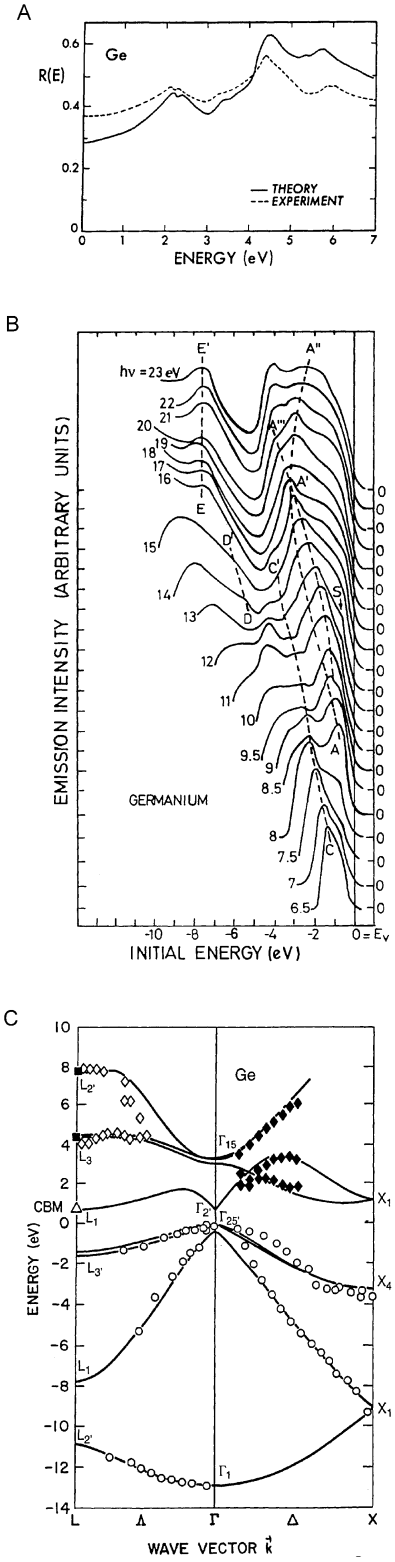


Figure 3. Comparison of various spectroscopies applied to the band structure of germanium. (**A**) The optical reflectivity $R(E)$ determines critical points near the band gap, (**B**) angle-integrated photoemission adds critical points farther away, and (**C**) angle-resolved photoemission and inverse photoemission provide the full $E(\mathbf{k})$ band dispersion (Phillip and Ehrenreich, 1963; Grobman et al., 1975; Chelikowski and Cohen, 1976; Wachs et al., 1985; Hybertsen and Louie, 1986; Ortega and Himpsel, 1993).

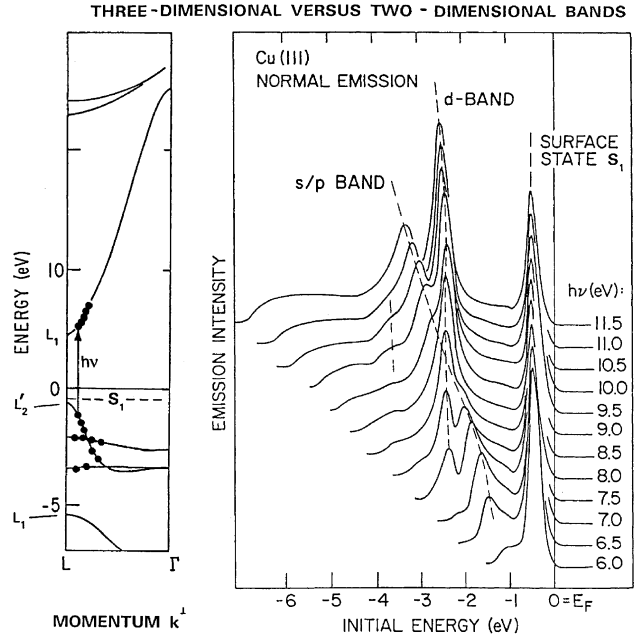


Figure 4. Mapping bulk and surface bands of copper by angle-resolved photoemission. Variable photon energy from monochromatized synchrotron radiation makes it possible to tune the $\mathbf{k}$ component perpendicular to the surface $k^{\perp}$ and to distinguish surface states from bulk states by their lack of $k^{\perp}$ dispersion (Knapp et al., 1979; Himpsel, 1983).

the surface, $\mathbf{k}^{\parallel}$, are derived from the polar and azimuthal angles ϑ and ϕ of the electron. The third momentum component, $k^{\perp}$, is varied by tuning the photon energy $h\nu$. This can be seen in Figure 4, in which angle-resolved photoemission spectra at different photon energies hν are plotted together with the relevant portion of the band structure. The parallel momentum component $\mathbf{k}^{\parallel}$ is kept zero by detecting photoelectrons in normal emission; the perpendicular component $k^{\perp}$ changes as the photon energy of the vertical interband transitions is increased. A complete band structure determination requires a tunable photon source, such as synchrotron radiation (or a tunable photon detector in inverse photoemission). Surface states, such as the state near E_F, labeled S_1 in Figure 4, do not exhibit a well-defined $k^{\perp}$ quantum number. Their binding energy does not change when the $k^{\perp}$ of the upper state is varied by changing $h\nu$. This lack of $k^{\perp}$ dispersion is one of the characteristics of a surface state. Other clues are that a surface state is located in a gap of bulk states with the same $\mathbf{k}^{\parallel}$ and the surface state is sensitive to contamination. To complete the set of quantum numbers, one needs the point group symmetry and the spin in ferromagnets. The former is obtained via dipole selection rules from the polarization of the photon, the latter from the spin polarization of the photoelectron.

For two-dimensional (2D) states in thin films and at surfaces, the determination of energy bands is almost trivial since only E and $\mathbf{k}^{\parallel}$ have to be determined. These quantities obey the conservation laws

$$E_1 = E_u - h\nu \qquad (1)$$

and

$$\mathbf{k}_1^{\parallel} - \mathbf{k}_u^{\parallel} - \mathbf{g}^{\parallel} \qquad (2)$$

where $\mathbf{g}^{\parallel}$ is a vector of the reciprocal surface lattice, u denotes the upper state, and l the lower state. These conservation laws can be derived from the invariance of the crystal with respect to translation in time and in space (by a surface lattice vector). For the photon, only its energy $h\nu$ appears in the balance because the momentum of a UV photon is negligible compared to the momentum of the electrons. The subtraction of a reciprocal lattice vector simply corresponds to plotting energy bands in a reduced surface Brillouin zone, i.e., within the unit cell in $\mathbf{k}^{\parallel}$ space. For a three-dimensional (3D) bulk energy band, the situation becomes more complicated since the momentum component perpendicular to the surface is not conserved during the passage of the photoelectron across the surface energy barrier. However, $k^{\perp}$ can be varied by changing the photon energy $h\nu$, and extremal points, such as the Γ and L points in Figure 4, can thus be determined. A discussion of the practical aspects and capabilities of various band mapping schemes is given below (see Practical Aspects of the Method) and in several reviews (Plummer and Eberhardt, 1982; Himpsel, 1983; Kevan, 1992). Experimental energy bands are compiled in Landolt-Börnstein (1989, 1994).

In ferromagnets the bands are split into two subsets, one with majority spin, the other with minority spin (see Fig. 5). The magnetic exchange splitting δE_{ex} between majority and minority spin bands is the key to magnetism. It causes the majority spin band to become filled more than the minority band, thus creating the spin imbalance that produces the magnetic moment. An overview of the band structure of ferromagnets and magnetic thin-film structures is given in Himpsel et al. (1998).

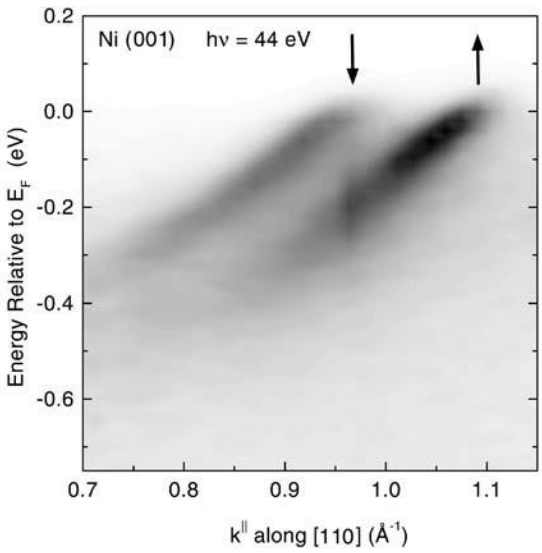


Figure 5. Using multidetection of energy and angle to map out the $E(\mathbf{k})$ band dispersion of Ni near the Fermi level E_F from a Ni(001) crystal. The majority and minority spin bands are split by the ferromagnetic exchange splitting δE_{ex}. High photoemission intensity is shown in dark (Petrovykh et al., 1998).

Competitive and Related Techniques

The relation of UPS to optical spectroscopy has been mentioned in the context of Figure 3 already. Several other spectroscopies involve core levels but provide information about valence electrons as well. Figure 6 shows the simplest processes, which involve transitions between two levels only, one of them a core level. More complex phenomena involve four levels, such as Auger electron spectroscopy (see AUGER ELECTRON SPECTROSCOPY) and appearance potential spectroscopy.

Core-level photoelectron spectroscopy determines the energy of a core level relative to the Fermi level E_F (Fig. 6A) and is known as x-ray photoelectron spectroscopy (XPS) or electron spectroscopy for chemical analysis (ESCA). A core electron is ionized, and its energy is obtained by subtracting the photon energy

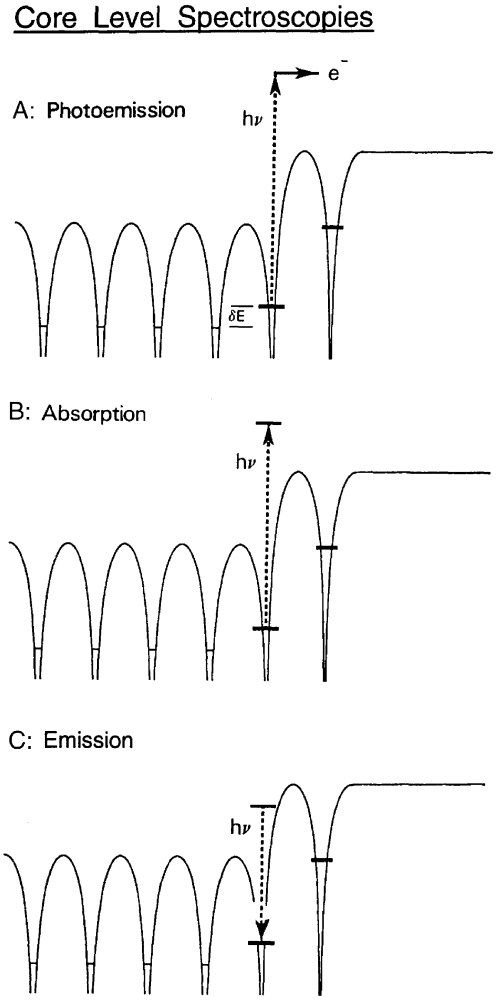


Figure 6. Spectroscopies based on core levels. (**A**) X-ray photoelectron spectroscopy measures the binding energy of core levels, (**B**) core-level absorption spectroscopy detects transitions from a core level into unoccupied valence states, and (**C**) core-level emission spectroscopy detects transitions from occupied valence states into a core hole. In contrast to UPS, these spectroscopies are element-specific but cannot provide the full momentum information.

from the kinetic energy of the photoelectron. The energy shifts of the substrate and adsorbate levels are a measure of the charge transfer and chemical bonding at the surface.

Core-level absorption spectroscopy determines the pattern of unoccupied orbitals by exciting them optically from a core level (Fig. 6B). It is also known as near-edge x-ray absorption fine structure (NEXAFS; see XAFS SPECTROSCOPY) or x-ray absorption near-edge structure (XANES). Instead of measuring transmission or reflectivity, the absorption coefficient is determined in surface experiments by collecting secondary products, such as photoelectrons, Auger electrons, and core-level fluorescence. The short escape depth of electrons compared to photons provides surface sensitivity. These spectra resemble the density of unoccupied states, projected onto specific atoms and angular momentum states. In a magnetic version of this technique, magnetic circular dichroism (MCD), the difference in the absorption between parallel and antiparallel alignment of the electron and photon spin is measured. Compared to UPS, the momentum information is lost in core-level absorption spectroscopy, but element sensitivity is gained. Also, the finite width of the core level is convolved with the absorption spectrum and limits the resolution.

Core-level emission spectroscopy can be viewed as the reverse of absorption spectroscopy (Fig. 6C). Here, the valence orbital structure is obtained from the spectral distribution of the characteristic x rays emitted during the recombination of a core hole with a valence electron. The core hole is created by either optical or electron excitation. As with photoemission and inverse photoemission, core-level emission spectroscopy complements core-level absorption spectroscopy by mapping out occupied valence states projected onto specific atoms.

PRINCIPLES OF THE METHOD

The Photoemission Process

The most general theory of photoemission is given by an expression of the golden rule type, i.e., a differential cross-section containing a matrix element between the initial and final states ψ_i and ψ_f and a phase space sum (see Himpsel, 1983; Dose, 1985; Smith, 1988):

$$\frac{d\sigma}{d\Omega} \sim \sqrt{E_{\text{kin}}} \sum_i |\langle \Psi_f | \mathbf{A} \cdot \mathbf{p} + \mathbf{p} \cdot \mathbf{A} | \Psi_i \rangle|^2 \times \delta(E_f - E_i - h\nu) \tag{3}$$

The factor $\sqrt{E_{\text{kin}}}$ containing the kinetic energy E_{kin} of the photoelectron represents the density of final states, the scalar product of the vector potential $\mathbf{A}$ of the photon and the momentum operator $\mathbf{p} = -i\hbar\partial/\partial\mathbf{r}$ is the dipole operator for optical excitation, and the δ function represents energy conservation. This is often called the one-step model of photoemission, as opposed to the three-step model, which approximates the one-step model by a sequence of three simpler steps. For practical purposes (see Practical Aspects of the Method; Himpsel, 1983), we have to consider mainly the various selection rules inherent in this expression, such as the conservation of energy (Equation 1), parallel momentum (Equation 2), and spin, together with the point group selection rules. Since there is no clear-cut selection rule for the perpendicular momentum, it is often determined approximately by using a nearly free electron final state.

While selection rules provide clear yes-no decisions, there are also more subtle effects of the matrix element that can be used to bring out specific electronic states. The atomic symmetry character determines the energy dependence of the cross-section (Yeh and Lindau, 1985), allowing a selection of specific orbitals by varying the photon energy. For example, the *s,p* states in transition and noble metals dominate the spectra near the photoelectric threshold, while the *d* states turn on at 10 eV above threshold. It takes photon energies of 30 eV above threshold to make the *f* states in rare earths visible. Resonance effects at a threshold for a core-level excitation can also enhance particular orbitals. Conversely, the cross-section for states with a radial node exhibits so-called Cooper minima, at which the transitions become almost invisible.

The wave functions of the electronic states in solids and at surfaces are usually approximated by the ground-state wave functions obtained from a variety of schemes, e.g., empirical tight binding and plane-wave schemes (Chelikowski and Cohen, 1976; Smith et al., 1982) and first-principles local density approximation (LDA; see Moruzzi et al., 1978; Papaconstantopoulos, 1986). Strictly speaking, one should use the excited-state wave functions and energies that represent the hole created in the photoemission process or the extra electron added to the solid in the case of inverse photoemission. Such quasiparticle calculations have now become feasible and provide the most accurate band dispersions to date, e.g., the so-called *GW* calculations, which calculate the full Green's function *G* of the electron/hole and the fully screened Coulomb interaction *W* but still neglect vertex and density gradient corrections (Hybertsen and Louie, 1986; see also BONDING IN METALS). Particularly in the case of semiconductors the traditional ground-state methods are unable to determine the fundamental band gap from first principles, with Hartree-Fock overestimating it and LDA underestimating it, typically by a factor of 2. The band width comes out within 10% to 20 % in LDA calculations.

Various types of wave functions can be involved in the photoemission process, as shown in Figure 7. The states that are propagating in the solid lead to vertical transitions that conserve all three momentum components and are being used for mapping out bulk bands. Evanescent states conserve the parallel momentum only and are more sensitive to the surface. At elevated temperatures one has to consider phonon-assisted transitions, which scramble the momentum information about the initial state completely.

The materials property obtained from angle-integrated UPS is closely related to the density of states and is often interpreted as such. Strictly speaking, one measures

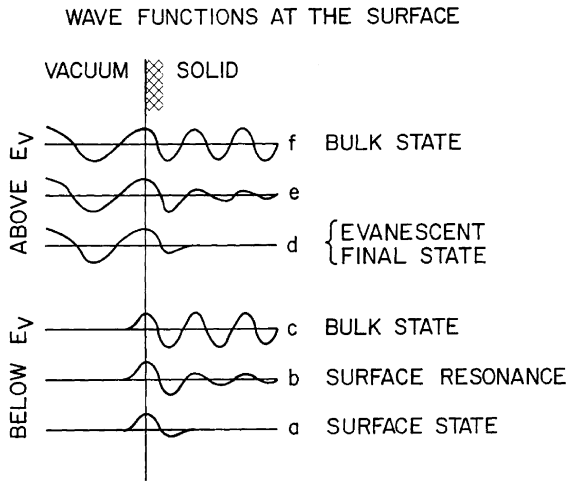


Figure 7. Various types of wave functions encountered in UPS. Propagating bulk states give rise to transitions that conserve all three $\mathbf{k}$ components; evanescent states ignore the $k^{\perp}$ component (Smith and Himpsel 1983).

the energy distribution of the joint density of initial and final states, which includes the optical matrix element (Grobman et al., 1975). Angle-resolved UPS spectra provide the $E(\mathbf{k})$ band dispersions that characterize electrons in a solid completely. Since the probing depth is several atomic layers in UPS, it is possible to determine these properties for the bulk as well as for the surface.

Energy Bands in Solids

The mapping of energy bands in solids and at surfaces is based on the conservation laws for energy and parallel momentum (Equations 1 and 2). Figures 3, 4, and 5 give examples. Presently, energy bands have been mapped for practically all elemental solids and for many compounds. Such results are compiled in Landolt-Börnstein (1989, 1994). For the ferromagnets, see Himpsel et al. (1998). Since energy band dispersions comprise the full information about electrons in solids, all other electronic properties can in principle be obtained from them. Such a program has been demonstrated for optical properties (Chelikowski and Cohen, 1976; Smith et al., 1982). For ferromagnetism, the most significant band parameters are the exchange splitting δE_{ex} between majority and minority spin bands and the distance between the top of the majority spin d bands and the Fermi level (Stoner gap). The exchange splitting is loosely related to the magnetic moment ($\sim$1 eV splitting per Bohr magneton), the Stoner gap to the minimum energy for spin flip excitations (typically from a few tenths of an electron volt down to zero).

Surface States

The electronic structure of a surface is characterized by 2D energy bands that give the relation between E and $\mathbf{k}^{\parallel}$. The momentum perpendicular to the surface, $k^{\perp}$, ceases to be a good quantum number. The states truly specific to the surface are distinguished by their lack of interaction with bulk states, which usually requires that they are located at points in the $E(\mathbf{k}^{\parallel})$ diagram where bulk states of the same symmetry are absent. This is revealed in band diagrams where the $E(\mathbf{k}^{\parallel})$ band dispersions of surface states are superimposed on the regions of bulk bands projected along $k^{\perp}$.

On metals, one finds fairly localized, d-like surface states and delocalized s,p-like surface states. The d states carry most of the spin polarization in ferromagnets. Their cross-section starts dominating relative to s,p states as the photon energy is increased. An example of an s,p-like surface state is given in Figure 4. On the Cu(111) surface a p_z-like surface state appears close to the Fermi level (S_1). It is located just above the bottom of the L_2'-L_1 gap of the bulk s, p band. A very basic type of s,p surface state is the image state in a metal. The negative charge of an electron outside a metal surface induces a positive image charge that binds the electron to the surface. In semiconductors, surface states may be viewed as broken bond orbitals.

Electronic Phase Transitions

The electronic states at the Fermi level are a crucial factor in many observed properties of a solid. States within a few kT of the Fermi level determine the transport properties, such as electrical conductance, where k is the Boltzmann constant and T the temperature in kelvin. They also drive electronic phase transitions, such as superconductivity, magnetism, and charge density waves. Typically, these phase transitions open a gap at the Fermi level of a few multiples of kT_C, where T_C is the transition temperature. Occupied states near the Fermi level move down in energy by half the gap, which lowers the total energy. In recent years, the resolution of UPS experiments has reached a level where sub-kT measurements have become nearly routine. The best-known examples are measurements of the gap in high-temperature superconductors (Olson et al., 1989; Shen et al., 1993; Ding et al., 1996). Compared to other techniques that probe the superconducting gap, such as infrared absorption and tunneling, angle-resolved photoemission provides its $\mathbf{k}$ dependence.

Atoms and Molecules

Atoms and molecules in the gas phase exhibit discrete energy levels that correspond to their orbitals. There is an additional fine structure due to transitions between different vibrational and rotational states in the lower and upper states of the photoemission process. This fine structure disappears when molecules are adsorbed at surfaces. Nevertheless, the envelope function of the molecular orbitals frequently survives and facilitates the fingerprinting of adsorbed species. A typical molecular photoelectron spectrum is shown in Figure 8 (Turner et al., 1970), a typical spectrum of an adsorbed molecular fragment in Figure 9 (Sutherland et al., 1997).

Ultraviolet photoelectron spectroscopy has been used to study surface reactions in heterogeneous catalysis and

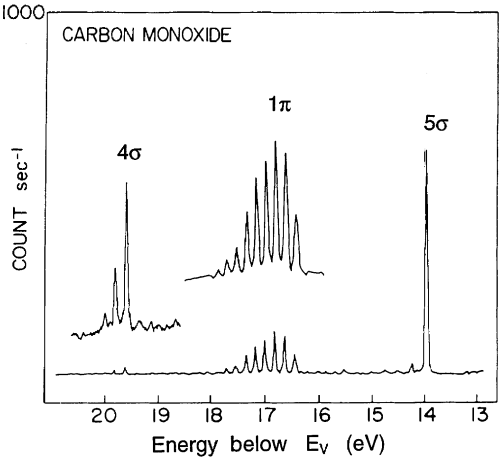


Figure 8. UPS spectrum of the free CO molecule, showing several molecular orbitals and their vibrational fine structure. Note that the reference level is the vacuum level E_V for atoms and molecules, whereas it is the Fermi level E_F for solids (Turner et al., 1970).

in semiconductor surface processing. For example, the methyl and ethyl groups adsorbed on a silicon surface in Figure 9 exhibit the same C 2*s* orbital splittings as in the gas phase (compare Pireaux et al., 1986), allowing a clear identification of the fragments that remain on the surface after reaction with dimethylsilane and diethylsilane. Generally, one distinguishes the weak adsorption on inert surfaces at low temperatures (physisorption regime) and strong adsorption on reactive surfaces at room temperature (chemisorption). During physisorption, molecular orbitals are shifted rigidly toward higher energies by 1 to 2 eV due to dielectric screening of the valence holes created in the photoemission process. Essentially, surrounding electrons in the substrate lower their energy by moving toward the valence hole on the adsorbate molecule, generating extra energy that is being transferred to the emitted photoelectron. In the chemisorption regime, certain molecular orbitals react with the substrate and experience an additional chemical shift. A well-studied example is the adsorption of CO on transition metals, where the chemically active 5σ orbital is shifted relative to the more inert 4σ and 1π orbitals. The orientation of the adsorbed molecules can be determined from the angle and polarization dependence of the UPS spectra using dipole selection rules. If the CO molecule is adsorbed with its axis perpendicular to the surface and photoelectrons are detected along that direction, the σ orbitals are seen with the electric field vector **E** of the light along the axis and the π orbitals with **E** perpendicular to it. Similar selection rules apply to mirror planes and make it possible to distinguish even and odd wave functions.

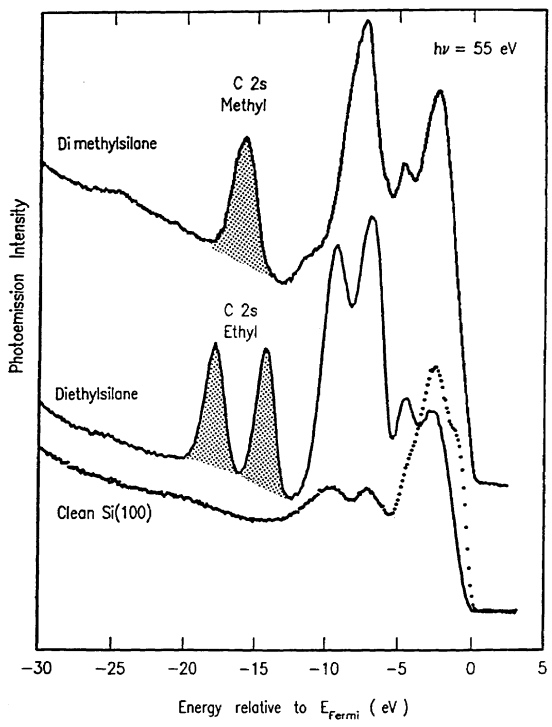


Figure 9. Fingerprinting of methyl and ethyl groups deposited on a Si(100) surface via reaction with dimethylsilane and diethylsilane. The lowest orbitals are due to the C 2*s* electrons. The number of carbon atoms in the chain determines the number of C 2*s* orbitals (Sutherland et al., 1997).

PRACTICAL ASPECTS OF THE METHOD

Light Sources

To be able to excite photoelectrons across a valence band that is typically 10 to 20 eV wide, the photon energy has to exceed the valence band width plus the work function (typically 3 to 5 eV). This is one reason conventional lasers have been of limited applicability in photoelectron spectroscopy, except for two-photon pump-probe experiments with short pulses (Haight, 1995). The most common source of UV radiation is based on a capillary glow discharge using the He(I) line. It provides monochromatic radiation at 21.2 eV with a line width as narrow as 1 meV (Baltzer et al., 1993; a powerful electron cyclotron resonance source is marketed by Gammadata). This radiation originates from the 2*p*-to-1*s* transition in neutral He atoms. Emission can also be produced from He ions, whose primary He(II) emission line is at an energy of 40.8 eV. Similarly, emission lines of Ne(I) at 16.8 eV and Ne(II) at 26.9 eV are being used in photoelectron spectroscopy.

During the last two decades, synchrotron radiation has emerged both as a powerful and convenient excitation source in photoelectron spectroscopy (Winick and Doniach, 1980; Koch, 1983; Winick et al., 1989, some synchrotron light sources are listed in Appendix A). Synchrotron radiation is emitted by relativistic electrons kept in a circular orbit by bending magnets. In recent years, undulators have been developed that contain 10 to 100 bends in a row. For a well-focused electron beam and a highly perfect magnetic field, photons emitted from all bends are superimposed coherently. In this case, the amplitudes add up, as opposed to the intensities, and a corresponding increase in spectral brilliance by 2 to 4 orders of magnitude is

achievable. The continuous synchrotron radiation spectrum shifts its weight toward higher energies with increasing energy of the stored electrons and increasing magnetic fields. A variety of electron storage rings exists worldwide that are dedicated to synchrotron radiation. Three spectral regions can be distinguished by their scientific scope: For the photon energies of 5 to 50 eV required by UPS, storage ring energies of 0.5 to 1 GeV combined with undulators are optimum. For core-level spectroscopies at 100- to 1000-eV photon energy, a storage ring energy of 1.5 to 2 GeV is the optimum. For studies of the atomic structure of solids, even shorter wavelengths are required, which correspond to photon energies of 5 to 10 keV and storage ring energies of 6 to 8 GeV combined with an undulator. Bending magnets emit higher photon energies than undulators due to their higher magnetic field. Synchrotron radiation needs to be monochromatized for photoelectron spectroscopy. The typical range of UPS is covered best by normal-incidence monochromators, which provide photon energies up to $\sim$30 eV with optimum intensity and up to $\sim$50 eV with reduced intensity, while suppressing higher energy photons from higher order diffracted light and harmonics of the undulators. Resolutions of better than 1 meV are routinely achievable now. Synchrotron radiation has a number of desirable properties. The most widely utilized properties are a tunable photon energy and a high degree of polarization. Tunable photon energy is necessary for reaching the complete momentum space by varying $k^{\perp}$, for distinguishing surface from bulk states, and for adjusting relative cross-sections of different orbitals. Linear polarization (in the plane of the electron orbit) is critical for determining the point group symmetry of electrons in solids. Circular polarization (above and below the orbit plane) allows spin-specific transitions in magnetic systems.

Some synchrotron facilities are listed in Appendix A, and figures of merit to consider regarding light sources and detectors for UPS are listed in Appendix B.

Electron Spectrometers

The most common type of photoelectron spectrometers have been the cylindrical mirror analyzer (CMA) for high-throughput, angle-integrated UPS and the hemispherical analyzer for high-resolution, angle-resolved UPS. Popular double-pass CMAs were manufactured by PHI. With hemispherical spectrometers, resolutions of $<$3 meV are achievable with photoelectrons. A recent high-resolution model is the Scienta analyzer developed in Uppsala (Martensson et al., 1994; marketed by Gammadata). As the energy resolution of spectrometers improves, the sample temperature needs to be reduced to take full advantage of that resolution for solid-state samples. For example, the thermal broadening of the Fermi edge is $\sim$0.1 eV at room temperature. Temperatures below 10 K are required to reach the limits of today's spectrometers. With increasing energy and angular resolution, the throughput of electron spectrometers drops dramatically. This can be offset by parallel detection. Energy multidetection in hemispherical analyzers has become a widely available option. It is possible to detect an angle together with the energy in hemispherical and toroidal analyzers. Such a 2D scheme consists of a channel plate electron multiplier with optical readout. An example is given in Figure 5, where the band structure of Ni is mapped using a hemispherical Scienta analyzer with multidetection of the energy and the polar angle (Petrovykh et al., 1998). Other designs detect two angles simultaneously and display the angular distribution of monochromatized photoelectrons directly on a channel plate using an ellipsoidal electron mirror combined with a spherical retarding grid (Eastman et al., 1980). Photoemission intensity distributions over an 85° emission cone can be acquired in a few seconds at an undulator synchrotron light source.

To detect the spin of photoelectrons, one typically measures the right/left asymmetry in the electron-scattering cross-section at heavy nuclei (Mott scattering) after accelerating photoelectrons to 20 to 100 keV (Burnett et al., 1994). Less bulky detectors use scattering at the electron cloud surrounding a heavy nucleus, typically at 100- to 200-eV electron energy. This can be achieved by random scattering or by diffraction at a single crystal. Such detectors lack the absolute calibration of a Mott detector. Generally, spin detection costs 3 to 4 orders of magnitude in detection efficiency.

Alignment Procedures

The electron spectrometer needs to be aligned with respect to the light source (such as a synchrotron beam) and the sample with respect to the electron spectrometer. These seemingly straightforward steps can take a substantial amount of time and may lead to artifacts by inexperienced users, such as a strongly energy- and angle-dependent transmission of the spectrometer. Residual electric and magnetic fields in the sample region make the alignment more difficult and prevent low-energy electrons of a few electron volts from being detected. Electrostatic deflection plates in the spectrometer lens speed up the alignment.

First, the light beam needs to intersect the spectrometer axis at the focal point of the spectrometer. A course alignment can be achieved by using a laser beam along the axis of the spectrometer and zero-order light at a synchrotron. Hemispherical spectrometers often have small holes in the hemispheres for that purpose. The two beams can be brought to intersect at a target sample. If they do not cross each other, there will not be any sample position for a reasonable spectrum. To move the intersection point to the focal point, it is useful to have mechanical alignment markers, such as wire cross-hairs for sighting the proper distance from the front lens of the spectrometer.

Second, the sample needs to be placed at the intersection of the light beam and the analyzer axis. Again, the laser and the light beam can be utilized for this purpose. If they are not available (e.g., when bringing in a new sample into a prealigned spectrometer), a systematic optimization of the photoemission intensity can be used: First, set the sample at the focal distance from the analyzer, as determined by the cross-hairs. Then translate the sample in the two directions orthogonal to the analyzer axis and determine the cutoff points for the photoemission intensity

at opposite edges of the sample and place the sample half way in between.

These alignment procedures provide a first approximation, but residual fields and an uncertain position of the ultraviolet photon beam on the sample require a symmetry check for acquiring accurate angle-resolved UPS data. Mirror symmetry with respect to crystallographic planes provides the exact normal emission geometry and the azimuthal orientation of the sample.

Useful Relations

Some relations occur frequently when processing UPS data. In the following we list the most common ones:

The energy $h\nu$ of a photon is given (in electron volts) by its wavelength λ (in angstroms) via

$$\lambda = \frac{12399}{h\nu} \tag{4}$$

The wave vector $k = |\mathbf{k}|$ of an electron is given (in reciprocal angstroms) by its kinetic energy E_{kin} (in electron volts) via

$$k = 0.51\sqrt{E_{\text{kin}}} \tag{5}$$

The inverse of this relation is

$$E_{\text{kin}} = 3.81 \times k^2 \tag{6}$$

For determining $k^{\perp}$, the upper energy bands are often approximated by the "empty lattice" bands, which are given by

$$E = V + 3.81 \times (k \pm |g|)^2 \tag{7}$$

where k is in reciprocal angstroms and E in electron volts, V is the inner potential (typically -10 eV relative to the vacuum level or -5 eV relative to the Fermi level), and $\mathbf{g}$ (also in reciprocal angstroms) is a reciprocal lattice vector determined by the spacing d_{plane} of equivalent atomic planes:

$$|\mathbf{g}| = \frac{2\pi}{d_{\text{plane}}} \quad \mathbf{g} \perp \text{planes} \tag{8}$$

Likewise, a reciprocal surface lattice vector $\mathbf{g}^{\parallel}$ is determined by the spacing d_{row} of equeivalent atomic rows:

$$|\mathbf{g}| = \frac{2\pi}{d_{\text{row}}} \quad \mathbf{g} \perp \text{rows} \tag{9}$$

The empty lattice approximation used in Figure 10 and Equation 7 improves at higher photon energies (>30 eV), where the kinetic energy of the photoelectrons becomes large compared to the Fourier components of the lattice

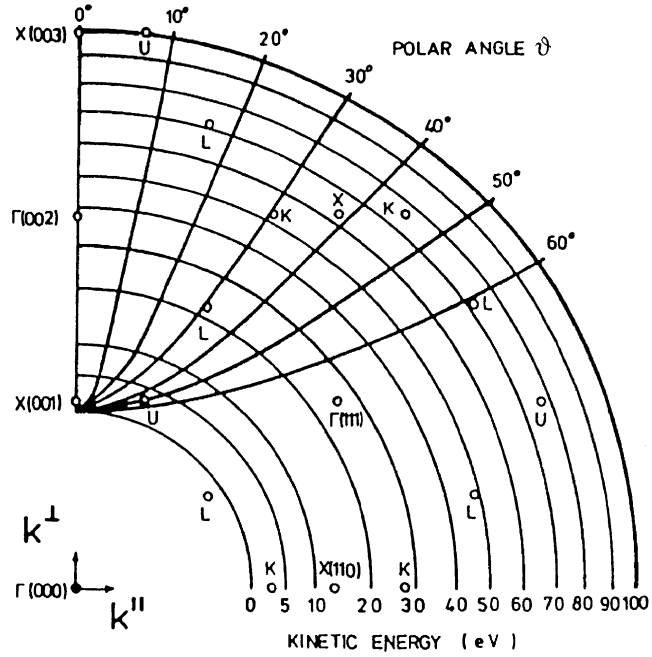


Figure 10. Location of direct transitions in **k** space, shown for a Ni(001) surface along the [110] emission azimuth. The circular lines represent the location of possible upper states at a given kinetic energy using empty lattice bands. The lines fanning out toward the top represent different emission angles from the [001] sample normal (Himpsel, 1983).

potential.Various "absolute" methods are available to refine the **k** values. These methods determine the $k^{\perp}$ component by triangulation from different crystal faces or by mirror symmetry properties of high-symmetry lines (see Plummer and Eberhardt, 1982; Himpsel, 1983; Kevan, 1992).

The thermal broadening of the Fermi function that cuts off photoelectron spectra at high energy is proportional to kT. Taking the full width at half-maximum of the derivative of the Fermi function, one obtains a width $3.5kT$, which is 0.09 eV at room temperature and 1 meV at liquid He temperature. The half-height points of the derivative correspond to the 15% and 85% levels of the Fermi function.

Sensitivity Limits

A typical measurement volume in UPS is an area 1 mm^2 times a depth of 1 nm. In photoelectron microscopes, it is possible to reduce the sampling area to dimensions of <0.1 μm. Several types of such microscopes have been tested in recent years. Imaging devices accelerate the photoelectrons to ~20 keV and use regular electron microscope optics to form an image. In scanning microscopes, the sample is scanned across a finely focused ultraviolet light and soft x-ray source. In both cases, the photoelectron energy analysis has been rather rudimentary so far. Most contrast mechanisms utilized to date involved core levels or work function differences, and not fine structure in the valence band.

The sensitivity to dilute species is rather limited in UPS due to the substantial background of secondary electrons (see Fig. 1). A surface atom density of 10^{14} cm^{-2} (about a tenth of a monolayer) is relatively easy to detect, a concentration of 10^{13} cm^{-2} is the limit. A much higher signal-to-background ratio can be achieved by core-level emission spectroscopy (see Fig. 6C).

The time resolution can be as short as tens of femtoseconds in pump-probe two-photon photoemission experiments using fast lasers (Haight, 1995).

METHOD AUTOMATION

While the sample preparation is highly individual in UPS, the actual data acquisition can be automated easily. A few voltages are sufficient to control a photoelectron spectrometer, such as the kinetic energy and the pass energy. The photon energy is controlled by a monochromator at a synchrotron light source, which in most cases is computer controlled. Labview has become a widely used control system. The angular parameters can be set by stepping motors that drive goniometers for the sample and the spectrometer, with a maximum of five axes, three for the sample and two for the spectrometer. Typically, the acquisition of a series of spectra is preprogrammed. They start out with a wide kinetic energy range at high photon energy, which gives an overview of all valence states, plus some shallow core levels for elemental characterization. Subsequent spectra zero in on specific states, such as adsorbate orbitals or Fermi-level crossings. For each of them, one wants to step through a variety of emission angles and photon energies to vary the parallel and perpendicular momenta, respectively.

DATA ANALYSIS AND INITIAL INTERPRETATION

The first look at UPS spectra from a new sample is typically reserved for making sure that the surface is not contaminated. For example, oxygen contamination generates a characteristic O 2*p* peak at ~6 eV below the Fermi level in metals. Then one might want to compare the spectrum to a set of published reference spectra from related materials or molecules. This "fingerprinting" often gives a quick overview of the energies of specific orbitals, such as *d* states in transition metals, *f* states in rare earths, and adsorbate orbitals. When in doubt, the photon energy dependence of the cross-section can be utilized to enhance certain orbitals by resonant photoemission at core-level thresholds (e.g., *f* levels) or to quench them at Cooper minima (e.g., orbitals with radial nodes).

The next step is often a determination of the Fermi-level position from the high-energy cutoff of a metallic sample, such as Au or Pt. A fit of the Fermi-Dirac function convoluted by a Gaussian resolution function over the uppermost few tenths of an electron volt gives an accurate Fermi-level position.

After characterizing the sample by angle-integrated UPS, the full $E(\mathbf{k})$ band dispersion is obtained from the variation with angle and photon energy. Usually, the starting point is normal emission at various photon energies, which provides the $E(k^{\perp})$ relation along a high-symmetry line. This geometry features the strictest selection rules and, therefore, is the easiest to interpret. An example is given in Figure 4, where a series of normal emission spectra at various photon energies (on the right side) is converted into a set of energy bands (on the left side). The k values of the data points on the left are obtained from the peak positions on the right (dashed lines) after adding the photon energies and converting the resulting upper state energies E_u into $k^{\perp}$ using the calculated upper band on the left (cf. Equation 1). Alternatively, one could use the somewhat cruder empty lattice bands from Equation 7.

By switching the polarization of the electric field vector of the photons between parallel and perpendicular to the surface (the latter asymptotically), different orbital orientations can be selected, such as $p_{x,y}$ and p_z, respectively. In normal emission, for example, the $p_{x,y}$ bands are excited by the component of the electric field vector in the x,y directions (parallel to the sample surface), and the p_z states are excited by the z component normal to the surface.

Having covered normal emission, it is useful to vary the polar angle within a mirror plane perpendicular to the surface, where the wave functions still exhibit even/odd symmetry properties. Here, orbitals that are even with respect to the mirror plane are excited by the in-plane electric field component, and odd orbitals are excited by the component perpendicular to the mirror plane.

Off-normal photoemission peaks can be assigned to transitions at specific $\mathbf{k}$ points by using Equations 5 to 7, as in Figure 10. Taking the data in Figure 5 as example, let us follow a typical processing sequence. The data in Figure 5 are taken from a Ni(001) crystal in the (110) emission plane using a photon energy $h\nu = 44$ eV. In the raw data, the Fermi-level crossings occur at a voltage $V_{\text{spectr}} \approx 40$ V on the electron spectrometer and at a polar angle $\vartheta \approx 18°$ from the sample normal. The spectrometer voltage yields a kinetic energy $E_{\text{kin}} \approx 40$ eV when discounting work function differences between spectrometer and sample (see below). From E_{kin} and ϑ, one obtains a parallel wave vector $k^{\parallel} = k \sin\vartheta \approx 1$ Å^{-1} via Equation 5. Going with E_{kin} and $k^{\parallel}$ into Figure 10, one finds the location of the transition in $\mathbf{k}$ space as the intersection of the circle for $E_{\text{kin}} \approx 40$ eV and the vertical line for $k^{\parallel} = 1$ Å^{-1}. The k scale in Figure 10 is in units of $2\pi/a = 1.8$ Å^{-1} (a = cubic lattice constant of Ni). Consequently, the transition happens about half way on a horizontal line between $\Gamma(002)$ and K, where a line for $\vartheta = 18°$ would intersect the circle for $E_{\text{kin}} = 40$ eV.

It should be noted that the energy scale provided by the voltage reading of an electron spectrometer is referred to the vacuum level of the spectrometer and not to that of the sample. A contact potential develops spontaneously between the sample and the spectrometer that compensates for the difference in work functions and accelerates or retards the photoelectrons slightly toward the analyzer. Such a contact potential needs to be compensated for accurate angular measurements of slow photoelectrons (a few electron volts). A typical spectrometer work function is 4 eV, so samples with smaller work functions require

negative bias and samples with larger work function positive bias. The optimum bias is determined such that the vacuum-level cutoff coincides with a zero voltage reading on the spectrometer. A negative sample bias of a few volts is often used for moving the spectrum up in energy when determining the sample work function from the difference between the photon energy and the width of the photoelectron spectrum (Fig. 2).

With two-dimensional detectors, such as the E, ϑ detection of the hemispherical analyzer (Martensson et al., 1994; Petrovykh et al., 1998; Fig. 5) and the ϑ, φ detection of the ellipsoidal mirror analyzer (Eastman et al., 1980), one needs to divide by the two-dimensional transmission function of the spectrometer to obtain quantitative data. An approximate transmission function can be achieved by increasing the photon energy at identical spectrometer settings, such that secondary electrons are detected whose angular distribution is weak. Careful alignment of sample and light source with respect to the spectrometer and thorough shielding from magnetic and electric fields minimize transmission corrections. After obtaining the 2D intensity distributions $I(E, \vartheta)$ and $I(\vartheta, \varphi)$ and converting ϑ, φ into $\mathbf{k}^{\|}$, one is able to generate $I(\mathbf{k}^{\|})$ distributions. These are particularly useful at the Fermi level, where the features are sharpest and the the electronic states relevant for electronic transport, magnetism, and superconductivity are located.

SAMPLE PREPARATION

Well-characterized surfaces are essential for reliable UPS data. In general, data from single crystals (either in bulk form or as epitaxial films) are more reliable than from polycrystalline materials. Single-crystal surfaces can be cleaned of adsorbed molecules and segregated bulk impurities more thoroughly and can be characterized by diffraction methods, such as low-energy electron diffraction (LEED) and reflection high-energy electron diffraction (RHEED). Surface impurities are monitored by Auger electron spectroscopy or by various core-level spectroscopies, including photoemission itself (see Fig. 6). A wide variety of methods have been developed for preparing single-crystal surfaces, with treatments specific to individual materials and even specific crystal faces. Many of them are compiled by Musket et al. (1982; see also SAMPLE PREPARATION FOR METALLOGRAPHY).

Metals are typically sputter annealed, whereby impurities are removed from the surface by Ar^+ ion bombardment (typically 1 keV at a few microamperes of current and an Ar pressure of 5×10^{-5} Torr). Bulk impurities continue to segregate to the surface and are sputtered away one monolayer at a time, a process that can take months to complete. Therefore, many chemical methods have been developed to deplete the bulk more quickly of impurities. For example, the typical impurities C, N, O, and S can be removed from a catalytically active surface, such as Fe, by heating in 1 atm of hydrogen, which forms volatile CH_4, NH_3, H_2O, and H_2S. Careful electropolishing can also go a long way in removing damage from mechanical polishing and producing a highly perfect surface after just a few sputter anneals, particularly for the soft noble metals. Highly reactive metals such as rare earths, actinides, and early transition metals are prepared in purest form as epitaxial films on a W(110) surface, where they can also be removed by simple heating to make room for a new film.

Semiconductors can be cleaved or heated above the desorption point of the native oxide. Only the crystal orientations with the lowest surface energy cleave well, such as Ge(111), Si(111), and GaAs(110). On silicon, the native oxide film of $\sim$10 Å thickness desorbs at $\sim$1050°C. Dipping into ultrapure 10% HF solution removes the oxide and leaves a mostly H-terminated surface that can be cleaned at 850°C and below. If carbon is present at the surface, it forms SiC particles upon heating that require a flash to 1250°C for diffusing carbon away from the surface into the bulk.

Insulator surfaces are most perfect if they can be obtained by cleavage, which produces the electrically neutral surfaces, e.g., (100) for the NaCl structure and (111) for the CaF_2 structure. Sputter annealing tends to deplete ionic insulators of the negative ion species.

Layered compounds such as graphite, TaS_2, or certain high-temperature superconductors are easiest to prepare. A tab is glued to the front surface with conductive, vacuum-compatible epoxy and knocked off in vacuum. A Scotch tape will do the same job but needs to be kept out of the ultrahigh-vacuum chamber.

If only polycrystalline samples are available, one sometimes needs to resort to cruder cleaning methods, such as filing in vacuum with a diamond file. It is not clear, however, how much of the surface exposed by filing consists of grain boundaries with segregated contamination.

Sample mounting needs to be matched to the preparation method. Samples can be clamped down by spring clips or spotwelded to wires. The material for the springs or wires should not diffuse along the sample surface. Molybdenum is a widely useful spring material that remains elastic even at high temperatures. Tantalum spotwelds very well and diffuses little. Heating can be achieved ohmically or by electron bombardment. The former is better controllable; the latter makes it easier to reach high temperatures. It is particularly difficult to mount samples that require high-temperature cleaning (such as W) but have to be cooled and insulated electrically as well. In such a case, the sample is attached to a cryostat via a sapphire plate that insulates electrically but conducts the heat at low temperature. Most of the temperature drop during heating occurs at W heating wires that attach the sample to a copper plate on top of the sapphire. For high-resolution measurements, it is helpful to minimize potential fluctuations around the sample by painting the area surrounding the sample with graphite using an Aquadag spray. This method is also used to coat the surfaces of electron spectrometers seen by the electrons. Insulating samples need special mounting to avoid charging. Often, the sample potential stabilizes itself after a few seconds of irradiation and leads to a shifted spectrum with reduced kinetic energies. Mounting an electron flood gun near the sample can provide zero-kinetic-energy electrons that compensate the charging. Alternatively, mild heating of the sample may create thermal carriers to provide sufficient conductivity.

Molecules for adsorption experiments are admitted through leak valves that allow gas dosing with submonolayer control. A dosage of 1 s at 10^{-6} Torr (= 1 Langmuir = 1 L) corresponds to every surface atom being exposed to about one gas molecule. The probability of an arriving molecule to stick to the surface (the sticking coefficient) ranges from unity down to 10^{-3} in typical cases.

PROBLEMS

Potential problems can arise mainly due to insufficient sample preparation and mounting. Ultraviolet photoelectron spectroscopy is a surface-sensitive technique that probes only a few atomic layers and thus detects surface contamination and structural defects with uncanny clarity. A thorough literature search is advisable on preparation methods for the particular material used. Insulating samples need charging compensation. Certain semiconductor surfaces produce a significant photovoltage at low temperature, which leads to an energy shift. The determination of the emission angle from purely geometrical sample alignment often is not accurate due to deflection of the photoelectrons by residual fields. A symmetry check is always advisable.

On the instrumentation side, there are a variety of lesser pitfalls. If the sample and spectrometer are not aligned properly to the light beam, the spectra can be highly distorted by energy- and angle-dependent transmission. A typical example is the loss of intensity for slow electrons near the vacuum level. It is important to have a clearly thoughtout alignment procedure, such as the one described above. Residual magnetic and electric fields have the same effect, e.g., due to insufficient magnetic shielding and charging insulators in the sample region, respectively. Another artifact is the flattening of intense photoemission peaks due to a saturation of the electron detector. The channeltron or channelplate multipliers used for electron detection have a rather clearly defined saturation point where the voltage drop induced by the signal current reduces the applied high voltage and thus reduces the gain. Low-resistance "hot" channel plates have higher saturation count rates, which can be as large as 10^7 counts/s for large plates. The high-resolution capability of hemispherical analyzers is easily degraded by incorrect electrical connections, such as ground loops or capacitive pickup. It may be necessary to use a separation transformer for the electronics.

Finally, the photoemission process itself is not always straightforward to interpret. Many different types of wave functions involved are in UPS (see Fig. 7). Some of them conserve all $\mathbf{k}$ components (direct transitions between bulk states), others conserve $\mathbf{k}^{\|}$ only (evanescent states), and phonon-assisted transitions do not conserve $\mathbf{k}$ at all. That leaves quite a few options for the assignment of spectral features that have led to long-standing controversies in the past, such as direct versus indirect transitions. Many years of experience with well-defined single-crystal surfaces and support from calculations of photoemission spectra have made it possible to sort out the various processes, as long as a full set of data at various photon energies, angles, and surface orientations is available. Such complexity is inherent in a process that is capable of determining all the details of an electron wave function.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Award No. DMR-9531009 (the Synchrotron Radiation Center at the University of Wisconsin at Madison) and by the Department of Energy under Contract No. DE-AC03-76SF00098 (the Advanced Light Source at the Lawrence Berkeley National Laboratory).

LITERATURE CITED

Baltzer, P., Karlsson, L., Lundqvist, M., and Wannberg, B. 1993. *Rev. Sci. Instrum.* 62:2174.

Burnett, G. C., Monroe, T. J., and Dunning, F. B. 1994. *Rev. Sci. Instrum.* 65:1893.

Cardona, M. and Ley, L. 1978. Photoemission in Solids I, General Principles, Springer Topics in Applied Physics, Vol. 26. Springer-Verlag, Berlin.

Chelikowski, J. R. and Cohen, M. L. 1976. *Phys. Rev. B* 14:556.

Ding, H., Yokoya, T., Campuzano, J. C., Takahashi, T., Randeria, M., Norman, M. R., Mochiku, T., Hadowaki, K., and Giapintzakis, J. 1996. *Nature* 382:51.

Dose, V. 1985. *Surf. Sci. Rep.* 5:337.

Eastman, D. E., Donelon, J. J., Hien, N. C., and Himpsel, F. J. 1980. *Nucl. Instrum. Methods* 172:327.

Grobman, W. D., Eastman, D. E., and Freeouf, J. L. 1975. *Phys. Rev. B* 12:4405.

Haight, R. 1995. *Surf. Sci. Rep.* 21:275.

Himpsel, F. J. 1983. *Adv. Phys.* 32:1–51.

Himpsel, F. J. 1986. *Comments Cond. Mater. Phys.* 12:199.

Himpsel, F. J. and Lindau, I. 1995. Photoemission and photoelectron spectra. *In* Encyclopedia of Applied Physics, Vol. 13 (G. L. Trigg and E. H. Immergut, eds.) pp. 477–495. VCH Publishers, New York.

Himpsel, F. J., Ortega, J. E., Mankey, G. J., and Willis, R.F. 1998. *Adv. Phys.* 47:511–597.

Hybertsen, M. S. and Louie, S. G. 1986.. *Phys. Rev. B* 34:5390.

Kevan, S. D. (ed.) 1992. Angle-Resolved Photoemission. Elsevier, Amsterdam, The Netherlands.

Knapp, J. A., Himpsel, F. J., and Eastman, D. E. 1979. *Phys. Rev. B* 19:4952.

Koch, E. E. (ed.). 1983. Handbook on Synchrotron Radiation. North-Holland Publishing, Amsterdam, The Netherlands.

Landolt-Börnstein, 1989, 1994. Electronic structure of solids: Photoemission spectra and related data. *In* Numerical Data and Functional Relationships in Science and Technology, New Series, Group III, Vol. 23 a,b (A. Goldmann and E.-E. Koch, eds.). Springer-Verlag, Berlin.

Martensson, N., Baltzer, P., Brühwiler, P. A., Forsell, J.-O., Nilsson, A., Stenborg, A., and Wannberg, B. 1994. *J. Electron. Spectrosc.* 70:117.

Moruzzi, V. L., Janak, J. F., and Williams, A. R. 1978. Calculated Electronic Properties of Metals. Pergamon, New York.

Musket, R. G., McLean, W., Colmenares, C. A., Makowiecki, S. M., and Siekhaus, W. J. 1982. *Appl. Surf. Sci.* 10:143.

Olson, C. G., Liu, R., Yang, A.-B., Lynch, D. W., Arko, A. J., List, R. S., Veal, B. W., Chang, Y. C., Jiang, P. Z., and Paulikas, A. P. 1989. *Science* 245:731.

Ortega, J. E. and Himpsel, F. J. 1993. *Phys. Rev. B* 47:2130.

Papaconstantopoulos, D. A. 1986. Handbook of the Band Structure of Elemental Solids. Plenum, New York.

Petrovykh, D. Y., Altmann, K. N., Höchst, H., Laubscher, M., Maat, S., Mankey, G. J., and Himpsel, F. J. 1998. *Appl. Phys. Lett.* 73:3459.

Phillip, H. R. and Ehrenreich, H. 1963. *Phys. Rev.* 129:1550.

Pireaux, J. J., Riga, J., Thiry, P. A., Caudano, R., and Verbist, J. J. 1986. *Phys. Scr. T* 13:78.

Plummer, E. W. and Eberhardt, W. 1982. *Adv. Chem. Phys.* 49: 533.

Shen, Z.-X., Dessau, D. S., Wells., B. O., King, D. M., Spicer, W. E., Arko, A. J., Marshall, D., Lombardo, L. W., Kapitulnik, A., Dickinson, P., Doniach, S., DiCarlo, J., Loeser, A. G., and Park, C. H. 1993. *Phys. Rev. Lett.* 70:1553.

Smith, N. V. 1988. *Rep. Prog. Phys.* 51:1227.

Smith, N. V. and Himpsel, F. J. 1983. Photoelectron spectroscopy. *In* Handbook on Synchrotron Radiation, Vol. 1 (E.E. Koch, ed.) pp. 905–954. North-Holland Publishing, Amsterdam, The Netherlands.

Smith, N. V., Lässer, R., and Chiang, S. 1982. *Phys. Rev. B* 25:793.

Sutherland, D. G. J., Himpsel, F. J., Terminello, L. J., Baines, K. M., Carlisle, J. A., Jimenenz, I., Shuh, D. K., and Tong, W. M. 1997. *J. Appl. Phys.* 82:3567.

Turner, D. W., Baker, C., Baker, A. D., and Brundle, C. R. 1970. Molecular Photoelectron Spectroscopy. Wiley, New York.

Wachs, A. L., Miller, T., Hsieh, T. C., Shapiro, A. P., and Chiang, T. C. 1985. *Phys. Rev. B* 32:2326.

Winick, H. and Doniach, S. (eds.). 1980. Synchrotron Radiation Research. Plenum, New York.

Winick, H., Xian, D., Ye, M.-H., and Huang, T. (eds.). 1989. Applications of Synchrotron Radiation. Gordon and Breach, New York.

Yeh, J. J. and Lindau, I. 1985. *Atomic Data Nucl. Data Tables* 32:1.

KEY REFERENCES

Cardona and Ley, 1978. See above.

An overview of UPS, provides chapters on selected areas of research.

Gaarenstroom, S. W. and Hecht, M.H. (eds.) *Surface Science Spectra*. American Vacuum Society, Research Triangle Park, N.C.

A journal that publishes UPS spectra of specific materials, including documentation about the measurement.

Himpsel, 1983. See above.

Presents principles of angle-resolved photoemission with emphasis on band structure determination.

Kevan 1992. See above.

Comprehensive book containing articles on various topics in angle-resolved photoemission.

Landolt-Börnstein 1989, 1994. See above.

Compilations of UPS data about the electronic structure of solids, combining photoemission spectra with optical data and band structure calculations.

Plummer and Eberhardt, 1982. See above.

Principles of angle-resolved photoemission with emphasis on adsorbate states.

Smith and Himpsel, 1983. See above.

Brief introduction to UPS and its applications

APPENDIX A: SYNCHROTRON LIGHT SOURCES

A complete list of the dozens of synchrotron radiation sources worldwide would go beyond the scope of this unit. The following list gives typical machines that are optimized for either valence states (10- to 100-eV photon energy, "VUV") or shallow core levels (100- to 1000-eV photon energy, "Soft X-ray"):

United States: ALS, Berkeley, CA (Soft X-ray); NSLS UV ring, Brookhaven, NY (VUV); SRC, Madison, WI (VUV).

Europe: BESSY, Berlin, Germany (VUV, Soft X-ray); ELETTRA, Trieste, Italy (Soft X-ray); MAX-Lab, Lund, Sweden (VUV, Soft X-ray).

Asia: Photon Factory, Tsukuba, Japan (Soft X-ray); SRRC, Hsinchu, Taiwan (Soft X-ray).

Some Companies that Carry UPS Equipment

In general, companies that specialize in equipment for ultra-high-vacuum and surface science carry UPS equipment as well, such as He lamps and spectrometers. Companies that provide equipment at the high end of the line are (1) Gammadata (formerly Scienta), PO Box 15120, S-750 15 Uppsala, Sweden, http://www.gammadata.se (also has the unique ECR He lamp) and (2) Omicron (bought out VSW, a previous supplier of hemispherical analyzers), Idsteiner Str. 78, D-65232 Taunusstein, Germany, http:// www.omicron.de. Companies that have equipment in the middle range are (3) PHI (in the United States, used to sell CMAs, also selling hemispherical analyzers) and (4) Staib (in Germany, sells a CMA).

The problem with naming particular companies in this field is the volatile nature of this business. Companies are bought out and spun off frequently. Up-to-date information can usually be found at trade shows associated with conferences such as the AVS (American Vacuum Society) fall conference and the APS (American Physical Society) March meeting.

APPENDIX B: FIGURE OF MERIT FOR LIGHT SOURCES AND DETECTORS

Liouville's theorem of the conservation of the volume in phase space, $(\delta x\ \delta y\ \delta z)(\delta k_x\ \delta k_y\ \delta k_z)$, dictates the ultimate limitations of light sources and electron spectrometers. Applying this theorem to electron and photon beams, one realizes that reducing the cross-section of a beam causes its angular divergence to increase (at fixed energy). Likewise, when electrons are retarded to a low-pass energy in a spectrometer to obtain better energy resolution, the effective source size goes up (with a hemispherical retarding grid) or the divergence increases (in a retarding lens).

Applied to light sources, one can define spectral brilliance (or brightness) as the photon flux divided by the

emission solid angle, source area, and the energy band width $\delta E/E$. Typical bending magnet values are 10^{13} photons/(s $\times$ mrad^2 $\times$ mm^2 $\times$ 0.1%). Undulators provide 10^2 to 10^4 times the spectral brilliance of bending magnets. Applied to electron storage rings for synchrotron radiation, one uses the emittance (cross-section times angular spread) of the stored electron beam as a measure of beam quality, with the lowest emittance electron beam providing the highest gain in spectral brilliance at an undulator. Brilliance is important for microresolved UPS. Flux is often the relevant quantity for ordinary UPS, particularly when spin resolved, in which case only a good photon energy resolution is required. To achieve this, light needs to be squeezed only in one dimension through the entrance slit of a monochromator.

For electron spectrometers, the relevant figure of merit is étendue, i.e., the source area times the solid angle accepted by the spectrometer at a given resolving power $E/\delta E$. Normal-incidence electron optics provides a large étendue (see Eastman et al., 1980). Likewise, large physical dimensions increase the étendue (Baltzer et al., 1993).

F. J. HIMPSEL
University of Wisconsin
Madison, Wisconsin

ELLIPSOMETRY

INTRODUCTION

Ellipsometry, also known as reflection polarimetry or polarimetric spectroscopy, is a classical and precise method for determining the optical constants, thickness, and nature of reflecting surfaces or films formed on them. Ellipsometry derives its name from the measurement and tracking of elliptically polarized light that results from optical reflection. Typical experiments undertaken with this technique consist of measuring the changes in the state of polarization of light upon reflection from the investigated surfaces. Since this technique is very sensitive to the state of the surface, many experiments that involve such surface changes can be detected and followed. Hence ellipsometry finds applicability in a wide variety of fields, such as physical and chemical adsorption; corrosion; electrochemical or chemical formation of passive layers on metals, oxides, and polymer films; dissolution; semiconductor growth; and immunological reactions (Hayfield, 1963; Archer, 1964; Arsov, 1985; Hristova et al., 1997).

The main strength of this technique lies in its capability not only to allow *in situ* measurements and provide information about the growth kinetics of thin films but also simultaneously to allow the determination of many or all of the optical parameters necessary to quantify the system (Martens et al., 1963; Zaininger and Revesz, 1964; Schmidt, 1970). Ellipsometry can be used to detect and follow film thickness in the monoatomic range, which is at least an order of magnitude smaller than what can be studied by other optical methods (e.g., interferometry and reflection spectroscopy). The basic theory and principles of ellipsometry have been described by a number of authors. The original equations of ellipsometry given by Drude (1889) and various other authors have been reviewed by Winterbottom (1946), who has described straightforward procedures for determining the thickness and optical constants of films on reflecting surfaces. Heavens (1955) and Vasicek (1960) have also discussed the exact and approximate solution procedures for measuring optical constants.

When linearly polarized light having an arbitrary electric field direction is reflected from a surface, the resultant light is elliptically polarized due to the difference in phase shifts for the components of the electric field that are parallel and perpendicular to the plane of incidence. According to the principle of reversibility, light of this resultant ellipticity, when incident on a reflecting surface, should produce linearly polarized light. This is the principle under which the ellipsometer operates. The angle of incidence of light, its wavelength, and the ellipticity of the incident light are related by means of certain theoretical relations to the optical parameters of the substrate and any film that might exist on it. By knowing the optical properties of the substrate and by carrying out ellipsometric measurements, one can find the thickness of a film and its refractive index with relative ease based on certain well-known relations that will be described in the next section.

Ellipsometry can be broadly divided into two categories. One class of ellipsometry, called null ellipsometry, concerns itself with performing a zero signal intensity measurement of the light beam that is reflected from the sample. In the other class of ellipsometry, measurements of the ellipticity and intensity of light reflected from the sample are performed and correlated to the sample properties. This technique is referred to as photometric ellipsometry. The techniques and methodology described in this unit will concentrate mainly on the principles of null ellipsometry. However, references to the photometric techniques will be provided in those cases where they would provide an advantage over nulling methods.

Two parameters that are measured in ellipsometry are the change in relative amplitude and relative phase of two orthogonal components of light due to reflection. The relative, rather than absolute, measurement is one reason for the high resolution of ellipsometry. The other arises from the measured quantities usually being azimuth angles (rotation around an optical axis) that can easily be measured with high resolution (0.01°). The availability of high-speed computers and comprehensive software algorithms have also generated interest in ellipsometry measurements (Cahan and Spainer, 1969; Ord, 1969).

PRINCIPLES OF THE METHOD

Reflecting Surfaces

Let us consider a thin-film reflecting surface as shown in Figure 1. Consider that a monochromatic light beam is incident on the substrate at an angle of 45°. This incident beam can be resolved into two orthogonal components, one

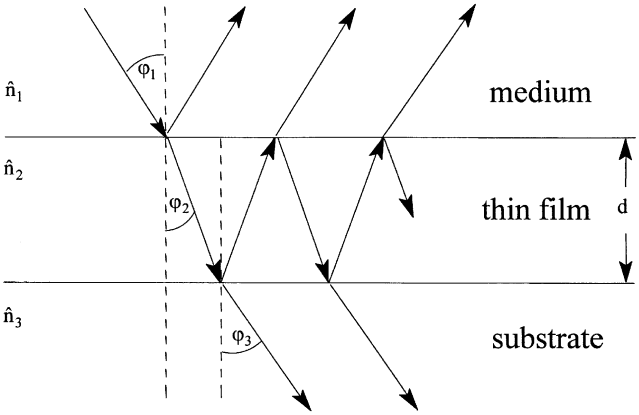


Figure 1. Reflection and refraction of light at a planar thin-film covered metal surface (d = film thickness).

in the plane of incidence, denoted by p, and the other perpendicular to it, denoted by s. The beam obtained after the reflection from the substrate (and film) will be shifted in phase and amplitude in both the p and s components. If we denote the phase shift in the parallel component by δ^s p and the perpendicular component by δ^S, the phase shift between the p and s components can then be written as

$$\Delta = \delta^p - \delta^s \tag{1}$$

The amplitudes of these two waves will also be different from their initial amplitudes, and their ratio can be written as

$$\tan\psi = \frac{E_{\mathrm{r}}^{\mathrm{p}}}{E_{\mathrm{r}}^{\mathrm{s}}} \tag{2}$$

It is these two parameters, the relative phase difference (Δ) and the ratio of the relative amplitudes ($\tan\psi$) between the parallel and perpendicular components of an electromagnetic wave that is reflected from a film-covered surface, that one obtains from ellipsometric experiments. Both the phase and amplitude shifts occur at each boundary between regions of different optical properties. The final amplitude of the wave that is observed by an analyzer is a sum from multiple reflecting layers. Hence, the reflected wave will depend on the properties of both the thin film and the substrate.

The sum of the difference in amplitudes between the parallel and perpendicular components of a multiply reflected beam is obtained using the well-known Fresnel reflection coefficients (Azzam and Bashara, 1977).

For Figure 1, the Fresnel coefficients can be calculated as follows. From the known value of the incidence angle φ_1 at the boundary between air (or a suitable medium in which the sample is immersed during the experiment) and the film, the cosine of refraction angle φ_2 can be calculated by Snell's law (Menzel, 1960):

$$\cos\varphi_2 = \left[1 - \left(\frac{\hat{n}_1}{\hat{n}_2}\sin\varphi_1\right)^2\right]^{0.5}, \tag{3}$$

where $\hat{n}_1$ and $\hat{n}_2$ are the complex refractive indices of media 1 and 2, respectively (Brown, 1965). One can then determine a complex reflection coefficient at each interface, 1-2 and 2-3, as the ratio of the magnitude of the electric vectors in a single orientation (either parallel or perpendicular) before and after reflection. It can be seen that there will be two reflection coefficients at each interface, one for the parallel and another for the perpendicular component of the reflected and incident light. These two components for the reflection between interfaces 1 and 2 are

$$r_{1,2,\mathrm{p}} = \frac{E_{\mathrm{r}}^{\mathrm{p}}}{E_1^{\mathrm{p}}} = \frac{\hat{n}_2\cos\varphi_1 - \hat{n}_1\cos\varphi_2}{\hat{n}_2\cos\varphi_1 + \hat{n}_1\cos\varphi_2} \tag{4}$$

$$r_{1,2,s} = \frac{E_{\mathrm{r}}^{\mathrm{s}}}{E_1^{\mathrm{s}}} = \frac{\hat{n}_1\cos\varphi_1 - \hat{n}_1\cos\varphi_2}{\hat{n}_1\cos\varphi_1 + \hat{n}_1\cos\varphi_2} \tag{5}$$

In a similar manner, the refraction angle at the boundary between the film and the substrate can be calculated as

$$\cos\varphi_3 = \left[1 - \left(\frac{\hat{n}_1}{\hat{n}_3}\sin\varphi_1\right)^2\right]^{0.5} \tag{6}$$

and

$$r_{2,3,\mathrm{p}} = \frac{\hat{n}_3\cos\varphi_2 - \hat{n}_2\cos\varphi_3}{\hat{n}_3\cos\varphi_2 + \hat{n}_2\cos\varphi_3} \tag{7}$$

$$r_{2,3,\mathrm{s}} = \frac{\hat{n}_2\cos\varphi_2 - \hat{n}_3\cos\varphi_3}{\hat{n}_2\cos\varphi_2 + \hat{n}_3\cos\varphi_3} \tag{8}$$

The four reflection coefficients given above are called Fresnel's reflection coefficients, after Augustin Fresnel. These reflection coefficients can be combined to give two total reflection coefficients, one for the parallel and another for the perpendicular component:

$$r_p = \frac{r_{1,2,p} + r_{2,3,p}\exp(-2i\delta)}{1 + r_{1,2p}r_{2,3,p}\exp(-2i\delta)} \tag{9}$$

and

$$r_s = \frac{r_{1,2,s} + r_{2,3,s}\exp(-2i\delta)}{1 + r_{1,2s}r_{2,3,s}\exp(-2i\delta)} \tag{10}$$

where

$$\delta = 2\pi\hat{n}_2\frac{d_2}{\lambda}\cos\varphi_2 \tag{11}$$

These total reflection coefficients r_p and r_s also include the contributions of partial reflections from the immersion medium/film and film/substrate. The ratio of these two reflection coefficients is denoted as ρ and is also related to the relative phase shift parameter Δ and the relative amplitude attenuation parameter ψ by the following fundamental equation in ellipsometry:

$$\rho = \frac{r_p}{r_{\mathrm{s}}} = \tan\psi\exp(i\Delta) \tag{12}$$

Hence, combining the above two equations, one can relate the parameters ψ and δ into the optical parameters of the surface and the substrate as

$$\tan\psi\exp(i\Delta) = \frac{r_{1,2,\mathrm{p}} + r_{2,3,\mathrm{p}}\exp(-2i\delta)}{1 + r_{1,2,\mathrm{p}}r_{2,3,\mathrm{p}}\exp(2i\delta)} \frac{1 + r_{1,2,\mathrm{s}}r_{2,3,\mathrm{s}}\exp(-2i\delta)}{r_{1,2,\mathrm{s}} + r_{2,3,\mathrm{s}}\exp(2i\delta)} \tag{13}$$

The values of Δ and ψ are determined from the ellipsometric experiments and can hence be used to determine the ratio of the reflection coefficients. Subsequently, they can be used through Equation 13 to determine the optical properties of the surface or film of interest.

The following paragraphs provide some equations that can be used to determine various physical parameters that are of interest in a sample, such as reflectivity, refractive index, dielectric constants, and optical conductivity. The derivations for these properties are not presented since it is beyond the scope of this unit.

Reflectivity

Reflectivity is defined as the quotient of reflected energetic flux to incident energetic flux, and is given as

$$R_{\mathrm{s}} = \frac{(p - n_0\cos\varphi)^2 + q^2}{(p + n_0\cos\varphi)^2 + q^2} \tag{14}$$

and

$$R_{\mathrm{p}} = R_{\mathrm{s}}\tan^2\psi \tag{15}$$

where

$$p = n_0\tan\varphi\sin\varphi\frac{\cos 2\psi}{1 + \sin 2\varphi\cos\Delta} \tag{16}$$

and

$$p = n_0\tan\varphi\sin\varphi\frac{\sin 2\psi\sin\Delta}{1 + \sin 2\psi\cos\Delta} \tag{17}$$

Here, R_p and R_s denote the reflectivity along the parallel and perpendicular components, respectively, φ is the angle of incidence of the light beam, and n_0 is the refractive index of the surrounding media in which the reflected surface is located (for air $n_0 = 1$).

Optical Constants

The complex refractive index $\hat{n}$ is represented as

$$\hat{n} = n - ik \tag{18}$$

Here n and k represent the index of refraction (real part) and the index of extinction (imaginary part), respectively. The following equations can be used for the calculation of n and k through ellipsometric parameters Δ, ψ and the angle of incidence φ for bare surfaces without the presence of any film or impurities at the surface:

$$n^2 - k^2 = n_0^2\sin^2\varphi\left[1 + \frac{\tan^2\varphi(\cos^2 2\psi - \sin^2\psi\sin^2\Delta)}{(1 + \sin 2\psi\cos\Delta)^2}\right] \tag{19}$$

and

$$2nk = \frac{n_0^2\sin^2\varphi\tan^2\varphi\sin 4\psi\sin\Delta}{(1 + \sin 2\psi\cos\Delta)^2} \tag{20}$$

Dielectric Constants

The complex dielectric function $\hat{\varepsilon}$ is given as

$$\hat{\varepsilon} = \varepsilon_1 + i\varepsilon_2 \tag{21}$$

where, again, ε_1 and ε_2 are real and imaginary parts, respectively. These values can be calculated by the equations

$$\varepsilon_1 = n^2 - k^2 \tag{22}$$

$$\varepsilon_2 = 2nk \tag{23}$$

or

$$\varepsilon_2 = 2pq \tag{24}$$

where

$$pq = nk \tag{25}$$

Optical Conductivity

The complex optical conductivity is

$$\hat{\sigma} = \sigma_1 + i\sigma_2 \tag{26}$$

where σ_1 and σ_2 are real and imaginary parts, respectively, and are calculated by the values of dielectric constants:

$$\sigma_1 = \frac{\varepsilon_2\omega}{4\pi} \tag{27}$$

$$-\sigma_2 = \frac{(\varepsilon_1 - 1)\omega}{4\pi} \tag{28}$$

The angular velocity $\omega = 2\pi\nu$ and $v = c/\lambda$ is the frequency (in inverse seconds).

PRACTICAL ASPECTS OF THE METHOD

A schematic of a nulling ellipsometer is shown in Figure 2. It consists of two arms, the first containing the components

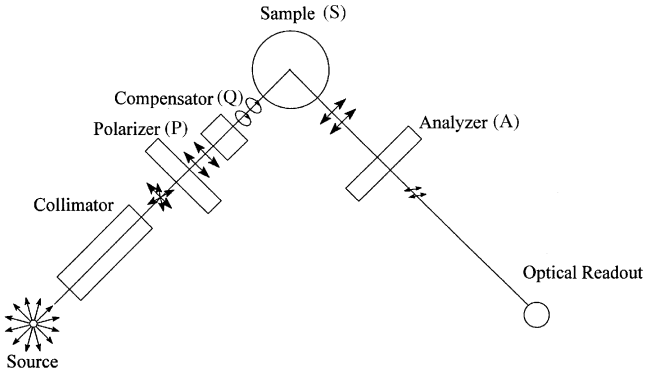


Figure 2. Schematic representation of a nulling ellipsometer.

necessary for sending polarized light into the sample and the second for analyzing the light reflected from the sample. There are two fundamental sets of component arrangements that are prevalent in ellipsometers. The first and the most widely used configuration consists of a polarizer (P), a compensator (Q), the sample (S), and an analyzer (A) arranged in that order, also referred to as a PQSA or PCSA arrangement. The other set consists of an arrangement in which the compensator follows the sample, thereby giving the name PSQA or PSCA. There are significant differences in the principle and the equations for computing the surface properties between these two arrangements. The most widely used PQSA alignment will be treated here.

In the PQSA alignment, light from a source passes through a polarizer. The polarizer typically converts the unpolarized incident light into linearly polarized light. This light then passes through the compensator that converts it into elliptically polarized light. The light then falls on the sample at a known angle of incidence φ with respect to the sample normal. The light reflected from the sample then passes through an analyzer, which is set at an angle so as to extinguish the reflected light. The criterion of light extinction is verified by either the naked eye or a photosensor coupled with a photomultiplier. For a specific incident wavelength and angle of incidence, there are only two readings that are obtained from the ellipsometer, the angle of the polarizer and that of the analyzer. These two readings can then be used to calculate ψ and Δ and subsequently calculate the optical properties of either a bare surface or a coating on top of it. The following paragraphs provide a more detailed description of the components in a generic ellipsometer.

Light Source

In theory, any wave that can be polarized, i.e., any electromagnetic radiation, can be used as a source for ellipsometric detection. However, most experimental systems use a light source in the visible range because of the availability of lenses and detectors for efficiently polarizing and detecting light. The source can be either monochromatic by itself, such as a low-power helium-neon laser or a gas discharge lamp that can be filtered using an external monochromator.

Polarizer and Analyzer

The function of the polarizer is to produce a linearly polarized light as the output to an incident light of arbitrary polarization. A birefringent polarizer can be used to obtain linear polarization to analyze. A typical birefringent polarizer converts the incident light into two beams, both of which are linearly polarized and one of which is rejected by making it undergo total internal reflection. A Glan-Thompson prism is an example of a birefringent polarizer that is commonly used.

Compensator

The compensator is typically a quarter-wave plate that converts the incident linearly polarized light into elliptically polarized light of known orientation. The quarter-wave plate introduces a relative phase shift of 1/4 between the fast- and slow-axis components of the light. A Babinet-Soleil compensator, the principle of which is explained by Azzam and Bashara (1977), is common in modern-day instruments.

Alignment

A vertical alignment of the polarizer and analyzer arm is typically the first step in ensuring that the system is properly aligned before the experiment is conducted. Manual ellipsometers often have some adjustable height mechanism such as adjustable feet or simple screws. It should also be verified that the geometric axes of the two arms are perpendicular to the central axis of the ellipsometer system.

Once the three crucial axes are aligned, the next step is to align and calibrate the ellipsometer angles. Conventionally, the polarizer and analyzer angles are measured as being positive counterclockwise from the plane of incidence when looking into the light beam. The alignment of the polarizer and analyzer angles in the ellipsometer is a very critical step in determining the optical properties of the system. There are several methods for arriving at a null point in order to calibrate the polarizer and analyzer angles (McCrackin et al., 1963; Archer, 1968; Azzam and Bashara, 1977). The method for step-by-step alignment for a manual ellipsometer given by McCrackin et al. (1963) has been used extensively by many manual ellipsometer users. The alignment must be made such that the polarizer and analyzer scales read zero when the plane of transmission of the films are parallel to the plane of incidence. For the case where there is ellipticity in either the polarizer or the analyzer, the calibrations account for both true calibration errors and imperfection parameters in the polarizer or the analyzer. The reader is referred to specific literature for the effect and the means to counteract these imperfections (Ghezzo, 1969; Graves, 1971; Steel, 1971).

After the alignment, the actual readings that are to be taken are the values of the polarizer and the analyzer angles P and A, respectively. Initially, the quarter-wave plate is set either at $Q = +45^\circ$ or $Q = -45^\circ$. There is a multiplicity of readings for P and A for both these settings for the compensator (McCrackin et al., 1963). It has been shown that for two positions of the compensator lens, at

$\pm 45^\circ$ there exist 32 readings for the polarizer and analyzer combination, for which there is a minimum obtained in the photomultiplier intensity. This whole data range falls into four zones, two with the compensator at $+45^\circ$ and two at -45°. Experiments are carried out typically in two out of the four zones and averaged in order to reduce azimuth errors. Although a complete zonal average as mentioned by McCrackin et al. (1963) need not be done for each experiment, it is advisable to be aware of the zone in which the experimental angles are being recorded.

METHOD AUTOMATION

It takes several minutes to perform the manual nulling of an ellipsometer reading by changing the values of the polarizer and analyzer angles P and A. However, in most cases, as when *in situ* studies are being carried out or when there are multiple states that are to be studied in a very short time, it is necessary to have some form of automation in order to give accurate and fast results. Hence there has been interest in developing automatic ellipsometers. These ellipsometers can be divided into two broad categories, those that still use the principle of nulling the light exiting from the sample and others that measure the intensity of light coming out of the analyzer and correlate this intensity and ellipticity into the surface properties.

Automatic Null Ellipsometers

The first ing automatic ellipsometers used motors to change the polarizer and analyzer angles P and A to drive them to null. Initially, the systems that used motor-driven nulling also required visual reading of the P and A values. This process was time consuming and very inefficient. This led to instruments that avoided mechanical nulling. One of the most widely used automatic nulling ellipsometers uses Faraday coils between the polarizer-compensator and sample-analyzer. One such automatic null ellipsometer using Faraday coil rotators was built by Mathieu et al. (1974). One Faraday coil rotates the polarization of light before it enters the compensator. The Faraday coil that is placed before the analyzer compensates for the polarized light by rotating to a certain degree, until a null point is reached. The DC current levels in the two Faraday coils provide an accurate measure of the degree of rotation of these coils. These values can be directly used for the calculation of the physical properties or can be fed to a computer program that computes the necessary optical parameters. The rate of data acquisition is very high since there are no mechanically moving parts. This technique has seen widespread use in systems in which *in situ* measurements are done. In such a case, the polarizer or the analyzer is kept at a fixed azimuth and the other is driven to null by use of the Faraday coil, and the current is recorded as a function of time.

Automatic Intensity-Measuring Ellipsometer

Another major class of ellipsometers measures the intensity and the polarization state of light after reflection from the sample. These instruments fall into two categories: (1) rotating-analyzer ellipsometer and (2) polarization-modulated ellipsometer. The rotating-analyzer ellipsometer (Hauge and Dill, 1973), as the name indicates, uses a continuously rotating analyzer to carry out the measurement. Light passes through a fixed monochromator, a fixed polarizer, an adjustable sample, and the rotating analyzer before being detected by the detector as a beam of cyclically varying intensity. The output light intensity is digitized as a function of discrete angular positions of the analyzer. From the digital output light intensity, the output polarization state of the light is obtained. This is subsequently used along with other experimental parameters, such as the angle of incidence and polarizer angle, and fed into a numerical computational algorithm that yields the film thickness and refractive index.

The polarization-modulated ellipsometer uses a piezobirefringent element to produce a periodic, known relative phase shift in the amplitudes of the perpendicular incident light components. This light input is also converted into digital data typically using an analog-to-digital converter (ADC), and the output polarization state is determined as a function of each discrete input state. This is subsequently fed into a numerical algorithm to calculate the required optical parameters. Many different variations of these polarization-modulated ellipsometers have been developed over the years (Jasperson and Schnatterly, 1969; Drevillon et al., 1982) and have found widespread use because of their accuracy and high-speed measurement capabilities.

There have also been many other combinations used to improve the speed and accuracy of the ellipsometric technique. Some of these ellipsometers are the rotating-polarizer ellipsometer (Faber and Smith, 1968), the rotating-detector ellipsometer (Nick and Azzam, 1989), and the comparison ellipsometer.

DATA ANALYSIS AND INITIAL INTERPRETATION

Calculation of Δ and ψ from Ellipsometer Readings P and A

The experimental data typically consist of two angles, those of the polarizer, P, and the analyzer, A. When using a manual ellipsometer, one should be careful to cover the zone in which the readings are taken. Since there are errors in azimuth values that might arise out of measurements in different zones, it is preferred to either take the whole sequence of data from a single zone or average each data point in two or more zones. For an example of multiple zones, consider the perpendicular Fresnel reflection coefficient for one interface in Equation 10. The value for this expression is the same for δ, $\delta + 2\pi$, $\delta + 4\pi, \ldots, \delta + 2n\pi$. Hence, the curve for Δ vs. ψ repeats when $\delta = 2\pi$. It is therefore helpful to have an advance estimate of the thickness in ellipsometric experiments. The reader is referred to the original paper by McCrackin et al. (1963) for a detailed procedure for the zonal analysis of a manual ellipsometer.

Once confirmed that the whole data set is in the same zone or that it is properly averaged, the calculation of the phase and amplitude change parameters Δ and ψ for

the system is straightforward. The relative phase change between the parallel and perpendicular polarized light, Δ, and the relative amplitude change between these two components, ψ, are related to P and A by

$$\Delta = 2P + \frac{1}{2}\pi \tag{29}$$

$$\psi = A \tag{30}$$

when the compensator is kept at $+45^\circ$. Recall that the above relations assume the polarizer, compensator, and analyzer are ideal instruments arranged in a PCSA configuration. If the compensator follows the sample, appropriate adjustments must be made.

Smith (1969) has derived a set of exact and approximate equations for determining the values for the shift parameters Δ and ψ that is independent of the compensator transmission properties. For such a set of calculations, one has to obtain readings in at least two of the four zones and then use the following formulas to obtain Δ and ψ. For a PSCA arrangement,

$$\tan^2\psi = -\tan p_1 \tan p_2 \tag{31}$$

$$\cos\Delta = \frac{1 - \cot p_1 \tan p_2}{2 - (\cot p_1 \tan p_2)^{1/2}} \frac{\cos(A_1 + A_2)}{\sin(A_2 - A_1)} \tag{32}$$

The subscripts 1 and 2 indicate the readings obtained in zones 1 and 2 for P and A, respectively. Similar exact equations for the PCSA arrangement are

$$\tan^2\psi = -\tan A_1 \tan A_2 \tag{33}$$

$$\tan\Delta = Y \sin\Delta_e \tan(P_1 + P_2) \tag{34}$$

where

$$Y = \frac{2[\cot(p_1 - \pi/4)\cot(\pi/4 - p_2)]^{1/2}}{1 + \cot(P_1 - \pi/4)\cot(\pi/4 - P_2)} \tag{35}$$

$$\sin\Delta_e = \left[\frac{1 + [1 - \tan(P_1 - \pi/4)\cot(\pi/4 - P_2)]^2}{4\tan(P_1 - \pi/4)\cot(\pi/4 - P_2)} \frac{\sin^2(A_2 + A_1)}{\sin^2(A_2 - A_1)}\right]^{1/2} \tag{36}$$

Modern automatic ellipsometers use this or other similar formulas to account for the nonidealities while determining the values of these two ellipsometer parameters.

Calculation of Optical Properties from Values of Δ and ψ

The optical properties of interest in the system can be determined using values of Δ and ψ.

The parameters are functions of many physical and material properties. Some of these dependencies can be written in the form

$$\Delta_i = f_i(n_1, k_1, n_2, k_2, n_3, k_3, d, \varphi, \lambda) \tag{37}$$

$$\psi_i = f_i(n_1, k_1, n_2, k_2, n_3, k_3, d, \varphi, \lambda) \tag{38}$$

Most often, some form of previous measurements can determine some of these unknown parameters. For example, φ can be adjusted beforehand, and λ can be fixed by knowing or determining the wavelength of the incident light. If the surrounding medium is air, then $n_1 = 1$ and $k_1 = 0$, or if the medium is some totally transparent liquid, $k_1 = 0$ and n_1 can be determined by refractometry.

The equations used to compute these optical properties are very complicated. Hence, some sort of computer algorithm is usually necessary. One of the first computer programs for ellipsometric calculations was written by McCrackin (1969). In this program, the calculations can be made for various options pertaining to the physical properties of the measured specimen. The program calculates the properties of, e.g., multiple films, inhomogenous films, and films composed of a mixture of materials. For the case of a substrate with a very thin surface oxide film, Shewchun and Rowe (1971) provided a method for calculation of the apparent (substrate-plus-film) and true (substrate-only) values of Δ and ψ by varying the incident angle. These authors have shown a flow chart for a portion of the computer program used to determine the substrate and film optical constants and thickness. Joseph and Gagnaire (Gagnaire, 1983) utilized a variance analysis using the least-squares method for application to anodic oxide growth on metallic surfaces. They have developed a method of analysis that permits the simultaneous calculation of complex refractive indices of the film and the substrate and the film thickness.

By fixing the values of the refractive indices and all other properties in a system, one can theoretically compute a Δ-ψ curve. Figure 3 gives such a theoretical Δ-ψ curve for $\varphi = 70^\circ$, $\lambda = 546.1$ nm, $n_1 = 1.337$, and $\hat{n}_3 = 2.94 - 3.57i$, $k_1 = k_2 = 0$, for various n_2 and δ values. Each curve in this figure shows the locus of points for increasing film

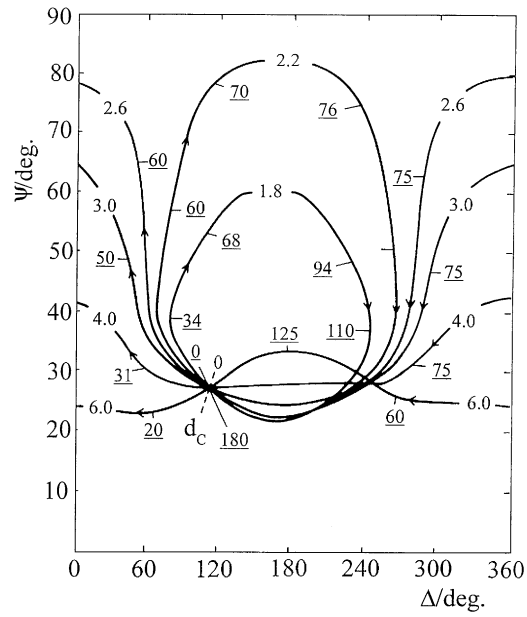


Figure 3. Theoretically computed δ–ψ curves for fixed values of $n_1 = 1.337$, $k_1 = 0$, $\hat{n}_3 = 2.94 - 3.57i$, and $k_2 = 0$ and various values of n_2 from 1.8 to 6. Underlined numbers correspond to values of δ.

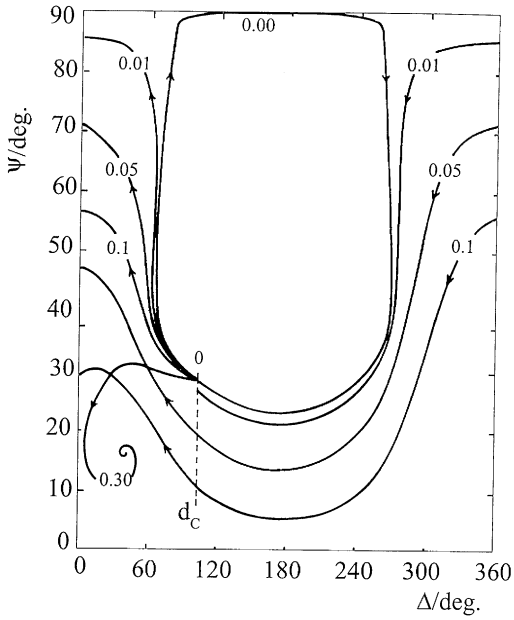


Figure 4. Theoretically computed $\delta - \psi$ curves for fixed values of $n_1 = 1.337$, $k_1 = 0$, $n_2 = 2.38$, and $\hat{n}_3 = 2.94 - 3.57i$ for various values of k_2 from 0 to 0.3.

thickness of fixed n_2 or k_2. The arrows show the direction of increasing thickness and the nonunderlined numbers are the values of the indices of refraction of the film.

For a nonabsorbing film, $k_2 = 0$, and Δ and ψ are cyclic functions of thickness for each n_2 value. The curves repeat periodically with every 180° in δ. Figure 4 shows a similar figure obtained for an absorbing film, where $k_2 > 0$. It can be seen that the curves do not quite have a periodic form. The curve front moves toward a lower ψ value with increasing cycle number. In the case where k_2 is high (0.3), the curves show a spiral shape. To determine the value of the refractive index n_2 for a nonabsorbing film $k_2 = 0$, a large number of theoretical curves have to be drawn and compared with the experimental values. The data-fitting procedure is much more complicated for absorbing films that have a complex refractive index. In such a case, both n and k values have to be changed, and families of curves have to be drawn for the relationship between Δ and ψ. From these curves, the one that best fits the experimentally observed curves can be used to determine n_2 and k_2. One can use any standard fitting tool for determining the minimum error between the experimental and predicted values or use some of the ellipsometry-specific methods in the literature (Erfemova and Arsov, 1992).

SAMPLE PREPARATION

Ellipsometric measurements can be carried out only with specimens that have a high degree of reflectivity. The degrees of smoothness and flatness are both factors that must be considered in the selection of a substrate. An indication of the extent of regularity of the whole surface is obtained by determination of P and A values at several points on the specimen. Various techniques have been used for the preparation and cleaning of the substrate surface, such as evaporation in high vacuum, mechanical polishing, chemical polishing, electrochemical polishing, and cathodic cleaning of surfaces by dissolution of natural oxide films. Other relatively less common methods include sputtering with argon atoms and thermal annealing.

Evaporation in Vacuum

Metal specimens with high reflectivity and purity at the surface can be prepared as films by vapor deposition in vacuum. The substrates generally used are glass microscope slides, which are first cleaned by washing and vapor phase degreasing before being introduced into the vacuum chamber, where they are cleaned by ion bombardment with argon. Adjusting parameters such as the rate and time of evaporation regulates the thickness of the evaporated film. Many other variables have an impact on the properties of evaporated films. They include the nature and pressure of residual gas in the chamber, intensity of the atomic flux condensing on the surface, the nature of the target surface, the temperature of the evaporation source and energy of the impinging atoms, and contamination on the surface by evaporated supporting material from the source.

Mechanical Polishing

Mechanical polishing is one of the most widely used methods in surface preparation. Generally, mechanical polishing consists of several steps. The specimens are initially abraded with silicon carbide paper of various grades, from 400 to 1000 grit. After that, successive additional fine mechanical polishing on silk disks with diamond powders and sprays (up to 1/10 μm grain size) is practiced to yield a mirror finish. This results in a bright and smooth surface suitable for optical measurements. But, it is well known that mechanical treatments change the structure of surface layers to varying depths depending on the material and the treatment. Mechanically polished surfaces do not have the same structural and optical properties as the bulk material. As a result of mechanical polishing, surface crystallinity is impaired. In addition, there exists a thin deformed layer extending from the surface to the interior due to the friction and heating caused by mechanical polishing. Such changes are revealed by the deviations in the optical properties of the surface from those of the bulk. Hence mechanical polishing is utilized as a sole preparation method, primarily if the specimen will be used as a substrate for ellipsometric study of some other surface reactions, e.g., adsorption, thermal, or electrochemical formation of oxide films and polymer electrodeposition.

Chemical Polishing

This method of surface preparation is generally used after mechanical polishing in order to remove the damaged layer produced by the mechanical action. A specimen is immersed in a bath for a controlled time and subsequently ultrasonically washed in an alcohol solution. The final cleaning is extremely important to remove bath components that may otherwise be trapped on the surface.

Various baths of different chemical compositions and immersion conditions (temperature of bath and time of treatment depending on the nature of the metal specimen) are described in the literature (Tegart, 1960). During chemical polishing, oxygen evolution is possible along with the chemical dissolution of the metal substrate. In such cases, the polished surfaces may lose their reflectivity, becoming less suitable for optical measurements. A major problem associated with chemical polishing is the possible formation of surface viscous layers that result in impurities on the surface. Care must be taken to identify and remove impurities like S, Cl, N, O, and C that might be formed by the use of chemicals.

Electrochemical Polishing

Electrochemical polishing is a better way to prepare metal surfaces than mechanical and chemical polishing. The metal is placed in an electrochemical cell as an anode, while a Pt or carbon electrode with a large surface area is used as the cathode. Passage of current causes the dissolution of parts of the metal surface, the peaks dissolving at a greater rate than the valleys, which leads to planarization and a subsequently reflecting surface.

Electrochemical Dissolution of Natural Oxide Films

Dissolution of the naturally formed oxide films can be carried out by cathodic polarization. It is very important to determine an optimum value of the cathodic potential for each metal in order to minimize the thickness of the oxide layer. For potentials more anodic than the optimum value, the dissolution is very slow and the natural oxide film is not dissolved completely. Potentials higher than this optimum increase the risk for hydrogen contamination at the surface, its penetration in the substrate, and its inclusion in the metal lattice. The formation of thin hydrated layers is also possible.

PROBLEMS

There are various sources, some from the equipment and others from the measuring system or the measurement process itself, that can introduce errors and uncertainties during an ellipsometric measurement. Most of these errors are correctable with certain precautions and corrections during or after the experiment.

A frequent source of error is the presence of a residual oxide layer on the surface of the film under study. The presence of even a very thin film (50 Å) can cause the refractive index to be markedly different than the actual substrate value (Archer, 1962). Hence, in the case of an absorbing substrate and a nonabsorbing thin film, four optical parameters need to be determined: the refractive index and the thickness of the film, n_2 and d_2, and the complex refractive index coefficients of the substrate, n_3 and k_3. Since only two parameters are determined, Δ and ψ, it is not possible to determine all four parameters using a single measurement. To estimate the optical properties of the substrate, multiple measurements are made (So and Vedam, 1972) using (1) various film thickness values, (2) various angles of incidence, or (3) various incident media. One set of substrate optical properties consistent with all the multiple readings can then be found.

Errors in measurement also occur due to surface roughness in the substrate (or the film). Predictions of the optical parameters for surfaces with very minor surface roughness (50 Å) have shown large errors in the determined refractive indices (Festenmaker and McCrackin, 1969). Hence, it is imperative that one has a good knowledge of the degree of impurities or abnormalities that a sample might have. Sample preparation is crucial to the accuracy of the results, and care must be taken to determine a method most suited for the particular application.

The polarizer, compensator, and analyzer are all subject to possibly incorrect settings and hence azimuth errors. The analysis can be freed from azimuth errors by performing the analysis and averaging the results with components set at angles that are 90° apart, described as zone averaging by McCrackin et al. (1963).

The magnitudes of Δ and ψ vary as a function of the angle of incidence. There are several types of systematic errors encountered during the angle-of-incidence measurements that produce errors in Δ and ψ. Many studies suggest that the derived constants are not true constants, but are dependent on the incidence angle (Vasicek, 1960; Martens et al., 1963). The beam area divided by the cosine of the angle of incidence gives the area of the surface under examination. Hence, there is a significant increase in the area under examination for a small change in the incident angle. A change in the angle of incidence simultaneously changes the area of the surface under examination. As a matter of convenience, it is suggested that the ellipsometric measurements be made with an angle of incidence near the principal angle, i.e., the angle at which the relative phase shift between the parallel and perpendicular components of reflected light, Δ, is 90°. The value of this principal angle depends mainly on the refractive index of the substrate, and is between 60° and 80° for typical metal substrates.

Other sources of error include those arising from the deflection of the incident light beam due to an imperfection in the polarizer, multiply reflected beams, and incompletely polarized incident light. The importance of reliable and accurate optical components should therefore be stressed, since most of these errors are connected to the quality of the equipment one procures.

It should also be noted that the steps after the experimental determination of Δ and ψ are also susceptible to error. One typically has to input the values of Δ and ψ into a numerical algorithm to compute the optical properties. One should have an idea of the precision of the computation accuracy of the input data. There are also cases in which the approximations in the relations used to compute the physical parameters are more extensive than necessary and introduce more uncertainty than a conservative approach would cause.

LITERATURE CITED

Archer, R. J. 1962. Determination of the properties of films on silicon by the method of ellipsometry. *J. Opt. Soc. Am.* 52:970.

Archer, R. J. 1964. Measurement of the physical adsorption of vapors and the chemisorption of oxygen and silicon by the method of ellipsometry. *In* Ellipsometry in the Measurements of Surfaces and Thin Films (E. Passaglia, R.R. Stromberg, and J. Kruger, eds.) pp. 255–272. National Bureau of Standards, Washington, DC, Miscellaneous Publication 256.

Archer, R. J., 1968. Manual on Ellipsometry. Gaertner Scientific, Chicago.

Arsov, Lj. 1985. Dissolution electrochemique des films anodiques du titane dans l'acide silfurique. *Electrochim. Acta* 30:1645–1657.

Azzam, R. M. A. and Bashara, N. M. 1977. Ellipsometry and Polarized Light. North-Holland Publishing, New York.

Brown, E. B. 1965. Modern Optics. Reinhold Publishing, New York.

Cahan, B. D. and Spainer, R. F. 1969. A high speed precision automatic ellipsometer. *Surf Sci*. 16:166.

Drevillon, B., Perrin, J., Marbot, R., Violet, A., and Dalby, J. L. 1982. Fast polarization modulated ellipsometer using a microprocessor system for digital Fourier analysis. *Rev. Sci. Instrum.* 53(7):969–977.

Drude, P. 1889. Ueber oberflachenschichten. *Ann. Phys.* 272:532–560.

Erfemova, A. T. and Arsov, Lj. 1992. Ellipsometric in situ study of titanium surfaces during anodization. *J. Phys. France* II:1353–1361.

Faber, T. E. and Smith, N. V. 1968. Optical measurements on liquid metals using a new ellipsometer. *J. Opt. Soc. Am.* 58(1):102–108.

Festenmaker, C. A. and McCrackin, F. L. 1969. Errors arising from surface roughness in ellipsometric measurement of the refractive index of a surface. *Surf. Sci.* 16:85–96.

Gagnaire, J. 1983. Ellipsometric study of anodic oxide growth: Application to the titanium oxide systems. *Thin Solid Films* 103:257–265.

Ghezzo, M. 1960. Method for calibrating the analyser and the polarizer in an ellipsometer. *Br. J. Appl. Phys. (J. Phys. D)* 2: 1483–1485.

Ghezzo, M. 1969. Method for calibrating the analyzer and polarizer in an ellipsometer. *J. Phys. Ser. D* 2:1483–1490.

Graves, R. H. W. 1971. Ellipsometry using imperfect polarizers. *Appl. Opt.* 10:2679.

Hauge, P. S. and Dill, F. H. 1973. Design and operation of ETA, and automated ellipsometer. *IBM J. Res. Dev.* 17:472–489.

Hayfield, P. C. S. 1963. American Institute of Physics, Handbook. McGraw-Hill, New York.

Heavens, O. S., 1955. Optical Properties of Thin Solid Films. Dover, New York.

Hristova, E., Arsov, Lj., Popov, B., and White, R. 1997. Ellipsometric and Raman spectroscopic study of thermally formed films on titanium. *J. Electrochem. Soc.* 144:2318–2323.

Jasperson, S. N. and Schnatterly, S. E. 1969. An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. *Rev. Sci. Instrum.* 40(6):761–767.

Martens, F. P., Theroux, P., and Plumb, R. 1963. Some observations on the use of elliptically polarized light to study metal surfaces. *J. Opt. Soc. Am.* 53(7):788–796.

Mathieu, H. J., McClure, D. E., and Muller, R. H. 1974. Fast self-compensating ellipsometer. *Rev. Sci. Instrum.* 45:798–802.

McCrackin, F. L. 1969. A FORTRAN program for analysis of ellipsometer measurements. Technical note 479. National Bureau of Standards, pp. 1–76.

McCrackin, F. L., Passaglia, E., Stromberg, R. R., and Steinberg, H. L. 1963. Measurement of the thickness and refractive index of very thin films and the optical properties of surfaces by ellipsometry. *J. Res. Natl. Bur. Stand.* A67:363–377.

Menzel, H. D. (ed.). 1960. Fundamental Formulas of Physics, Vol. 1. Dover Publications, New York.

Nick, D. C. and Azzam, R. M. A. 1989. Performance of an automated rotating-detector ellipsometer. *Rev. Sci. Instrum.* 60(12):3625–3632.

Ord, J. L. 1969. An elliposometer for following film growth. *Surf. Sci.* 16:147.

Schmidt, E. 1970. Precision of ellipsometric measurements. *J Opt. Soc.* 60(4):490–494.

Shewchun, J. and Rowe, E. C. 1971. Ellipsometric technique for obtaining substrate optical constants. *J. Appl. Phys.* 41(10): 4128–4138.

Smith, P. H. 1969. A theoretical and experimental analysis of the ellipsometer. *Surf. Sci.* 16:34–66.

So, S. S. and Vedam, K. 1972. Generalized ellipsometric method for the absorbing substrate covered with a transparent-film system. Optical constants for silicon at 3655 Å. *J. Opt. Soc. Am.* 62(1):16–23.

Steel, M. R. 1971. Method for azimuthal angle alignment in ellipsometry. *Appl. Opt.* 10:2370–2371.

Tegart, W. J. 1960. Polissage electrolytique et chemique des metaux au laboratoire et dans l'industry. Dunod, Paris.

Vasicek, A. 1960. Optics of Thin Films. North-Holland Publishing, New York.

Winterbottom, A. W. 1946. Optical methods of studying films on reflecting bases depending on polarization and interference phenomena. *Trans. Faraday Soc.* 42:487–495.

Zaininger, K. H. and Revesz, A. G. 1964. Ellipsometry—a valuable tool in surface research. *RCA Rev.* 25:85–115.

KEY REFERENCES

Azzam, R. M. A. (ed.). 1991. Selected Papers on Ellipsometry; SPIE Milestone Series, Vol. MS 27. SPIE Optical Engineering Press, Bellingham, Wash.

A collection of many of the path-breaking publications up to 1990 in the field of ellipsometry and ellipsometric measurements. Gives a very good historical perspective of the developments in ellipsometry.

Azzam and Bashara, 1977. See above.

Gives an excellent theoretical basis and provides an in-depth analysis of the principles and practical applications of ellipsometry.

McCrackin et al., 1963. See above.

Provides a good explanation of the practical aspects of measuring the thickness and refractice indices of thin films. Includes development of a notation for identifying different null pairs for the polarizer/analyzer rotations. Also provides a method to calibrate the azimuth scales of the ellipsometer divided circles.

Passaglia, E., Stromberg, R.R., and Kruger, J. (eds.). 1964. Ellipsometry in the Measurement of Surfaces and Thin Films. Symposium Proceedings, Washington, DC, 1963. National Bureau of Standards Miscellaneous Publication 256.

Presents practical aspects of ellipsometric measurements in various field.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

A	Analyzer settings with respect to the plane of incidence, deg
C	Compensator settings with respect to the plane of incidence, deg
E_i	Amplitude of the incident electric wave
E_r	Amplititude of the reflected electric wave
d	Film thickness, cm
E_x	Amplitude of the electric wave in the x axis
i	$\sqrt{-1}$
k	Absorption index (imaginary part of the refractive index)
n	Real part of the refractive index
$\hat{n}$	Complex refractive index
P	Polarizer settings with respect to the plane of incidence, deg
Q	Compensator settings with respect to the plane of incidence, deg
R	Reflectivity
r_p	Fresnel reflection coefficient for light polarized parallel to the plane of incidence
r_s	Fresnel reflection coefficient for light polarized perpendicular to the plane of incidence
Δ	Relative phase change, deg
δ	Change of phase of the beam crossing the film
$\hat{\varepsilon}$	Complex dielectric function
λ	Wavelength, cm
ν	Frequency, s^{-1}
ρ	Ratio of complex reflection coefficients
$\hat{\sigma}$	Complex optical conductivity
φ	Angle of incidence, deg
$\tan\psi$	Relative amplitude attenuation, deg
ω	Angular velocity, rad

Lj. Arsov
University Kiril and Metodij
Skopje, Macedonia
M. Ramasubramanian
B. N. Popov
University of South Carolina
Columbia, South Carolina

IMPULSIVE STIMULATED THERMAL SCATTERING

INTRODUCTION

Impulsive stimulated thermal scattering (ISTS) is a purely optical, non-contacting method for characterizing the acoustic behavior of surfaces, thin membranes, coatings, and multilayer assemblies (Rogers et al., 2000a), as well as bulk materials (Nelson and Fayer, 1980). The method has emerged as a useful tool for materials research in part because: (1) it enables accurate, fast, and nondestructive measurement of important acoustic (direct) and elastic (derived) properties that can be difficult or impossible to evaluate in thin films using other techniques; (2) it can be applied to a wide range of materials that occur in microelectronics, biotechnology, optics, and other areas of technology; and (3) it does not require specialized test structures or impedance-matching fluids which are commonly needed for conventional mechanical and acoustic tests. Further, recent advances in experimental design have simplified the ISTS measurement dramatically, resulting in straightforward, low-cost setups, and even in the development of a commercial ISTS photoacoustic instrument that requires no user adjustments of lasers or optics. With this tool, automated single-point measurements, as well as scanning-mode acquisition of images of acoustic and other physical properties, are routine.

ISTS, which is based on transient grating (TG) methods (Nelson and Fayer, 1980; Eichler et al., 1986), is a spectroscopic technique that measures the acoustic properties of thin films over a range of acoustic wavelengths. It uses mild heating produced by crossed picosecond laser pulses to launch coherent, wavelength-tunable acoustic modes and thermal disturbances. The time-dependent surface ripple produced by these motions diffracts a continuous-wave probing laser beam that is overlapped with the excited region of the sample. Measuring the temporal variation of the intensity of the diffracted light yields the frequencies and damping rates of acoustic waves that propagate in the plane of the film. It also determines the arrival times of acoustic echoes generated by subsurface reflections of longitudinal acoustic wavepackets that are launched at the surface of the film. The wavelength dependence of the acoustic phase velocities (i.e., product of the frequency and the wavelength) of in-plane modes, which is known as the acoustic dispersion, is determined either from a single measurement that involves the excitation of acoustic waves with a well defined set of wavelengths, or from the combined results of a series of measurements that each determine the acoustic response at a single wavelength. Interpreting this dispersion with suitable models of the acoustic waveguide physics yields viscoelastic (e.g., Young's modulus, Poisson's ratio, acoustic damping rates, stress, etc.) and/or other physical properties (e.g., density, thickness, presence or absence of adhesion, etc.) of the films. The acoustic echoes, which are recorded in the same measurements, provide additional information that can simplify this modeling. The ISTS data also generally contain information from nonacoustic (e.g., thermal, electronic, etc.) responses. This unit, however, focuses only on ISTS measurement of acoustic motions in thin films. It begins with an overview of other related measurement techniques. It then describes the ISTS acoustic data and demonstrates how it can be used to determine: (1) the stress and flexural rigidity in thin membranes (Rogers and Bogart, 2000; Rogers et al., 2000b; Rogers and Nelson, 1995); (2) the elastic constants of membranes and supported films (Rogers and Nelson, 1995; Duggal et al., 1992; Shen et al., 1996; Rogers and Nelson, 1994; Rogers et al., 1994a); and (3) the thicknesses of single or multiple films in multilayer stacks (Banet et al., 1998; Gostein et al., 2000).

Competitive and Related Techniques

Acoustic properties of thin films are most commonly evaluated with conventional ultrasonic tests, which involve a

source of ultrasound (e.g., a transducer), a propagation path, and a detector. The ability to excite and detect acoustic waves with wavelengths that are short enough for thin film evaluation (i.e., wavelengths comparable to or smaller than the film thickness) requires high-frequency transducers/detectors fabricated directly on the sample, or coupled efficiently to it with impedance-matching liquids or gels. Both of these approaches restrict the range of structures that can be examined; they limit the usefulness of conventional acoustic tests for thin film measurement. Photoacoustic methods overcome the challenge of acoustic coupling by using laser light to excite and probe acoustic disturbances without contacting the sample. In addition to ISTS, there are two other general classes of photoacoustic techniques for measuring acoustics in thin films. In the first, a single excitation pulse arrives at the surface of the sample and launches, through mild heating, a longitudinal (i.e., compressional) acoustic wavepacket that propagates into the depth of the structure (Thomsen et al., 1986; Eesley et al., 1987; Wright and Kawashima, 1991). Parts of this acoustic disturbance reflect at buried interfaces, such as the one between the film and its support or between films in a complex multilayer stack. A variably delayed probe pulse measures the time dependence of the optical reflectivity or the slope of the front surface of the sample in order to determine the time of arrival of the various acoustic echoes. Data from this type of measurement are similar in information content to the acoustic echo component of the ISTS signal. In both cases, the data reveal the out-of-plane longitudinal acoustic velocities when the thicknesses of the films are known. The measured acoustic reflectivity can also be used to determine properties (e.g., density) that are related to the change in acoustic impedance that occurs at the interface. This technique has the disadvantage that it requires the excitation pulses to be strongly absorbed by the sample. It also typically relies on expensive and complex femtosec laser sources and relatively slow detection schemes that use probe pulses. Although it can be used to measure thicknesses accurately, it does not yield information on transversely polarized (i.e., shear) acoustic waves or on modes that propagate in the plane of the film. As a result, the only elastic property that can be evaluated easily is the out-of-plane compressional modulus.

Another method uses a cylindrically-focused excitation pulse as a broadband line source for surface propagating waves (Neubrand and Hess, 1992; Hess, 1996). Examining the changes in position, intensity, or phase of at least one other laser beam that strikes the sample at a location spatially separated from the excitation region provides a means for probing these waves. The data enable reliable measurement of surface acoustic wave velocities over a continuous range of acoustic wavelengths (i.e., the dispersion) when the separation between the excitation and probing beams (or between the probing beams themselves) is known precisely. The measured dispersion can be used, with suitable models for the acoustic physics, to extract other properties (e.g., density, thickness, elastic properties) of the samples. This technique has the disadvantage that out-of-plane acoustic properties are not probed directly. Also, data that include multiple acoustic velocities at a single wavelength (e.g., multiple modes in an acoustic waveguide) can be difficult to interpret. We note that this method and the one described in the previous paragraph share the "source–propagation path–receiver" approach of conventional acoustic testing techniques. They are similar also in their generation of an essentially single-cycle acoustic pulse or "wavepacket" that includes a wide range of wavevectors and corresponding frequencies. The combined information from these two techniques is present in the ISTS data, in a separable and more easily analyzed form.

Although it is not strictly a photoacoustic method, surface Brillouin scattering (SBS; Nizzoli and Sandercock, 1990) is perhaps more closely related to ISTS than the techniques described above. An SBS experiment measures the spectral properties of light scattered at a well-defined angle from the sample. The spectrum reveals the frequencies of incoherent, thermally populated acoustic modes with wavelengths that satisfy the associated phase-matching condition for scattering into the chosen angle. The information obtained from SBS and ISTS measurements is similar. An important difference is that the ISTS technique uses coherent, laser-excited phonons rather than incoherent, thermally populated ones. ISTS signals are therefore much stronger than those in SBS and they can be detected rapidly in the time domain. This form of detection enables acoustic damping rates, for example, to be evaluated accurately without the deconvolution procedures that are necessary to interpret spectra collected with the sensitive Fabry-Perot filters that are commonly used in SBS. Also, with ISTS it is possible simultaneously to excite and monitor acoustic modes with more than one wavelength, to determine their phases, and to measure them in real-time as they propagate across the sample. These and other capabilities, which are useful for accurately evaluating films or other structures with dimensions comparable to the acoustic wavelength, are absent from traditional forms of SBS.

Finally, in some cases, certain properties that can be derived from ISTS measurements (e.g., elastic constants, density, thickness, stress, etc.) can be determined with other, nonacoustic methods. Although complete descriptions of all of the possible techniques is beyond the scope of this unit, we list a few of the more established methods.

1. Elastic constants can be determined with uniaxial pull-testers, nanoindenters (Pharr and Oliver, 1992), and specialized micromechanical test structures (Allen et al., 1987).
2. Stress, and sometimes elastic constants, are typically measured with tests that use deflections or vibrations of drumhead membranes (Maden et al., 1994; Vlassak and Nix, 1992) or cantilevered beams (Mizubayashi et al., 1992), or these are inferred from strain evaluated using X-ray diffraction (Clemens and Bain, 1992; also see X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS).
3. Thicknesses of transparent films are often determined with reflectometers or ellipsometers (See ELLIPSOMETRY). For opaque films, thickness is evaluated

using stylus profilometry or grazing incidence x-ray reflection; in the case of conducting films, it is determined indirectly from measurements of sheet resistance. Vinci and Vlassak (1996) present a review of techniques for measuring the mechanical properties of thin films and membranes.

PRINCIPLES OF THE METHOD

As mentioned above (see Introduction), ISTS uses short (relative to the time-scale of the material response of interest) pulses of light from an excitation laser to stimulate acoustic motions in a sample. The responses are measured with a separate probing laser. Figure 1 schematically illustrates the mechanisms for excitation and detection in the simplest case. Here, a single pair of excitation pulses crosses at an angle θ at the surface of a thin film on a substrate. The optical interference of these pulses produces a sinusoidal variation in intensity with a period Λ, given by

$$\Lambda = \frac{\lambda_e}{2\sin(\theta/2)} \tag{1}$$

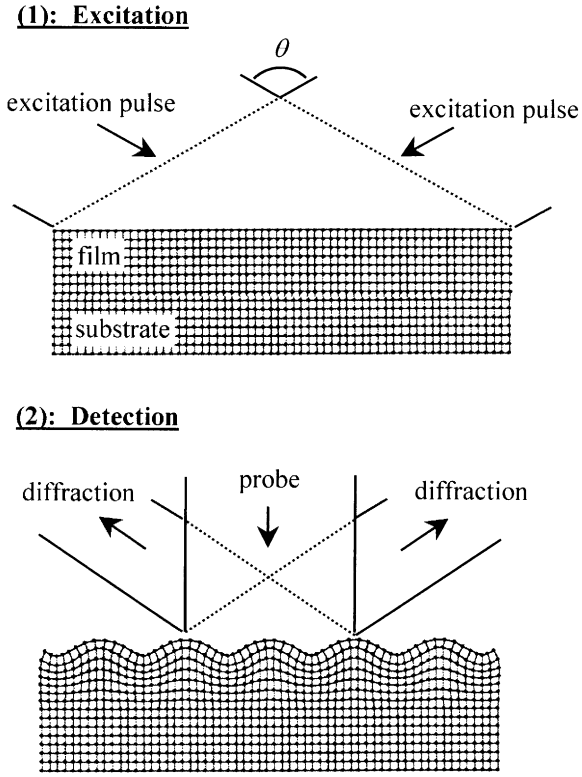


Figure 1. Schematic illustration of the ISTS measurement. Crossed laser pulses form an optical interference pattern and induce coherent, monochromatic acoustic and thermal motions with wavelengths that match this pattern. The angle between the pulses determines the wavelength of the response. A continuous wave probing laser diffracts from the ripple on the surface of the sample. Measuring the intensity of the diffracted light with a fast detector and transient recorder reveals the time dependence of the motions.

The wavelength of the excitation light (λ_e) is chosen so that it is partly absorbed by the sample; this absorption induces heating in the geometry of the interference pattern. The resulting spatially periodic thermal expansion launches coherent, monochromatic counter-propagating surface acoustic modes with wavelength Λ. It also simultaneously generates acoustic waves that propagate into the bulk of the sample and produce the acoustic echoes mentioned previously (Shen et al., 1996; Crimmins et al., 1998). In thin films, the former and latter responses typically occur on nanosecond and picosecond time scales, respectively. On microsecond time scales, the thermally induced strain slowly relaxes via thermal diffusion. The temporal nature of the surface motions that result from these responses can be determined in real time with each shot of the excitation laser by measuring the intensity of diffraction of a continuous wave probe laser with a fast detector and transient recorder. For motions that exceed the bandwidth of conventional detection electronics (typically ~1 to 2 GHz), it is possible to use either variably delayed picosecond probing pulses to map out the response in a point-by-point fashion (Shen et al., 1996; Crimmins et al., 1998) or a continuous wave probe with a streak camera for ultrafast real-time detection. It is also possible to use, instead of the streak camera, a scanning spectral filter to resolve the motions in the frequency domain (Maznev et al., 1996). This unit focuses primarily on the in-plane acoustic modes, partly because they can be excited and detected rapidly in real-time using commercially available, low-cost laser sources and electronics. Also, these modes provide a route to measuring a wide range of thin film elastic constants and other physical properties.

The following functional form describes the response, *R(t)*, when the thermal and the in-plane acoustic motions both contribute

$$R(t) \propto A_{\text{therm}} e^{-\Gamma_t} + \sum_i B_i e^{-\gamma_i t} \cos\omega_i t \tag{2}$$

where A_{therm} and B_i are the amplitudes of the thermal and acoustic responses, respectively, and the ω_i are the frequencies of the acoustic modes. The summation extends over the excited acoustic modes, i, of the system. The thermal and acoustic decay rates are, respectively, Γ_t and γ_i. In the simplest diffraction-based detection scheme, the measured signal is proportional to the diffraction efficiency, which is determined by the square of the *R(t)*. These responses, as well as the acoustic echoes, are all recorded in a single measurement of the area of the sample that is illuminated by the excitation and probing lasers. The ISTS measurement provides spatial resolution that is comparable to this area, which is typically a circular or elliptical region with a characteristic dimension (e.g., diameter or major axis) of 50 to 500 μm. In certain situations, the effective resolution can be significantly better than this length scale (Gostein et al., 2000).

Figure 2 shows data from a polymer film on a silicon substrate. The onset of diffraction coincides with the arrival of the excitation pulses at $t = 0$. The slow decay of signal is associated with thermal diffusion (Fig. 2B); the oscillations in Fig. 2A are due to acoustic modes that propagate

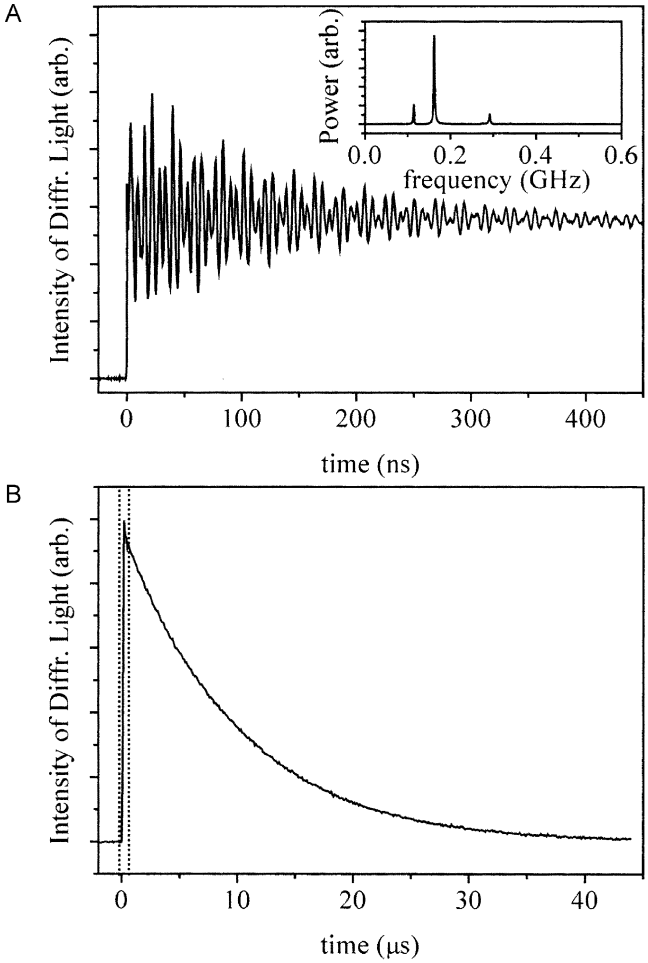


Figure 2. Typical ISTS data from a thin polymer film (thickness ~4 μm) on a silicon substrate. Part (**A**) shows the in-plane acoustic response, which occurs on a nanosec time scale and has a single well defined wavelength (8.32 μm) determined by the color and crossing angle of the excitation pulses. The oscillations in the signal reveal the frequencies of the different acoustic waveguide modes that are excited in this measurement. The inset shows the power spectrum. The frequencies are determined by the acoustic wavelength and the mechanical properties of the film, the substrate, and the nature of the interface between them. The acoustic waves eventually damp out and leave a nonoscillatory component of signal that decays on a microsec time scale (**B**). This slow response is associated with the thermal grating. Its decay rate is determined by the wavelength of the response and the thermal diffusivity of the structure. The dashed lines in (B) indicate the temporal range displayed in (A).

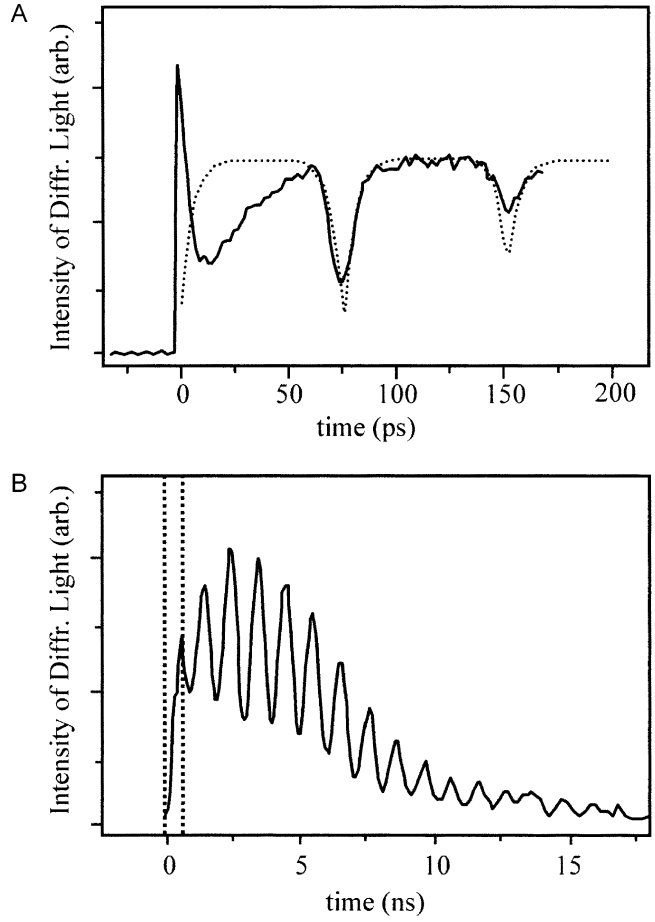


Figure 3. Typical ISTS data from an ultrathin metal film (thickness ~200 nm) on a silicon substrate. Part (**A**) shows the arrival of two acoustic echoes produced by longitudinal acoustic wavepackets that are launched at the surface of the film and reflect at the interface between the film and the substrate and at the interface between the film and the surrounding air. Propagation of acoustic modes in the plane of the film causes oscillations in the signal on a nanosec time scale (**B**). Thermal diffusion produces the overall decay in signal that occurs on a nanosec time scale. The dashed lines in (B) indicate the temporal range displayed in (A).

in the plane of the film. The frequencies and damping rates of the acoustic modes, along with the acoustic wavelength determined from Equation 1, define the real and imaginary parts of the phase velocities. On a typically faster (picosecond) time scale it is also possible to resolve responses due to longitudinal waves that reflect from the film/substrate interface. Figure 3 shows both types of acoustic responses evaluated in a single measurement on a thin metal film on a silicon substrate.

Although the measured acoustic frequencies and echoes themselves can be important (e.g., for filters that use surface acoustic waves or thin film acoustic resonances, respectively), the intrinsic elastic properties of the films are often of interest. The acoustic echoes yield, in a simple way, the out-of-plane compressional modulus, c_o, when the density, ρ, and thickness, h, are known. The measured roundtrip time in this case defines, with the thickness, the out-of-plane longitudinal acoustic velocity, v_o; the modulus is $c_0 = \rho v_0^2$.

Extracting moduli from the in-plane acoustic responses is more difficult because thin films form planar acoustic waveguides that couple in- and out-of-plane compressional and shearing motions (Farnell and Adler, 1972; Viktorov, 1976). An advantage of this characteristic is that, in principle, the dispersion of the waveguide modes can be used to determine a set of anisotropic elastic constants as well as film thicknesses and densities. Determining these properties requires an accurate measurement of the dispersion of the waveguide and a detailed understanding of how

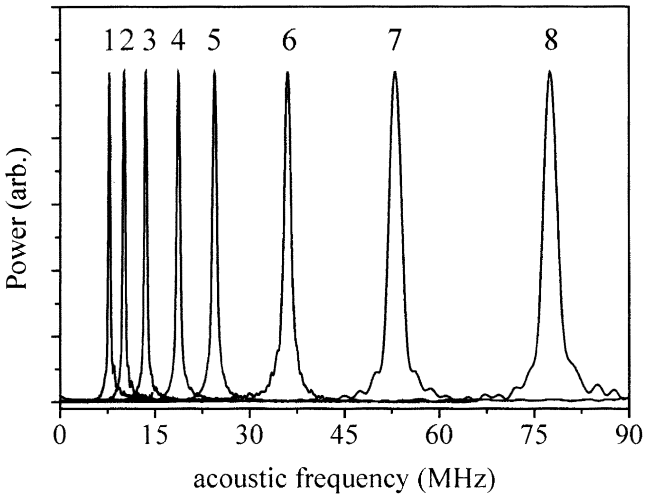


Figure 4. Power spectra from data collected in ISTS measurements on a thin membrane at several different acoustic wavelengths between ~30 μm and ~6 μm and labeled 1 to 8. The variation of the frequency with wavelength defines the acoustic dispersion. Analyzing this dispersion with physical models of the waveguide acoustics of the membrane yields the elastic constants and other properties.

acoustic waves propagate in layered systems. There are two approaches for determining the dispersion. One involves a series of measurements with different angles between the excitation pulses to determine the acoustic response as a function of wavelength. Figure 4 shows the results of measurements that determine the dispersion of a thin unsuppported membrane using this method (Rogers and Bogart, 2000). The other approach uses a single measurement performed with specialized beam-shaping optics that generate more than two excitation pulses (Rogers, 1998). Figure 5 shows data that were collected in an

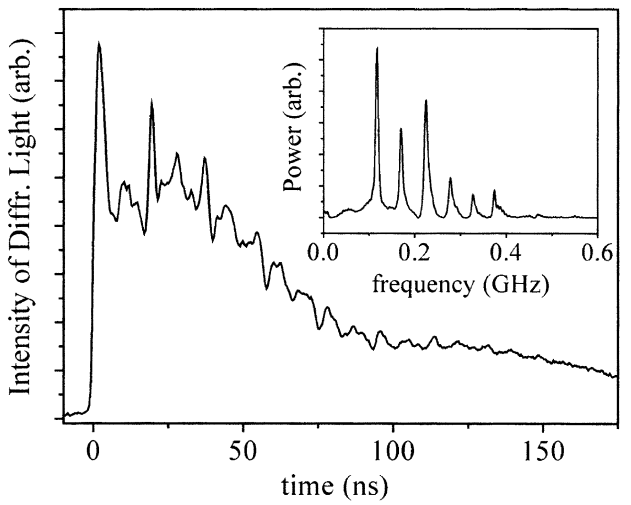


Figure 5. ISTS data from a thin film of platinum on a silicon wafer. This measurement used a system of optics to generate and cross six excitation pulses at the surface of the sample. The complex acoustic disturbance launched by these pulses is characterized by six different wavelengths. The power spectrum reveals the frequency components of the signal. A single measurement of this type defines the acoustic dispersion.

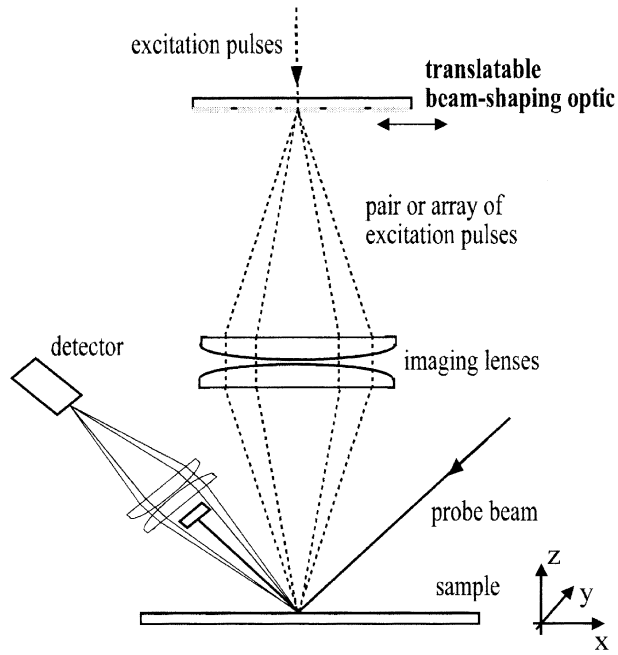


Figure 6. Optical system for ISTS measurements. Pulses from an excitation laser pass through a beam-shaping optic that splits the incident pulse into an array of diverging pulses. This optic contains a set of patterns, each of which produces a different number and distribution of pulses. A pair of imaging lenses crosses some of the pulses at the surface of a sample. The optical interference pattern produced in this way defines the wavelengths of acoustic motions stimulated in the sample. Diffraction of a probe laser is used to monitor the time dependence of these motions. A pair of lenses collects the diffracted signal light and directs it to a fast photodetector. The wavelengths of the acoustic waves can be changed simply by translating the beam-shaping optic.

ISTS measurement with six crossed excitation pulses. The response in this case contains contributions from acoustic modes with six different wavelengths, which are defined by the geometry of the six-pulse interference pattern. In both cases, the size of the excitation region is chosen, through selection of appropriate lenses, to be at least several times larger than the longest acoustic wavelength. This geometry ensures precise definition of the spatial periods of the interference patterns and, therefore, of the acoustic wavelengths.

The simplicity and general utility of the modern form of ISTS derives from engineering advances in optical systems that enable rapid measurement of acoustic dispersion with the two approaches described above (Rogers et al., 1997; Maznev et al., 1998; Rogers and Nelson, 1996; Rogers, 1998). Figure 6 schematically illustrates one version of the optical setup. Specially designed, phase-only beam shaping optics produce, through diffraction, a pair or an array of excitation pulses from a single incident pulse. A selected set of the diffracted pulses pass through a pair of imaging lenses and cross at the sample surface. Their interference produces simple or complex intensity patterns with geometries defined by the beam shaping optic and the imaging lenses. For the simple two-beam interference described by Equation 1, this optic typically

consists of a binary phase grating optimized for diffraction at the excitation wavelength. Roughly 80% of the light that passes through this type of grating diffracts into the $+1$ and -1 orders. The imaging lenses cross these two orders at the sample to produce a sinusoidal intensity pattern with periodicity Λ when the magnification of the lenses is unity and the period of the grating is 2Λ. More complex beam-shaping optics produce more than two excitation pulses, and, therefore, interference patterns that are characterized by more than one period. In either case, a useful beam-shaping optic contains many (20 to 50) different spatially separated diffracting patterns. The excitation geometry can then be adjusted simply by translating this optic so that the incident excitation pulses pass through different patterns. With this approach, the response of the sample at many different wavelengths can be determined rapidly, without moving any of the other elements in the optical system.

Detection is accomplished by imaging one or more diffracted probe laser beams onto the active area of a fast detector. When the intensity of these beams is measured directly, the signal is proportional to the product of the square of the amplitude of the out-of-plane acoustic displacements (i.e., the square of Equation 2) and the intensity of the probing light. Heterodyne detection approaches, which measure the intensity of the coherent optical interference of the signal beams with collinear reference beams generated from the same probe laser, provide enhanced sensitivity. They also simplify data interpretation since the heterodyne signal, *S(t)*, is linear in the material response. In particular, in the limit that the intensity of the reference beam, I_r, is large compared to the diffracted signal

$$S(t) \propto |\sqrt{I_p}R(t) + \sqrt{I_r}e^{i\varphi}| \approx I_r + 2\sqrt{I_p I_r}R(t)\cos\varphi \qquad (3)$$

where φ is the phase difference between the reference and diffracted beams and I_p is the intensity of the probing beam. A general scheme for heterodyne detection that uses the beam-shaping optic to produce the excitation pulses, as well as to generate the reference beam, is a relatively new and remarkably simple approach that makes this sensitive detection method suitable for routine use (Maznev et al., 1998; Rogers and Nelson, 1996). Figure 7 schematically illustrates the optics for measurements on transparent samples; a similar setup can be used in reflection mode. With heterodyne and nonheterodyne detection, peaks in the power spectrum of the signal define frequencies of the acoustic responses. In the case of nonheterodyne signals, sums and differences and twice these frequencies (i.e., cross terms that result from squaring Equation 2) also appear. Figure 8 compares responses measured with and without heterodyne detection.

The measured dispersion can be combined with models of the waveguide acoustics to determine intrinsic material properties. The general equation of motion for a nonpiezoelectric material is given by

$$\frac{\partial^2 \mathbf{u}_j}{\partial t^2} = \frac{\mathbf{c}_{ijkl}}{\rho}\frac{\partial^2 \mathbf{u}_k}{\partial x_i \partial x_l} + \frac{\partial}{\partial x_i}\left(\boldsymbol{\sigma}_{ik}^{(r)}\frac{\partial \mathbf{u}_j}{\partial x_k}\right) \qquad (4)$$

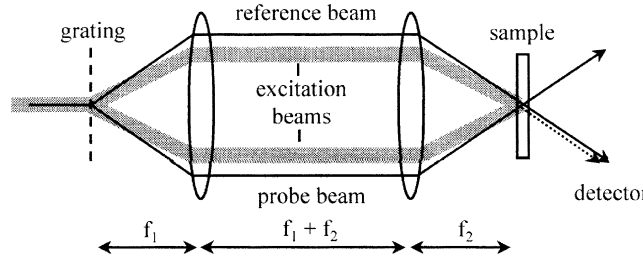


Figure 7. Optical system for ISTS measurements with heterodyne detection. Both the probe and the excitation laser beams pass through the same beam-shaping optic. The excitation pulses are split and recombined at the surface of the sample to launch the acoustic waves. The probe light is also split by this optic into more than one beam. For this illustration, the beam shaping optic produces a single pair of probing and excitation pulses. One of the probe beams is used to monitor the material response. The other emerges collinear with the signal light and acts as a reference beam for heterodyne amplification of signal generated by diffraction of the beam used for probing. The beam-shaping optic and imaging lenses ensure overlap of the signal and reference light. An optical setup similar to the one illustrated here can be used for measurements in reflection mode.

where $\mathbf{c}$ is the elastic stiffness tensor, $\mathbf{u}$ is the displacement vector, ρ is the density of the material and $\boldsymbol{\sigma}$ is the residual stress tensor. Solutions to this equation, with suitable boundary conditions at all material interfaces, define the dispersion of the velocities of waveguide modes in arbitrary systems (Farnell and Adler, 1972;Viktorov, 1976). Inversion algorithms based on these solutions can be used to determine the elastic constants, densities, and/or film thicknesses from the measured dispersion. For thickness determination, the elastic constants and densities are typically assumed to be known; they are treated as fixed parameters in the inversion. Similarly, for elastic constant evaluation, the thicknesses and densities are treated as fixed parameters. It is important to note, however, that only the elastic constants that determine the velocities of modes with displacement components in the vertical (i.e., sagittal) plane can be determined, because ISTS probes only these modes. Elastic constants that govern the propagation of in-plane shear acoustic waves polarized in the plane of the film, for example, cannot be evaluated. Figure 9 illustrates calculated distributions of displacements in the lowest six sagittal modes for a polymer film that is strongly bonded to a silicon substrate. As this figure illustrates, the modes involve coupled in- and out-of-plane shearing and compressional motions in the film and the substrate. The elastic constants and densities of the film and substrate materials and the thickness of the film determine the spatial characters and velocities of the modes. Their relative contributions to the ISTS signal are determined by their excitation efficiency and by their diffraction efficiency; the latter is dictated by the amount of surface ripple that is associated with their motion.

PRACTICAL ASPECTS OF THE METHOD

The setup illustrated in Figure 6 represents the core of the experimental apparatus. The entire system can either

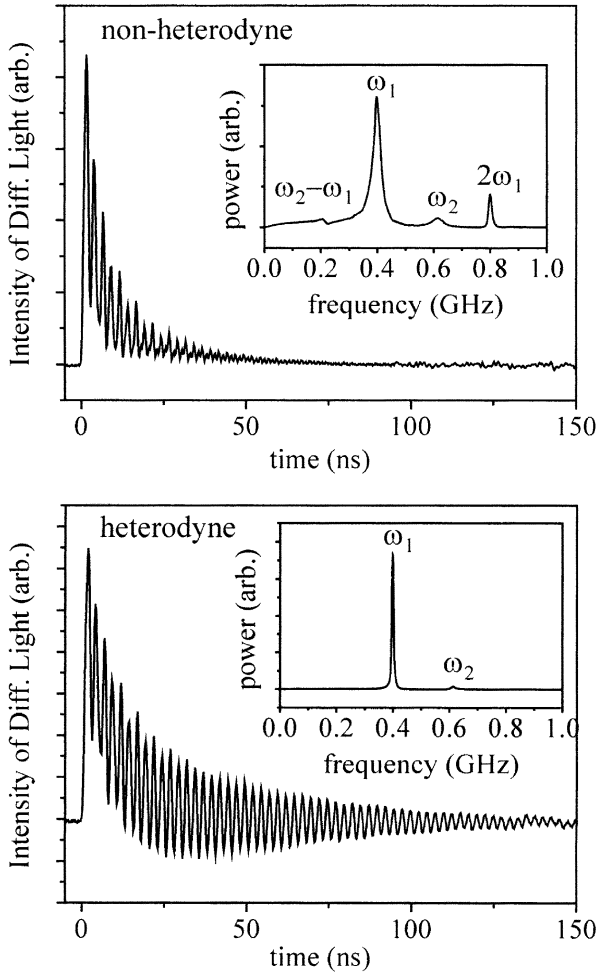


Figure 8. ISTS signal and power spectrum of the signal from an electroplated copper film (thickness ~1.8 μm) on a layer of silicon dioxide (thickness ~0.1 μm) on a silicon substrate at an acoustic wavelength ~7 μm. Part (**A**) shows the signal obtained without heterodyne amplification. The fact that the signal is quadratic with respect to the material displacement results in what appears as a fast decay of the acoustic oscillations. This effect is caused by the decay of the background thermal response rather than a decay in the acoustic component of the signal itself. Part (**B**) shows the heterodyne signal; the spectrum contains two peaks corresponding to Rayleigh and Sezawa waves, i.e., the two lowest-order modes of the layered structure. The signal without heterodyning shows these frequencies and combinations of them, due to the quadratic dependence of the signal on the response. The peak width in the spectrum of the signal in (**A**) is considerably larger than that in (**B**). This effect is due to the contribution of the thermal decay to the peak width in the nonheterodyne case. Note also that the heterodyne signal appears to drop below the baseline signal measured before the excitation pulses arrive. This artifact results from the insensitivity of the detector to DC levels of light.

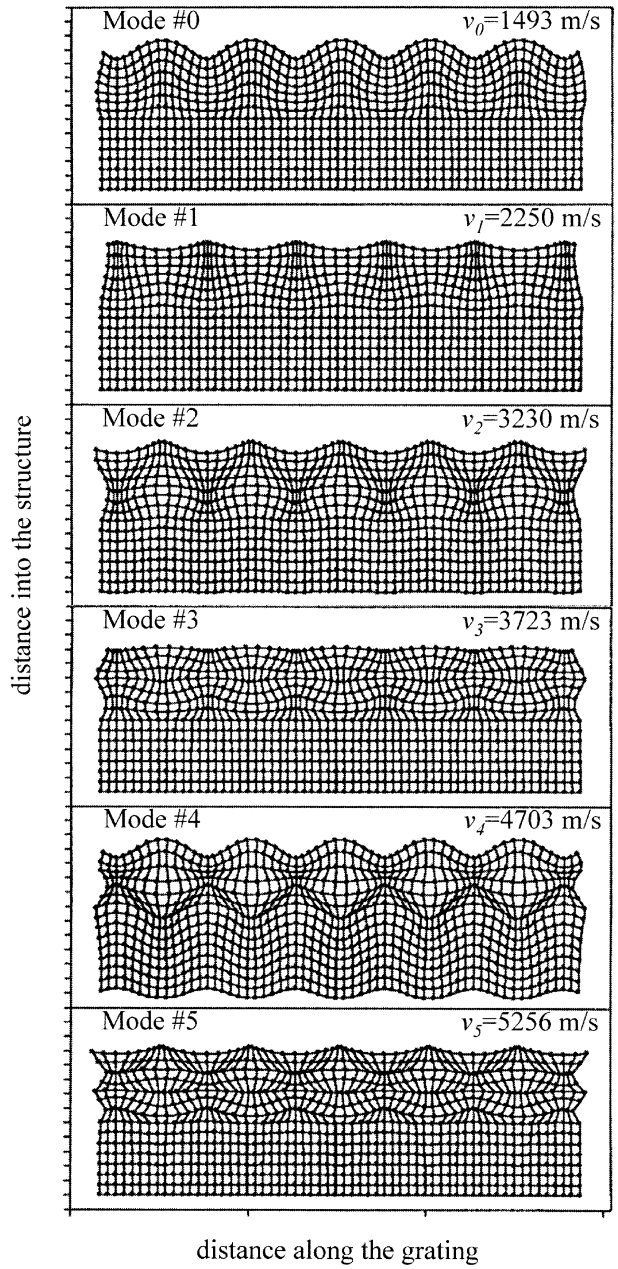


Figure 9. Computed displacement distributions for the lowest six waveguide modes in a polymer film supported by a silicon substrate. These modes are known as Rayleigh or Sezawa modes. They are excited and probed efficiently in ISTS measurements because they involve sagittal displacements that couple strongly both to the thermal expansion induced by the excitation pulses and to surface ripple, which is commonly responsible for diffracting the probe laser. The spatial nature of these modes and their velocities are defined by the mechanical properties of the film and the substrate, the nature of the interface between them, and the ratio of the acoustic wavelength to the film thickness.

be obtained commercially or most of it can be assembled from conventional, off-the-shelf components (lenses, lasers, etc.). The beam-shaping optics can be obtained as special order parts from commercial digital optics vendors. Alternatively, they can be custom built using relatively simple techniques for microfabrication (Rogers, 1998).

The excitation pulses are typically derived from a flash lamp or diode-pumped Nd:YAG laser, although any source of laser pulses with good coherence and mode properties can be used. The pulse duration must be short compared to the temporal period of the excited acoustic wave. In many instances, pulses shorter than ~300 psec

are suitable. Pulse energies in the range of one or a few μJ provide adequate excitation for most samples. Nonlinear crystals can be used to double, triple, quadruple, or even quintuple the output of the Nd:YAG (wavelength ~1064 nm) to produce 532, 355, 266, or 213 nm light. The color is selected so that a sufficient fraction of the incident light is absorbed by the sample. Ultraviolet (UV) light is a generally useful wavelength in this regard, although the expense and experimental complexity of using multiple nonlinear crystals represent a disadvantage of UV light when a Nd:YAG laser is used. Nonlinear frequency conversion can be avoided by depositing thin absorbing films onto samples that are otherwise too transparent or reflective to examine directly at the fundamental wavelength of the laser. The probe laser can be pulsed or continuous wave. For most of the data presented in this unit, we used a continuous-wave infrared (850 nm) diode laser with a power of ~200 mW. Its output is electronically gated so that it emits only during the material response, which typically lasts no longer than 100 μsec after the excitation pulses strike the sample.

Alignment of the optics (which is performed at the factory for commercial instruments) ensures that the probing beam overlaps the crossed excitation beams and that signal light reaches the detector. The size of the beams at the crossing point is typically in the range of one or several hundred microns. The alignment of the probe laser can be aided by the use of a pinhole aperture to locate the crossing point of the excitation beams. Routine measurement is accomplished by placing the surface of the sample at the intersection of the excitation and probing laser beams, moving the beam-shaping optic to the desired pattern(s), recording data, and interpreting the data. The sample placement is most easily achieved by moving the sample along the *z* direction (Fig. 6) to maximize the measured signal, which can be visualized in real time on a digitizing oscilloscope. The surface normal of the sample is adjusted to be parallel to the bisector of the excitation beams.

Recorded data typically consist of an average of responses measured for a specified number of excitation-probing events. The strength of the signal, the required precision, and the necessary measurement time determine the number of averages: between one hundred and one thousand is typical. This averaging requires <1 sec with common excitation lasers (repetition rates ~1 kHz). Translating this optic so that the excitation laser beam passes through different diffracting patterns changes the acoustic wavelength(s). Single or multiple measurements of this type define acoustic response frequencies as a function of wavelength, which, in most cases, varies from 2 to 200 μm. The optics and the wavelengths of the lasers determine the range that is practical.

The measurement procedures described above apply to a wide variety of samples. Figures 10 to 13 show, respectively, data from a film of lead zirconium titanate (PZT) on silicon, a thin layer of nanoporous silica glass on silicon, a layer of paint on a plastic automobile bumper part, and a film of cellulose. In the last example, the data were collected with a single shot of the laser. In each case, the oscillations in the signal are produced by surface acoustic waveguide modes that involve coupled displacements in the outer film, and, in general, in all material layers beneath it. Their frequencies are therefore functions of the material properties of each component of the structure and the nature of the boundary conditions between them. These and other waveguide properties can be calculated

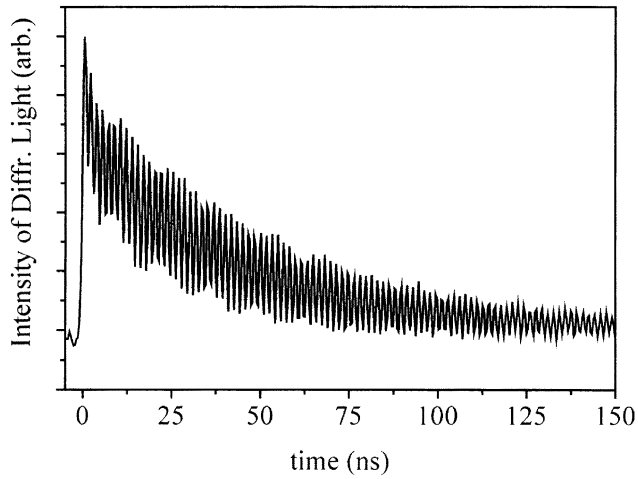


Figure 10. ISTS data from a thin film of lead zirconium titanate on a layer of platinum on a silicon wafer. Crossed laser pulses coherently stimulate the lowest-order Rayleigh waveguide acoustic mode, with a wavelength defined by the optical interference pattern. The motions induce ripple on the surface of the structure; this ripple diffracts a probing laser beam that is overlapped with the excited region of the sample. The acoustic waveguide motions involve coupled displacements in the film and underlying substrate. In this case, the overall decay of the signal caused by thermal diffusion occurs on a time scale comparable to the acoustic decay time.

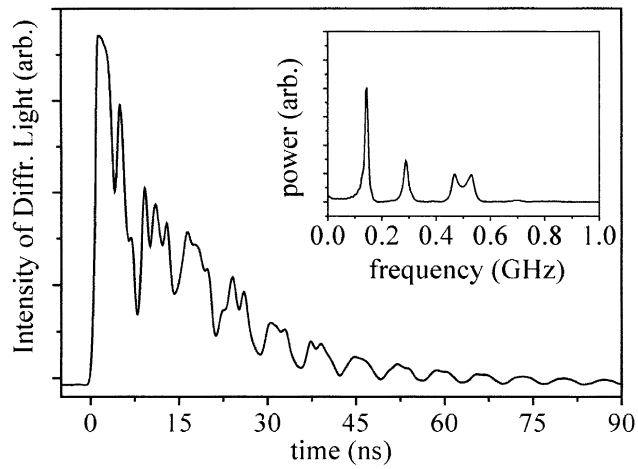


Figure 11. ISTS data from a thin film of a nanoporous glass on a silicon substrate. A thin overlayer of aluminum was added in this case to induce optical absorption at the wavelength of the excitation laser (1064 nm). The measurement stimulates numerous acoustic waveguide modes, each with a wavelength equal to the period of the optical interference pattern formed by the crossed laser pulses. The power spectrum in the inset shows the various frequency components of the signal. In this case, the overall decay of the signal caused by thermal diffusion occurs on a time scale comparable to the decay time of the lowest order acoustic waveguide mode.

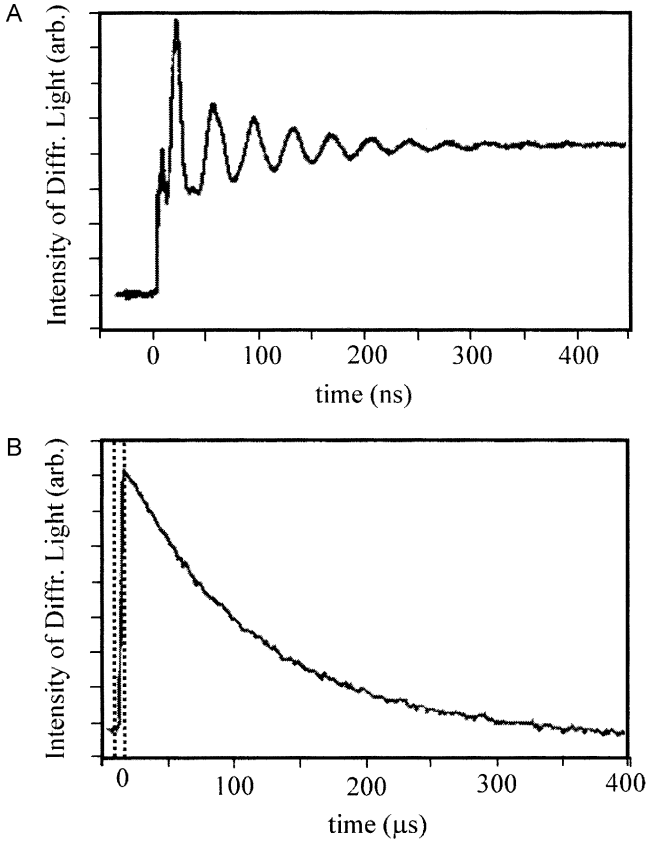


Figure 12. ISTS data from a layer of paint on a plastic automobile bumper part. Part (**A**) shows the acoustic component of the signal. Part (**B**) shows the thermal decay, which occurs on a microsec time scale. The vertical dashed lines in the bottom frame indicate the temporal window of the signal shown in the upper frame. The acoustic and thermal properties of paint can be relevant to developing materials that resist delamination and maintain their mechanical flexibility even after exposure to harsh environmental conditions. The dashed lines in (B) indicate the temporal range illustrated in (A).

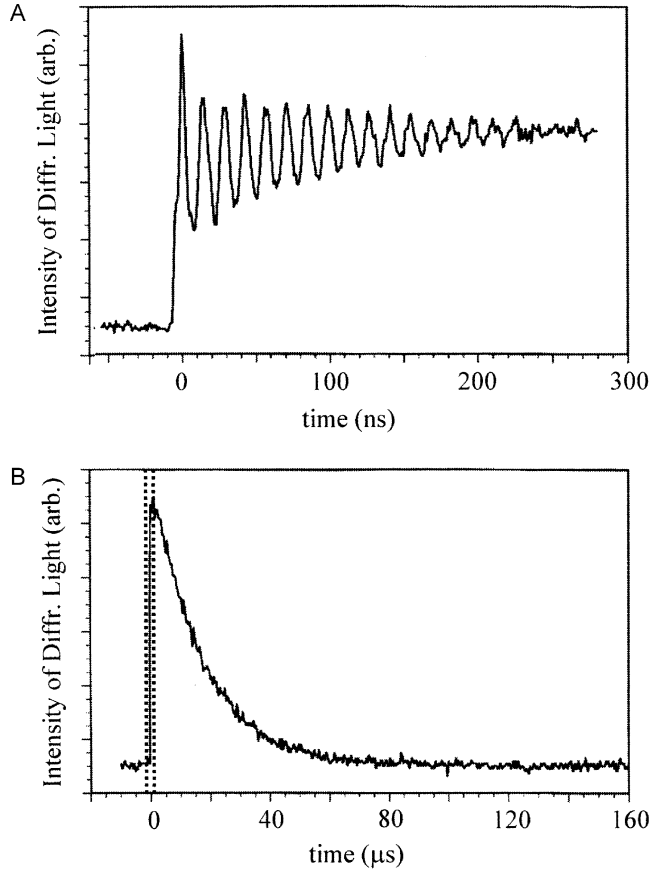


Figure 13. ISTS data collected in a single shot of the laser from a film of cellulose. Parts (**A**) and (**B**) show the acoustic and thermal components of the signal that was recorded in this measurement. The dashed lines in (B) indicate the temporal range illustrated in (A).

from first principles using standard procedures (Farnell and Adler, 1972; Viktorov, 1976).

The results in Figures 10 to 13 are representative of the range of acoustic responses that are typically encountered. The PZT film on silicon (Fig. 10) only supports a single waveguide mode at this acoustic wavelength, because its thickness is small compared to the wavelength. The response consists, therefore, of a single frequency whose value is determined primarily by the density and thickness of the film and by the mechanical properties of the silicon. In contrast, the nanoporous glass on silicon (Fig. 11) supports many acoustic waveguide modes because the acoustic wavelength is comparable to the film thickness. The data, as a result, contain many frequency components. Several of these modes involve motions confined largely to the film; their characteristics are determined primarily by the film mechanical properties and its film thickness. The layer of paint on the plastic bumper part (Fig. 12) is thick compared to the acoustic wavelength and it therefore supports many waveguide modes, like the nanoporous glass. The excitation in this case, however, is localized to the very top surface of the sample. Only the lowest waveguide mode, which is known as the Rayleigh mode, is stimulated and detected efficiently. The frequency of this mode is determined entirely by the mechanical properties of the film. It is independent of the film thickness and of the mechanical properties of the substrate. Finally, the cellulose sample (Fig. 13) is a thick free-standing film. The excited mode, in this case, is similar, in its physical characteristics and in its sensitivity to film properties, to the Rayleigh mode observed in the paint sample.

The precision (i.e., repeatability) of the frequencies measured from data like those shown in Figs. 10 to 13 can be better than 0.01% to 0.001%. The acoustic wavelengths are determined by the experimental apparatus: the geometry of the beam-shaping optic, its position relative to the imaging lenses, and the magnification of these lenses. This information, or a direct measurement of the excitation beam crossing angles, can be used to evaluate the wavelengths. Measurements with high accuracy are typically achieved by measuring response frequencies from standard samples or by using a microscope to view the interference pattern directly. With these latter two techniques the wavelengths can be determined to an accuracy of better than 0.1%.

METHOD AUTOMATION

Many aspects of the measurement can be automated. The beam-shaping optic in a commercial machine, for example, is mounted on a computer-controlled motorized translation stage. This motor, along with the transient recorder (i.e., digitizing oscilloscope), a variable neutral density filter for adjusting the intensity of the excitation light, and a precision stage that supports the sample, are all controlled by a single computer interface. Robotic handlers are also available to load and unload samples into and from the machine. An automated focusing operation positions the sample along the z direction to maximize the signal. The computer controls data collection at user-defined acoustic wavelengths and also computes the power spectrum from the time domain data. It fits peaks in this spectrum to define the acoustic frequencies, and, in the case of film thickness, uses these frequencies (and the known acoustic wavelength, elastic constants, and densities) to evaluate the unknown layer(s). This fully automated procedure can be coordinated with motion of the sample stage to yield high-resolution (better than 10-μm in certain cases) images of the acoustic frequency, other aspects of the ISTS signal (e.g., amplitude, acoustic damping rate, thermal decay time, etc.), and/or computed material characteristics (e.g., film thickness). In a noncommercial ISTS apparatus for general laboratory use, the stages for the sample, the beam-shaping optic, the neutral density filter, etc., are manually controlled. Data downloaded from an oscillosope in such a setup are analyzed using a separate computer equipped with software for the fitting and modeling procedures.

DATA ANALYSIS AND INITIAL INTERPRETATION

As discussed in the previous sections, the ISTS measurements provide information on the acoustic properties of the sample. The acoustic frequencies can be obtained directly by fitting the time domain ISTS data to expressions given in Equation 2 or Equation 3. They can also be determined by Fourier transformation. In the latter case, the positions of the peaks in the power spectrum are typically defined by fitting them to Lorentzian or Gaussian line shapes. The experimental apparatus defines the crossing angles of the excitation pulses and, therefore, the acoustic wavelength(s).

Inverse modeling with the acoustic frequencies and wavelengths determines elastic and other properties. In this procedure, the elastic constants, densities and thicknesses of the "unknown" layer(s) are first set to some approximate initial values. The dispersion is then calculated with Equation 4 and the known properties of the other layers. The sum of squared differences, χ^2, between the computed and measured modal phase velocities provides a metric for how well the modeling reproduces the observed dispersion. In a generally nonlinear iterative search routine, the properties of the unknown layers are adjusted and χ^2 is calculated for each case. The properties that minimize χ^2 represent best fit estimates for the intrinsic characteristics of the "unknown" layers. For the simple case of determining the thickness of a single layer in a film stack whose other properties are known, this procedure reduces to a straightforward non-iterative calculation that yields the thickness from a single measured phase velocity. Clearly, for these types of fitting procedures to be successful, the number of parameters to be determined cannot exceed the number of measured velocities. The uncertainties in the fitted parameters will typically decrease as the number of measured velocities, and the range of wavelengths increase. These uncertainties can be quantitatively estimated using established statistical analysis such as the F test (Beale, 1960). Below we examine this inverse modeling in some detail, with illustrations for rigidity and stress determination in thin membranes, elastic moduli determination in supported and unsupported films, and thickness evaluation of films in multilayer stacks.

We begin with the simple case of a thin unsupported membrane. When the thickness (h) of the membrane is much smaller than the acoustic wavelength, then it is possible to model the dispersion of the drumhead mode (i.e., the lowest-order waveguide mode, which is typically easy to observe in an ISTS measurement) with small-deflection plate theory. We also assume that the ISTS excitation region is large enough to ignore the dimension perpendicular to the interference fringes (i.e., the y axis is Fig. 6). In these limits, Equation 4 reduces to (Rogers and Bogart, 2000)

$$\frac{Eh^2}{12(1-\upsilon^2)}\frac{\partial^4 \mathbf{u}}{\partial x^4} + \rho\frac{\partial^2 \mathbf{u}}{\partial t^2} = \sigma\frac{\partial^2 \mathbf{u}}{\partial x^2} \tag{5}$$

where E is Young's modulus, ν is the Poisson ratio, ρ is the density of the film, and σ is the residual stress. The coordinate x lies along the interference fringes. The dispersion of the phase velocity (ν_φ) determined with Equation 5 can be written

$$\nu_\varphi = \sqrt{\frac{E}{12\rho(1-\nu^2)}(kh)^2 + \frac{\sigma}{\rho}} = \sqrt{\frac{D}{\rho h}k^2 + \frac{\sigma}{\rho}} \tag{6}$$

where D is the flexural rigidity. This simple equation can be used with ISTS measurements of the dispersion of the lowest-order Lamb mode to determine, for example, the flexural rigidity and stress when the thickness and density of the membrane are known. Figure 14 shows ISTS data from an unsupported bilayer membrane of tungsten (W; ~25 nm)/silicon nitride (SiN; ~150 nm) and best fit curves that use Equation 6. The results confirm the validity of the plate mode approximations for this sample. The stress and rigidity measured for this sample are 241 ± 1 MPa and 253 ± 1 GPa respectively. The uncertainties in this case are dominated by uncertainties in the density of the membrane. We note that the ISTS results agree with independent evaluation of these quantities by the resonant frequency (RF) method and the bulge test, respectively (Rogers and Bogart, 2000; Rogers et al., 2000b). Unlike ISTS, however, both of these methods require specialized test structures. Also, they do not offer spatial resolution

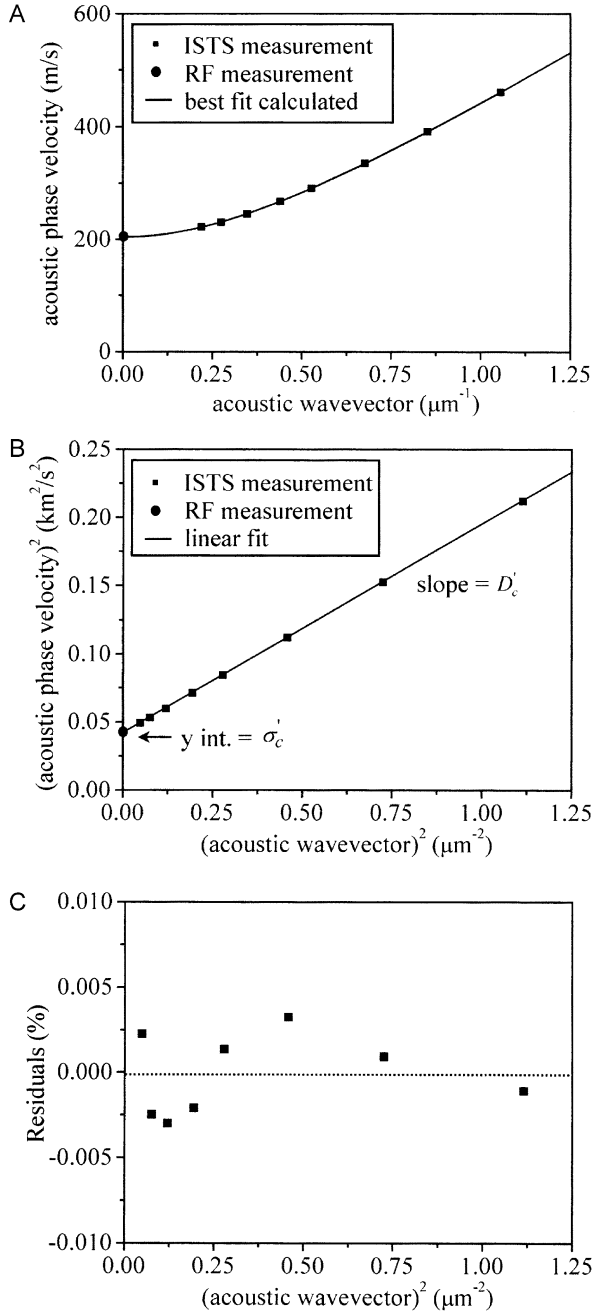


Figure 14. Dispersion in a bilayer membrane of tungsten (25 nm) and silicon nitride (150 nm) measured by ISTS. Part (**A**) shows the measured (symbols) and best fit calculated (line) variation of the acoustic phase velocity with wavevector. Part (**B**) shows the expected linear behavior of the square of the velocity with the square of the wavevector. The slope of the line that passes through these data points defines the ratio of the composite flexural rigidity to the product of the composite density and the membrane thickness. Its intercept determines the ratio of the composite residual stress (tensile) to the composite density. Part (**C**) illustrates the percent deviation of the data shown in part (**B**) from the best fit line. The small size of these residuals provides some indication of (1) the accuracy of the acoustic frequency measured by ISTS, (2) the accuracy of the measured wavelengths, and (3) the validity of the plate theory approximations. The large circular symbols in parts (**A**) and (**B**) are data from independent resonant frequency measurements on this sample.

and they cannot easily evaluate in-plane anisotropies in the stress or rigidity.

As the acoustic wavelength approaches the thickness of the membrane, the plate theory ceases to be valid, and a complete waveguide description based on Equation 4 is required. In this regime, multiple dispersive modes known as Lamb modes are possible (Farnell and Adler, 1972; Viktorov, 1976). Their velocities, v_i, at an acoustic wavelength, Λ, are defined by positions of zeroes in the determinant of a matrix defined by the boundary conditions appropriate for the system. If this determinant is denoted by the function G, then for a mechanically isotropic, stress-free membrane we can write

$$G(v_i, \Lambda, h, v_{tr}^{(f)}, v_{lg}^{(f)}) = 0 \tag{7}$$

where $v_{tr}^{(f)}$ and $v_{lg}^{(f)}$ are the intrinsic transverse and longitudinal velocities of the film (i.e., the velocities that would characterize acoustic propagation in a bulk sample of the film material). Simple equations relate these intrinsic velocities to the Young's modulus and Poisson ratio (Fetter and Walecka, 1980). Figure 15 shows the measured dispersion of a set of Lamb modes in an unstressed polymer membrane and calculations based on Equation 7 with best fit estimates for the intrinsic velocities: $v_{tr}^{(f)} = 1126 \pm 4$ m/sec, and $v_{lg}^{(f)} = 2570 \pm 40$ m/sec (Rogers et al., 1994b). A general feature of the dispersion is that the mode velocities scale with the product of the acoustic wavevector, $k = 2\pi/\Lambda$, and the thickness, h, as expected based on the plate mode result of Equation 6. Plate-like behavior of the lowest-order mode results when the acoustic wavelength is long compared to the film thickness (i.e., kh is small). In this limit, the second-lowest-order mode acquires a velocity slightly smaller than the intrinsic longitudinal velocity of the film, $v_{lg}^{(f)}$. When kh is large, the velocities of the two lowest modes approach the Rayleigh wave velocity of the film (i.e., the Rayleigh mode is a surface localized wave that propagates on a semi-infinite substrate). All

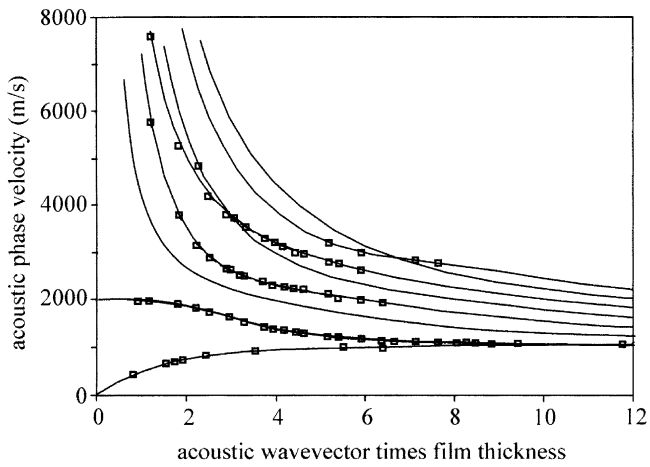


Figure 15. Dispersion in a thin unsupported polymer film measured by ISTS (symbols). The lines represent best fit calculations that allowed the intrinsic longitudinal and transverse acoustic velocities of the film to vary. With the density, simple expressions can be used to relate these velocities to the elastic moduli.

other mode velocities approach the intrinsic transverse velocity of the film, $v_{tr}^{(f)}$. These and other characteristic features of the dispersion illustrate how the elastic properties (i.e., intrinsic velocities) of a thin film can be determined from the measured variation of waveguide mode velocities with wavelength. Iterative fitting routines achieve this determination in an automated fashion that also allows for statistical estimation of uncertainties. Similar procedures can be used to evaluate membranes with anisotropic elastic constants (Rogers et al., 1994a).

When the film is supported by a substrate or when it is part of a multilayer stack, the associated boundary conditions are functions of the elastic properties, densities, and thicknesses of all constituent layers and the supporting substrate. In the general case, the boundary condition determinant can be written

$$G(v_i, \Lambda, h^{(f1)}, v_{tr}^{(fl)}, v_{lg}^{(f1)}, \rho^{(f1)}, h^{(f2)}, v_{tr}^{(f2)}, v_{lg}^{(f2)}, \rho^{(f2)}, \ldots, v_{tr}^{(s)}, v_{lg}^{(s)}, \rho^{(s)}) = 0 \quad (8)$$

where $v_{tr}^{(fi)}$ and $v_{lg}^{(fi)}$ are the intrinsic transverse and longitudinal velocities of the films, $h^{(i)}$ are their thicknesses, and $\rho^{(fi)}$ are their densities. The corresponding quantities for the substrate (assumed to be semi-infinite) are $v_{tr}^{(s)}$, $v_{lg}^{(s)}$ and $\rho^{(s)}$. Figure 16 shows the best fit computed and measured dispersion for a thin film of nanoporous silica on a silicon wafer (Rogers and Case, 1999). The intrinsic velocities determined from this fitting are $v_{lg} = 610 \pm 50$ m/sec and $v_{tr} = 400 \pm 30$ m/sec. The uncertainties in this case are dominated by uncertainties in the thickness and density of a thin overlayer of aluminum that was used to induce slight absorption at 1064 nm, the wavelength of the excitation pulses in these measurements. Note that the overall behavior of the dispersion in this case is much different from that for an unsupported film (Fig. 15). The velocity of the lowest mode approaches the Rayleigh velocities of the film and substrate at large and small values of kh, respectively. All other modes approach $v_{tr}^{(f)}$ at large kh. As kh decreases, each of these modes is, at a certain value of kh, "cut off" at $v_{tr}^{(s)}$. Above this velocity they are no longer strictly guided modes. In this regime, some of their energy "leaks" into bulk transverse waves in the semi-infinite substrate.

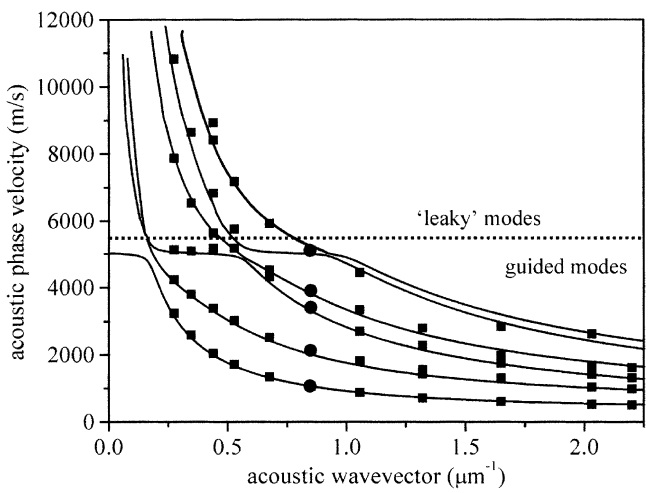


Figure 16. Dispersion in a thin film of nanoporous glass on a silicon substrate measured by ISTS (symbols). In this case, a thin film of aluminum (thickness ~75 nm) was deposited on top of the nanoporous glass in order to induce absorption at the wavelength (1064 nm) of the excitation laser used for these measurements. The lines represent best fit calculations that allowed the intrinsic longitudinal and transverse acoustic velocities of the nanoporous glass to vary. The fitting used literature values for the densities and elastic properties of the aluminum and the silicon. The thickness of the aluminum film was fixed to its nominal value. The circular symbols represent data illustrated in Figure 8b.6.11.

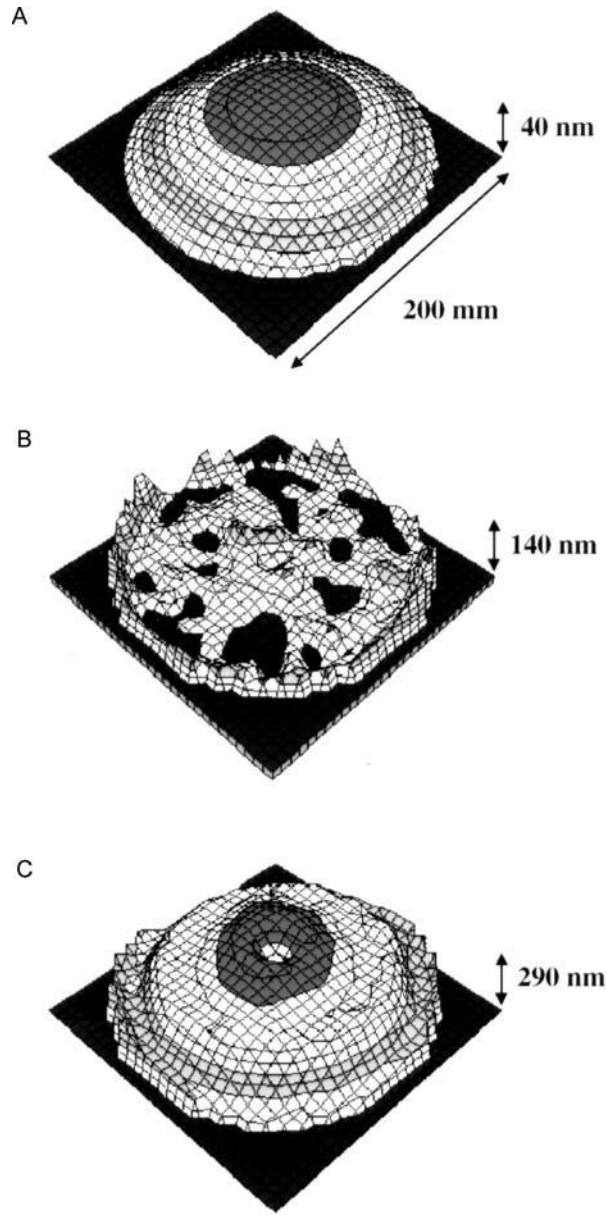


Figure 17. ISTS measured maps of thickness in thin films of copper on silicon wafers with 200 mm diameters. Parts (**A**), (**B**), and (**C**) show, respectively, copper deposited by sputter deposition (average thickness ~180 nm), electroplating (average thickness ~1.6 μm), and an electroplated copper film after chemical-mechanical polishing (average thickness ~500 nm).

The dependence of the waveguide mode velocities on film thickness provides the basis for thickness determination with ISTS (Banet et al., 1998). Often, in practice a single measurement of a mode velocity coupled with the properties of the other components of the structure determines the thickness. The accuracy of the measurement is, of course, highest when the velocity of the waveguide mode depends strongly on the thickness and when the properties of the other components of the structure are known accurately. In a silicon supported system, the velocity of the lowest-order mode (typically excited and probed efficiently in an ISTS measurement) in most cases exhibits sufficient variation with thickness when kh is not large. For films with thicknesses less than ~100 μm, and with properties that are reasonably dissimilar to those of the substrate, it is possible to tune k into a range that allows good sensitivity. Figures 17 and 18 show several examples of thickness evaluation in thin metal films used in microelectronics. Note that the accuracy of the measured thickness depends on the assumed elastic properties, densities, and thicknesses of the substrate and other films in the stack. For many systems, literature values for the physical properties yield acceptable accuracy. In other cases, procedures that calibrate these properties through ISTS analysis of samples with known thicknesses are necessary. For complex structures, such as copper damascene features in microelectronics (i.e., lines of copper embedded in layers of oxide), effective medium models are often preferred to the complex calculations that are required to simulate these systems using first-principles physics based on Equation 4. Figure 19 shows the results of line scans across copper damascene structures that use these types of models for computing thicknesses. With calibrated tools for measuring these and other systems for microelectronics, thickness

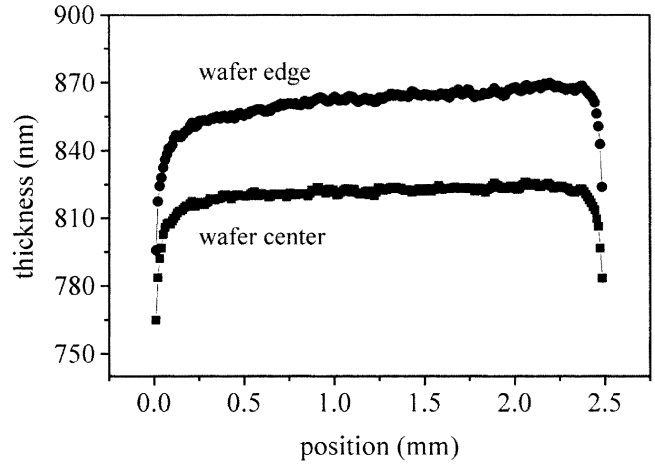


Figure 19. Thickness profiles of copper damascene arrays consisting of copper lines (width 0.5 μm, length ~1 mm and average thickness ~800 nm) separated by 0.5 μm and embedded in a layer of silicon dioxide. Acoustic wave propagation direction is along the trenches (i.e., fringes of the excitation interference pattern lie perpendicular to the long dimension of the copper lines). In this case, effective elastic properties of the copper damascene structure needed for the thickness calculation can be determined by averaging elastic constants of copper and silicon dioxide using a parallel spring model. Thickness values determined using these effective properties were in agreement with SEM measurements. The data illustrate a major problem in the copper interconnect technology, i.e., nonuniformities in the chemical-mechanical polishing process used to form these structures.

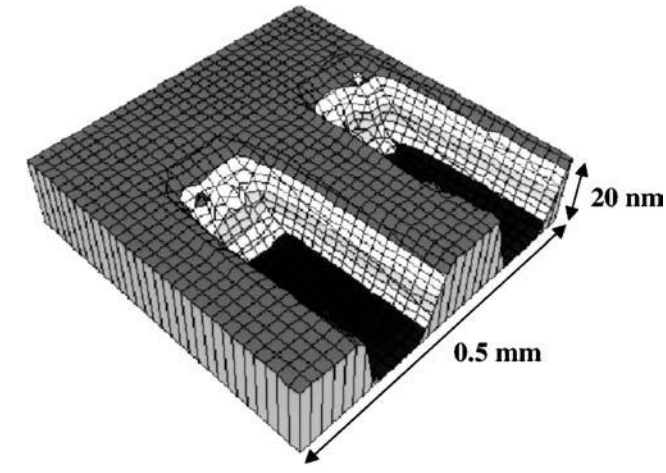


Figure 18. Detailed contour map of an 0.8 × 0.8−mm region showing variations in the thickness of a layer of tantalum in a stack structure of 200 nm copper/Ta (patterned)/300 nm silicon dioxide/silicon. The measurements clearly resolve the regions of the structure where there is no tantalum. The thickness of tantalum in the other areas is ~25 nm. In order to measure the thickness of this buried layer, thicknesses and properties of the other layers, and the mechanical properties and density of the tantalum, were assumed to be known.

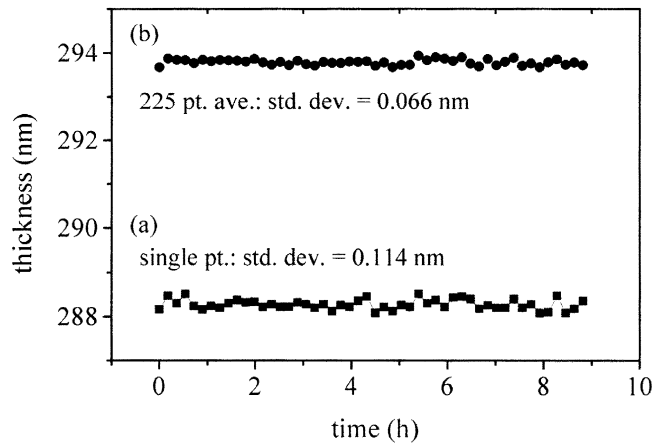


Figure 20. Reproducibility data of ISTS metal film thickness measurement obtained on a commercially available system (Philips Analytical's PQ Emerald). The sample (a silicon wafer with 200 mm diameter and coated with 300 nm tantalum on 300 nm silicon dioxide) was loaded and unloaded from the system for each measurement. The signal was averaged over 300 laser shots (measurement time ~1 sec per point). (**A**) Single-point data. (**B**) Average of a 225-point wafer map. In both cases, the reproducibility is on the order or less than 0.1 nm, roughly a single atomic diameter.

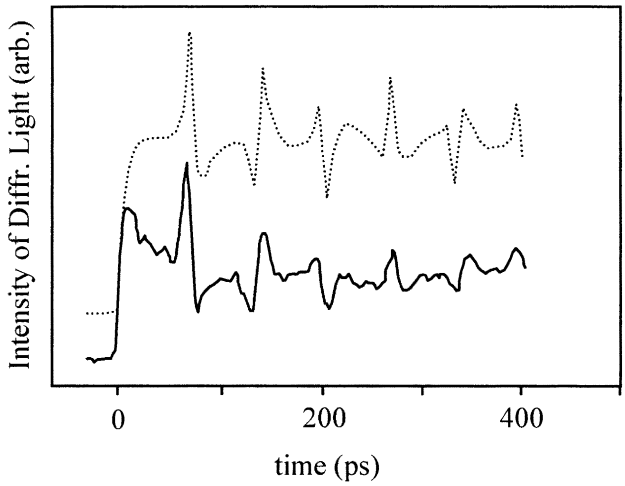


Figure 21. ISTS data on a picosec time scale from a multilayer stack of aluminum (thickness ~100 nm) on titanium-tungsten (thickness ~75 nm) on a silicon wafer. The dashed line shows a simulated response. The features in the data correspond to reflections of longitudinal acoustic wavepackets generated at the surface of the sample by the crossed excitation pulses. The timing of these echoes, with the out-of-plane longitudinal acoustic velocities of the materials, yields the thicknesses of both of the films.

precision and accuracy are both in the range of a few angstroms. Figure 20 illustrates the repeatability for a typical film.

When the thicknesses of more than one film in a multilayer stack are required, additional data, such as (1) velocities of more than one waveguide mode, (2) the velocity of a single mode at more than one k, (3) information other than the acoustic frequency (e.g., acoustic damping, thermal or electronic response, etc.), or (4) the out-of-plane acoustic echo component of the ISTS signal can be employed. Figure 21 illustrates the last approach. The precision and accuracy (again, with proper calibration) are both in the angstrom range with this method. Commercial ISTS tools use a combination of the thermal and acoustic components of the data to determine thicknesses of two films in a stack (Gostein et al., 2000). Here, the precision and accuracy are in the range of 0.05 to 0.15 nm. The particular technique that is most convenient for a given measurement application depends on the sample.

SAMPLE PREPARATION AND SPECIMEN MODIFICATION

The primary constraints on the sample are that (1) it be sufficiently smooth that it does not scatter enough light to interfere with measurement of the diffracted signal, and (2) it absorb a sufficient fraction of the excitation light to yield a measurable response. From a practical standpoint, the intensity of scattered light that is spatially close to or overlapping the signal should not be large compared to the intensity of the signal itself. Typical diffraction efficiencies from acoustic motions in an ISTS experiment are in the range of 10^{-6}. Heterodyne detection schemes significantly amplify the signal levels, thereby reducing the effects of parasitic scatter. Spatial filtering can also help to reduce the amount of scatter that reaches the detector. From a sample-preparation point of view, scatter can be reduced in many cases by adding an index-matching fluid, or by mechanical polishing.

As mentioned above (see Practical Aspects of the Method), in order to ensure adequate ISTS signals, excitation light must produce some mild amount of heating in the sample. This heating can be achieved by using an excitation wavelength that is absorbed directly by the sample, or by adding a thin absorbing overlayer. Generally, the materials used for these overlayer films and their thicknesses are chosen to minimize their effects on the acoustic response of the sample under test. Alternatively, the acoustic influence of these films can be accounted for directly through calculation. Thin aluminum coatings are useful because they provide sufficient absorption to allow high quality ISTS measurements with excitation pulse energies in the range of a few μJ and wavelengths in the infrared (e.g., 1064 nm). Aluminum also has the advantage that it is a low-density material (small acoustic loading for thin films) that can be easily deposited in thin film form by thermal evaporation. Coatings with ~50 to 75 nm thickness provide sufficient absorption, and they also only minimally affect the acoustic response of most samples.

PROBLEMS

Problems that can arise with ISTS evaluation fall into two classes, those associated with acquiring data and those related to interpreting it. For interpretation, multilayer samples can provide the biggest challenge, because uncertainties in the properties of any layer in the stack can significantly affect the accuracy for evaluating the "unknown" layer(s). Accurate measurements of the dispersion of multiple modes over a wide range of k reduces the significance of this problem because it decreases the number of assumptions necessary for the properties of the other layers in the stack.

For data collection, the sample must absorb some reasonable fraction of the excitation light. Convenient laser sources for excitation (e.g., pulsed Nd:YAG lasers) provide direct output in the near-IR range (e.g., 1064 nm). Because of the relatively high peak power, nonlinear optical crystals can be used to generate harmonics of this fundamental frequency. Gold, for example, reflects too efficiently at 1064 nm to allow for ISTS measurement with the fundamental output of a Nd:YAG laser. It absorbs, however, enough frequency-doubled light at 532 nm to allow for easy measurement. Generally, the fourth harmonic of these types of lasers (266 nm) provides a useful wavelength for measurements on a wide range of samples.

With a suitable wavelength for excitation, high-quality signals can be obtained for most samples that do not scatter strongly at the wavelength of the probing laser. Scattered light that reaches the detector can interfere with data collection either by simply adding noise or, when it coherently interferes with diffracted signal light, by distorting the shape of the measured waveform. The former problem can, in most cases, be eliminated by carefully

blocking the scattered light so that it does not reach the detector. The latter problem cannot be handled in the same way because these effects are produced by scattered light that spatially overlaps the signal light. This interference essentially causes uncontrolled heterodyne amplification. It is often difficult to simulate accurately this interference and its subtle effects on the accuracy of procedures for extracting the response frequency (e.g., Fourier transformation followed by peak fitting), because the phase and intensity of the scattered light are not generally known. These effects can be minimized, however, by employing heterodyne detection and adjusting the intensity of the reference beam so that its interference with the signal dominates the measured response. In any case, the precision of the measurement can be easily evaluated by repeated measurement of a single spot on a sample. The standard deviation of a set of frequencies measured in this fashion should be less than ~0.1% for most samples.

LITERATURE CITED

Allen, M. G., Mehregany, M., Howe, R. T., and Senturia, S. D. 1987. Microfabricated structures for the in situ measurement of residual stress, Young's modulus, and ultimate strain of thin films. *Appl. Phys. Lett.* 51:241–243.

Banet, M. J., Fuchs, M., Rogers, J. A., Reinold, J. H., Knecht, J. M., Rothschild, M., Logan R., Maznev, A. A., and Nelson, K. A. 1998. High-precision film thickness determination using a laser-based ultrasonic technique. *Appl. Phys. Lett.* 73:169–171.

Beale, E. M. L. 1960. Confidence regions in non-linear estimation. *J. Roy. Stat. Soc. B*. 22:41–76.

Clemens, B. M., and Bain, J. A. 1992. Stress determination in textured thin films using X-ray diffraction. *M. R. S. Bulletin* July:46–51.

Crimmins, T. F., Maznev, A. A., and Nelson, K. A. 1998. Transient grating measurements of picosecond acoustic pulses in metal films. *Appl. Phys. Lett.* 74:1344–1346.

Duggal, A. R., Rogers, J. A., and Nelson, K. A. 1992. Real-time optical characterization of surface acoustic modes of polyimide thin-film coatings. *J. Appl. Phys.* 72:2823–2839.

Eesley, G. L., Clemens, B. M., and Paddock, C. A. 1987. Generation and detection of picosecond acoustic pulses in thin metal films. *Appl. Phys. Lett.* 50:717–719.

Eichler, H. J., Gunter, P., and Pohl, D. W. 1986. Laser-Induced Dynamic Gratings. Springer-Verlag, New York.

Farnell, G. W. and Adler, E. L. 1972. Elastic wave propagation in thin layers. *In* Physical Acoustics, Principles and Methods, Vol. IX (W. P. Mason and R. N. Thurston, eds.) pp. 35–127. Academic Press, New York.

Fetter, A. L. and Walecka, J. D. 1980. Theoretical Mechanics of Particles and Continua. McGraw-Hill, New York.

Gostein, M., Banet, M. J., Joffe, M., Maznev, A., Sacco, R., Rogers, J. A. and Nelson, K. A. 2000. Opaque film metrology using transient-grating optoacoustics (ISTS). *In* Handbook of Silicon Semiconductor Metrology (A. Diebold, ed.) Marcel-Dekker, New York.

Hess, P. 1996. Laser diagnostics of mechanical and elastic properties of silicon and carbon films. *Appl. Surf. Sci.* 106:429–437.

Maden, M. A., Jagota, A., Mazur, S., and Farris, R. J. 1994. Vibrational technique for stress measurement in thin films. 1. Ideal membrane behavior. *J. Am. Ceram. Soc.* 77:625–635.

Maznev, A. A., Nelson, K. A., and Yagi, T. 1996. Surface phonon spectroscopy with frequency-domain impulsive stimulated light scattering. *Sol. St. Comm.* 100:807–811.

Maznev, A. A., Rogers, J. A., and Nelson, K. A. 1998. Optical heterodyne detection of laser-induced gratings. *Opt. Lett.* 23:1319–1321.

Mizubayashi, H., Yoshihara, Y., and Okuda, S. 1992. The elasticity measurements of aluminum-nm films. *Phys. Status Solidi A* 129:475–481.

Nelson, K. A. and Fayer, M. D. 1980. Laser induced phonons: A probe of intermolecular interactions in molecular solids. *J. Chem. Phys.* 72:5202–5218.

Neubrand, A. and Hess, P. 1992. Laser generation and detection of surface acoustic-waves: Elastic properties of surface layers. *J. Appl. Phys.* 71:227–238.

Nizzoli, F. and Sandercock, J. R. 1990. Surface Brillouin scattering from phonons. *In* Dynamical Properties of Solids Vol. 6 (G. K. Horton and A. A. Maradudin, eds.) pp. 281–335. Amsterdam, North-Holland.

Pharr, G. M. and Oliver, W. C. 1992. Measurement of thin film mechanical properties using nanoindentation. *M. R. S. Bulletin* July, 28–33.

Rogers, J. A. 1998. Complex acoustic waveforms excited with multiple picosecond transient gratings formed using specially designed phase-only beam-shaping optics. *J. Acoust. Soc. Amer.* 104:2807–2813.

Rogers, J. A. and Nelson, K. A. 1994. Study of Lamb acoustic waveguide modes in unsupported polyimide films using real-time impulsive stimulated thermal scattering. *J. Appl. Phys.* 75: 1534–1556.

Rogers, J. A. and Nelson, K. A. 1995. Photoacoustic determination of the residual stress and transverse isotropic elastic moduli in thin films of the polyimide PMDA/ODA. *IEEE Trans. UFFC* 42:555–566.

Rogers, J. A. and Nelson, K. A. 1996. A new photoacoustic/photothermal device for real-time materials evaluation: An automated means for performing transient grating experiments. *Physica B* 219–220:562–564.

Rogers, J. A. and Case, C. 1999. Acoustic waveguide properties of a thin film of nanoporous silica on silicon. *Appl. Phys. Lett.* 75: 865–867.

Rogers, J. A. and Bogart, G. R. 2000. Optical evaluation of the flexural rigidity and residual stress in thin membranes: picosecond transient grating measurements of the dispersion of the lowest order Lamb acoustic waveguide mode. *J. Mater. Res.* 16:217–225.

Rogers, J. A., Dhar, L., and Nelson, K. A. 1994a. Noncontact determination of transverse isotropic elastic moduli in polyimide thin films using a laser based ultrasonic method. *Appl. Phys. Lett.* 65:312–314.

Rogers, J. A., Yang, Y., and Nelson, K. A. 1994b. Elastic modulus and in-plane thermal diffusivity measurements in thin polyimide films using symmetry selective real-time impulsive stimulated thermal scattering. *Appl. Phys. A* 58:523–534.

Rogers, J. A., Fuchs M., Banet, M. J., Hanselman, J. B., Logan, R., and Nelson, K. A. 1997. Optical system for rapid materials characterization with the transient grating technique: Application to nondestructive evaluation of thin films used in microelectronics. *Appl. Phys. Lett.* 71:225–227.

Rogers, J. A., Maznev, A. A., Banet, M. J., and Nelson, K. A. 2000a. Optical generation and characterization of acoustic waves in thin films: fundamentals and applications. *Ann. Rev. Mat. Sci.* 30:117–157.

Rogers, J. A., Bogart, G. R., and Miller, R. E. 2000b. Noncontact quantitative spatial mapping of stress and flexural rigidity in thin membranes using a picosecond transient grating photoacoustic technique. *J. Acoust. Soc. Amer.* 109:547–553.

Shen, Q., Harata, A., and Sawada, T. 1996. Theory of transient reflecting grating in fluid/metallic thin film/substrate systems for thin film characterization and electrochemical investigation. *Jpn. J. Appl. Phys.* 35:2339–2349.

Thomsen, C., Grahn, H. T., Maris, J. H., and Tauc, J. 1986. Surface generation and detection of phonons by picosecond light pulses. *Phys. Rev. B* 34:4129–4138.

Viktorov, I. A. 1976. Rayleigh and Lamb Waves. Plenum, New York.

Vinci, R. P. and Vlassak, J. J. 1996. Mechanical behavior of thin films. *Ann. Rev. Mater. Sci.* 26:431–462.

Vlassak, J. J. and Nix, W. D. 1992. A new bulge test technique for the determination Youngs modulus and Poisson ratio of thin films. *J. Mater. Res.* 7:3242–3249.

Wright, O. B., and Kawashima, K. 1991. Coherent phonon detection from ultrafast surface vibrations. *Phys. Rev. Lett.* 69: 1668–1671.

KEY REFERENCES

Eichler et al., 1986. See above.

Provides a review of transient grating methods and how they can be used to measure nonacoustic properties.

Rogers et al., 2000a. See above.

Provides a review of ISTS methods with applications to acoustic evaluation of thin films, membranes and other type of microstructures.

APPENDIX: VENDOR INFORMATION

Currently there is a single vendor that sells complete ISTS instruments for metrology of thin metal films found in microelectronics:

Philips Analytical
Worldwide headquarters:
Lelyweg 1, 7602 EA Almelo,
The Netherlands
Tel.: +31 546 534 444
Fax: +31 546 534 598

In the U.S.:
Philips Analytical
12 Michigan Dr.
Natick, Mass. 01760
Tel.: (508) 647–1100
Fax: (508) 647–1111
http://www.analytical.philips.com

The names of the two free-standing tools are Impulse 300 and PQ Emerald.

There are also integrated metrology solutions for Cu interconnect process tools (e.g.: Paragon electroplating system from Semitool).

Semitool
655 West Reserve Dr.
Kalispell, Montana 59901
Tel.: (406) 752-2107
Fax: (406) 752-5522

Philips Analytical's Impulse 300 and PQ Emerald are fully automated all-optical metal thin film metrology systems for metal interconnect process control in semiconductor integrated circuit manufacturing. Impulse 300 is a versatile system for wafers up to 300 mm in size. PQ Emerald is a small-footprint (1.2 $\times$ 1.53−m) system for wafer size up to 200 mm. Both systems have measurement spot size 25 $\times$ 90 microns and pattern-recognition capabilities enabling measurements on product wafers. Measurements can be done on test pads as well as on high-density arrays of submicron structures, e.g., damascene line arrays.

Typical reproducibility: 1 to 2 Å (plasma vapor deposited Cu, Ta), 10 to 50 Å (thick electroplated Cu). Measurement time 1 sec /site, throughput up to 70 wafers per hr.

Key applications are copper interconnect process control, including seed deposition, electroplating and chemical-mechanical polishing.

Philips Analytical also offers integrated metrology solutions for Cu PVD, electroplating and polishing process tools.

JOHN A. ROGERS
Bell Laboratories, Lucent Technologies
Murray Hill, New Jersey
ALEX MAZNEV
Philips Analytical
Natick, Massachusetts
KEITH A. NELSON
Massachusetts Institute of Technology
Cambridge, Massachusetts

This page intentionally left blank

RESONANCE METHODS

INTRODUCTION

This chapter shows how nuclear and electron resonance spectroscopies can help solve problems in materials science. The concept of a probe, located centrally within a material, is common to all resonance spectroscopy techniques. For example, the nucleus serves as the probe in nuclear magnetic resonance (NMR), nuclear quadrupole resonance (NQR), and Mössbauer spectrometry, while a paramagnetic atom (i.e., one containing unpaired electrons) often serves as a probe in electron spin resonance (ESR). Sometimes the interest is in measuring the total numbers of probes within a material, or the concentration profiles of probe nuclei as in the case for basic NMR imaging. More typically, details of the measured energy spectra are of interest. The spectra provide the energies of photons that are absorbed by the probe, and these are affected by the electronic and/or nuclear environment in the neighborhood of the probe. It is often a challenge to relate this local electronic information to larger features of the structure of materials.

The units in this chapter describe how local electronic and magnetic structure can be studied with resonance techniques. The energy levels of the probe originate with fundamental electrostatic and magnetic interactions. These are the interaction of the nuclear charge with the local electron density, the electric quadrupole moment of the nucleus with the local electric field gradient, and the spin of the nucleus or spin of the electrons with the local magnetic field. For magnetic interactions in NMR, for example, the experimental spectra depend on how the energy of the probe differs for different orientations of the nuclear spin in the local magnetic field. In general these interactions involve some constant factors of the probe itself, such as its spin and gyromagnetic ratio. The energy level of the probe is the product of these constant factors and a local field. The resonance condition is such that a photon is absorbed, which causes the probe to undergo a transition to a state that differs in energy from its initial state by the energy of the absorbed photon. Since the parameters intrinsic to the probe are known for both states, the experimental spectrum provides the spectrum of the local field quantity characteristic of the material. Continuous wave methods excite probes with specific local fields. Pulse techniques excite all available probes, and differentiate them later by the phases of their precession in their local fields.

The energy differences of the probe levels are small. Radio transmitters provide the excitation photons in the case of NMR and NQR, and microwave generators provide the photons in ESR. The energy from these excitations is eventually converted into heat in the material. For Mössbauer spectrometry and other nuclear methods such as perturbed angular correlation spectroscopy (PACS), the photon, a γ ray, is provided by a radioisotope source. The photon energy is absorbed by exciting internal nuclear transitions. The subsequent decay occurs through the emission of ionizing radiation.

The ultimate energy resolution of these resonance spectroscopies is the energy precision, ΔE, that is set by the uncertainty principle

$$\Delta E \cong \frac{\hbar}{\Delta t} \tag{1}$$

where Δt is the lifetime of the excited state. [The Δt can also be imposed experimentally by exposing the sample to an additional radiofrequency pulse(s).] It is fortunate for the nuclear techniques of NMR, NQR, and Mössbauer spectrometry that Δt is relatively long, so the energy resolution, ΔE, is high. This makes nuclear spectra useful for studies of materials, because the weak hyperfine interactions between the nucleus and the surrounding electrons provide information about the electrons in the solid. The sampling of the local field is modified when the probe atom undergoes diffusive motions during the time Δt. Local probe methods provide unique information on the jumping frequencies of the probe atoms, and sometimes on the jump directions.

In spite of the high energy resolution of resonance spectroscopies, they are not typically known for their spatial resolution. Microscopies, as those for electrons or light, have not been developed for resonance spectroscopies because their photons cannot be focused. Especially for NMR and NQR, however, control over magnetic fields and the imposed pulse sequences can be used to make images of the concentrations of resonant probes within the material. These methods are under rapid development, driven largely by market forces outside materials science.

The units in this chapter enumerate some requirements for samples, but an obvious requirement is that the material must contain the probe itself. A set of allowed nuclei is listed in the units of NMR, NQR, and Mössbauer spectrometry, and the unit on ESR discusses unpaired electrons in materials. It is rare for the sample to have too high a concentration of probes; usually the difficulty is that the sample is too dilute in the nuclear or electron probe. Sensitivity is often a weakness of resonance spectroscopies, and measurements may require modest sample sizes and long data acquisition times to achieve statistically reliable spectra. An increased data acquisition rate is possible if the resonance frequencies are known in advance, and data need be acquired at only one point of the spectrum. Such specialized methods can be useful for routine characterizations of materials, but are usually not the main mode of operation for laboratory spectrometers.

Sometimes the probe is peripheral to the item of interest, which may be an adjacent atom, for example. Useful data can sometimes be obtained from measurements where the probe is a nearby spectator. As a rule of thumb, however, resonant probe experiments tend to be most informative when the probe is itself at the center of the

action, or at least is located at the atom of interest in the material.

Nuclear and electron resonance spectroscopies are specialized techniques. They are not methods that will be used by many groups in materials science. Especially when capital investment is high, as in the case of NMR imaging, it makes sense to contact an expert in the field to assess feasibility and reliability of measurements. Even when the capital investment in equipment is low, as for Mössbauer spectrometry, novices can be misled in several aspects of data interpretation. Collaborations are fairly commonplace for practitioners of resonance spectroscopies. The viewpoint of a material from a resonant probe is a small one, so resonance spectroscopists are accustomed to obtaining complementary experimental information when solving problems in materials science. Nevertheless, the viewpoint of a material from a resonant probe is unique, and can provide information unavailable by other means.

BRENT FULTZ

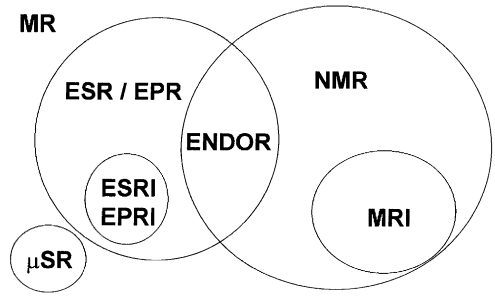


Figure 1. A Venn diagram of the field of magnetic resonance (MR). Abbreviations: ENDOR, electron nuclear double resonance; EPR, electron paramagnetic resonance; EPRI, electron paramagnetic resonance imaging; ESR, electron spin resonance; ESRI, electron spin resonance imaging; μSR, muon spin resonance; MRI, magnetic resonance imaging; NMR, nuclear magnetic resonance.

NUCLEAR MAGNETIC RESONANCE IMAGING

INTRODUCTION

Nuclear magnetic resonance imaging is a tomographic imaging technique that produces maps of the nuclear magnetic resonance signal in a sample. The signal from a volume element (voxel) in the sample is represented as an intensity of a picture element (pixel) in an image of the object. Nuclear magnetic resonance imaging has developed as an excellent noninvasive medical imaging technique, providing images of anatomy, pathology, circulation, and functionality that are unmatched by other medical imaging techniques. Nuclear magnetic resonance imaging is also an excellent nondestructive analytical tool for studying materials. Applications of nuclear magnetic resonance imaging in the field of materials science have also been developed, but to a lesser extent. The most probable reason why the development of materials applications of nuclear magnetic resonance imaging has lagged behind that of clinical applications is the cost-to-information ratio. The 1998 price of a clinical magnetic resonance imager is between one and two million U.S. dollars.

Before beginning a discussion of nuclear magnetic resonance imaging, it is useful to review the field of magnetic resonance (MR). Magnetic resonance is the general term that encompasses resonance processes associated with magnetic moments of matter. Please refer to Figure 1 throughout this introductory discussion. The two major branches of magnetic resonance are electron paramagnetic resonance (EPR), or electron spin resonance (ESR), as it is sometimes called (see ELECTRON PARAMAGNETIC RESONANCE SPECTROSCOPY), and nuclear magnetic resonance (NMR). A much smaller third component of magnetic resonance is muon spin resonance (μSR). EPR and ESR measure properties of the magnetic moment of the electrons or electron orbitals, while NMR is concerned with the magnetic moment of nuclei with unpaired protons and neutrons. Electron nuclear double resonance (ENDOR) examines the interactions between the nuclear and electron moments. All of the above techniques are used to measure properties of atoms and molecules in materials.

Within the fields of EPR and NMR are imaging disciplines. The term magnetic resonance imaging (MRI) should be used to describe imaging techniques based on both NMR and EPR or ESR. For historical reasons, MRI has been used for nuclear magnetic resonance imaging, the mapping of the spatial distribution of the NMR signals in matter. EPR imaging (EPRI) and ESR imaging (ESRI) have been used to describe imaging of the spatial distribution of electron and electron orbital magnetic moments in matter. This unit will concentrate on MRI, as previously defined, of materials.

MRI is used to image both large- and small-diameter objects. When imaging small objects, <1 cm in diameter, the imaging process is typically called NMR microscopy. When imaging larger objects, the procedure is generally called MRI. In reality, the imaging procedures are the same. The distinction arises from the type of instrument used, rather than the dimensions of the material. NMR microscopy is performed on an NMR spectrometer equipped with imaging hardware. Samples that do not fit into the microscope are imaged on a small-bore imager or clinical whole-body imager. No distinction will be made in this unit between MRI and NMR microscopy of materials.

MRI has been applied in many disciplines to study materials. MRI, as a nondestructive analytical technique, has very few competing techniques. X-ray and ultrasound imaging techniques come close to providing similar information in a nondestructive manner, but do not have the sensitivity to many of the fundamental properties of matter that MRI has. MRI may be used to measure distance, flow, diffusion coefficients, concentration, microscopic viscosity, partial pressure, pH (Metz et al., 1998), temperature (Aime et al., 1996; Zuo et al., 1996), magnetic susceptibility, and the dielectric constant. A few good MRI reviews are (Komoroski, 1993; Callaghan, 1991; Blumich et al., 1994; Hall et al., 1990; Gibbs and Hall, 1996; and Attard et al., 1991).

In food science, MRI has been used to study the consistency (Duce et al., 1990), baking (Duce and Ablett, 1995), cooking, curing (Guiheneuf et al., 1996; Guiheneuf, Gibbs, and Hall, 1997), storage, spoilage (Guiheneuf et al., 1997), hydration (Duce and Hall, 1995), and gelling (Holme and Hall, 1991) of food. MRI can be used to examine the ripening of fruits and vegetables, as well as damage by disease, insects, and bruising. MRI can determine the moisture and oil content, distribution, and transport in harvested grains and seeds (Duce et al., 1994). Lumber grading, or the mapping of knots, defects, and disease in lumber (Pearce et al., 1997), can be achieved by MRI. MRI has been used in botany to study disease and the flow and diffusion of water in plants.

In soil science, MRI can be used to map the water content and transport in soil. Soil (Amin et al., 1994, 1996), rock cores (Horsfield et al., 1990; Fordham et al., 1993a; Fordham et al., 1995), and oil well cores have been imaged to determine hydrocarbon and water content, as well as porosity, diffusion (Fordham et al., 1994; Horsfield, 1994), capillary migration (Carpenter et al., 1993), and flow characteristics.

In civil engineering, MRI has been used to study the curing and porosity of cement as well as water transport in concrete (Link et al., 1994) and other building materials (Pel et al., 1996). MRI can determine the effectiveness of paints and sealers in keeping out moisture.

In chemistry, MRI has been used to image chemical waves associated with certain chemical reactions (Armstrong et al., 1992), hydrogen adsorption into palladium metal (McFarland et al., 1993), processes in chromatographic columns (Ilg et al., 1992), as well as in reactions and crystals (Komoroski, 1993). MRI has been used to characterize ceramics (Wang et al., 1995)—flaws (Karunanithy and Mooibrook, 1989), cracks, voids (Wallner and Ritchey, 1993), and binder distribution (Wang et al., 1993) in ceramics have been imaged. MRI has been used to noninvasively study the homogeneity of solid rocket propellants (Maas et al., 1997; Sinton et al., 1991).

Polymers have been the subject of many MRI investigations (Blumich and Blumler, 1993). MRI has been used to study the uptake of water into the polymer coating used on silicon chips (Hafner and Kuhn, 1994) and polyurethane dispersion coating (Nieminen and Koenig, 1990). MRI has been used to study curing (Jackson, 1992), vulcanization (Mori and Koenig, 1995), solvent ingress (Webb and Hall, 1991), inhomogeneities (Webb et al., 1989), diffusion (Webb and Hall, 1990), multicomponent diffusion (Grinsted and Koenig, 1992), cyclic sorption-desorption (Grinsted et al., 1992), aging (Knorgen et al., 1997), filler inhomogeneities (Sarkar and Komoroski, 1992), adhesion (Nieminen and Koenig, 1989), and water uptake (Fyfe et al., 1992) in polymers.

MRI is also used to image flow (Amin et al., 1997; Seymour and Callaghan, 1997), diffusion (Callaghan and Xia, 1991), and drainage (Fordham et al., 1993b) in materials, and hence gain a better understanding of the internal microscopic structure of the materials.

There have been two ways in which MRI has been used to study materials. One approach images the material, and the other images the absence of material. The first approach is to directly image the NMR signal from a signal-bearing substance in the sample. The second is to introduce a signal-bearing substance into voids within the material and image the signal from the signal-bearing substance. The latter procedure is good for porous samples that do not have a measurable NMR signal of their own. The latter approach is more common because it can be performed on clinical magnetic resonance imagers with routinely available clinical imaging sequences. The former approach typically requires either faster imaging sequences, or different hardware and imaging sequences capable of looking at a solid-state NMR signal.

PRINCIPLES OF THE METHOD

Nuclear magnetic resonance imaging is based on the principles of nuclear magnetic resonance. The choice of imaging parameters necessary to produce a high quality MR image requires knowledge of many of the microscopic properties of the system of spins being imaged. It is therefore necessary to understand the principles of NMR before those of MRI can be introduced.

Theory of NMR

NMR is based on a property of the nucleus of an atom called spin. The material presented in this unit pertains to spin 1/2 nuclei. Spin can be thought of as a simple magnetic moment. When a spin 1/2 nucleus is placed in an external magnetic field, the magnetic moment can take one of two possible orientations, one low-energy orientation aligned with the field and one high-energy orientation opposing the field. A photon with an amount of energy equal to the energy difference between the two orientations or states will cause a transition between the states. The greater the magnetic field, the greater the energy difference, and hence the greater the frequency of the absorbed photon. The relationship between the applied magnetic field B and the frequency of the absorbed photon ν is linear.

$$\nu = \gamma B \tag{1}$$

The proportionality constant γ is called the gyromagnetic ratio. The gyromagnetic ratio is a function of the magnitude of the nuclear magnetic moment. Therefore, each isotope with a net nuclear spin possesses a unique γ. The γ value of some of the more commonly imaged nuclei are listed in Table 1. In NMR, B fields are typically 1 to 10 tesla (T); therefore ν is in the MHz range.

Table 1. Gyromagnetic Ratio of Some Commonly Imaged Nuclei

Nucleus	Gyromagnetic Ratio (MHz/T)
^{1}H	42.58
^{13}C	10.71
^{17}O	5.77
^{19}F	40.05
^{23}Na	11.26
^{31}P	17.24

NMR has the ability to distinguish between nuclei in different chemical environments. For example, the resonance frequency of hydrogen in water is different than that for the hydrogen nuclei in benzene. This difference is called the chemical shift of the nucleus, and is usually reported in ppm relative to a reference. In hydrogen NMR this reference is tetramethylsilane (TMS). The chemical shift for a freely tumbling nucleus i in a nonviscous solution is

$$\delta_i = \frac{(\nu_{TMS} - \nu_i)}{\nu_{TMS}} \times 10^6 \qquad (2)$$

In solids and some viscous liquids, δ may be anisotropic due to the fixed distribution of orientations of the nuclei in the molecule.

Molecules with interacting spins will display spin-spin coupling, or a splitting of the energy levels due to interactions between the nuclei. The spin-spin interaction is orientation-dependent. Once again, in nonviscous liquids, this orientation-dependence is averaged out and seen as one average interaction. In liquids, multiple resonance lines may be observed in an NMR spectrum, depending on the magnitude of the spin-spin coupling. In solids and viscous liquids, the distribution of orientations results in broad spectral absorption peaks. Imaging solids will require different and less routine techniques than are used in imaging nonviscous liquids.

The remainder of this description of NMR will adopt a more macroscopic perspective of the spin system in which a static magnetic field, B_0, is applied along the z axis. Groups of nuclei experiencing the exact same B_0 are called spin packets. When placed in a B_0 field, spin packets precess about the direction of B_0 just as a spinning top precesses about the direction of the gravitational field on earth. The precessional frequency, also called the Larmor frequency, ω, is equal to $2\pi\nu$. Adapting the conventional magnetic resonance coordinate system, the direction of the precession is clockwise about B_0, and the symbol ν_0 is reserved for spin packets experiencing exactly B_0 (Fig. 2). It is often helpful in NMR and MRI to adopt a rotating frame of reference to describe the motion of magnetization vectors. This frame of reference rotates about the z axis at ν_0. The coordinates in the rotating frame of reference are referred to as z, x', and y'. An excellent animated depiction of the motion of a magnetization vector in a magnetic field can be found in the Web-based hypertext book, *The Basics of MRI*, by J. P. Hornak (*http://www.cis.rit.edu/htbooks/mri/*).

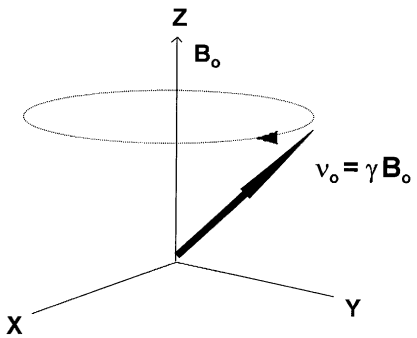


Figure 2. Clockwise precession of the net magnetization vector in an *xyz* coordinate system.

An NMR sample contains millions of spin packets, each with a slightly different Larmor frequency. The magnetization vectors from all these spin packets form a cone of magnetization around the z axis. At equilibrium, the net magnetization vector **M** from all the spins in a sample lies in the center of the cone along the z axis. Therefore, the longitudinal magnetization M_z equals M and the transverse magnetization M_{xy} equals zero at equilibrium. Net magnetization perturbed from its equilibrium position will want to return to its equilibrium position. This process is called spin relaxation.

The return of the z component of magnetization to its equilibrium value is called spin-lattice relaxation. The time constant that describes the exponential rate at which M_z returns to its equilibrium value M_{z0} is called the spin-lattice relaxation time, T_1. Spin-lattice relaxation is caused by time-varying magnetic fields at the Larmor frequency. These variations in the magnetic field at the Larmor frequency cause transitions between the spin states and hence change M_z. Time-varying fields are caused by the random rotational and translational motions of the molecules in the sample possessing nuclei with magnetic moments. The frequency distribution of random motions in a liquid varies with temperature and viscosity. In general, relaxation times tend to get longer as B_0 increases. T_1 lengthens because there are fewer relaxation-causing frequency components present in the random motions of the molecules as ν increases.

At equilibrium, the transverse magnetization, M_{xy}, equals zero. A net magnetization vector rotated off of the z axis creates transverse magnetization. This transverse magnetization decays exponentially with a time constant called the spin-spin relaxation time T_2. Spin-spin relaxation is caused by fluctuating magnetic fields that perturb the energy levels of the spin states and dephase the transverse magnetization. T_2 is inversely proportional to the number of molecular motions less than and equal to the Larmor frequency. Specialists in NMR break down T_2 further into a pure T_2 due to molecular interactions and one due to inhomogeneities in the B_o field. The overall T_2^* is referred to as "T_2 star."

$$\frac{1}{T_2^*} = \frac{1}{T_{2\,\text{molec}}} + \frac{1}{T_{2\,\text{inhomogeneous}}} \qquad (3)$$

In pulsed NMR and MRI, radiofrequency (RF) energy is put into a spin system by sending RF into a resonant LC circuit, the inductor of which is placed around the sample. The inductor must be oriented with respect to the B_0 magnetic field so that the oscillating RF field created by the RF flowing through the inductor is perpendicular to B_0. The RF magnetic field is called the B_1 magnetic field. When the RF inductor, or coil as it more often called, is placed around the x axis, the B_1 field will oscillate back and forth along the x axis.

In pulsed NMR spectroscopy, it is the B_1 field that is pulsed. Turning on a B_1 field for a period of time, τ, will

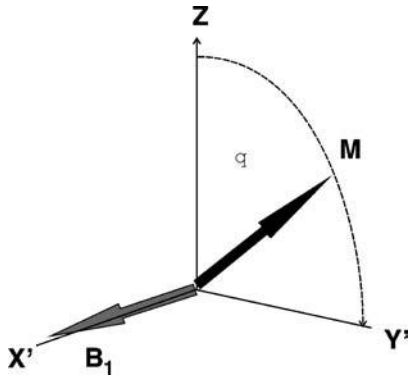


Figure 3. Rotation of net magnetization vector M by B_1 field in a rotating frame of reference.

cause the net magnetization vector to precess in ever widening circles around the z axis. Eventually the vector will reach the xy plane. If B_1 is left on longer, the net magnetization vector will reach the negative z axis. In the rotating frame of reference this vector appears to rotate away from the z axis, as depicted in Figure 3. The rotation angle θ, which is measured clockwise about the direction of B_1 in radians, is proportional to γ, B_1, and τ

$$\theta = 2\pi\gamma B_1 \tau \tag{4}$$

Any transverse magnetization, M_{xy}, will precess about the direction of B_0. An NMR signal is generated from transverse magnetization rotating about the z axis. This magnetization will induce a current in a coil of wire placed around the x or y axis. As long as there is transverse magnetization that is changing with respect to time, there will be an induced current in the coil. For a group of nuclei with one identical chemical shift, the signal will be an exponentially decaying sine wave. The sine wave decays with time constant T_2^*. It is predominantly the inhomogeneities in B_0 that cause the spin packets to dephase.

Net magnetization, which has been rotated away from its equilibrium position along the z axis by exactly 180°, will not create transverse magnetization and hence not give a signal. The time-domain signal from a net magnetization vector in the xy plane is called a free induction decay (FID). This time-domain signal must be converted to a frequency-domain spectrum to be interpreted for chemical information. The conversion is performed using a Fourier transform. The hardware in most NMR spectrometers and magnetic resonance imagers detects both M_x and M_y simultaneously. This detection scheme is called quadrature detection. These two signals are equivalent to the real and imaginary signals; therefore the input to the Fourier transform will be complex. Sampling theory tells us that one need only digitize the FID at a frequency of f complex points per second in order to obtain a spectrum of frequency width f, in Hz.

In pulsed Fourier transform NMR spectroscopy, short bursts of RF energy are applied to a spin system to induce a particular signal from the spins within a sample. A pulse sequence is a description of the types of RF pulses used and the response of the magnetization to the pulses. The simplest and most widely used pulse sequence for routine NMR spectroscopy is the 90-FID pulse sequence. As the name implies, the pulse sequence is a 90° pulse followed by the acquisition of the time-domain signal called an FID. The net magnetization vector, which at equilibrium is along the positive z axis, is rotated by 90° down into the xy plane. The rotation is accomplished by choosing an RF pulse width and amplitude so that the rotation equation for a 90° pulse is satisfied. At this time the net magnetization begins to precess about the direction of the applied magnetic field B_0. Assuming that the spin-spin relaxation time is much shorter than the spin-lattice relaxation time ($T_2 \ll T_1$), the net magnetization vector begins to dephase as the vectors from the individual spin packets in the sample precess at their own Larmor frequencies. Eventually, M_{xy} will equal zero and the net magnetization will return to its equilibrium value along z. The signal that is detected by the spectrometer is the decay of the transverse magnetization as a function of time. The FID or time-domain signal is Fourier transformed to yield the frequency domain NMR spectrum.

Theory of MRI

The basis of MRI is Equation 1, which states that the resonance frequency of a nucleus is proportional to the magnetic field it is experiencing. The application of a spatially varying magnetic field across a sample will cause the nuclei within the sample to resonate at a frequency related to their position.

For example, assume that a one-dimensional linear magnetic field gradient G_z is set up in the B_0 field along the z axis. The resonant frequency, ν, will be equal to

$$\nu = \gamma(B_0 + zG_z) \tag{5}$$

The origin of the xyz coordinate system is taken to be the point in the magnet where the field is exactly equal to B_0 and spins resonate at ν_0. This point is referred to as the isocenter of the magnet. Equation 6 explains how a simple one-dimensional imaging experiment can be performed. The sample to be imaged is placed in a magnetic field B_0. A 90° pulse of RF energy is applied to rotate magnetization into the xy plane. A one-dimensional linear magnetic field gradient G_z is turned on after the RF pulse and the FID is immediately recorded. The Fourier transform of the FID yields a frequency spectrum that can be converted to a spatial, z, spectrum as a function

$$z = \frac{(\nu - \nu_0)}{\gamma G_z} \tag{6}$$

This simple concept of a one-dimensional image can be expanded to a two-dimensional image by employing the concept of back-projection imaging similar to that used in computed tomography (CT) imaging. (Hornak, 1995). If a series of one-dimensional images, or projections of the signal in a sample, are recorded for linear one-dimensional magnetic field gradients applied along several different trajectories in a plane, the spectra can be transformed into a two-dimensional image using an inverse

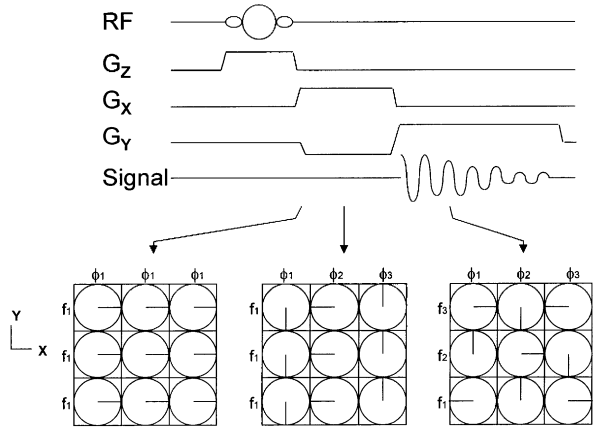


Figure 4. Timing diagram for a 90-FID imaging sequence and the behavior of nine spin vectors in the imaged plane at three times in the imaging sequence.

radon transform (Sanz, 1988) or back projection algorithm (Lauterbur, 1973). This procedure is often used in materials imaging where short T_2 values prevent the use of the more popular Fourier-based method (Kumar and Ernst, 1975; Smith, 1985). A Fourier-based imaging technique collects data from the k space of the imaged object. This k space is the two-dimensional Fourier transform of the image. Figure 4 depicts a timing diagram for a Fourier imaging sequence using simple 90-FID sequence. The timing diagram describes the application of RF and three magnetic-field gradients called the slice selection (G_s), phase encoding (G_p), and frequency encoding (G_f) gradients. The first step in the Fourier imaging procedure is slice selection of those spins in the object for which the image is to be generated.

Slice selection is accomplished by the application of a magnetic field gradient at the same time as the RF pulse is applied. An RF pulse, with frequency width $\Delta\nu$ centered at ν, will excite, when applied in conjunction with a field gradient G_z, spins centered at $z = (\nu - \nu_0)/\gamma G_z$ with a spread of spins at z values $\Delta z = \Delta\nu/\gamma G_z$. Spins experiencing a magnetic field strength not satisfying the resonance condition will not be rotated by the RF pulses, and hence slice selection is accomplished. The image slice thickness (Thk) is given by Δz. For a clean slice, i.e., where all spins along the slice thickness are rotated by the prescribed rotations, the frequency content of the pulse must be equal to a rectangular-shaped function. Therefore, the RF pulse must be shaped as a sinc function ($\sin x/x$) in the time domain.

The next step in the Fourier imaging procedure is to encode some property of the spins as to the location in the selected plane. One could easily encode spins as to their x position by applying a gradient G_x after the RF pulse and during the acquisition of the FID. The difficulty is in encoding the spins with information as to their y location. This is accomplished by encoding the phase of the precessing spin packets with y position (Fig. 4). Phase encoding is accomplished by turning on a gradient in the y direction immediately after the slice-selection gradient is turned off and before the frequency-encoding gradient is turned on. The spins in the excited plane now precess at a frequency dependent on their y position. After a period of time, the gradient is turned off and the spins have acquired a phase equal to

$$\phi = 2\pi\gamma\tau y G_y \tag{7}$$

Figure 4 describes, for nine magnetization vectors, the effect of the application of a phase-encoding gradient, G_y, and a frequency-encoding gradient, G_x. The phase-encoding gradient assigns each y position a unique phase. The frequency-encoding gradient assigns each x position a unique frequency. If one had the capability of independently assessing the phase and frequency of a spin packet, one could assign its position in the xy plane. Unfortunately, this cannot be accomplished with a single pulse and signal. The amplitude of the phase-encoding gradient must be varied so a 2π radian phase variation between the isocenter and the first resolvable point in the y direction can be achieved, as well as a $256\,\pi$ radian variation from center to edge of the imaged space for 256-pixel resolution in the phase-encoding direction. The result is to traverse, line by line, the k space of the image. The negative lobe on the frequency-encoding gradient (Fig. 4.), which was not previously described, shifts the center of k space to the center of the signal acquisition window.

The B_0 field is created by a large-diameter, solenoidal-shaped, superconducting magnet. The gradient fields are created by room temperature gradient coils located within the bore of the magnet. These coils are driven by high-current audiofrequency amplifiers. The B_1 field is introduced into the sample by means of a large LC circuit that surrounds the object to be imaged. The same or a separate LC circuit is used to detect the signals from the precessing spins in the body.

The field of view (FOV) is dependent on the quadrature sampling rate, R_s during the G_f and on the magnitude of G_f.

$$\mathrm{FOV} = R_s/G_f \tag{8}$$

The two-dimensional k space data set is Fourier transformed and the magnitude image generated from the real and imaginary outputs of the Fourier transform.

The number of pixels across the field of view, N, is typically 256 in both directions of the image. Assuming a 20-cm FOV and 3-mm slice thickness, the in-plane resolution is $\sim$0.8 mm. The volume of a voxel is therefore $\sim$2 mm^2. When imaging features occupying fractions of a voxel, each voxel comprises more than one substance. As a consequence, the NMR signal from a voxel is a sum of the NMR signals from the substances found in the voxel. Variations in the signal from a voxel, due to the relative amounts of the components found in a voxel, is referred to as a partial volume effect.

A magnetic resonance image can be thought of as a convolution of a two-dimensional NMR spectrum of the spin-bearing substance with the spin-concentration map from the imaged object. This is better visualized in one dimension, where x is the imaged dimension. If $f_i(\nu G_x/\gamma)$ is the NMR spectrum of the spin type converted to distance units, and $g_i(x)$ is the distribution of the spins,

the one-dimensional image is the convolution of these two functions. Since there may be more than one component in a voxel, the one-dimensional image, $h(x)$, becomes

$$h(x) = \sum_i g_i(x) \otimes f_i(\nu G_x/\gamma) \tag{9}$$

Defining Γ_i as the full line width at half height of component i in Hz, $\Gamma_i G_x/\gamma$ for the largest Γ_i must be less than FOV/N for optimum resolution in $h(x)$.

The most commonly imaged spin-bearing nucleus is hydrogen. The two most commonly imaged molecules containing hydrogen are hydrocarbons and water. These hydrogens yield one signal in the image, as chemical shift and spin-spin splitting information is generally not utilized. Occasionally the different chemical shifts for water and hydrocarbon hydrogens can lead to an artifact in an image called a chemical-shift artifact. Other hydrogens associated with many samples have very short T_2 values and do not contribute directly to the signal.

In the example of Figure 4, the slice-selection gradient was applied along the z axis and the phase and frequency-encoding gradients along the y and x axes, respectively. In practice, the gradients can be applied along any three orthogonal directions, with the only restrictions being that the slice selection gradient be perpendicular to the imaged plane.

The most routinely used imaging sequence is the spin-echo. Its popularity is attributable to its ability to produce images that display variations in T_1, T_2, and spin concentration of samples. This sequence consists of 90° and 180° RF pulses repeated every TR seconds (Fig. 5). These pulses are applied in conjunction with the slice selection gradients.

The phase-encoding gradient is applied between the 90° and 180° pulses. The frequency-encoding gradient is turned on during the acquisition of the signal. The signal is referred to as an echo because it comes about from the refocusing of the transverse magnetization at time TE after the application of the 90° pulse. The signal, when TE $\ll$ TR, from a voxel will be equal to a sum over all the different types of spins, i, in the voxel.

$$S(\mathrm{TE}, \mathrm{TR}) = k \sum_i \rho_i (1 - e^{-\mathrm{TR}/T_{1i}}) e^{-\mathrm{TE}/T_{2i}} \tag{10}$$

Spin density, ρ, is the number of spins per voxel, and k is a proportionality constant.

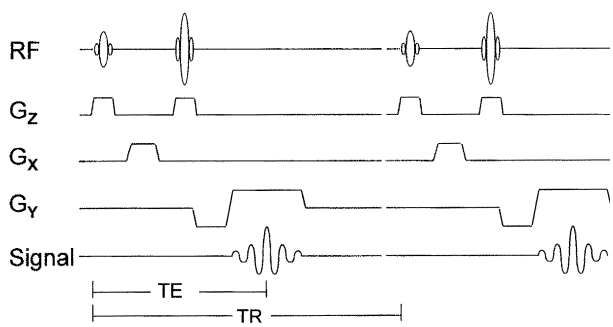


Figure 5. Timing diagram for a spin-echo imaging sequence.

PRACTICAL ASPECTS OF THE METHOD

The signal in a magnetic resonance image is a function of T_1, T_2, T_2^*, ρ, the diffusion coefficient (D), the velocity (V), and chemical shift (δ) of the spins. The various imaging sequences are designed to produce a signal that is proportional to one of these parameters. A weighted image is one in which the signal intensity is dependent on one of these properties. For example, a T_1-weighted image is one in which the image intensity displays differences in T_1 of the sample.

All images are generated with the aid of an imaging pulse sequence similar to those presented in Figures 4 and 5. There are hundreds of imaging sequences, but only a fraction of these are routinely used. A few are mentioned here, and others can be found in the literature. Because the implementation of these is imager-dependent, readers are encouraged to refer to their instrument manuals for specific details. When searching the scientific literature for information on pulse sequences, the reader is reminded that much MRI development work has been published by scientists in the clinical literature.

There are two goals in MRI of materials. One is to visualize structure, texture, or morphology in samples. The second is to visualize spatial or temporal variations in some property of the sample. In the former, the MRI researcher needs to maximize the contrast-to-noise ratio between features in the sample. The imaging parameters are chosen to maximize the difference in signal intensity of two adjacent regions in the image of the sample. For example, Figures 6 and 7 indicate how contrast can be achieved based on the choice of TE and differences in T_2, and the choice of TR and variations in T_1. If the visualization of spatial or temporal variations is the goal, the absolute intensity of a pixel is important and the choice of imaging parameters is generally chosen to maximize the signal-to-noise ratio.

Before any of the following imaging sequences may be implemented, T_1 and T_2 of the sample at the field strength of the imager must be known. These values are needed to determine the optimum values of TR, TE, FOV, Thk, and so on.

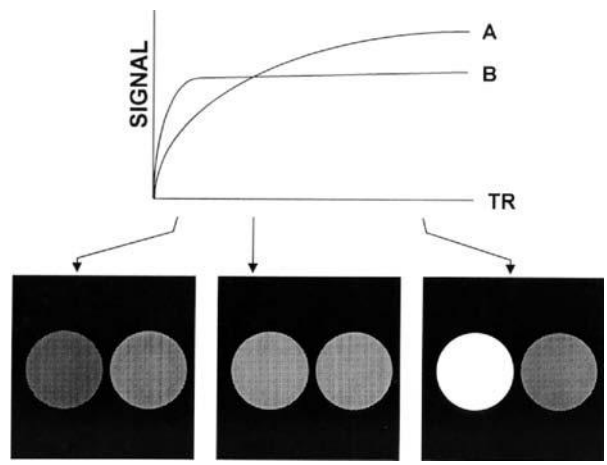


Figure 6. Signal of two substances, A and B, as a function of TR, and their corresponding simulated images at the indicated TR values.

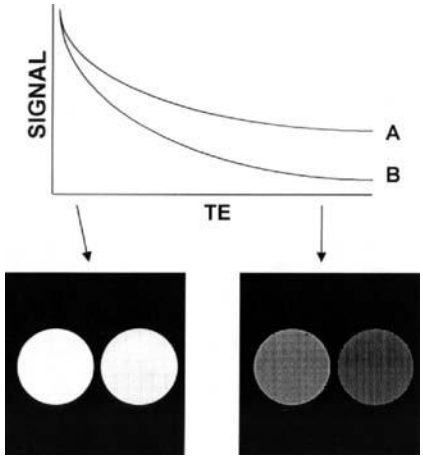


Figure 7. Signal of two substances, A and B, as a function of TE, and their corresponding simulated images at the indicated TE values.

Spin-Echo Sequence

The spin-echo sequence is one of the most common imaging sequences. This sequence was described earlier (see discussion of Theory of MRI), and its timing diagram can be found in Figure 5. This sequence is used to produce images whose signal intensity is a function of the T_1, T_2, and ρ of the sample as described by Equation 10.

A spin-echo sequence will produce a T_1 weighted image when TR $< T_1$ and TE $\ll T_2$. A T_2-weighted image can be produced by a spin-echo sequence when TR $\gg T_1$ and TE $> T_2$. A spin-echo sequence will produce a ρ-weighted image when TR $\gg T_1$ and TE $\ll T_2$. Due to time constraints, these conditions are rarely precisely met, so weighted images will have some dependence on the remaining two properties.

Gradient Recalled Echo Sequence

A gradient recalled echo sequence is essentially the 90-FID sequence described earlier in the theory section. The signal from the gradient-recalled echo sequence is given by the following equation, where θ is the rotation angle of the RF pulse

$$S(\mathrm{TE},\mathrm{TR}) = k\sum_i \rho_i \frac{(1-e^{\mathrm{TR}/T_{1i}})\sin\theta\, e^{-\mathrm{TE}/T^*_{2i}}}{(1-\cos\theta\, e^{-\mathrm{TR}/T_{1i}})} \qquad (11)$$

The repetition time, TR, has the same definition as in the spin-echo sequence. A maximum in the signal is produced because the frequency-encoding gradient is first turned on negative so as to position the start of the signal at the edge of k space. Therefore, the echo time, TE, is the time between the RF pulse and the maximum signal. The signal has a T_2^* dependence instead of a T_2 dependence. The rotation angle θ is often set to a value less than 90° so that equilibrium is reached more quickly in the period TR. T_1-, T_2^*-, and ρ-weighted images can be made with a gradient-recalled sequence.

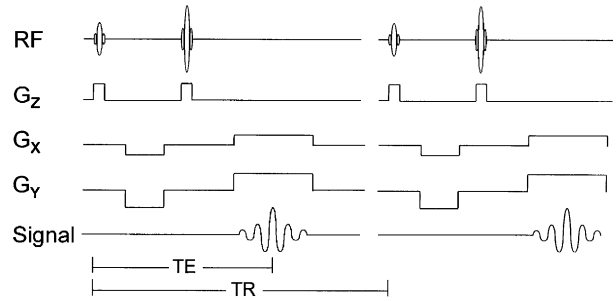


Figure 8. Timing diagram for a back-projection imaging sequence.

Back-Projection Imaging Sequence

The back-projection imaging sequence is useful when T_2 values are shorter than 10 ms. The principle is based on the acquisition of several one-dimensional images of the sample (Lauterbur, 1973). The direction of the one-dimensional image is varied by 360° around the imaged object. Figure 8 depicts the timing diagram for a spin-echo back-projection sequence. A slice selection gradient is applied in conjunction with the RF pulses so as to select the desired image plane. In this example an xy plane is imaged.

A readout gradient, G_R, is applied during the acquisition of the signal such that the angle (α) subtended by its direction relative to the x-axis varies between 1° and 360° as given by the following equations

$$G_x = G_R\cos(\alpha) \qquad (12)$$
$$G_y = G_R\sin(\alpha) \qquad (13)$$

Each of these one-dimensional images represents the projection of the NMR signal across the image perpendicular to G_R. G_R is also applied between the 90° and 180° pulses so as to position the echo properly in k space. This set of projection images is back-projected through space using an inverse radon transform to produce an image (Sanz et al., 1988). The advantage of this sequence is that TE can be made less than the corresponding TE in a conventional spin-echo sequence. Therefore samples with shorter T_2 values can be imaged. Since this back-projection sequence utilizes a spin-echo pulse sequence, S(TR,TE) is given by Equation 10. Since

$$\Gamma_i = \frac{1}{\pi T_{2i}} \qquad (14)$$

and being mindful of Equation 9, gradient strengths must be greater for samples with shorter T_2 values.

Echo-Planar Imaging Sequence

Echo-planar imaging is a rapid magnetic resonance imaging technique, capable of producing images at video rates. The technique records an entire image in a single TR period. Echo-planar imaging therefore measures all lines of k space in a single TR period. A timing diagram for an echo-planar imaging sequence is depicted in Figure 9.

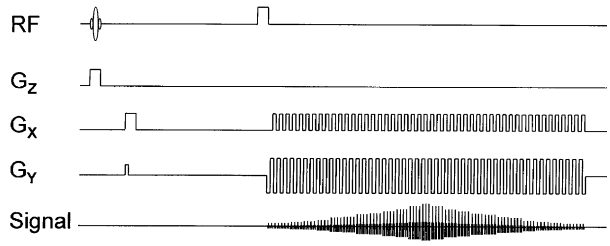


Figure 9. Timing diagram for an echo-planar imaging sequence.

A 90° slice-selective RF pulse is applied in conjunction with a slice-selection gradient. Initial phase-encoding gradient and frequency-encoding gradient pulses position the spins at the corner of k space. A 180° pulse follows, which is not slice selective. Next, the phase- and frequency-encoding gradients are cycled so as to raster through k space. This is equivalent to putting 256 phase- and frequency-encoding gradients in the usual period when the echo is recorded.

The rate at which k space is traversed is so rapid that it is possible to obtain 30 images per second. This implies that echo-planar imaging is suitable for studying dynamic processes in materials such as reaction kinetics. Echo-planar images require large and faster-than-normal gradients. As a consequence, echo-planar images are prone to severe magnetic susceptibility artifacts.

Diffusion Imaging Sequence

Diffusion imaging is used to produce images whose intensity is related to the diffusion coefficient, D, of the spins being imaged. The pulsed-gradient spin-echo imaging sequence is used to produce diffusion-weighted images. The timing diagram for this sequence is presented in Figure 10. The signal from a pulsed-gradient spin-echo imaging sequence, S, as a function of TR, TE, diffusion encoding gradient (G_j) in the j direction, width of the diffusion gradient (δ), and separation of the diffusion gradients (Δ) is given by Equation 15.

$$S(\mathrm{TR}, \mathrm{TE}, G_j, \delta, \Delta) = k \sum_i \rho_i (1 - e^{-\mathrm{TR}/T_{1i}}) \times e^{-\mathrm{TE}/T_{2i}^{-(G_j \gamma \delta)^2 D_{j,i}(\Delta - \delta/3)}} \quad (15)$$

The summation is over the i components in the voxel. The function of the diffusion gradient pulses is to dephase magnetization from spins that have diffused to a new location in the period Δ. These pulses have no effect on stationary spins, as can be seen by Equation 15 reducing to Equation 10 when the diffusion coefficient is zero.

An image map of the diffusion in one dimension (j) may be calculated from diffusion images recorded at different G_j values. Data of $S(G_\mathrm{j})$ as a function of G_j^2 is fit on a pixel-by-pixel basis to obtain D_j for each pixel in the image space. Diffusion coefficient images can be measured for $10^{-9} < D_\mathrm{j} < 10^{-4}$ cm^2/s.

Flow Imaging Sequence

A flow imaging sequence is similar to that used for diffusion imaging, except that the motion-encoding gradients are tuned to pick up very fast motion. Figure 11 depicts the timing diagram for a phase-contrast imaging sequence implemented with a spin-echo sequence for imaging flow in the z direction. The basis of the phase-contrast sequence is a bipolar magnetic field gradient (G_BP) pulse. A bipolar gradient pulse is one in which the gradient is turned on and off in one direction for a period of time, and then turned on and off in the opposite direction for an equivalent amount of time. A positive bipolar gradient pulse has the positive lobe first and a negative bipolar gradient pulse has the negative lobe first. The area under the first lobe of the gradient pulse must equal that of the second. A bipolar gradient pulse has no net effect on stationary spins. The bipolar gradient pulse will affect spins that have a velocity component in the direction of the gradient.

If a bipolar gradient pulse is placed in an imaging sequence, it will not affect the image since all we have done is impart a phase shift to the moving spins. Since an image is a magnitude representation of the transverse magnetization, there is no effect. However, if two imaging sequences are performed in which the first has a positive bipolar gradient pulse and the second a negative bipolar gradient pulse, and the raw data from the two are subtracted, the signals from the stationary spins will cancel and the flowing spins will add. The vectors from the stationary spins cancel and those from the moving spins have a net magnitude. The net effect is an image of the flowing spins.

The direction of the bipolar gradient yields signal only from those spins with a flow component in that direction. The timing diagram in Figure 11 will measure flow in the z

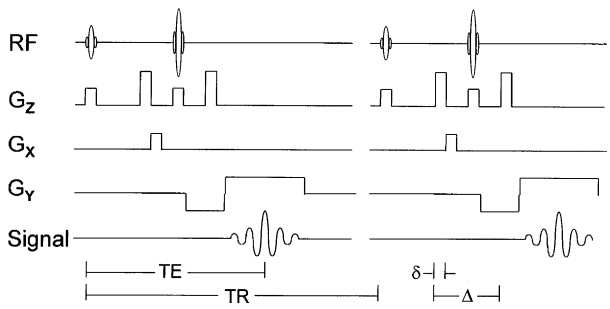


Figure 10. Timing diagram for a diffusion imaging sequence.

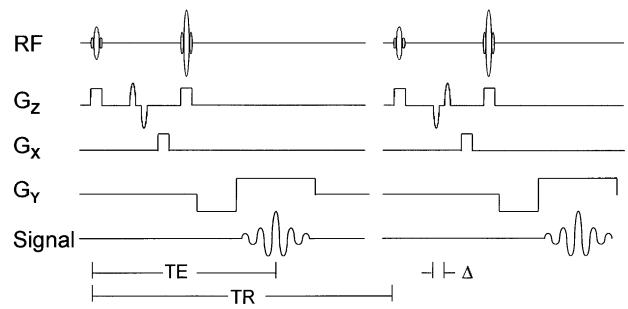


Figure 11. Timing diagram for a flow imaging sequence.

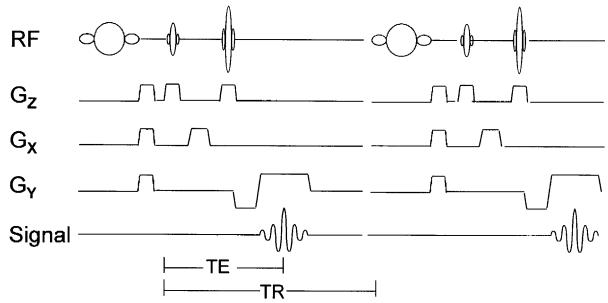


Figure 12. Timing diagram for a chemical shift imaging sequence that saturates the spins at a specific chemical shift.

direction because the bipolar gradient pulses are applied by G_z. Images of the flow rate between 5 and 400 cm/s may be recorded.

Chemical Shift Imaging Sequence

A chemical shift imaging (CSI) sequence is one that can map the spatial distribution of a narrow band of chemical shift components found in a sample. There are several possible CSI sequences. The one presented here consists of the modified spin-echo sequence in Figure 12. This sequence has the usual slice-selective 90° and 180° pulses, and gradients. It has an additional one or more saturation pulses before the normal spin-echo sequence designed to saturate the spins at specific chemical shift values. The frequency of the 90° and 180° pulses are tuned to the resonance frequency of the desired chemical shift component. The frequency of the saturation pulse is tuned to the resonance frequency of the chemical shift component that is to be eliminated. The frequency, bandwidth, and amplitude of the saturation pulse must be set to effectively saturate the undesired chemical shift components. This technique works best when the T_1 of the saturated component is long, thus preventing the generation of transverse magnetization before the application of the spin-echo sequence. Magnetic field gradients are occasionally applied between the saturation pulse and the spin echo sequence to dephase any residual transverse magnetization after the saturation pulse. Images of chemical shift differences as low as 1 ppm are attainable.

Three-Dimensional Imaging

With the development of fast desktop computers with high-resolution video or holographic monitors, three-dimensional images are readily viewed and manipulated. MRI can readily produce three-dimensional images by acquiring contiguous tomographic slices through an object. To minimize the time necessary to acquire this amount of data, a fast imaging sequence, such as a gradient-recalled echo or echo-planar is typically used. Alternatively, a multislice sequence, which utilizes the dead time between repetitions to acquire data from other slices, can be utilized.

Solid-State Imaging

The NMR spectra of solids typically consist of very broad absorption lines due to anisotropic chemical shifts and spin-spin (or dipole-dipole) interactions. Recalling the limitations imposed by Equation 14 and Equation 9, this would require extremely large and fast gradients, as well as fast digitizers. Solid-state magic-angle spinning NMR techniques can be used to decrease the line width (Fukushima, 1981). This task is accomplished by spinning the sample about an axis oriented at 54.7° to the applied B_0 field. This angle is referred to as the magic angle because at this angle the dipole interaction contribution to the line width equals zero. Unfortunately, two-dimensional imaging becomes difficult due to the need to spin the sample at the magic angle. A one-dimensional image may be recorded of a sample spinning at the magic angle by applying a frequency-encoding gradient along the direction of the magic-angle spinning axis.

The dipolar interaction may also be minimized using line-narrowing pulse sequences. The specifics of these sequences are beyond the scope of this unit. The reader is directed to the relevant literature (Chingas et al., 1986; McDonald et al., 1987; Rommel et al., 1990) and instrument manuals for specifics.

Instrument Setup and Tuning

Certain aspects of the imager should be tuned for each sample. These aspects include the sample probe or imaging coil, the power required to produce a 90° pulse, the receiver gain, and the homogeneity of the B_0 magnetic field. The reader is directed to a more detailed work (Fukushima and Roeder, 1981) for more information than is presented here.

The imaging coil sends an RF magnetic field into the sample and detects an RF response from the sample. The homogeneity of the transmitted RF field as well as that of the sensitivity will directly influence image quality. A coil with maximal field homogeneity, or a technique for correcting the inhomogeneity in the measured property (Li, 1994), is needed.

The imaging coil is an LC circuit which is designed to resonate at the operating frequency of the imager. This resonance frequency changes when the sample is placed inside the imaging coil. A tuning capacitor on many imaging coils allows the user to change the resonance frequency of the coil. A second capacitor allows the user to adjust the impedance of the coil to that of the imager, typically 50 Ω.

The amount of power necessary to produce a 90° pulse, P_{90}, will vary from sample to sample and from sample coil to sample coil. P_{90} is proportional to the coil volume, V_c, and the bandwidth, BW, of the imaging coil. Therefore the more closely matched the coil is to the size of the sample, the less power is required to produce a 90° pulse. The BW of the imaging coil will depend on the electrical quality of the coil and the dielectric constant and conductivity of the sample. A salt water solution sample will cause the imaging coil to have a larger BW than a hydrocarbon sample, and hence require more transmitter power to produce a 90° pulse. The amount of power necessary to produce a 90° pulse is determined by observing the signal from the sample as the transmitter power is increased from zero. The first maximum in the signal corresponds to the

amount of power necessary to produce a 90° pulse. Optimizing V_c and BW is also important because the signal is proportional to the fraction of the imaging coil occupied by the imaged object, and inversely proportional to the BW of the coil.

The receiver gain of the imager must be adjusted such that the amplitude of the greatest signal is less than that of the dynamic range of the digitizer. Each imager has a slightly different procedure for achieving this step. In general, the amplitude of the time-domain signal from the center of k space must be less than the dynamic range. To set this, the image acquisition parameters, such as TR, TE, FOV, and Thk, are first set and a time-domain signal is recorded from the imaging sequence with the phase-encoding gradient turned off. Amplifier gains are adjusted to keep this signal from saturating the amplifier.

The quality of the images produced by an MR imaging device is directly related to the homogeneity of the B_0 magnetic field across the sample. The B_0 field from the imaging magnet may be extremely homogeneous in the absence of a sample. When a sample is placed in the magnetic field, it distorts the fields around and within the sample due to its magnetic susceptibility and geometry. The homogeneity of the B_0 field in the sample is optimized using a set of shim coils found on the imager. The coils are set up to superimpose small, temporally static, spatial magnetic field gradients on the B_0 field. There are often ten or twenty different functional gradient forms (i.e., z, z^2, x, x^2, y, y^2, xy, etc.) which can be superimposed on B_0. These fields can be adjusted to cancel out spatial variations in the B_0 field across the sample. The FID from a 90-FID pulse is monitored as the gradients are varied. Maximizing the height and length of the FID optimizes the B_0 field.

Safety

Imaging magnets are typically 1 to 5 tesla (T). These magnets, especially those with a large-diameter bore, tend to have large magnetic fields that extend out around the magnet. These magnetic fields can cause pacemakers to malfunction, erase magnetic storage media, and attract ferromagnetic objects into the bore of the magnet. Persons with pacemakers must not be allowed to stray into a magnetic field of greater than 5×10^{-4} T. A magnetic field of approximately 5×10^{-3} T will erase the magnetically encoded material on credit cards and floppy disks. Ferromagnetic materials should be kept clear of the magnet, for these objects can be attracted to the magnet and damage experimental apparatus or injure persons in their path. They will also cause damage to the internal support structure of the superconducting magnet and could cause the magnet to quench.

DATA ANALYSIS AND INITIAL INTERPRETATION

The raw data from a magnetic resonance imager is the k space data of the image. These data are converted to the image with the use of a two-dimensional Fourier transform. The raw k space data are occasionally multiplied by an exponential with a time constant less than T_2^* to reduce noise in the image. This is equivalent to convolving the image with a Lorentzian function, but less time-consuming. This procedure for reducing the noise in an image will also reduce resolution if the time constant for the exponential multiplier is less than T_2^*.

Most magnetic resonance imaging data are 16-bit unsigned integer numbers and many images have signals that span this range. Displaying these data in the form of a pixel in an image for interpretation by the human eye necessitates the windowing and leveling of the data. This process takes the useable portion of the 16-bit data and converts it to 8-bit data for interpretation by the human eye-brain system. All imagers have software to perform this conversion. The level is translated to a brightness, and the window to the contrast, on a monitor.

Often the goal is to determine the concentration of a signal-bearing substance in the sample as a function of distance. When TE is much less than the minimum T_2 value and TR is much greater than the maximum T_1 in the sample, the image intensity from a spin-echo sequence is approximately equal to the concentration. Often it is not possible or practical to record an image under these conditions. In these cases, the spatial variation in T_1 and T_2 must be measured before the spin density can be determined.

There are several MR imagers on the market ranging in price from one to two million U.S. dollars. As with any instrument, there is a range of available sophistication. An expensive unit may be purchased that has much of the hardware and software necessary to image any material. Less expensive instruments are available that require the user to write some specific pulse sequence computer code or build specific application imaging coils. Details on imaging coil construction can be found in the literature (Hornak et al., 1986, 1987, 1988; Marshall et al., 1998; Szeglowski and Hornak, 1993).

When imaging a pure substance, T_1, T_2, and ρ images may be calculated from spin-echo data. T_1 is best calculated from S(TR) at constant TE using a fast least-squares technique (Gong, 1992). T_2 is calculated from S(TE) at constant TR using a nonlinear least-squares approach (Li and Hornak, 1993). The spin density may then be calculated from S(TR,TE), T_1, T_2, and Equation 10. Samples with multiple components require multiexponential techniques for calculating T_{1i}, T_{2i}, and ρ_i (Windig et al., 1998).

The concentration of some substances may be imaged indirectly through their effect on T_1 and T_2. For example, paramagnetic substances will shorten T_1 and T_2 of water. The relaxation rates ($1/T_1$ and $1/T_2$) of water are inversely proportional to the concentration of the paramagnetic species. Therefore, if images of T_1 or T_2 can be generated, the concentration of the paramagnetic substance can be determined. The concentration profiles of paramagnetic substances that have been studied by this technique include oxygen and transition metal ions, such as Ni, Cu, and Mn (Antalek, 1991). Other properties of materials, such as pH and temperature, may be measured if they cause a change in resonant frequency of the imaged nucleus. A change in the frequency will cause a change in the phase in the image. Phase information is typically discarded when a magnitude image is taken, so this information must be extracted before the magnitude calculation.

SAMPLE PREPARATION

In nuclear MRI, sample preparation is minimal, as the goal is to image samples in their natural state. Samples must be of a size that will fit inside of the RF sample coil of the imager. Sample containers must be made of materials that are transparent to RF. The dielectric constant of the sample container must be between that of air and that of the sample, so as to minimize reflection of RF. Polyvinyl chloride or polyethylene work well for most aqueous and hydrocarbon-based signals. The magnetic susceptibility of the sample container, and the sample, must be approximately equal to that of air to minimize susceptibility distortions. Glass containers produce large susceptibility distortions.

MRI of materials under extreme conditions of pressure and temperature are possible, but sample containers and hardware for containing the extreme pressures and temperatures are not commercially available and must be made by the researcher.

PROBLEMS

The major reason for not seeing a magnetic resonance image or for seeing one with a poor signal-to-noise ratio (SNR) is insufficient signal intensity. Occasionally, other hardware problems may be the cause. Both of these causes are discussed in this section.

Recall from Equation 8 that the signal intensity in an imaging sequence is proportional to the spin density. If the number of spins in a voxel is insufficient to produce a measurable signal, a poor SNR will result. The spin density may be increased by increasing the concentration of the signal-bearing nuclei or by increasing the voxel size. The voxel size may be increased by increasing the Thk, decreasing the number of pixels across the image, or by increasing the FOV. Increasing the voxel size will, however, result in a loss of image resolution.

Equation 10 also tells us that if T_2 is small compared to TE, or if T_1 is long compared to TR, we will see a small signal. These conditions are often the case in many solids. A short T_2 will cause the transverse magnetization to decay before a signal can be detected. A long T_1 will not allow the net magnetization to recover to its equilibrium value along the positive z axis before the next set of RF pulses.

Occasionally, a sample may contain ferromagnetic particles that destroy the homogeneity of the B_0 field in the sample and either distort the image or destroy the signal entirely. Certain concrete samples contain iron filings, which prevent one from seeing an image of free water in concrete. Some samples contain high concentrations of paramagnetic substances that shorten the T_2 of water to value less than TE, thus destroying the signal.

Two additional properties that may perturb the NMR signal are conductivity and dielectric constant (Roe et al., 1996). A high conductivity will cause a skin effect artifact in the image. A high conductivity causes the RF magnetic fields to diminish with depth into the sample. Therefore, the outside of the object appears brighter than the inside of the object. A sample with a high dielectric constant causes the opposite artifact. Standing waves are set up within the imaged object when the dimension of the object approaches a half wavelength of the RF in the object. If concentration maps are desired, it may be possible to compensate for the effects of the spatially inhomogeneous B_1 field (Li, 1994).

The magnetic resonance imager is a multicomponent, complex imaging system consisting of RF, gradient, magnet, computer, etc. subsystems. All of these subsystems must be operating properly for optimal signal. As with any complex instrument, it is recommended that test standards be imaged before each run to test the operation of the subsystem components individually and as a whole, integrated system. The subcomponent that the user has the most control over is the imaging coil. The imaging coil sends out the oscillating B_1 field into the sample and detects the oscillating field from the magnetization in the sample. The imaging coil is an electrical LC circuit that must be tuned to the resonant frequency of the spins in the sample at the magnetic field strength of the imager, and impedance matched to that of the transmission line from the imager. The conductivity and dielectric constant of the sample affect the resonant frequency of the imaging coil and the matching. When high-Q imaging coils are used, the resonant frequency of the coil must be tuned for and the impedance matched for each sample. Failure to tune and match the coil for each sample may result in large amounts of reflected power from the imaging coil, which will diminish signal and possibly damage the RF transmission and detection circuitry. Tuning and matching the coil is typically accomplished with tuning and matching variable capacitors located on the imaging coil.

Differences in the magnetic susceptibility of the sample may distort the image. Adjusting the shim settings may minimize the susceptibility artifact.

LITERATURE CITED

Aime, S., Botta, M., Fasano, M., Terreno, E., Kinchesh, P., Calabi, L., and Paleari, L. 1996. A new chelate as contrast agent in chemical shift imaging and temperature sensitive probe for MR spectroscopy. *Magn. Reson. Med.* 35:648–651.

Amin, M. H. G., Hall, L. D., and Chorley, R. J. 1994. Magnetic resonance imaging of soil-water phenomena. *Magn. Reson. Imag.* 12:319.

Amin, M. H. G., Richards, K. S., and Hall, L. D. 1996. Studies of soil-water transport by MRI. *Magn. Reson. Imag.* 14:879.

Amin, M. H. G., Gibbs, S. J., and Hall, L. D. 1997. Study of flow and hydrodynamic dispersion in a porous medium using pulsed-field-gradient magnetic resonance. *Proc. Math. Phys. Eng. Sci.* 453:489.

Antalek, B. J. 1991. MRI studies of porous materials. M. S. Thesis, Rochester Institute of Technology.

Armstrong, R. L., Tzalmona, A., Menzinger, M., Cross, A., and Lemaire, C. 1992. *In* Magnetic Resonance Microscopy (B. Blumich and W. Kuhn, eds.) pp. 309–323. VCH Publishers, Weinheim, Germany.

Attard, J., Hall, L., Herrod, N., and Duce, S. 1991. Materials mapped with NMR. *Physics World.* 4:41.

Blumich, B. and Blumler, P. 1993. NMR imaging of polymer materials. *Makromol. Chem.* 194:2133.

Blumich, B., Blumler, P., and Weigand, F. 1994. NMR imaging and materials research. *Die Makromolekulare Chemie Macromolecular Symp.* 87:187.

Callaghan, P. T., 1991. Principles of Nuclear Magnetic Resonance Microscopy, Clarendon, Oxford.

Callaghan, P. T. and Xia, Y. 1991. Velocity and diffusion imaging in dynamic NMR microscopy. *J. Magn. Reson.* 91:326.

Carpenter, T. A., Davies, E. S., and Hall, C. 1993. Capillary water migration in rock: Process and material properties examined by NMR imaging. *Mater. Struct.* 26:286.

Chingas, G. C., Miller, J. B., and Garroway, A. N. 1986. NMR images of solids. *J. Magn. Reson.* 66:530–535.

Duce, S. L. and Hall, L. D. 1995. Visualisation of the hydration of food by nuclear magnetic resonance imaging. *J. Food Engineering.* 26:251.

Duce, S. L., Carpenter, T. A., and Hall, L. D. 1990. Nuclear magnetic resonance imaging of chocolate confectionery and the spatial detection of polymorphic states of cocoa butter in chocolate. *Lebensm. Wiss. Technol.* 23:545.

Duce, S. L., Ablett, S., and Hall, L. D. 1994. Quantitative determination of water and lipid in sunflower oil and water and meat/fat emulsions by nuclear magnetic resonance imaging. *J. Food Sci.* 59:808.

Duce, S. L., Ablett, S., and Hall, L. D. 1995. Nuclear magnetic resonance imaging and spectroscopic studies of wheat flake biscuits during baking. *Cereal Chem.* 72:05.

Fordham, E. J., Horsfield, M. A., and Hall, L. D. 1993a. Depth filtration of clay in rock cores observed by one-dimensional ^{1}H NMR imaging. *J. Colloid Interface Sci.* 156:253.

Fordham, E. J., Hall, L. D., and Ramakrishnan, T. S. 1993b. Saturation gradients in drainage of porous media: NMR imaging measurements. *Am. Inst.. Chem. Eng. J.* 39:1431.

Fordham, E. J., Gibbs, S. J., and Hall, L. D. 1994. Partially restricted diffusion in a permeable sandstone: Observations by stimulated echo PFG NMR. *Magn. Reson. Imag.* 12:279.

Fordham, E. J., Sezginer, A., and Hall, L. D. 1995. Imaging multiexponential relaxation in the (y,log 3 T1) plane, with application to clay filtration in rock cores. *J. Magn. Reson.* A 113:139.

Fukushima, E. and Roeder, S. B. W. 1981. Experimental Pulse NMR: A Nuts and Bolts Approach. Addison-Wesley, Reading, Mass.

Fyfe, C. A., Randall, L. H., and Burlinson, N. E. 1992. Observation of heterogeneous trace (0.4% w/w) water uptake in bisphenol a polycarbonate by NMR imaging. *Chem. Mater.* 4:267.

Grinsted, R. A. and Koenig, J. L. 1992. Study of multicomponent diffusion into polycarbonate rods using NMR imaging. *Macromolecules* 25:1229.

Grinsted, R. A., Clark, L., and Koenig, J. L. 1992. Study of cyclic sorption-desorption into poly(methyl methacrylate) rods using NMR imaging. *Macromolecules* 25:1235.

Gibbs, S. J. and Hall, L. D. 1996. What roles are there for magnetic resonance imaging in process tomography? *Meas. Sci. Technol.* 7:827.

Gong, J. and Hornak, J. P. 1992. A fast T_1 algorithm. *J. Magn. Reson. Imag.* 10:623–626.

Guiheneuf, T. M., Tessier, J.-J., and Hall, L. D. 1996. Magnetic resonance imaging of meat products: Automated quantitation of the NMR relaxation parameters of cured pork, by both "bulk" NMR and MRI methods. *J. Sci. Food Agric.* 71:163.

Guiheneuf, T. M., Couzens, P. J., and Hall, L. D. 1997. Visualisation of liquid triacylglycerol migration in chocolate by magnetic resonance imaging. *J. Sci. Food Agric.* 73:265.

Guiheneuf, T. M., Gibbs, S. J., and Hall, L. D. 1997. Measurement of the inter-diffusion of sodium ions during pork brining by one-dimensional ^{23}Na magnetic resonance imaging (MRI). *J. Food Eng.* 31:457.

Hafner, S. and Kuhn, W. 1994. NMR-imaging of water content in the polymer matrix of silicon chips. *Magn. Reson. Imag.* 12:1075–1078.

Hall, L. D., Hawkes, R. C., and Herrod, N. J. 1990. A survey of some applications of NMR chemical microscopy. *Philos. Trans. Phys. Sci.* 333:477.

Holme, K. R. and Hall, L. D. 1991. Chitosan derivatives bearing c10-alkyl glycoside branches: a temperature-induced gelling polysaccharide. *Macromolecules* 24:3828.

Hornak, J. P. 1995. Medical imaging technology. *In* Kirk-Othmer Encyclopedia of Chemical Technology, 4th ed, John Wiley & Sons, New York.

Hornak, J. P., Ceckler, T. L., and Bryant, R. G. 1986. Phosphorus-31 NMR spectroscopy using a loop-gap resonator. *J. Magn. Reson.* 68:319–322.

Hornak, J. P., Szumowski, J., and Bryant, R. G. 1987. Elementary single turn solenoids used as the transmitter and receiver in magnetic resonance imaging. *J. Magn. Res. Imag.* 5:233–237.

Hornak, J. P., Marshall, E., Szumowski, J., and Bryant, R. G. 1988. MRI of extremities using perforated single turn solenoids. *Magn. Reson. Med.* 7:442–448.

Horsfield, M. A., Hall, C., and Hall, L. D. 1990. Two-species chemical-shift imaging using prior knowledge and estimation theory: Application to rock cores. *J. Magn. Reson.* 87:319.

Ilg, M., Maier-Rosenkrantz, J., Muller, W., Albert, K., and Bayer, E. 1992. Imaging of the chromatographic process. *J. Magn. Reson.* 96:335–344.

Jackson, P. 1992. Curing of carbon-fiber reinforced epoxy resin; non-invasive viscosity measurement by NMR imaging. *J. Mater. Sci.* 27:1302.

Karunanithy, S. and Mooibroek, S. 1989. Detection of physical flaws in alumina reinforced with SiC fibers by NMR imaging in the green state. *J. Mater. Science.* 24:3686.

Komoroski, R. A. 1993. NMR imaging of materials. *Anal. Chem.* 65:1068A.

Knorgen, M., Heuert, U., and Kuhn, W. 1997. Spatially resolved and integral NMR investigation of the aging process of carbon black filled natural rubber. *Polymer Bull.* 38:101.

Kumar, A., Welti, D., and Ernst, R. E. 1975. NMR Fourier zeugmatography. *J. Magn. Reson.* 18:69–83 (also *Naturwiss.* 62:34).

Lauterbur, P. G. 1973. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature* 242:190–191.

Li, X. 1994. Tissue parameter determination with MRI in the presence of imperfect radiofrequency pulses. M. S. Thesis, Rochester Institute of Technology.

Li, X. and Hornak, J. P. 1993. Accurate determination of T_2 images in MRI. *Imag. Sci. Technol.* 38:154–157.

Link, J., Kaufmann, J., and Schenker, K. 1994. Water transport in concrete. *Magn. Reson. Imag.* 12:203–205.

Maas, W. E., Merwin, L. H., and Cory, D. G. 1997. Nuclear magnetic resonance imaging of solid rocket propellants at 14.1 T. *J. Magn. Reson.* 129:105.

Marshall, E. A., Listinsky, J. J., Ceckler, T. L., Szumowski, J., Bryant, R. G., and Hornak, J. P. 1998. Magnetic resonance imaging using a ribbonator: Hand and wrist. *Magn. Reson. Med.* 9:369–378.

McDonald, P. J., Attard, J. J., and Taylor, D. G. 1987. A new approach to NMR imaging of solids. *J. Magn. Reson.* 72:224–229.

McFarland, E. W. and Lee, D. 1993. NMR microscopy of absorbed protons in palladium. *J. Magn. Reson.* A102:231–234.

Metz, K. R., Zuo, C. S., and Sherry, A. D. 1998. Rapid simultaneous temperature and pH measurements by NMR through thulium complex. 39th Experimental NMR Conference, Asilomar, Calif., 1998.

Mori, M. and Koenig, J. L. 1995. Solid-State C-13 NMR studies of vulcanized elastomers XIII: TBBS accelerated, sulfur-vulcanization of carbon black filled natural rubber. *Rubber Chem. Technol.* 68:551.

Nieminen, A. O. K. and Koenig, J. L. 1989. NMR imaging of the interfaces of epoxy adhesive joints. *J. Adhesion* 30:47.

Nieminen, A. O. K. and Koenig, J. L. 1990. Evaluation of water resistance of polyurethane dispersion coating by nuclear magnetic resonance imaging. *J. Adhesion* 32:105.

Pearce, R. B., Fisher, B. J., and Hall, L. D. 1997. Water distribution in fungal lesions in the wood of sycamore, *Acer pseudoplatanus*, determined gravimetrically and using nuclear magnetic resonance imaging. *New Phytol.* 135:675.

Pel, L., Kopinga, K., and Brocken, H. 1996. Determination of moisture profiles in porous building materials by NMR. *Magn. Reson. Imag.* 14:931.

Roe, J. E., Prenttice, W. E., and Hornak, J. P. 1996. A multipurpose MRI phantom based on a reverse micelle solution. *Magn. Reson. Med.* 35:136–141.

Rommel, E., Hafner, S., and Kimmich, R. 1990. NMR imaging of solids by jenner-broekaert phase encoding. *J. Magn. Reson.* 86: 264–272.

Sanz, J. L. C., Hinkle, E. B., and Jain, A. K. 1988. Radon and Projection Transform-Based Computer Vision, Springer-Verlag, Berlin.

Sarkar, S. N. and Komoroski, R. A. 1992. NMR imaging of morphology, defects, and composition of tire composites and model elastomer blends. *Macromolecules*, 25:1420–1426.

Seymour, J. D. and Callaghan, P. T. 1997. Generalized approach to NMR analysis of flow and dispersion in porous media. *Am. Inst.. Chem. Eng. J.* 43:2096.

Sinton, S. W., Iwamiya, J. H., Ewing, B., and Drobny, G. P. 1991. NMR of solid rocket fuel. *Spectroscopy* 6:42–48.

Smith, S. L. 1985. Nuclear magnetic resonance imaging. *Anal. Chem.* 57:A595–A607.

Szeglowski, S. D. and Hornak, J. P. 1993. Asymmetric single-turn solenoid for MRI of the wrist. *Magn. Reson. Med.* 30:750–753.

Wang, P.-S., Malghan, S. G., and Raman, R. 1995. NMR characterization of injection-moulded alumina green compact. Part II. T2-weighted proton imaging. *J. Mater. Sci.* 30:1069.

Wang, P.-S.., Minor, D. B., and Malghan, S. G. 1993. Binder distribution in Si3N4 ceramic green bodies studied by stray-field NMR imaging. *J. Mater. Sci.* 28:4940.

Wallner, A. S. and Ritchey, W. M. 1993. Void distribution and susceptibility differences in ceramic materials using MRI. *J. Mater. Res.* 8:655.

Webb, A. G. and Hall, L. D. 1990. Evaluation of the use of nuclear magnetic resonance imaging in the study of Fickian diffusion in rubbery polymers. 1. Unicomponent solvent ingress. *Polymer Commun.* 31:422.

Webb, A. G. and Hall, L. D. 1990. Evaluation of the use of nuclear magnetic resonance imaging in the study of Fickian diffusion in rubbery polymers. 2. Bicomponent solvent ingress. *Polymer Commun.* 31:425.

Webb, A. G. and Hall, L. D. 1991. An experimental overview of the use of nuclear magnetic resonance imaging to follow solvent ingress into polymers. *Polymer Bull.* 32:2926.

Webb, A. G., Jezzard, P., and Hall, L. D. 1989. Detection of inhomogeneities in rubber samples using NMR imaging. *Polymer Commun.* 30:363.

Windig, W., Hornak, J. P., and Antalek, B. J. 1998. Multivariate image analysis of magnetic resonance images with the direct exponential curve resolution algorithm (DECRA). Part 1: algorithm and model study. *J. Magn. Reson.* 132:298–306.

Zuo, C. S. Bowers, J. L., Metz, K. R., Noseka, T., Sherry, A. D., and Clouse, M. E. 1996. $TmDOTP^{5-}$: a substance for NMR temperature measurements. *In Vivo Magn. Reson. Med.* 36:955–959.

KEY REFERENCES

Stark, D. D. and Bradley, W. G. 1988. Magnetic Resonance Imaging. Mosby, St. Louis, Mo.

A comprehensive source of information on MRI as applied to clinical magnetic resonance imaging.

Callaghan, P. T. 1991. Principles of Nuclear Magnetic Resonance Microscopy, Clarendon, Oxford.

INTERNET RESOURCES

http://www.cis.rit.edu/htbooks/mri/

J. P. Hornak, 1996. The Basics of MRI. An excellent hypertext resource describing the theory of magnetic resonance imaging with animated diagrams.

http://www.cis.rit.edu/htbooks/nmr/

J. P. Hornak, 1998. The Basics of NMR. An excellent hypertext resource describing the theory of nuclear magnetic resonance spectroscopy with animated diagrams.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

α	angle between readout gradient and x axis
B_0	static magnetic field
B_1	magnitude of the RF magnetic field
BW	bandwidth
C	capacitance
CSI	chemical shift imaging
D_i	diffusion coefficient of component i
δ	width of the diffusion encoding gradient in a pulse sequence
Δ	separation of the diffusion encoding gradients in a pulse sequence
δ_i	chemical shift of component i
ESR	electron spin resonance
EPR	electron paramagnetic resonance
FOV	field of view
f	frequency width of an NMR spectrum
ϕ	phase angle
γ	gyromagnetic ratio
G_BP	bipolar magnetic field gradient pulse
G_i	magnetic field gradient in the i direction
G_R	readout gradient
G_s	slice selection gradient
G_p	phase encoding gradient
G_f	frequency encoding gradient

Γ_i	full width at half height of a spectral absorption line for component i
k	proportionality constant used in NMR signal equations to include amplifier gains
L	inductance
MRI	magnetic resonance imaging
$\mathbf{M}$	net magnetization vector
$\mathbf{M}_{xy}$	transverse component of magnetization
$\mathbf{M}_z$	longitudinal component of magnetization
μSR	muon spin resonance
NMR	nuclear magnetic resonance
ν_i	frequency of i
P_{90}	RF power necessary to produce a 90° pulse
ρ_i	spin density of i
RF	radio frequency
S	signal
T_{1i}	spin-lattice relaxation time or component i
T_{2i}	spin-spin relaxation time of component i
T_{2i}	$^*T_2$ star of component i
Thk	slice thickness
TR	repetition time
TE	echo time
τ	width of an RF or phase encoding gradient pulse
TMS	the NMR standard tetramethylsilane
θ	rotation angle of magnetization by the RF pulse
V	velocity
V_c	imaging coil volume
$\otimes$	convolution symbol

JOSEPH P. HORNAK
Rochester Institute of Technology
Rochester, New York

NUCLEAR QUADRUPOLE RESONANCE

INTRODUCTION

Nuclear quadrupole resonance (NQR) was once written off as a "dead" field, but has recently had a modest rebirth, and has been applied with success to several areas of materials science, most notably the high-T_c superconductors. The name itself is a misnomer—NQR is really (NMR) at zero field. For that reason, it has many of the same disadvantages as the more familiar NMR spectroscopy (NUCLEAR MAGNETIC RESONANCE IMAGING)—e.g., poor sensitivity and complications arising from molecular dynamics, but because it does not require a superconducting magnet, NQR is generally cheaper and can be applied to a far greater number of nuclei. It is inherently noninvasive and nondestructive, and can be applied both to pure substances and to materials in situ. It is currently, for example, being used as an antiterrorist technique to screen persons and baggage for the presence of explosives, using receiver coils that are of the order of 1 m in radius. The principal disadvantage of NQR, other than sensitivity, is that, as a radiofrequency (RF) technique, it is incompatible with conducting or ferromagnetic materials; however, for the somewhat narrow range of materials to which it can be applied, NQR gives information which is otherwise unavailable.

The NQR properties of a material are primarily defined by a transition frequency or set of transition frequencies, which can be used to determine two parameters—the quadrupole coupling constant or q.c.c., and the asymmetry parameter, or η. These quantities are often quite characteristic of particular materials—as when they are used to detect drugs or explosives—but are also sensitive to temperature and pressure, and can therefore be used as a noninvasive probe of these properties. Moreover, by applying external spatially varying magnetic or radiofrequency fields, the NQR signal can be made position-dependent, allowing the distribution of the NQR-active species in the material to be imaged. Finally, the relaxation times of NQR nuclei are highly sensitive to dynamics, and so can be used as a local probe of molecular mobility.

In general, NQR imaging gives the same information about the distribution of NMR active species as NMR imaging, but applies to nuclei with spin $>1/2$ rather than spin 1/2 nuclei.

The internal structure of materials can also be imaged by x-ray tomography and ultrasonic methods, which depend on the existence of inhomogeneities in the distribution of x-ray absorbers, or acoustic properties, respectively. The stress and pressure dependences of the NQR frequency are generally much stronger than those of NMR imaging, and therefore NQR is a superior method for mapping the temperature or stress distribution across a sample. Such stress distributions can be determined for optically transparent material by using polarized light; for opaque material, there may be no alternative methods.

This unit aims to be a general and somewhat cursory review of the theory and practice of modern one-dimensional (1D) and two-dimensional (2D) NQR and NQR imaging. It will certainly not be comprehensive. For a more general grounding in the theory of the method, one may best go back to the classic works of Abragam (1961) and Das and Hahn (1958); more modern reviews are referenced elsewhere in the text.

PRINCIPLES OF THE METHOD

Nuclear Moments

Elementary electrostatics teaches us that the nucleus, like any electromagnetic distribution, can be physically treated from the outside as a series of electric and magnetic moments. The moments of the nucleus follow a peculiar alternating pattern; nuclei in general have an electric monopole (the nuclear charge), a magnetic dipole moment, electric quadrupole moment, magnetic octopole moment, and electric hexadecapole moment. The last two, while they have been observed (Liao and Harbison, 1994) have little practical significance, and will not be examined further. The magnetic dipole moment of the nucleus enables NMR, which is treated elsewhere (NUCLEAR MAGNETIC RESONANCE IMAGING). It is the electric quadrupole moment of the nucleus that enables NQR.

Whether or not a nucleus has an electric quadrupole moment depends on its total spin angular momentum, $\mathbf{I}$, a property that is quantized in half-integer multiples of the quantum of action, h (Planck's constant), and often

referred to in shorthand form as the nuclear spin. Nuclei with spin zero, such as ^{12}C, have neither a magnetic dipole nor an electric quadrupole moment. Nuclei with spin 1/2 have magnetic dipole moments but not electric quadrupole moments, while nuclei with spin 1 or higher have both magnetic dipole moments and electric quadrupole moments. For reasons connected with the details of nuclear structure and particularly the paucity of odd-odd nuclei (those with odd numbers of both protons and neutrons) there are few nuclei with integer spin: the naturally occurring spin 1 nuclei are ^{2}H, ^{6}Li, and ^{14}N, while ^{10}B, ^{40}K, and ^{50}V are the single examples of spin 3, 4, and 6 respectively. The quadrupole moments of ^{2}H and ^{6}Li are extremely small, making direct nuclear quadrupole resonance impossible except through the use of SQUIDs (superconducting quantum interference devices, see Practical Aspects of the Method and (TECHNIQUES TO MEASURE MAGNETIC DOMAIN STRUCTURES). The higher integer spin nuclei are of no experimental importance, and therefore the only integer spin nucleus that has been studied much by NQR is ^{14}N.

In contrast, half-integer spin nuclei are liberally spread across the periodic table, and they have therefore been the subject of most NQR work. They range from the very common spin 3/2, for which there are scores of different stable isotopes, through the comparatively rare spin 9/2, which is present in a few, such as ^{87}Sr and ^{209}Bi.

Nuclear electric quadrupole moments are typically about eight orders of magnitude smaller than molecular electric quadrupole moments. In Figure 1, they are listed in SI units of $C \cdot m^2$, though it is common to divide out the charge of the electron and give them in units of area. They vary from the tiny moment of -1.6×10^{-50} $C \cdot m^2$ for ^{6}Li to the comparatively huge value of 6.1×10^{-47} $C \cdot m^2$ in ^{179}Hf. They can be directly measured by atomic beam methods, but are frequently known only with very rough precision; for example, recent measurements of the moment of ^{137}Ba have varied from $Q = 0.228$ barns (1 barn $= 10^{-28}$ m^2)to $Q = 0.35$ barns (to convert Q in barns to eQ in $C \cdot m^2$ multiply by 1.602×10^{-47}). An extensive tabulation of measurements of nuclear moments is available in Raghavan (1989). As will be discussed below, such moments cannot easily be independently determined by NQR.

Nuclear Couplings

The Electric Field Gradient. Group theory dictates that nuclear moments couple only with the derivative of the local electric or magnetic field that possesses the same symmetry. This means that the nuclear magnetic dipole couples only with the magnetic field, and the electric quadrupole couples with the electric field gradient. In NMR, the magnetic field is usually externally applied and in most cases dwarfs the local field from the paramagnetism of the sample. However, in NQR, the local electric field gradient dwarfs any experimentally feasible external field gradient. As a ball park estimate of the local field gradient ($\nabla \mathbf{E}$) in an ionic crystal, we can compute the magnitude of the field gradient at a distance of $r = 0.2$ nm from a unipositive ion (q_r), using

$$\nabla \mathbf{E} = -\frac{(3\hat{\mathbf{r}}\hat{\mathbf{r}} - 1)q_r}{4\pi\varepsilon_0 r^3} \tag{1}$$

where ε_0 is the permitivity of the vacuum and $\hat{\mathbf{r}}$ is a unit vector. This yields a gradient of 3.6×10^{20} $V \cdot m^{-2}$, or between 8 and 9 orders of magnitude greater than the external field gradients that can currently be produced in the laboratory. NQR is therefore carried out using only the local field gradient of the sample.

The electric field gradient $\nabla \mathbf{E}$ is a traceless symmetric second-rank tensor with five independent elements that can be expressed in terms of two magnitudes and three

IA												IIIA	IVA	VA	VIA	VIIA	0
^{2}D 1 4.58(2)	IIA																He
^{6}Li 3/2 -1.3(1) ^{7}Li 3/2 -64.2(2)	Be											^{10}B 3 135.5(1) ^{11}B 3 65.0(4)	C	^{14}N 1 32.36(16)	^{17}O 5/2 -41.3(2)	F	Ne
^{23}Na 3/2 161(3)	^{25}Mg 5/2 322(5)											^{27}Al 5/2 224(10)	Si	P	^{33}S 3/2 -102(16)	^{35}Cl 3/2 -132.16(3) ^{37}Cl 3/2 -104.03(3)	Ar
		IIIB	IVB	VB	VIB	VIIB	——	VIIIB	——	IB	IIB						
^{39}K 3/2 96.2(6)	^{43}Ca 7/2 -78(8)	^{45}Sc 7/2 -346(2)	^{47}Ti 5/2 466(16) ^{49}Ti 7/2 384(16)	^{51}V 7/2 -83(16)	^{53}Cr 3/2 -240(80)	^{55}Mn 5/2 528(16)	Fe	^{59}Co 5/2 672(48)	^{61}Ni 3/2 260(24)	^{63}Cu 3/2 -352(24) ^{65}Cu 3/2 -312(6)	^{67}Zn 5/2 240(24)	^{69}Ga 3/2 269(2) ^{71}Ga 3/2 169(2)	^{73}Ge 9/2 -277(42)	^{75}As 3/2 503(10)	Se	^{79}Br 3/2 530(6) ^{81}Br 3/2 442(6)	Kr
^{85}Rb 5/2 439(3) ^{87}Rb 3/2 211(2)	^{87}Sr 9/2 536(32)	Y	^{91}Zr 5/2 -330(16)	^{93}Nb 3/2 -512(32)	^{95}Mo 5/2 -320(128) ^{97}Mo 5/2 408(21	Tc	^{99}Ru 3/2 126.57 ^{101}Ru 5/2 732.19	Rh	^{105}Pd 5/2 1057(18)	Ag	Cd	^{115}In 9/2 1298(2)	Sn	^{121}Sb 5/2 -577(64) ^{123}Sb 7/2 -785(80)	Te	^{127}I 7/2 -1264(2)	^{131}Xe 3/2 -192(19)
^{133}Cs 7/2 -5.94(22)	^{135}Ba 3/2 256(5) ^{137}Ba 3/2 392(6)	^{139}La 7/2 320(16)	^{177}Hf 7/2 5378(46) ^{179}Hf 9/2 6077(53)	^{181}Ta 7/2 5079(32)	W	^{185}Re 5/2 3492(32) ^{187}Re 5/2 3316(32)	^{189}Os 3/2 1371(45)	^{191}Ir 3/2 1307(14) ^{193}Ir 3/2 1203(14)	Pt	^{197}Au 3/2 876(26)	^{201}Hg 3/2 616(64)	Tl	Pb	^{209}Bi 9/2 -801(16)	Po	At	Rn

Figure 1. Spin and quadrupolar moments (in $C \cdot m^2$ \times 10^{50}) of quadrupolar nuclei.

Euler angles defining its directionality. The magnitude of the field gradient tensor along its *zz* direction in its principal axis system $(\nabla E)_{zz}$ is often given the shorthand notation *eq*, with *e*, the electric charge, factored out. The asymmetry parameter η is defined by

$$\eta = \frac{(\nabla E)_{yy} - (\nabla E)_{xx}}{(\nabla E)_{zz}} \tag{2}$$

Electric field gradients can be computed for nuclear sites in ionic lattices and in covalently bound molecules. For lattice calculations, point-charge models are rarely adequate, and in fact often give results that are entirely wrong. In most cases, we find that fixed and induced dipole and quadrupole moments need to be introduced for each atom or residue in order to get accurate values. For example, in calculations of the field gradient at the ^{14}N and cationic sites in $Ba(NO_3)_2$ and $Sr(NO_3)_2$ (Kye, 1998), we include the charges, the quadrupole moment of the nitrate ion (treated as a single entity and calculated by *ab initio* methods), the anisotropic electric polarizability of the nitrate ion (obtained from the refractive indices of a series of nitrates via the Clausius-Mosotti equation), and the quadrupole polarizability of the cations (calculated by other authors *ab initio*). We omit the electric polarizability of the cations only because they sit at a center of crystal symmetry and the net dipole moment must inevitably be zero; otherwise, this would be an important contribution. Briefly, after calculating the field and field gradient for each site using point charges, we compute the induced dipole at the nitrate as a vector using the nitrate polarizability tensor, and recalculate the field iteratively until convergence is obtained. We repeat this procedure again after introducing the quadrupole moment of the nitrate and quadrupole polarizabilities of the cations. Convergence of the lattice sums may be facilitated by an intelligent choice of the unit cells. The contributions of the various terms to the total field gradient at the nitrate in $Ba(NO_3)_2$ and $Pb(NO_3)_2$ are shown in Figure 2.

In covalent molecules, field gradients at the nucleus may be adequately obtained using computational modeling methods.

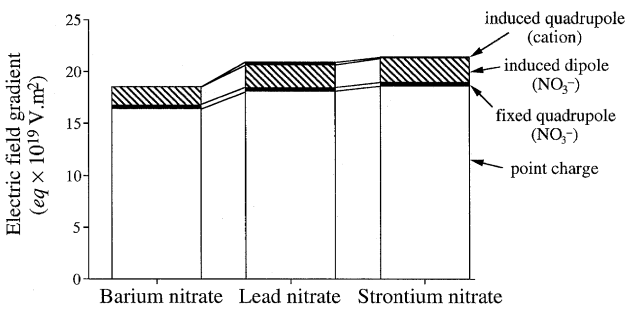


Figure 2. Contributions of the point charges, induced dipole moments, fixed quadrupolar and induced quadrupole moments to the total electric field gradient in ionic nitrates (from Kye and Harbison, 1998).

The Electric Quadrupole Hamiltonian. The coupling between the electric quadrupole moment and the ambient electric field gradient is expressed mathematically by the electric quadrupole coupling Hamiltonian.

$$H = \frac{e^2qQ}{4hI(2I-1)}\left[3I_z^2 - I(I+1) + \frac{\eta}{2}(I_+^2 + I_-^2)\right] \tag{3}$$

The quantity e^2qQ/h is the so-called electric quadrupole coupling constant, and is a measure of the strength of the coupling, dimensioned, as is usual in radiofrequency spectroscopy, in frequency units. I_z, I_+, and I_- are nuclear spin operators defined in the eigenbasis of the field gradient tensor.

The Sternheimer Effect and Electron Deformation Densities. One major complication of lattice calculations of electric field gradients is the Sternheimer antishielding phenomenon. Briefly, an atom or ion in the presence of a field gradient deforms into an induced quadrupole. The induced quadrupole does not merely perturb other atoms in the lattice; it also changes the field gradient within the electron distribution, usually increasing it, and often by a large amount. The field experienced by the nucleus is related to the external field by

$$\nabla \mathbf{E}_{\text{nuclear}} = (1 - \gamma_\infty)\nabla \mathbf{E}_{\text{external}} \tag{4}$$

where γ_∞ is the Sternheimer antishielding factor. Antishielding factors can be calculated by *ab initio* methods, and are widely available in the literature. They vary from small negative values for ions isoelectronic with helium, to positive values of several hundred for heavy ions. A representative selection of such factors for common ions is given in Table 1.

Energy Levels at Zero Field

Spin 1. At zero field, the electric quadrupole coupling is generally the only important component of the spin Hamiltonian. If the asymmetry parameter is zero, the Hamiltonian is already diagonal in the Cartesian (field gradient) representation, and yields nuclear energy levels of

$$E_{\pm 1} = \frac{e^2qQ}{4h}; \qquad E_0 = -\frac{e^2qQ}{2h} \tag{5}$$

There are two degenerate transitions, between $m = -1 \rightarrow m = 0$ and $m = 0 \rightarrow m = +1$, appearing at a frequency

$$\omega_Q = E_{\pm 1} - E_0 = \frac{3e^2qQ}{4h} \tag{6}$$

If $\eta \neq 0$, the degeneracy is lifted and the Hamiltonian is no longer diagonal in any Cartesian representation, meaning that the nuclear spin states are no longer pure eigenstates of the angular momentum along any direction, but rather are linear combinations of such eigenstates. The three energies are given by the equation

$$E_\pm = \frac{e^2qQ}{4h}(1 \pm \eta); \qquad E_0 = -\frac{e^2qQ}{2h} \tag{7}$$

Table 1. Sternheimer Quadrupole Antishielding Factors for Selected Ions

Ion	Free Ion	Crystal Ion
Li^+	0.261[a], 0.257[b,c], 0.256[d], 0.248[e]	0.282[a], 0.271[f]
B^{3+}	0.146[a], 0.145[b,c,d], 0.142[e]	0.208[a], 0.189[b]
N^{5+}	0.101[a]	0.110[a]
Na^+	−5.029[a], −5.072[b], 4.5±0.1[c,d,e]	−7.686[a], −4.747[f]
Al^{3+}	−2.434[a], −2.462[b], −2.59[d], −2.236[c]	−5.715[a], −3.217[f]
Cl^-	−82.047[a], −83.5[b], −53.91[c], −49.28[g], −63.21[h]	−38.915[a], −27.04[f]
K^+	−18.768[a], −19.16[b], −17.32[i], −12.17[c], −12.84[g], −18.27[h]	−28.701[a], −22.83[f]
Ca^{2+}	−14.521[a], −13.95[b], −12.12[b], −13.32[h]	−25.714[a], −20.58[f]
Br^-	−195.014[a], −210.0[b], −99.0[g], −123.0[i]	−97.424[a]
Rb^+	−51.196[a], −54.97[b], −49.29[g], −47.9[j]	−77.063[a]
Nb^{5+}	−22.204[a]	−28.991[a]
I^-	−331.663[a], −396.60[b], −178.75[g], −138.4[k]	−317.655[a]

[a] Sen and Narasimhan (1974).
[b] Feiock and Johnson (1969).
[c] Langhoff and Hurst (1965).
[d] Das and Bersohn (1956).
[e] Lahiri and Mukherji (1966).
[f] Burns and Wikner (1961).
[g] Wikner and Das (1958).
[h] Lahiri and Mukherji (1967).
[i] Sternheimer (1963).
[j] Sternheimer and Peierls (1971).
[k] Sternheimer (1966).

And there are now three observable transitions, two of which correspond to the degenerate transitions for $\eta = 0$, and one which generally falls at very low frequency.

$$\omega_{\pm} = \frac{3e^2qQ}{4h}\left(1 \pm \frac{\eta}{3}\right); \qquad \omega_0 = \frac{e^2qQ\eta}{2h} \tag{8}$$

Observation of a single NQR transition at zero field for ^{14}N is an indication that the asymmetry parameter is zero (or that one transition has been missed). Observation of a pair of transitions for any species permits the q.c.c. and η to be extracted, and thence all of the magnitude information for the tensor to be obtained. The low-frequency transition is seldom observed, and in any case contains redundant information.

Spin 3/2. The energy levels of spin 3/2 nuclei are given by

$$E_{\pm 1/2} = -\frac{e^2qQ}{4h}\left(1 + \frac{\eta^2}{3}\right)^{1/2} : E_{\pm 3/2} = \frac{e^2qQ}{4h}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{9}$$

As can be seen, the 1/2 and −1/2 states are degenerate, as are the 3/2 and −3/2 states. This degeneracy is not lifted by a nonzero asymmetry parameter, which mixes the degenerate states without splitting them. The single quadrupolar frequency is

$$\omega_Q = \frac{e^2qQ}{2h}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \tag{10}$$

Since this frequency is a function of both $e^{2qQ/h}$ and η, those quantities cannot be separately measured by simple NQR, though they can be obtained by a variety of two-dimensional methods (see Practical Aspects of the Method).

Spin 5/2. At $\eta = 0$, the energy levels of the spin 5/2 nucleus are given by

$$E_{\pm 1/2} = \frac{e^2qQ}{5h}; \quad E_{\pm 3/2} = -\frac{e^2qQ}{20h}; \quad e_{\pm 5/2} = \frac{e^2qQ}{4h} \tag{11}$$

Transitions between the $m = \pm 1/2$ and $\pm 3/2$ states, and between the $\pm 3/2$ and $\pm 5/2$ states are allowed. These appear at

$$\omega_1 = \frac{3e^2qQ}{20h}; \qquad \omega_2 = \frac{3e^2qQ}{10h} \tag{12}$$

and obviously lead to a 2:1 ratio of frequencies.

If the asymmetry parameter is nonzero, there are still two transition frequencies, but these deviate from a 2:1 ratio. Solving for them requires finding the roots of a cubic secular equation. The algebraic solutions have been published (Creel et al., 1980; Yu, 1991); more useful are the q.c.c. and η in terms of the measured frequencies

$$\begin{aligned} \frac{e^2qQ}{h} &= \frac{20}{3}\sqrt{\frac{(\nu_2^2 + \nu_2\nu_1 + \nu_1^2)}{7}}\cos(\phi/3) \\ \eta &= \sqrt{3}\tan(\phi/3) \\ \cos\phi &= \left(\frac{7}{(\nu_2^2 + \nu_2\nu_1 + \nu_1^2)}\right)^{3/2}\frac{(\nu_2 + 2\nu_1)(\nu_2 - \nu_1)(2\nu_2 + \nu_1)}{20} \end{aligned} \tag{13}$$

with $\nu_{1,2} = \omega_{1,2}/2\pi$. Thus, measurement of both frequencies for a spin 5/2 nucleus such as ^{127}I gives a complete set of magnitude information about the quadrupolar tensor.

In addition to these two transitions, for $\eta \neq 0$ the $\pm 1/2 \rightarrow \pm 5/2$ transition becomes weakly allowed, and can occasionally be observed. However, if the two single-quantum frequencies are already known, this nearly forbidden transition gives only redundant information.

Higher-Spin Nuclei. Spin 7/2 nuclei have three, and spin 9/2 four observable transitions at zero field. At $\eta = 0$, these frequencies are in simple algebraic ratio, and can be easily obtained from the Hamiltonian given in Abragam (1961). If $\eta \neq 0$, the situation is more complicated, although analytical expressions for the spin 7/2 frequencies are still available by solution of a quartic secular equation (Creel, 1983). Any pair of frequencies (if assigned) is sufficient to give e^2qQ/h and η.

First-Order Zeeman Perturbation and Zeeman Line Shapes

Most NQR spectra are collected not at zero field, but in the earth's field. In practice, in most cases, the effect of this small field is unobservable, save for certain systems with extremely narrow NQR lines. In these cases, most frequently highly ordered ionic crystals, Zeeman effects may be observed.

An external field introduces a preferential external reference from of the system, and for a single crystal (the simplest case) it therefore makes the spectrum dependent on the relative orientation of magnetic field and $\nabla\mathbf{E}$. For spin 3/2 systems (which are most frequently studied) the effect is to split the single transition into a pair of doublets whose splittings depend on the size of the magnetic field and its orientation with respect to the field-gradient tensor. Analytical expressions for the positions of these lines have been determined (Creel, 1983); in addition, certain preferred orientations of the tensor relative to field give no splitting, and the loci of these cones of zero splitting can be used to get the field-gradient tensor orientation, which is not directly available from NQR at zero field (Markworth et al., 1987).

In powders, Zeeman-perturbed spectra of spin 3/2 nuclei have a characteristic lineshape with four maxima; these line shapes can be used to determine the asymmetry parameter (Morino and Toyama, 1961).

Figure 3 shows the Zeeman-perturbed spectra of potassium chlorate ($KClO_3$), an excellent test sample for NQR, which has a ^{35}Cl resonance at 28.08 MHz. The spectra were obtained in a probe with an external field coil oriented parallel to the Earth's field, and shown for a series of coil currents. As can be seen, the minimum linewidth is obtained with a current of 0.2 A, where the external field presumably nulls out the Earth's field precisely. At higher or lower values of the current, characteristic Zeeman-perturbed line shapes are obtained.

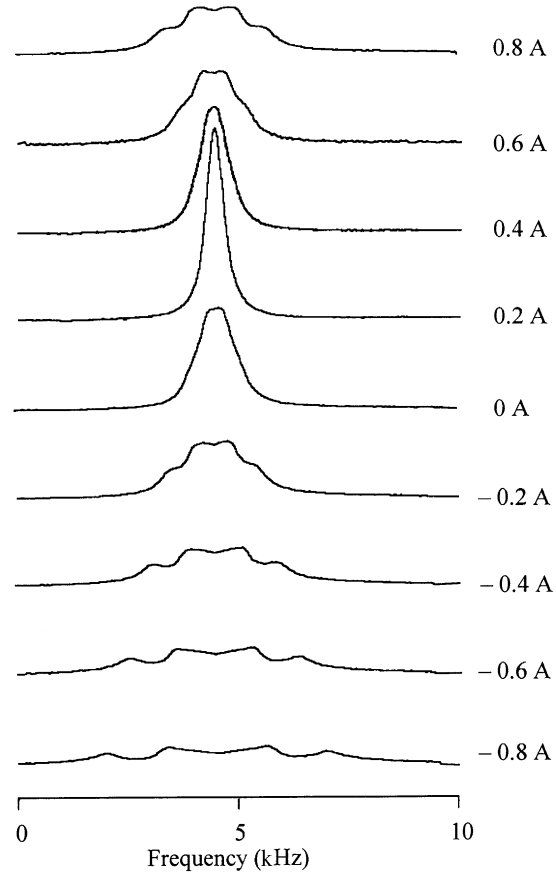


Figure 3. Effect of small external magnetic fields (expressed in terms of external coil current) on the lineshape of potassium chlorate resonance recorded at 28.85 MHz. A negative current corresponds to an applied field in the direction of the terrestrial magnetic field.

Spin Relaxation in NQR

As in NMR, NQR relaxation is generally parameterized in terms of two relaxation times: T_1, the longitudinal relaxation time, which is the time constant for relaxation of the populations of the diagonal elements of the electric quadrupole coupling tensor in its eigenbasis, and T_2, the transverse relaxation time, which governs the relaxation of off-diagonal elements. Except in the case of spin 3/2, more than one T_1 and T_2 are generally observed.

The longitudinal relaxation time is probably in most cases dominated by relaxation via very high-frequency motions of the crystal, typically librations of bonds or groups. Phenomenologically, it is generally observed to have an Arrhenius-like dependence on temperature, even in the absence of a well defined thermally activated motional mechanism.

The transverse relaxation time, T_2, must be distinguished from T_2^*, the reciprocal of the NQR line width. The latter is generally dominated by inhomogeneous broadening, which is a significant factor in all but the most perfectly crystalline systems. The processes giving rise to T_2 in NQR have, to our knowledge, not been systematically investigated, but it has been our observation that, in many cases, T_2 is similar in value to T_1, and is therefore likely a result of the same high-frequency motions.

T_1 and T_2 may be measured by the same pulse sequences used in high-field NMR; T_1 by an inversion recovery pulse sequence ($\pi - t - \pi/2 -$ acquire); T_2 using a single spin echo or train of spin echoes. T_1 values at room

temperature range from hundreds of microseconds to tens of milliseconds for half-integer spin nuclei with appreciable quadrupole coupling constants, such as ^{35}Cl, ^{63}Cu, ^{79}Br, or ^{127}I; for nuclei with smaller quadrupole moments, such as ^{14}N, they can range from a few milliseconds to (in rare cases) hundreds of seconds. T_1 values generally increase by one or two orders of magnitude on cooling to 77 K.

One phenomenon in which spin-relaxation times in NQR can be used to get detailed dynamical information is the "bleach-out" effect seen for trichloromethyl groups in the solid state. At low temperatures, three chlorine resonances are usually seen for such groups; however, as the temperature is raised, the resonances broaden and weaken, eventually disappearing, as a result of reorientation of the group about its three-fold axis. Depending on the environment of the grouping, "bleach-out temperatures" can be as low as $-100°C$ or as high at $70°C$, as is observed in the compound *p*-chloroaniline trichloroacetate (Markworth et al., 1987). Such exchange effects are quantum-mechanically very different from exchange effects in NMR, since reorientation of the group dramatically reorients the axis of quantization of the nuclear spin Hamiltonian, and is therefore in the limit of the "strong collision" regime. Under these circumstances, the relaxation time T_1 is to a very good approximation equal to the correlation time for the reorientation, making extraction of that correlation time trivial.

Similar phenomena are expected for other large-angle, low-frequency motions, but have otherwise been reported only for ^{14}N.

PRACTICAL ASPECTS OF THE METHOD

Spectrometers for Direct Detection of NQR Transitions

Frequency-Swept Continuous Wave Detection. For its first 30 years, the bulk of NQR spectroscopy was done using superregenerative spectrometers, either home-built or commercial (Decca Radar). The superregenerative oscillator was a solution to the constraint that an NMR spectrometer must be capable of sweeping over a wide range of frequencies, while in the 1950s and 1960s most radiofrequency devices (e.g., amplifiers, phase modulators and detectors, and receivers) were narrow-band and tunable. Briefly, the superregenerative circuit is one where the sample coil forms the inductive element in the tank circuit of the primary oscillator, which is frequency-swept using a motor-driven variable capacitor. Absorption of radiofrequency by the NQR spins causes a damping of the oscillator, which can be detected by a bridge circuit. The result is a frequency-swept spectrum in which NQR transitions are detected as the second derivative of their lineshape. The superregenerative circuit has the virtues of cheapness and robustness; however it has all the disadvantages of continuous wave detection—it does not lend itself easily to signal averaging, or to the more sophisticated 2D approaches discussed below. It is probably useful only for pure samples of molecular weights < 1000 Da.

Frequency-swept Fourier-transform Spectrometers. Most modern NQR spectroscopy likely is done on commercial solid-state NMR instruments, as the equipment requirements (save for the magnet) are very similar to those of NMR. Critical elements are a fast digitizer (<2 μs/point) and high-power amplification (preferably capable of 2 kW over a range of 5 to 200 MHz, although this depends on the sample).

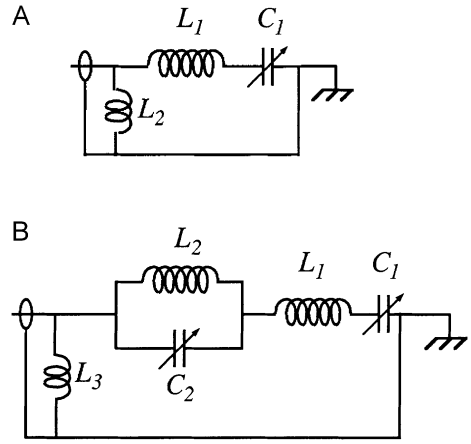


Figure 4. Probes for Fourier transform NQR. (**A**) Inductively matched series tank single-resonance probe; (**B**) inductively matched series/parallel tank double-resonance probe.

Standard static single-resonance solid-state NMR probes can be used for NQR; however, since there is no necessity that the probe be inserted in a magnet bore, some efficiencies can be achieved by building a dedicated probe. The circuit in Figure 4, panel A, has been successfully used for 10 years in the authors' laboratory. L_1, the sample coil, is typically a 20-turn solenoid wound from 26 AWG (American Wire Gauge) magnet wire, with dimensions of 4 cm long by 1.2 cm diameter; it holds a sample volume of ~4 ml. It is tuned by a series capacitance C_1; we generally employ a 1.5 to 30 pF Jennings vacuum capacitor rated at 15 kV. Matching to 50 Ω is accomplished by a 3 turn, 1 cm diameter inductor L_2. In general, inductive matches, while somewhat more cumbersome to adjust, are much more broad-banded than capacitive matches.

The probe gives a 1 to 2 μs $\pi/2$ pulse for most nuclei over a frequency range of ~20 to 45 MHz at a pulse power level of 500 W; additional series capacitance can be used to lower the frequency range, while going higher requires a smaller sample coil. The probe can simply be built inside an aluminum box, though somewhat better ringdown times are achieved with stainless steel. Acoustic ringing is however much less of a problem than in solid-state NMR, since the magnetoacoustic contribution is not present. If low temperatures are desired, the sample coil can be immersed in a coolant such as liquid nitrogen.

For double-resonance experiments, such as are required to correlate connected NQR transitions, we use the probe design in Figure 4, panel B. These double-resonance techniques involve irradiation of a spin $>3/2$ at two frequencies corresponding to distinct but connected NQR resonances. Double-resonance methods have not yet found significant application in materials research, and so are not described at length here; the interested reader is referred to Liao and

Harbison (1994). The circuit consists of a series and a parallel tank in series. In principle, either coil could contain the sample; in practice we use the series inductor L_1. A typical probe configuration uses coil inductances of $L_1 = 2.3$ μH and $L_2 = 1.2$ μH, with a small two- or three-turn matching inductor L_3. The circuit is essentially a pair of tank circuits, whose coupling varies with the separation of their resonance frequencies. Again, high-voltage 1.5 to 30 pF variable capacitors are used for C_1 and C_2. The two-fold difference between L_1 and L_2 ensures that the coupling between the two tanks is only moderate, so that resonance ν_1 primarily resides in tank circuit L_1C_1 and ν_2 in L_2C_2. With this configuration, ν_1 tunes between 17 MHz and 44 MHz and ν_2 between 47 and 101 MHz, although the two ranges are not entirely independent.

L_3 acts as a matching inductor for both resonances; differential adjustment of the matching inductance can be achieved by adjusting the coupling between the two inductors, either by partially screening the two coils by an intervening aluminum plate, or even by adjusting their relative orientation with a pair of pliers.

One notable feature that the probe lacks is any isolation between the two resonances; in fact, it has but a single port. This is because, unlike high-field NMR, it is seldom necessary to pulse one resonance while observing a second. We can therefore combine the output of both radio frequency channels before the final amplifier (which is broad-band and linear) and send the single amplifier output to a standard quarter-wave duplexer, which in practice is broad-band enough to handle the typical 2 : 1 frequency ratio used in NQR double-resonance experiments. Using single output state amplification and omitting isolation makes it far easier to design a probe to tune both channels over a wide frequency range.

SQUID Detectors. To overcome the sensitivity problems of NQR at low frequency, a DC superconducting quantum interference device (SQUID)—based amplifier can be used (Clarke, 1994; TonThat and Clarke, 1996). Using this device allows one to detect magnetic flux directly rather than differentially. This makes the intensity of the signal a linear function of the frequency instead of a quadratic function, greatly improving sensitivity at low frequencies compared to conventional amplification. SQUID NMR spectrometers can operate in both the continuous wave (Connor et al., 1990) or pulsed (TonThat and Clarke, 1996) modes. The latest spectrometers can attain a bandwidth up to 5 MHz (TonThat and Clarke, 1996). One of the drawbacks of the technique is cost, a result of the more complex design of the probe and spectrometer. SQUID spectrometers are not yet commercially available.

Experiments are performed at 4.2 K in liquid helium. Examples of nuclei observed using this technique are ^{27}Al in $Al_2O_3[Cr^{3+}]$ at 714 kHz (TonThat and Clarke, 1996), ^{129}Xe in laser-polarized solid xenon (Ton That et al., 1997), and ^{14}N in materials such as amino acids (Werner-Zwanzinger et al., 1994), cocaine hydrochloride (Yesinowski et al., 1995), and ammonium perchlorate (Clarke, 1994). Note that the detection of nitrogen transitions can be performed by cross-relaxation to protons coupled to nitrogen by the dipole-dipole interaction (Werner-Zwanzinger et al., 1994; Yesinowski et al., 1995).

Indirect Detection of NQR Transitions by Field Cycling Methods

Where relaxation properties of a material are favorable (long intrinsic proton relaxation time, short quadrupole relaxation time), indirect detection methods give a significant improvement in signal-to-noise ratio, since detection is via the proton spins, which have a large gyromagnetic ratio and relatively narrow linewidth. Such methods are therefore often favored over direct detection methods for spins with low NQR frequencies and low magnetic moments, e.g., ^{14}N or ^{35}Cl.

The term "field cycling" comes from the fact that the sample experiences different magnetic fields during the course of the experiment. The field intensity is therefore cycled between transients. Field cycling techniques were first developed during the 1950s (Pound, 1951). During a simple cycle, the sample is first magnetized at a high field, B_0, then rapidly brought adiabatically to a low field or to zero field. During the evolution period, the magnetization is allowed to oscillate under local interactions. The sample is finally brought back to high field, where the signal is detected. The field strength during the detection period can be identical to that for the preparation period, which is convenient but it can also be lower, as in "soak field" techniques (Koening and Schillinger, 1969). Also, other cycles involving more than two different field strengths have been developed, such as the "zero-field technique" (Bielecki et al., 1983). Field switching can be implemented either by electronic switches or by mechanically moving the sample. The latter is more simple and can be easily implemented on any spectrometer, but is limited for experiments necessitating very rapid field switching. Electronic switches are limited by Faraday's induction law, which dictates the maximum value of (dB_0/dt) for a given static B_0 field.

Field cycling can be applied in the solution state and in liquid crystals as well as in the solid state. As a competitive technique to NQR, quadrupolar nuclei can be observed directly; some examples are, but are not limited to, deuterium (Thayer and Pines, 1987) and nitrogen-14 (Selinger et al., 1994). Indirect detection ("level crossing") is also feasible by observing the signal of the abundant nucleus (typically protons) dipolar-coupled to the quadrupolar nucleus (Millar et al., 1985). An example of level crossing is given in a latter section to record the nutation spectrum of *tris*-sarcosine calcium chloride at 600 kHz (Blinc and Selinger, 1992). The large number of applications of field cycling NMR is outside the scope of this unit, and more detailed information can be found in reviews by Kimmich (1980) and Noack (1986); see Literature Cited for additional information.

One-Dimensional Fourier Transform NQR: Implementation

The hardest task in Fourier transform NQR is to find the resonance. While the chemistry of the species being studied sometimes gives one insight about where the resonances might lie—e.g., an organochlorine can be expected

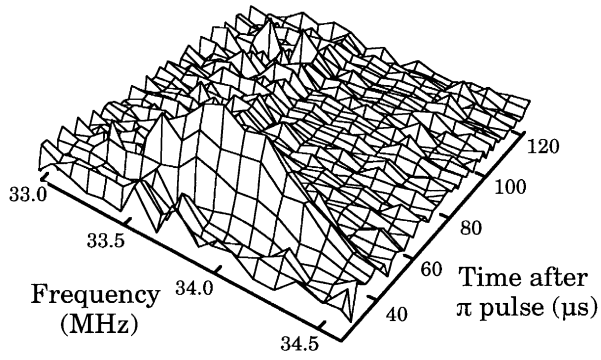


Figure 5. Frequency-swept spin-echo NQR. Each slice in the time dimension is a spin echo, recorded at a particular frequency; a slice in the frequency dimension parallel to the frequency axis gives the NQR spectrum.

to fall between 30 and 40 MHz—in general, this range will be far wider than the bandwidth of an NQR probe or the effective bandwidth of a high-power pulse. Some frequency sweeping will therefore have to be done. In a modern solid-state NMR spectrometer, the output power will usually be fairly constant over a range of 10 to 20 MHz, and the other components will generally be broad-band. Sweeping frequency, therefore, involves incrementing the spectrometer frequency by some fairly large fraction of the probe bandwidth, retuning the probe to that frequency, and acquiring a free-induction decay or preferably a spin echo at that frequency. Spectrometer frequencies are generally under computer control; the task of retuning the probe can be accomplished manually, or preferably automatically, by controlling the probe's tuning capacitor from the spectrometer via a stepper motor, and using this motor to minimize reflected power. The result of such a sweep is a series of spin echoes collected as a function of spectrometer bandwidth; these echoes, either phased or in magnitude mode, can be stacked in an array such as shown in Figure 5, which is a set of spin echoes collected for the ^{35}Cl signal of the polymer poly-4-chlorostyrene. A slice along the spine of the spin echo is the NQR spectrum collected at the resolution of the frequency sweep. If the line is narrower than this step frequency, the individual spin-echo slice in which the line appears may be Fourier transformed; if it is broad, the slice through the spin echoes suffices.

Two-dimensional Zero-Field NQR

Zeeman-Perturbed Nuclear Resonance Spectroscopy ZNQRS. For 3/2 spins, the $\pm 3/2 \rightarrow \pm 1/2$ transition depends on both the quadrupolar coupling constant, e^2qQ/h, and the asymmetry parameter η. The resonance frequency of the single line is given by

$$\nu_Q = \frac{1}{2}\left[\frac{e^2qQ}{h}\right]\left(1+\frac{\eta^2}{3}\right)^{1/2} \qquad (14)$$

That single-resonance frequency is insufficient to determine these two parameters separately. To overcome the problem, Ramachandran and Oldfield (1984) applied a small static Zeeman field, H_0, parallel to H_1, the radiofrequency field, as a perturbation to remove the degeneracy of the quadrupolar levels. The resulting splitting creates singularities, first described by Morino and Toyama (1961), at

$$\nu_Q \pm (1+\eta)\nu_0 \qquad (15)$$

and

$$\nu_Q \pm (1-\eta)\nu_0 \qquad (16)$$

where ν_0 is the Larmor frequency.

Note that applying the static field H_0 perpendicular to H_1 does not change the position of the singularities but affects the intensity distribution. By measuring the splitting, the asymmetry parameter can be determined.

Ramachandran and Oldfield's two-dimensional version of this experiment was implemented using both one pulse and spin echo (see Fig. 6, panels A and B). The field is turned on during the preparation period, during which the spins evolve under both the Zeeman and quadrupole interactions. The field is turned off during acquisition. The 2D spectrum is obtained by incrementing t_1 in regular intervals followed by 2D Fourier transform. The projection

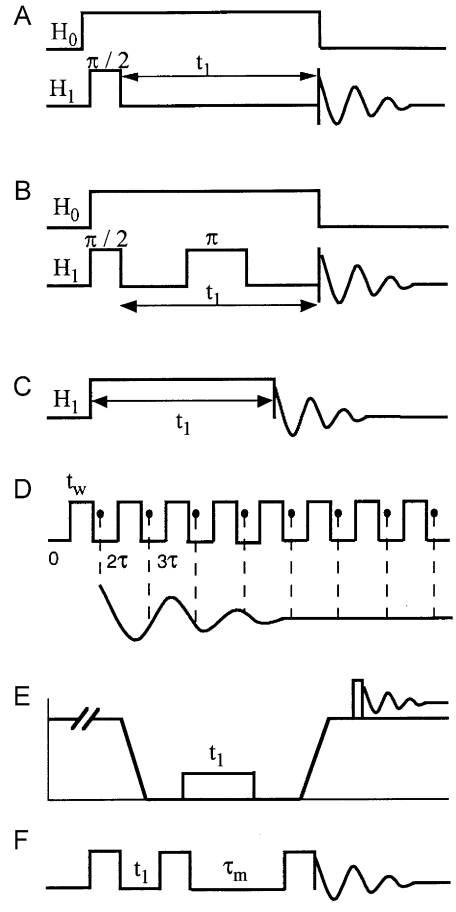


Figure 6. 2D NQR pulse sequences: (**A**) Zeeman-perturbed NQR one-pulse; (**B**) Zeeman-perturbed NQR spin echo; (**C**) zero-field nutation; (**D**) RF pulse train method; (**E**) level-crossing double resonance NQR nutation; (**F**) 2D exchange NQR.

on the ω_2 axis shows the zero-field NQR spectrum. The Zeeman NQR powder pattern is observed by projection along the ω_1 axis, which allows the determination of the asymmetry parameter η. The limitation of the method is that it requires a rapidly switchable homogeneous Zeeman field.

Zero-Field Nutation Nuclear Resonance Spectroscopy. Another method to determine the asymmetry parameter has been described (Harbison et al., 1989; Harbison and Slokenbergs, 1990) in which no Zeeman field is applied. The absence of a static field makes the frequency spectrum orientation independent. To determine η, it is necessary to obtain an orientation-dependent spectrum, but to do so, it is unnecessary to introduce an extra perturbation such as the Zeeman field. The sample radio frequency coil itself introduces an external preferential axis and thus an orientation dependence.

During an RF pulse in a zero-field NQR experiment, a 3/2 spin undergoes nutation about the unique axis of the EFG (Bloom et al., 1955). The strength of the effective H_1 field depends on the relative orientation of the coil and quadrupolar axes and goes to zero when the two axes are parallel. The nutation frequency is given by

$$\omega_N = (\sqrt{3}\omega_R \sin\theta)/2 \tag{17}$$

where $\omega_R = \gamma H_1$, θ is the angle between the coil axis and the unique axis of the electric field gradient tensor, and γ is the gyromagnetic ratio of the nucleus.

The voltage induced in the coil by the precessing magnetization after the pulse is proportional to $\sin\theta$. Figure 6, panel C, shows the 2D nutation NQR pulse sequence. For a single crystal, the NQR free precession signal is

$$F(t_1, t_2, \theta) \propto \sin\theta \sin(\sqrt{3}\omega_R t_1 \sin\theta/2) \sin(\omega_Q t_2) \tag{18}$$

where ω_Q is the quadrupolar frequency, t_1 is the time the RF pulse is applied and t_2 is the acquisition time.

For an isotropic powder, the nutation spectrum is obtained by powder integration over θ, followed by complex Fourier transformation in the second dimension and an imaginary Fourier transform in the first

$$G(\omega_1, \omega_2) \propto \int_0^{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sin^2\theta \sin(\sqrt{3}\omega_R t_1 \sin\theta/2) \times \sin(\omega_Q t_2) e^{i\omega_1 t_1} e^{i\omega_2 t_2}\, dt_2\, dt_1 d\theta \tag{19}$$

The angular factor $\sin\theta$ must be replaced by a factor $R(\theta, \phi)$ where θ and ϕ are the polar angles relating the coil axis and the quadrupolar tensor (Pratt et al., 1975). For an axially asymmetric tensor

$$R(\theta, \phi) = \sqrt{4\eta^2 \cos^2\theta + [9 + \eta^2 + 6\eta\cos(2\phi)]\sin^2\theta} \tag{20}$$

The 2D frequency-domain spectrum becomes

$$G(\omega_1, \omega_2) \propto \int_0^{\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sin\theta R(\theta, \phi) \sin(\omega_R t_1 R(\theta, \phi)/2\sqrt{3}\rho) \times \sin(\omega_Q t_2) e^{i\omega_1 t_1} e^{i\omega_2 t_2} dt_2\, dt_1\, d\theta\, d\phi \tag{21}$$

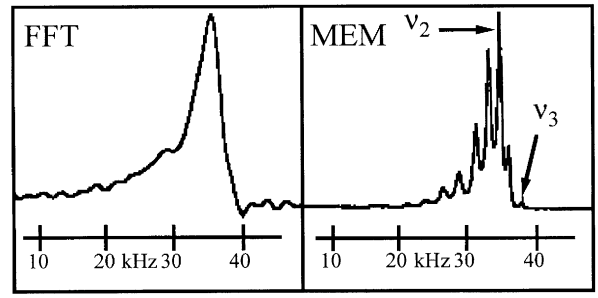


Figure 7. Fourier transform of the on-resonance nutation spectrum and corresponding MEM spectrum (the slices through the 2D spectrum parallel to the F_1 axis at $F_2 = \nu_{NQR}$) obtained at 77 K for $\nu_{NQR} = 36.772$ in $C_3N_3Cl_3$. The linear prediction filter used to obtain the MEM spectrum from $N = 162$ time-domain data points was $m = 0.95$ N (from Mackowiak and Katowski, 1996).

The on-resonance nutation spectrum in the ω_1 dimension (see Fig. 7) shows three singularities, ν_1, ν_2, and ν_3, whose frequencies are given by

$$2\pi\nu_1 = \frac{\eta\omega_R}{\sqrt{3}\sqrt{1 + \frac{1}{3}\eta^2}}, \quad 2\pi\nu_2 = \frac{(3 - \eta)\omega_R}{2\sqrt{3}\sqrt{1 + \frac{1}{3}\eta^2}},$$
$$\text{and} \quad 2\pi\nu_3 = \frac{(3 + \eta)\omega_R}{2\sqrt{3}\sqrt{1 + \frac{1}{3}\eta^2}} \tag{22}$$

The asymmetry parameter can be determined from the position of ν_2 and ν_3 alone (Harbison et al., 1989)

$$\eta = \frac{3(\nu_3 - \nu_2)}{\nu_3 + \nu_2} \tag{23}$$

Off-resonance effects can be calculated using a similar procedure (Mackowiak and Katowski, 1996). The nutation frequency is given by

$$\nu_N^{off} = \xi/\pi = \sqrt{(\nu_N^{on})^2 + (\Delta\nu)^2} \tag{24}$$

where ν_N^{on} is the on-resonance nutation frequency and $\Delta\nu$ is the resonance offset.

The asymmetry parameter is determined from the relation

$$\eta = \frac{3\left[\sqrt{(\nu_3^{off})^2 - (\Delta\nu)^2} - \sqrt{(\nu_2^{off})^2 - (\Delta\nu)^2}\right]}{\sqrt{(\nu_3^{off})^2 - (\Delta\nu)^2} + \sqrt{(\nu_2^{off})^2 - (\Delta\nu)^2}} \tag{25}$$

Another way to do an off-resonance nutation experiment is to start acquisition at some constant delay, t_o, after the RF pulse procedure (Sinyavski, 1991; Mackowiak and Katowski, 1996). The nutation spectrum consists of three lines. The line at $\Delta\nu$ is independent of the EFG parameters

and the asymmetry parameter can be calculated from any of the two other lines using

$$\eta = \frac{3\left[\sqrt{(\nu_3^{off})^2 \pm 2(\Delta\nu)^2} - \sqrt{(\nu_2^{off})^2 \pm 2(\Delta\nu)^2}\right]}{\sqrt{(\nu_3^{off})^2 \pm 2(\Delta\nu)^2} + \sqrt{(\nu_2^{off})^2 \pm 2(\Delta\nu)^2}} \quad (26)$$

The spectral resolution for an off-resonance experiment decreases with the frequency offset up to ~80% for a 100 kHz offset.

For experimental nutation spectra, the data are occasionally truncated and noisy. Therefore the position of the singularities is often poorly resolved, which makes an accurate determination of η difficult. As an alternative to the 2D Fourier transform, the maximum entropy method (MEM) can be used for processing the time-domain data (Sinjavski, 1991; Mackowiak and Katowski, 1993, 1996). The basis of the maximum entropy method is to maximize the entropy of the spectrum while maintaining correlation between the inverse Fourier transform of the spectrum and the free induction decay (FID). The process works by successive iterations in calculating the χ^2 correlation between the inverse Fourier transform of the trial spectrum and the FID. The spectrum is modified so its entropy increases without drastically increasing χ^2, until convergence is reached. One of the inherent advantages to this method is that the MEM spectrum is virtually noiseless and free of artifacts. Mackowiak and Katowski used the Burg algorithm to determine the lineshape that maximizes the entropy for the nutation experiment (Stephenson, 1988; Mackowiak and Katowski, 1993, 1996). In this unit we do not intend to discuss the mathematics involved in MEM, but we can point readers to a review by D.S. Stephenson about linear prediction and maximum entropy methods used in NMR spectroscopy (Stephenson, 1988; also see references therein). MEM has been very successfully applied in high-field NMR for both1D and 2D data reconstruction.

Figure 7 shows on-resonance simulated and experimental 2D Fourier transform (FT) and MEM spectra. The noise and resolution improvements for the MEM method are clearly visible. The singularities ν_2 and ν_3 can be easily measured from the MEM spectrum. The asymmetry parameter is determined as previously using Equation 23.

The quadrupolar frequency is very temperature sensitive. Rapid recycling can cause significant sample heating and introduce unwanted frequency shifts during the nutation experiment. Thus, recycle delays much longer than necessary for spin relaxation are used (Harbison et al., 1989; Harbison and Slokenbergs, 1990; Mackowiak and Katowski, 1996), together with active gas cooling. Changes in the RF excitation pulse length between slices can cause undesirable effects. Also, for weak NQR lines, acquiring the 2D NQR nutation experiment can take a long time. To reduce the experiment time, Sinyavski et al. (1996) reported a sequence of identical short RF pulses separated by time intervals τ.

The NQR signal induced in the RF coil just after turnoff of the nth pulse is

$$G(nt_w) = \langle I_x\rangle_n \sin\theta\cos\phi + \langle I_y\rangle_n \sin\theta\sin\phi + \langle I_z\rangle_n \cos\theta \quad (27)$$

where $\langle I_{x,y,z}\rangle_n$ are the expectation values of the magnetization along the coordinate axis and are calculated using a wave function approach (Pratt et al., 1975; Sinyavski et al., 1996).

In the general case, the expression in Equation 27 is a sum of signals with various phases. For the case when

$$-\omega_0\tau + \Delta\omega t_w = 2\pi k \quad \text{with} \quad k = 0, \pm1, \pm2\ldots \quad (28)$$

that is, when the phase accumulated due to resonance offset is an integer multiple of 2π, the free induction decay obtained is identical to that found in the two-dimensional nutation experiment. Here, t_w is the time delay between pulses in the nutation pulse train.

In addition, if the NQR signal is measured at some constant delay, t_d, after each pulse in the sequence, then the signal after synchronous detection is

$$G(nt_w) = \alpha R^2(\theta,\phi) \times \begin{bmatrix}\sin(2n\xi t_w)\cos(\Delta\omega t_d + \Psi) \\ +\Delta\omega[1-\cos(2n\xi t_w)]\sin(\Delta\omega t_d + \Psi)/2\xi\end{bmatrix} \Big/ 2\xi \quad (29)$$

$R(\theta,\phi)$ is defined in Equation 20, $\Delta\omega$ is the frequency offset, Ψ is a constant phase shift of the reference signal for the synchronous detector, and

$$2\xi = \sqrt{(\Delta\omega)^2 + 4m^2} \quad \text{with} \quad m = (\gamma H_1/4)R(\theta,\phi)/\sqrt{3+\eta^2} \quad (30)$$

Equation 29 describes the "discrete interferogram" of a single nutation line at the frequency 2ξ (Sinyavski et al., 1996). The pulse sequence for the train-pulse method is shown in Figure 6, panel D.

For the on-resonance condition ($\Delta\omega = 0$), t_w and τ can easily be chosen so that condition in Equation 28 is satisfied. For $\Delta\omega \neq 0$, Equation 28 can also be satisfied. If the sum of the intervals $t_w + \tau$ in the pulse train is an integer multiple of the data sampling period, the nutation interferogram $D(i)$ can be reconstructed from the raw data by the simple algorithm

$$D(i) = D[c + (i-1)(p\backslash n)] \quad (31)$$

where $i = 1, 2, \ldots, n, n$ is the number of pulses, p the number of data points, the backslash denotes an integer division (e.g., $a\backslash b$ is the largest integer less than a/b), and c is a constant that can be determined by trial (Sinyavski et al., 1996).

If the NQR line is broad, dead-time problems occur and a spin echo signal can be used. The equivalent of the expression in Equation 29 for the echo signal is

$$G(nt_w) = \alpha R^2(\theta,\phi)(1-[\Delta\omega/2\xi]^2)[1-\cos(2\xi t'_w)]\times [-\sin(2n\xi t_w)\cos\Psi + \Delta\omega[1-\cos(2n\xi t_w)]\sin\Psi/(4\xi)]/4\xi \quad (32)$$

where t'_w is the delay between nutation pulses.

The maximum time-saving factor can be roughly estimated as equal to the number of pulses contained in a pulse train (Sinyavski et al., 1996).

Level-Crossing Double Resonance NQR Nutation Spectroscopy. When the NQR signal of the 3/2 spin nuclei is weak (low natural abundance) or if the quadrupolar frequency is too low to be observable, a double-resonance technique known as level-crossing double resonance NQR nutation (Blinc and Selinger, 1992) can be used to retrieve the nutation spectrum. In this experiment, the quadrupolar nuclei are observed via their effect on signals from the other nuclei (typically protons). Dipolar coupling must be present between the two nuclei and spin-lattice relaxation times have to be relatively long. Figure 6, panel E, shows the field cycling and pulse sequence used for the experiment. The sample is first polarized in a strong magnetic field, H_0. The system is then adiabatically brought to zero magnetic field (either by decreasing H_0 or by moving the sample out of the magnet). At a specific field strength, level crossing occurs between the Zeeman levels of the 1/2 spin (protons) and the Zeeman-perturbed quadrupolar levels (Blinc et al., 1972; Edmonds and Speight, 1972; Blinc and Selinger, 1992). The level crossing polarizes the quadrupolar nuclei (decreasing the population difference $N_{\pm 1/2} - N_{\pm 3/2}$) and decreases the magnetization of the protons. An RF pulse of length t_1 is applied at zero field with frequency of $\Omega = \omega_Q + \delta$, where ω_Q is the pure quadrupolar frequency. The proton system is remagnetized and the proton FID acquired after a 90° pulse. The 2D spectrum is acquired by varying t_1 and Ω. The proton signal $S_H(t_1)$ is proportional to the population difference (Blinc and Selinger, 1992).

$$S_H(t_1) \propto (N_{\pm 1/2} - N_{\pm 3/2}) = A \cos[\omega(\theta, \varphi, \delta) t_1] \qquad (33)$$

where A is a constant and $\omega(\theta, \varphi, \delta) = \gamma_Q H_1 R(\theta, \varphi)/(2\sqrt{3} + \eta)$ is the nutation frequency of the quadrupolar nuclei. For a powder, Equation 33 is integrated over the solid angle Ω

$$S_H(t_1) = \frac{S_H(0)}{4\pi} \int \cos[\omega(\theta, \varphi, \delta) t_1] d\Omega \qquad (34)$$

The Fourier transform of $S_H(t_1)$ gives the nutation spectrum with singularities given in Equation 22. The ω_2 dimension shows the zero-field NQR spectra. This technique has been successfully applied to determine e^2qQ/h and η for NQR frequencies as low as 600 kHz (Blinc and Selinger, 1992) for *tris*-sarcosine calcium chloride.

2D Exchange NQR Spectroscopy. 2D exchange NMR spectroscopy was first suggested by Jeener et al. (1979) The technique has been widely used in high field. Rommel et al. (1992a) reported the first NQR application of the 2D exchange experiment.

The 2D exchange experiment (Rommel et al., 1992a; Nickel and Kimmich, 1995) consists of three identical RF pulses separated by interval t_1 and τ_m (Fig. 6, panel F). The first pulse produces spin coherence evolving during t_1; the second pulse partially transfers the magnetization components orthogonal to the rotating frame RF field into components along the local direction of the EFG. The mixing time, τ_m, is set to be much longer than t_1, so most of the exchange occurs during this interval. The last pulse generates the FID. The NQR signal oscillates along the axis of the RF coil. By incrementing t_1, and after double Fourier transform, the 2D spectrum $S(\nu_1, \nu_2, \tau_m)$ shows diagonal and cross-peaks, the latter corresponding to nuclei undergoing exchange.

The theoretical treatment is based on the fictitious spin 1/2 formalism (Abragam, 1961; Goldman, 1990; Nickel and Kimmich, 1995). The matrix treatment of the high field 2D exchange NMR (Ernst et al., 1987) has to be modified for the specifics of the NQR experiment. During the exchange process, in contrast to its high-field equivalent (where only the resonance frequency changes), the direction of quantization (given by the direction of the EFG) can also change. This reduces the cross-peak intensities by projection losses from the initial to the final direction of quantization of the exchange process (Nickel and Kimmich, 1995).

It is important for the 2D exchange that all resonances participating in the exchange be excited simultaneously. Phase-alternating pulses producing tunable sidebands can overcome this problem (Rommel et al., 1992a). If the exchanging resonances fall outside the feasable single resonance bandwidth, a double-resonance configuration could be used instead.

Spatially Resolved NQR and NQR Imaging

A logical step following 2D NQR experiments is to perform spatially resolved NQR and NQR imaging. This would be of particular value for the characterization of solid materials. Imaging is a very recent technique in NQR spectroscopy that only a handful of research groups are developing. The main restriction to performing NQR imaging is that even if the quadrupole resonances are narrow, it is practically impossible to shift the NQR frequency, since it only depends on the interaction between the quadrupole moment and the electric field gradient at the nuclei.

Magnetic Field Gradient Method. The first NQR experiment that allowed retrieval of spatial information was performed recently by Matsui et al. (1990). If the NQR resonance is relatively sharp, the half-width of a Zeeman-perturbed spectrum is proportional to the strength of the small static Zeeman field applied (Das and Hahn, 1958). Thus, the height of the pattern at the resonance frequency is reduced proportionally to the Zeeman field strength (Matsui et al., 1990). Figure 3 shows the changes in the spectrum lineshape of potassium chlorate due to different Zeeman field strengths. The sample coil is placed between two Helmholtz coils oriented perpendicular to the terrestrial magnetic field. Changing the current applied through the coil changes the field strength. Note that because the Zeeman field is parallel to the terrestrial field, a shimming effect appears for a current of 0.2 A, reducing the half-width by approximately a factor of two (unpublished results). This effect was not observed in the work of Matsui et al. (1990).

If a magnetic field gradient is applied, the reduction effect can be used for imaging since it depends on spatial location (Matsui et al., 1990). Given N discrete quadrupolar spin densities $\rho(X_n)$, the observed signal for the one-dimensional experiment will be the sum of all the powder

patterns affected by the field gradient. The spectral height of the spectrum at the resonance frequency ω_0 becomes

$$H(X_{n_0}) = \sum_{n=1}^{N} W(X_n - X_{n_0})\rho(X_n) \quad (35)$$

The function $W(X_n - X_{n_0})$represents the spectral height reduction at X_n. The reduction function can be measured by applying uniform Zeeman fields of known strength. The N discrete spin densities can be determined by solving the system of N linear equations given by Equation 21 at the resonance frequency (Matsui et al., 1990).

As the field gradient increases, the plot becomes similar to the projection of the sample. This is because the signal contribution of the neighboring locations decreases with increasing gradients. Determining the spin densities is a deconvolution, and small experimental errors (e.g., amplitude and location) are enhanced by the conversion.

Rotating Frame NQR Imaging. Rotating frame NQR imaging (ρ NQRI) is similar to the rotating frame zeugmatography proposed by Hoult (1979). This technique has the advantage of using pure NQR without any magnetic field or magnetic-field gradient, in contrast to the previous method. The rotating frame zeugmatography is a flip-angle encoding technique used in NMR imaging. Nonuniform radiofrequency fields are applied so that the flip angle of an RF pulse depends on the position with respect to the RF field gradient. The RF coils are designed to produce constant field gradients. For NQR, only the amplitude-encoding form of the method is applicable. In NQR, the transverse magnetization oscillates rather than precesses, making the phase encoding (Hoult, 1979) variant of the method inapplicable.

The first application of ρ NQRI was developed by Rommel et al. (1991b) in order to determine the one-dimensional profile of solid samples. In this experiment, an anti-Helmholtz coil (transmitter coil) produces the RF whose distribution is given by

$$B_1(z) = \frac{\mu_0 I}{2}\left[\frac{R^2}{(R^2 + (z+z_0)^2)^{3/2}} - \frac{R^2}{(R^2 + (z-z_0)^2)^{3/2}}\right] \quad (36)$$

where R is the coil radius, $z_0 = R\sqrt{3}/2$ is half the distance between the coils and $2I$ is the current amplitude through the coils. The receiver coil is placed coaxial to the transmitter coil (Rommel et al., 1991b). With this arrangement, transmitter and receiver coils are electrically decoupled. Later coil designs (Nickel et al., 1991) prefer the use of a surface coil acting as both a transmitter and a receiver coil. The sample is placed in only one-half of the transmitter coil.

The pulse sequence for the experiment is very similar to the 2D-nutation experiment (Harbison et al., 1989) described earlier in this unit. Free induction decay signals are excited with increasing RF pulse length.

Since ρ NQRI is restricted to solid samples, a simple way to access the second dimension is to rotate the sample by small angle increments, the rotation axis being perpendicular to the RF field gradient (Nickel et al., 1991; Kimmich et al., 1992; Rommel et al., 1992b). Using a surface coil, where the sample is placed outside the coil, creates the RF field gradient and allows two-dimensional spatial encoding. The 2D image is reconstructed using the projection/reconstruction procedure proposed by Lauterbur (1973). A stepping motor allows the sample to rotate, producing the 2D image. The surface coil is placed perpendicular to the rotation axis. The spatial information is amplitude encoded in the FID signals by the gradient of the RF pulse B_1. The RF gradient (G_1) is aligned along the coil z axis and considered as constant (Nickel et al., 1991).

$$G_1(z) = \frac{\partial B_1(z)}{\partial z} \quad (37)$$

The excitation pulse is characterized by the effective pulse length: $t_p = t_w\alpha$, where α is the transmitter attenuation factor and t_w is the proper pulse length. t_p can be varied by varying either t_w or α.

For 3/2 spins, the "pseudo" FID given in Equation 21 for the nutation experiment (Harbison et al., 1989) can be rewritten as

$$S(t_p) = \int_0^\infty \frac{1}{\pi}\tilde{\rho}(z)dz \int_0^\pi d\theta \int_0^{2\pi} \frac{\sqrt{3}}{2} R(\theta,\phi)\,\sin\theta \times \sin(\sqrt{3}t_p\omega_1(z)R(\theta,\phi)/2\sqrt{1+\eta^2/3}) \quad (38)$$

or

$$S(t_p) = c\int_0^\infty \frac{1}{\pi}\tilde{\rho}(z) f(z,t_p)dz \quad (39)$$

The proper distribution is $\tilde{\rho}(z) = g(\omega,z)\rho(z)$, where $g(\omega,z)$ is the local line shape acting as a weighting factor. $\omega_1(z) = \gamma B_1(z)$ and $R(\theta,\phi)$ are defined by Equation 6. For constant gradient, $\omega_1(z). = G_1$.

Introducing $t_p = t_d e^u, z = z_0 e^{-v}$ and $\xi = u - \nu$ leads to

$$\tilde{S}(u) \equiv S(t_d e^u) = c\int_0^\infty f_a(\nu) f_b(u-\nu)d\nu \quad (40)$$

The deconvolution can be performed using the following expression

$$f_a(\nu) = F_u^{-1}\left\{\frac{F_u\{S(t_d e^u)\}}{F_\xi\{f_b(\xi)\}}\right\} \quad (41)$$

$f_a(\nu)$ can be derived numerically, since both $f_b(\xi)$ and $S(t_d e^u)$ are known. The profile is given by $\tilde{\rho}\,(z) = f_a(v)/z$ and $v = \ln(z_o/z)$.

The 2D image is reconstructed from the profiles of all the orientations using the following steps (Nickel et al., 1991; Kimmich et al., 1992; Rommel et al., 1992b). The pseudo FIDs are baseline- and phase-corrected so that the imaginary part can be zeroed with no loss of information, and the profiles are determined. The profiles are

centered with respect to the rotation axis. If $G_1(z)$ is not constant, the profiles must be stretched or compressed by a coordinate transformation so that the abscissae are linearly related to the space coordinate. The profile ordinates must be corrected by a factor $G_1(z)/B_1(z)$ because the RF pickup sensitivity of the coil depends on z in the same way as B_1 and because the signal intensity is weighted by a factor $1/G_1(z)$. Finally, the image is reconstructed by the back-projection method (Lauterbur, 1973).

Robert et al. (1994) used the maximum entropy method (MEM) as an alternative to the Fourier transform to process the pseudo FIDs. The MEM procedure shows better resolution and avoids the noise and artifacts of the Fourier method. The drawback is the much longer time required for the MEM method to process the data (~30 s per FID compare to the millisecond scale of the fast Fourier transform or FFT). A Hankel transform can also be applied instead of FFT or MEM (Robert et al., 1994).

To reduce the acquisition time, a procedure similar to the 2D nutation NQR train pulse method (Sinyavski et al., 1996), discussed previously, can be used (Robert et al., 1996). Theoretical calculations show that the pseudo-FIDs recorded using this single experiment (SEXI) method (Robert et al., 1996) are identical to those obtained using the multiple-experiment procedure. Data processing and image reconstruction are performed as described previously using MEM processing and the back-projection algorithm. The SEXI method reduces acquisition time by a factor of ~50. So far, theoretical treatments and experiments using the SEXI method are only applied for 3/2 spins, while the multiple-experiment method also can be applied for 5/2 spins (Robert et al., 1996).

The combination of ρ NQRI and Matsui's Zeeman perturbed NQR technique, described earlier, allows adding the third dimension to the imaging process by adding slice selection (Robert and Pusiol, 1996a,b). The width of the Zeeman-perturbed NQR spectrum is proportional to the local Zeeman field (Matsui et al., 1990). In a zero-crossing magnetic field gradient, the nuclei in the zero-field region show sharp resonance while nuclei experiencing a field show weak and broad resonances. Therefore, the sample slice localized within the $B = 0$ plane will show an unchanged spectrum, while the other slices away from the zero-field plane show broad or even vanishing spectra. If the field gradient is large enough, signals outside the zero-field slice virtually disappear. Two Helmholtz coils of similar diameter, separated by a distance L and carrying current in opposite directions, produce the magnetic field gradient. Obviously, other coil designs can be used to produce the magnetic field gradient. Varying the electric current ratio between the Helmholtz coils shifts the zero-field plane. Each slice can be obtained as described previously using the SEXI method.

Robert and Pusiol (1996a,b) applied the ρ NQRI with slice-selection method on a test object filled with *p*-dichlorobenzene. Results are shown in Figure 8 for two different orientations of the test object. In both cases, the geometry of the test objects is well resolved and slice-selection appears to be a nice improvement of the ρ NQRI method.

Recently, a new imaging technique developed by Robert and Pusiol (1997) was reported. This method, instead of

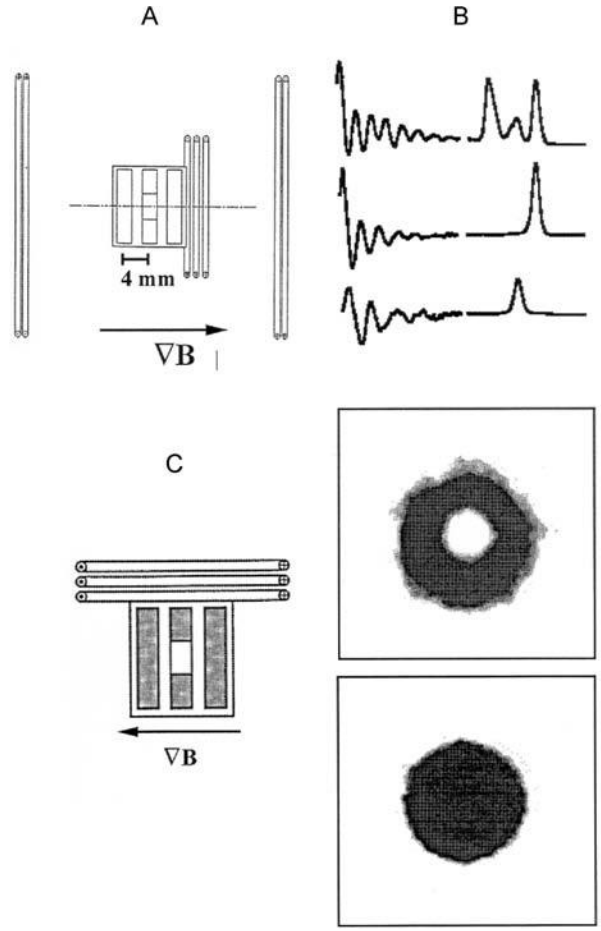


Figure 8. (**A**) Schematic arrangement of the surface RF and the selective magnetic gradient coils together with the cross-section of the object used for the test experiment. The selective static field gradient (∇B) is applied in a direction normal to the desired image plane. (**B**) Top: Pseudo-FID and profile along the cylindrical symmetry axis of the object imaged without external magnetic field. Middle: Magnetic gradient selecting the right hand cylinder. Bottom: After shifting the $B = 0$ plane towards the central cylinder. (**C**) Coil and test object arrangement for the 2D imaging experiment and cross image for the selection of the central cylinder (top) and of the external paradichlorobenzene disks (bottom) From Robert and Pusiol (1996a).

retrieving the second spatial dimension by rotating the sample, used a second surface RF coil perpendicular to the first one. In the "bidimensional rotating frame imaging technique 2D ρ NQRI," orthogonal RF gradients along the x and y axis are applied. A constant time interval T is introduced between the two pulses t_x and t_y. This is necessary to remove the transverse coherence created by the first pulse (Robert and Pusiol, 1997). The 2D experiment is carried out by incrementing t_x and t_y. Alternatively, the second pulse can be replaced by the SEXI train pulse described previously (Robert et al., 1996), and the complete 2D signal is obtained by incrementing only t_x. The resulting FIDs depend on both the x and y coordinates

$$F(t_x, t_y) \approx \int_0^\infty dx \int_0^\infty H_d \rho(x,y) S_1(t_x) S_2(t_y)\, dy \qquad (42)$$

where

$$S_1(t_x) = \int_0^{2\pi} d\phi_1 \int_0^{\pi} \sin\theta_1 \cos[\omega_1 \lambda(\theta_1, \phi_1) t_x]\, d\theta_1 \qquad (43)$$

and

$$S_2(t_y) = \int_0^{2\pi} d\phi_2 \int_0^{\pi} \sin\theta_2 \lambda(\theta_2, \phi_2) \sin[\omega_2 \lambda(\theta_2, \phi_2) t_y]\, d\theta_2 \qquad (44)$$

$\lambda(\theta, \phi)$ is proportional to $R(\theta, \phi)$ from Equation 20.

The corresponding 2D image is produced by determining the two-dimensional spin density function $\rho(t_x, t_y)$ The 2D image reconstruction is performed by a "true" 2D version of maximum entropy method and after RF field correction (Robert and Pusiol, 1997).

Temperature, Stress, and Pressure Imaging. As mentioned earlier, temperature strongly affects the quadrupolar frequency. Although this is an inconvenience for the spatially resolved imaging techniques, where the temperature must remain constant during acquisition to avoid undesired artifacts or loss in resolution, the temperature shift, as well as the effect of applying pressure or stress to the sample, can in fact be useful to detect temperature and pressure gradients, giving spectroscopic resolution in the second dimension. The spatial resolution in the first dimension can be performed using the initial ρ NQRI technique (Rommel et al., 1991a; Nickel et al., 1994) with no rotation of the sample and with a surface coil to produce the RF field gradient.

To study the effect of the temperature (Rommel et al., 1991b), a temperature gradient can be applied by two water baths at different temperature on each side of the sample. Data are collected in the usual way and are processed by 2D fast Fourier transform (FFT; Brigham, 1974). Figure 9, panel A, represents the ρ NQRI image of the test object and clearly shows the temperature gradients on the three sample layers through their corresponding frequency shifts. Using this technique, temperature gradients can be resolved to ~1°C/mm. Using MEM and deconvolution methods can highly improve the resolution, and experimental times can be greatly reduced by using the pulse-train method (Robert et al., 1996) instead of the multiple-experiment method.

To study pressure and stress, a similar procedure can be applied (Nickel et al., 1994). The probe contains two sample compartments, the first one serving as a reference while pressure is exerted on the second. Images are recorded for different applied pressures. If the sample is "partially embedded" in soft rubber or other soft matrices, broadening due to localized distribution of stress does not appear and only frequency shifts are observed in the image. If the pressure coefficient (in kHz/MPa) is known, pressures can be calculated from the corresponding frequency shifts. Figure 9, panel B, shows the shifting effect of applied pressure on the test sample. In the case of harder matrices or pure crystals, the frequency shift is accompanied by a line broadening caused by the local

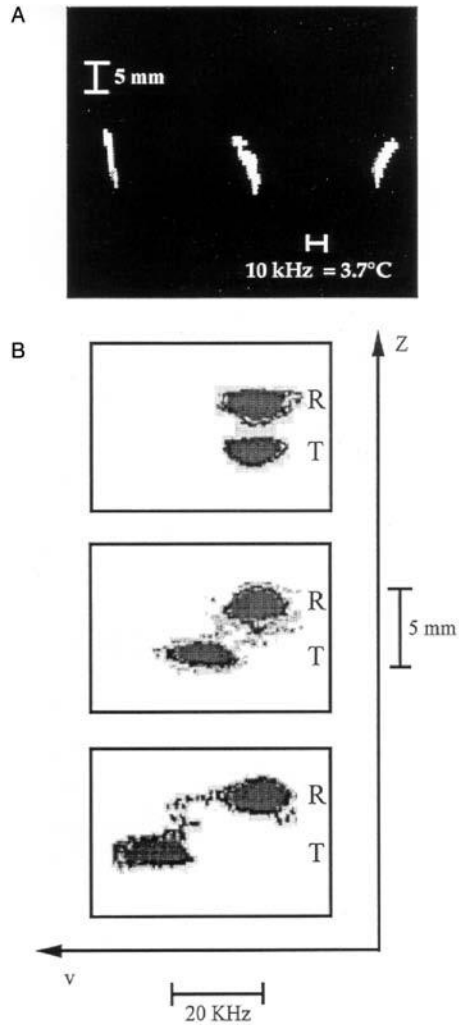


Figure 9. (**A**) Two-dimensional representation of temperature imaging experiments. The vertical axis represents the spacial distribution of the (cylindrical) sample, whereas the horizontal axis is the line shift corresponding to the local temperature (from Rommel et al., 1991a). (**B**) Spatial distribution of the ^{127}I NQR spectra of lithium iodate embedded in rubber in a two-compartment sample arrangement. Top: no applied pressure. Middle and bottom: applied pressure of 7 MPa and 14 MPa, respectively (from Nickel et al.,1994).

stress distribution within the sample. This broadening can even hide the shifting effect of the stress or pressure applied and the image only shows a decrease in intensity (Nickel et al., 1994).

Field Cycling Methods. Field cycling techniques are of potential interest for the imaging of materials containing quadrupolar nuclei dipolar-coupled to an abundant 1/2 nucleus (i.e., protons), and for detecting signals in the case where the quadrupolar frequency is too low or the signal too weak to be directly observed by pure NQR spectroscopy. Recently, Lee et al. (1993) applied field cycling to the one-dimensional ^{11}B imaging of a phantom of boric acid. The experiment consisted of a period at high field, during which the proton magnetization is built up, followed by a

rapid passage to zero field, during which the magnetization is transferred to the quadrupolar nuclei. RF irradiation at the quadrupolar frequency is applied and the remaining magnetization is transferred back to the proton by rapidly bringing the sample back to high field. Finally, the proton FID is obtained after solid echo. Spatial resolution is obtained by translating along the sample the small irradiation RF coil that produced the field gradient (Lee et al., 1993). The percentage of "recovered magnetization" is obtained from two reference experiments—maximum and minimum values of the "recovered magnetization" corresponding to "normal" and long residence times at zero field with no irradiation. Despite the low resolution, the image accurately represents the object and the experiment demonstrated the applicability of the technique. Using more elaborate coil designs, the same research group (Lee and Butler, 1995) produced ^{14}N 2D images spatially resolved in one dimension and frequency-resolved in the other. The image was produced by swapping both the RF coil position and the irradiation frequency at zero field.

It may be worth noting that field cycling imaging can possibly be achieved by performing the "level crossing double resonance NQR nutation spectroscopy" experiment (Blinc and Selinger, 1992) as described earlier, with a surface RF coil arrangement to access the spatial information as in the ρ NQRI techniques. If applicable, this would lead to improved image resolution.

So far, in contrast to its high field equivalent, and especially compared to medical imaging, the spatial resolution of the NQR imaging techniques is still fairly poor ($\sim$1 mm). Nevertheless, the recent advances are very convincing for the future importance of NQR imaging in materials science. Resolution improvements can be achieved by new coil arrangement designs and deconvolution algorithms. The limiting factor in resolution is ultimately the intrinsic linewidth of the signal. The future of NQR imaging may reside in the development of line-narrowing techniques.

DATA ANALYSIS AND INITIAL INTERPRETATION

Most modern NMR spectrometers contain a full package of one and two-dimensional data processing routines, and the interested reader is referred to the spectrometer documentation. Very briefly, processing time domain NMR data (free induction decays or FIDs) is done in five steps.

1. Baseline correction is applied to remove the DC offset in the FID.
2. Usually, the first few points of the free-induction decays are contaminated by pulse ringdown and give rise to artifacts. They must be removed by shifting the all data sets by several points.
3. Exponential multiplication is usually performed to increase the signal-to-noise ratio. This is performed by multiplying the real and imaginary part by the function $(-Ct)$, where t is the time in seconds and C is a constant in Hz. The larger the C, the faster the decay and the larger will be the broadening in the spectrum after the Fourier transform. This is equivalent to convolving the frequency domain spectrum with a Lorentzian function. Other apodization functions, such as sinebell and Gaussian multiplication, can also be used.
4. Spectral resolution can also be improved by doubling the size of the data set. This is done by filling the second half of the time-domain data with zeros.
5. Fourier transform allows transformation from the time to the frequency domain where the spectrum is analyzed. The theory of the fast Fourier transform (FFT) has been described in detail elsewhere and will not be discussed here. We refer the reader instead to Brigham (1974).

Two-dimensional data processing in general repeats this procedure in a second dimension, although there are some mathematical niceties involved. Any modern handbook of NMR spectroscopy, such as Ernst et al. (1987), deals with the details.

PROBLEMS

Sensitivity

As with most radiofrequency spectroscopy methods, the overriding problem is sensitivity. Boltzmann populations and magnetic moments of quadrupolar nuclei are small, and their direct detection either by nuclear induction or by SQUID methods often stretches spectrometers to the limit. Offsetting these sensitivity difficulties is the fact that longitudinal relaxation times are often short, and so very rapid signal averaging is often possible; in fact, typical data acquisition rates are limited not by relaxation but by thermal heating of the sample by the RF pulses.

Sensitivity depends on a host of factors—linewidth, relaxation times, the quadrupole constant and gyromagnetic ratio of the nucleus, the amount of sample, and, of course, the quality of the spectrometer. As an order-of-magnitude estimate, a "typical" NMR nucleus (^{63}Cu) with a quadrupole coupling constant in the range 20 to 25 MHz, can probably be detected in amounts of 1 to 10 μM.

Spurious Signals

As in any form of spectroscopy, it is essential to distinguish between signals arising from the sample and artifactual resonances from the sample container, probe, or elsewhere in the experimental apparatus. Artifacts extraneous to the sample can, generally, easily be detected by adequate controls, and with experience can be recognized and discounted; our probes, for example, have an impressive-looking NQR signal at 32.0 MHz, apparently due to a ceramic in the capacitors! A harder problem to deal with is identifying and assigning multiple NQR resonances in the same material; in contrast to NMR, at zero field the identity of a resonance cannot simply be determined from its frequency. The particular nucleus giving rise to an NQR signal can be identified with certainty if there exists another isotope of that nucleus. For example, natural-abundance copper has two isotopes—^{63}Cu and ^{65}Cu—with

a well-known ratio of quadrupole moments. Similar pairs are ^{35}Cl and ^{37}Cl, ^{79}Br and ^{81}Br, ^{121}Sb and ^{123}Sb, etc. Where a second isotope does not exist, or where additional confirmation is required, the gyromagnetic ratio of the nucleus can be measured with a nutation experiment (see Practical Aspects of the Method).

Two signals from the same nucleus and arising from different chemical species are the most difficult problem to deal with. Such signals generally cannot be assigned by independent means, but must be inferred from the composition and chemistry of the sample. This is an area where much more work needs to be done.

Dynamics

The broadening and disappearance of NQR lines caused by dynamical processes on a time scale comparable to the NQR frequency is a useful means of quantifying dynamics, if such processes are recognized; however, they may also cause expected NQR signals to be very weak or absent. We suspect the tendency of NQR signals to be very idiosyncratic in intensity to be largely due to this effect. If NQR lines are missing, or unaccountably weak, and dynamics is suspected to be the culprit, cooling the sample will often alleviate the problem.

Heterogeneous Broadening

Finally, another limitation on the detectability of NQR samples is heterogeneous broadening. In pure samples of perfect crystals, NQR lines can often be very narrow, but such systems are seldom of interest to the materials scientist; more often the material contains considerable physical or chemical heterogeneity. The spectrum of poly-4-chlorostyrene in Figure 5 is an example of chemically heterogeneous broadening; the coupling constants of the chlorine are distributed over $\sim$1 MHz due to local variations in the chemical environment of the aromatic chlorines. Similar broadening is observed, for example, in the copper NQR signals of high T_c superconductors. Physical heterogeneous broadening is induced by variations in temperature or strain over the sample. Simply acquiring NQR data too fast will generally induce inhomogenous RF heating of the sample, and since NQR resonances are exquisitely temperature dependent, the result is a highly broadened and shifted line. Strain broadening can be demonstrated simply by grinding a crystalline NQR sample in a mortar and pestle; the strain induced by grinding will often broaden the lines by tens or hundreds of kilohertz. Obviously, the former effect can be avoided by limiting the rate of data acquisition, and the latter by careful sample handling. When working with crystalline solids, we avoid even pressing the material into the sample tube, but pack it instead by agitation.

ACKNOWLEDGMENTS

We thank Young-Sik Kye for permission to include unpublished results. The data in Figure 5 were collected by Solomon Arhunmwunde, who died under tragic circumstances in 1996. This project was funded by NSF under grant number MCB 9604521.

LITERATURE CITED

Abragam, A. 1961. The Principles of Nuclear Magnetism. Oxford University Press, New York.

Bielecki, A., Zax, D., Zilm, K., and Pines, A. 1983. Zero field nuclear magnetic resonance. *Phys. Rev. Lett.* 50:1807–1810.

Blinc, R. and Selinger, J. 1992. 2D methods in NQR spectroscopy. *Z. Naturforsch.* 47a:333–341.

Blinc, R., Mali, M., Osredkar, R., Prelesnik, A., Selinger, J., Zupancic, I., and Ehrenberg, L. 1972. Nitrogen-14 NQR (nuclear quadrupole resonance) spectroscopy of some amino acid and nucleic bases via double resonance in the laboratory frame. *J. Chem. Phys.* 57:5087.

Bloom, M., Hahn, E. L., and Herzog, B. 1955. Free magnetic induction in nuclear quadrupole resonance. *Phys. Rev.* 97:1699–1709.

Brigham, E. O. 1974. The Fast Fourier Transform. Prentice-Hall, Englewood Cliffs, N.J.

Burns, G. and Wikner, E. G. 1961. Antishielding and contracted wave functions. *Phys. Rev.* 121:155–158.

Clarke, J. 1994. Low frequency nuclear quadrupole resonace with SQUID amplifiers. *Z. Naturforsch.* 49a:5–13.

Connor, C., Chang, J., and Pines, A. 1990. Magnetic resonance spectrometer with a D. C. SQUID detector. *Rev. Sci. Instrum.* 61:1059–1063.

Creel, R. B. 1983. Analytic solution of fourth degree secular equations: I=3/2 Zeeman-quadrupole interactions and I=7/2 pure quadrupole interaction. *J. Magn. Reson.* 52:515–517.

Creel, R. B., Brooker, H. R., and Barnes, R. G. 1980. Exact analytic expressions for NQR parameters in terms of the transition frequencies. *J. Magn. Reson.* 41:146–149.

Das, T. P. and Bersohn, R. 1956. Variational approach to the quadrupole polarizability of ions. *Phys. Rev.* 102:733–738.

Das, T. P. and Hahn, E. L. 1958. Nuclear Quadrupole Resonance Spectroscopy. Academic Press, New York.

Edmonds, D. T. and Speight, P. A. 1972. Nuclear quadrupole resonance of ^{14}N in pyrimidines purines and their nucleosides. *J. Magn. Reson.* 6:265–273.

Ernst, R. R., Bodenhausen, G., and Wokaun, A. 1987. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Clarendon, Oxford.

Feiock, F. D. and Johnson, W. R. 1969. Atomic susceptibilities and shielding factors. *Phys. Rev.* 187:39–50.

Goldman, M. 1990. Spin 1/2 description of spins 3/2. *Adv. Magn. Reson.* 14:59.

Harbison, G. S. and Slokenbergs, A. 1990. Two-dimensional nutation echo nuclear quadrupole resonance spectroscopy. *Z. Naturforsch.* 45a:575–580.

Harbison, G. S., Slokenbergs, A., and Barbara, T. S. 1989. Two-dimensional zero field nutation nuclear quadrupole resonance spectroscopy. *J. Chem. Phys.* 90:5292–5298.

Hoult, D. I. 1979. Rotating frame zeugmatography. *J. Magn. Reson.* 33:183–197.

Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. 1979. Investigation of exchange processes by two-dimensional NMR spectroscopy. *J. Chem. Phys.* 71:4546.

Kimmich, R. 1980. Field cycling in NMR relaxation spectroscopy: Applications in biological, chemical and polymer physics. *Bull. Magn. Res.* 1:195.

Kimmich, R., Rommel, E., Nickel, P., and Pusiol, D. 1992. NQR imaging. *Z. Naturforsch.* 47a:361–366.

Koening, S. H. and Schillinger, W. S. 1969. Nuclear magnetic relaxation dispersion in protein solutions. *J. Biol. Chem.* 244: 3283–3289.

Kye, Y.-S. 1998 The nuclear quadrupole coupling constant of the nitrate ion. Ph.D. Thesis. University of Nebraska at Lincoln.

Lahiri, J. and Mukherji, A. 1966. Self-consistent perturbation. II. Calculation of quadrupole polarizability and shielding factor. *Phys. Rev.* 141:428–430.

Lahiri, J. and Mukherji, A. 1967. Electrostatic polarizability and shielding factors for ions of argon configuration. *Phys. Rev.* 155:24–25.

Langhoff, P. W. and Hurst, R. P. 1965. Multipole polarizabilities and shielding factors from Hartree-Fock wave functions. *Phys. Rev.* 139:A1415–A1425.

Lauterbur, P. C. 1973. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature* 242:190–191.

Lee, Y. and Butler, L. G. 1995. Field cycling ^{14}N imaging with spatial and frequency resolution. *J. Magn. Reson.* A112:92–95.

Lee, Y., Michaels, D. C. and Butler, L. G. 1993. ^{11}B imaging with field cycling NMR as a line narrowing technique. *Chem. Phys. Lett.* 206:464–466.

Liao, M. Y. and Harbison, G. S. 1994 The nuclear hexadecapole interaction of iodine-127 in cadmium iodide measured using zero-field two dimensional nuclear magnetic resonance. *J. Chem. Phys.* 100:1895–1901.

Mackowiak, M. and Katowski, P. 1993. Application of maximum entropy methods in NQR data processing. *Appl. Magn. Reson.* 5:433–443.

Mackowiak, M. and Katowski, P. 1996. Enhanced information recovery in 2D on- and off-resonance nutation NQR using the maximum entropy method. *Z. Naturforsch.* 51a:337–347.

Markworth, A., Weiden, N., and Weiss, A. 1987. Microcomputer controlled 4-Pi-Zeeman split NQR spectroscopy of Cl-35 in single-crystal para-chloroanilinium trichloroacetate—crystal-structure and activation-energy for the bleaching-out process. *Ber. Bunsenges. Phys. Chem.* 91:1158–1166.

Matsui, S., Kose, K., and Inouye, T. 1990. An NQR imaging experiment on a disordered solid. *J. Magn. Reson.* 88:186–191.

Millar, J. M., Thayer, A. M., Bielecki, A., Zax, D. B., and Pines, A. 1985. Zero field NMR and NQR with selective pulses and indirect detection. *J. Chem. Phys.* 83:934–938.

Morino, Y. and Toyama, M. 1961. Zeeman effect of the nuclear quadrupole resonance spectrum in crystalline powder. *J. Chem. Phys.* 35:1289–1296.

Nickel, P. and Kimmich, R. 1995. 2D exchange NQR spectroscopy. *J. Mol. Struct.* 345:253–264.

Nickel, P., Rommel, E., Kimmich, R., and Pusiol, D. 1991. Two-dimensional projection/reconstruction rotating-frame NQR imaging (ρ NQRI). *Chem. Phys Lett.* 183:183–186.

Nickel, P., Robert, H., Kimmich R., and Pusiol D. 1994. NQR method for stress and pressure imaging. *J. Magn. Reson.* A111:191–194.

Noack, F. 1986. NMR field cycling spectroscopy: Principles and applications. *Prog. NMR Spectrosc.* 18:171–276.

Pound, R. V. 1951. Nuclear spin relaxation times in a single crystal of LiF. *Phys. Rev.* 81:156–156.

Pratt, J. C., Raganuthan, P., and McDowell, C. A. 1975. Transient response of a quadrupolar system in zero applied field. *J. Magn. Reson.* 20:313–327.

Raghavan, P. 1989. Table of Nuclear Moments. *Atomic Data and Nuclear Data Tables* 42:189–291.

Ramachandran, R. and Oldfield, E. 1984. Two dimensional Zeeman nuclear quadrupole resonance spectroscopy. *J. Chem. Phys.* 80:674–677.

Robert, H. and Pusiol, D. 1996a. Fast ρ -NQR imaging with slice selection. *Z. Naturforsch.* 51a:353–356.

Robert, H. and Pusiol, D. 1996b. Slice selection in NQR spatially resolved spectroscopy. *J. Magn. Reson.* A118:279–281.

Robert, H. and Pusiol, D. 1997. Two dimensional rotating-frame NQR imaging. *J. Magn. Reson.* 127:109–114.

Robert, H., Pussiol, D., Rommel, E., and Kimmich R. 1994. On the reconstruction of NQR nutation spectra in solid with powder geometry. *Z. Naturforsch.* 49a:35–41.

Robert, H., Minuzzi, A., and Pusiol, D. 1996. A fast method for the spatial encoding in rotating-frame NQR imaging. *J. Magn. Reson.* A118:189–194.

Rommel, E., Kimmich, R., and Pusiol, D. 1991a. Spectroscopic rotating-frame NQR imaging (ρ NQRI) using surface coils. *Meas. Sci. Technol.* 2:866–871.

Rommel, E., Nickel, P., Kimmich, R., and Pusiol, D. 1991b. Rotating-frame NQR imaging. *J. Magn. Reson.* 91:630–636.

Rommel, E., Nickel, P., Rohmer, F., Kimmich, R., Gonzales, C., and Pusiol, D. 1992a. Two dimensional exchange spectroscopy using pure NQR. *Z. Naturforsch.* 47a:382–388.

Rommel, E., Kimmich, R., Robert, H., and Pusiol, D. 1992b. A reconstruction algorithm for rotating-frame NQR imaging (ρ NQRI) of solids with powder geometry. *Meas. Sci. Technol.* 3:446–450.

Selinger, J., Zagar, V., and Blinc, R. 1994. ^{1}H-^{14}N nuclear quadrupole double resonance with multiple frequency sweeps. *Z. Naturforsch.* 49a:31–34.

Sen, K. D. and Narasimhan, P. T. 1974. Polarizabilities and antishielding factors in crystals. *In* Advances in Nuclear Quadrupole Resonance, Vol. 1 (J. A. S. Smith, ed.). Heyden and Sons, London.

Sinyavski, N., Ostafin, M., and Mackowiak, M. 1996. Rapid measurement of nutation NQR spectra in powder using an rf pulse train. *Z. Naturforsch.* 51a:363–367.

Stephenson, D. S. 1988. Linear prediction and maximum entropy methods in NMR spectroscopy. *Prog. Nucl. Magn. Reson Spectrosc.* 20:516–626.

Sternheimer, R. M. 1963. Quadrupole antishielding factors of ions. *Phys Rev.* 130:1423–1424.

Sternheimer, R. M. 1966. Shielding and antishielding effects for various ions and atomic systems. *Phys. Rev.* 146:140.

Sternheimer, R. M. and Peierls, R. F. 1971. Quadrupole antishielding factor and the nuclear quadrupole moments of several alkali isotopes. *Phys. Rev.* A3:837–848.

Thayer, A. M and Pines, A. 1987. Zero field NMR. *Acc. Chem. Res.* 20:47–53.

TonThat, D. M. and Clarke, J. 1996. Direct current superconducting quantum interference device spectrometer for pulsed magnetic resonance and nuclear quadrupole resonance at frequencies up to 5 MHz. *Rev. Sci. Instrum.* 67:2890–2893.

TonThat, D. M., Ziegeweid, M., Song, Y. Q., Munson, E. G., Applelt, S., Pines, A., and Clarke, J. 1997. SQUID detected NMR of laser polarized Xenon at 4.2 K and at frequencies down to 200 Hz. *Chem. Phys. Lett.* 272:245–249.

Werner-Zwanzinger, U., Zeigeweid, M., Black, B., and Pines, A. 1994. Nitrogen-14 SQUID NQR of L-Ala-L-His and of serine. *Z. Naturforsch.* 49a:1188–1192.

Wikner, E. G. and Das, T. P. 1958. Antishielding of nuclear quadrupole moment of heavy ions. *Phys. Rev.* 109:360–368.

Yesinowski, J. P., Buess, M. L., Garroway, A. N., Zeigeweid, M., and Pines A. 1995. Detection of ^{14}N and ^{35}Cl in cocaine base and hydrochloride using NQR, NMR and SQUID techniques. *Anal. Chem.* 67:2256–2263.

Yu, H.-Y. 1991. Studies of NQR spectroscopy for spin-5/2 systems. M. S. Thesis, SUNY at Stony Brook.

KEY REFERENCES

Abragam, 1961. See above.

Still the best comprehensive text on NMR and NQR theory.

Das, T. P. and Hahn, E. L. 1958. Nuclear Quadrupole Resonance Spectroscopy. Academic Press, New York.

A more specialized review of the theory of NQR: old but still indispensible.

Harbison et al., 1989. See above.

The first true zero-field multidimensional NMR experiment.

Robert and Pusiol, 1997. See above.

A good source for references on NQR imaging.

APPENDIX: GLOSSARY OF TERMS AND SYMBOLS

eq	the most distinct principal value of the electric field gradient at the nucleus
eQ	electric quadrupole moment of the nucleus
h	Planck's constant
H_1	applied radiofrequency field in tesla
I	nuclear spin
T_1, T_2	longitudinal and transverse relaxation times
γ	gyromagnetic ratio of the nucleus
γ_∞	Sternheimer antishielding factor of the atom or ion
η	asymmetry parameter of the electric field gradient at the nucleus
ν_Q	NQR resonance frequency in Hz
ω_Q	NQR resonance frequency in rad s^{-1}
ω_R	intrinsic precession frequency in rad s^{-1}; $\omega_R = \gamma H_1$

BRUNO HERREROS
GERARD S. HARBISON
University of Nebraska
Lincoln, Nebraska

ELECTRON PARAMAGNETIC RESONANCE SPECTROSCOPY

INTRODUCTION

Electron paramagnetic resonance (EPR) spectroscopy, also called electron spin resonance (ESR) or electron magnetic resonance (EMR), measures the absorption of electromagnetic energy by a paramagnetic center with one or more unpaired electrons (Atherton, 1993; Weil et al., 1994). In the presence of a magnetic field, the degeneracy of the electron spin energy levels is removed and transitions between the energy levels can be caused to occur by supplying energy. When the energy of the microwave photons equals the separation between the energy levels of the unpaired electrons, there is absorption of energy by the sample and the system is said to be at "resonance" (Fig. 1). The fundamental equation that describes the experiment for a paramagnetic center with one unpaired electron is $h\nu = g\beta B$, where h is Planck's constant, ν is the frequency of the microwaves, g is a characteristic of the sample, β is the Bohr magneton, and B is the strength of the magnetic field in which the sample is placed. Typically the experiment is performed with magnetic fields such that the energies are in the microwave region. If the paramagnetic center is tumbling rapidly in solution, g is a scalar quantity. When the paramagnetic center is immobilized, as in solid samples, most samples exhibit g anisotropy and g is then represented by a matrix. Hyperfine splitting of the signal occurs due to interaction with nuclear spins and can therefore be used to identify the number and types of nuclear spins in proximity to the paramagnetic center.

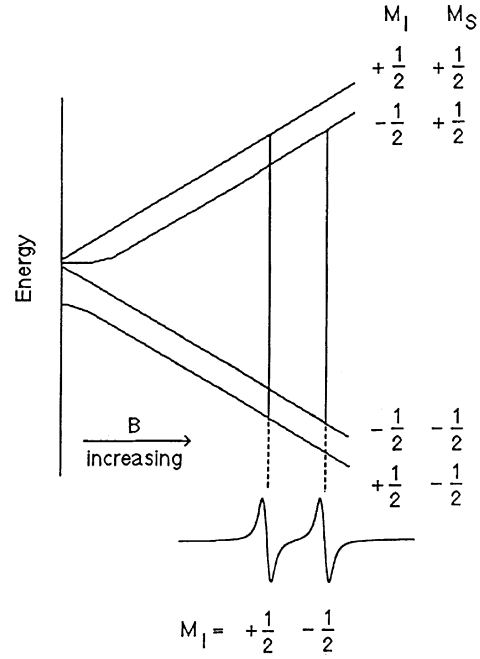


Figure 1. Energy-level splitting diagram for an unpaired electron in the presence of a magnetic field, interacting with one nucleus with $I = 1/2$. The energy separation between the two levels for the unpaired electron is linearly proportional to the magnetic field strength, B. Coupling to the nuclear spin splits each electron spin energy level into two. Transitions between the two electron energy levels are stimulated by microwave radiation when $h\nu = g\beta B$, where β is the electron Bohr magneton. Only transitions with $\Delta m_S = \pm 1, \Delta m_I = 0$ are allowed, so interaction with nuclear spins causes the signal to split into $2nI + 1$ lines, where n is the number of equivalent nuclei with spin I. The microwave magnetic field, B_1 is perpendicular to B. If the line shape is determined by relaxation, it is Lorentzian. The customary display in EPR spectroscopy is the first derivative of the absorption line shape. Sometimes the dispersion signal is detected instead of the absorption because the dispersion signal saturates less readily than the absorption. The dispersion signal is related to the absorption signal by the Kramers-Kronig transform.

The information content of EPR arises from the ability to detect a signal and from characteristics of the signal, including integrated intensity, hyperfine splitting by nuclear spins, *g* value, line shape, and electron spin relaxation time. The characteristics of the signal may depend on environmental factors, including temperature, pressure, solvent, and other chemical species. Various types of EPR experiments can optimize information concerning these observables. The following list includes some commonly asked materials science questions that one might seek to answer based on these observables and some corresponding experimental design considerations. Detailed background information on the physical significance and theoretical basis for the observable parameters is provided in the standard texts cited (Poole, 1967, 1983; Pake and Estle, 1973; Eaton and Eaton, 1990a, 1997a; Weil et al., 1994). Some general information and practical considerations are given in this unit.

Whenever unpaired electrons are involved, EPR is potentially the best physical technique for studying the system. Historically, the majority of applications of EPR have been to the study of organic free radicals and transition metal complexes (Abragam and Bleaney, 1970; Swartz et al., 1972; Dalton, 1985; Pilbrow, 1990). Today, these applications continue, but in the context of biological systems, where the organic radicals are naturally occurring radicals, spin labels, and spin-trapped radicals, and the transition metals are in metalloproteins (Berliner, 1976, 1979; Eaton et al., 1998). Applications to the study of materials are extensive, but deserve more attention than they have had in the past. Recent studies include the use of EPR to monitor the age of archeological artifacts (Ikeya, 1993), characterize semiconductors and superconductors, measure the spatial distribution of radicals in processed polymers, monitor photochemical degradation of paints, and characterize the molecular structure of glasses (Rudowicz et al., 1998).

One might seek to answer the following types of questions.

1. Are there paramagnetic species in the sample?

The signal could be due to organic radicals, paramagnetic metal ions, defects in materials (Lancaster, 1967) such as dangling bonds or radiation damage, or paramagnetic species intentionally added as probes. The primary concern in this case would be the presence or absence of a signal. The search for a signal might include obtaining data over a range of temperatures, because some signals can only be detected at low temperatures and others are best detected near room temperature.

2. What is the concentration of paramagnetic species? Does the concentration of paramagnetic species change as a function of procedure used to prepare the material or of sample handling?

Spectra would need to be recorded in a quantitative fashion (Eaton and Eaton, 1980, 1990a, 1992, 1997a).

3. What is the nature of the paramagnetic species? If it is due to a paramagnetic metal, what metal? If it is due to an organic radical, what is the nature of the radical? Is more than one paramagnetic species present in the sample?

One would seek to obtain spectra that are as well resolved as possible. To examine weak interactions with nuclear spins that are too small compared with line widths to be resolved in the typical EPR spectrum, but which might help to identify the species that gives rise to the EPR signal, one might use techniques such as ENDOR (electron nuclear double resonance; Atherton, 1993) or ESEEM (electron spin echo envelope modulation; Dikanov and Tsvetkov, 1992).

4. Does the sample undergo phase transitions that change the environment of the paramagnetic center?

Experiments could be run as a function of temperature, pressure, or other parameters that cause the phase transition.

5. Are the paramagnetic centers isolated from each other or present in pairs, clusters, or higher aggregates?

EPR spectra are very sensitive to interactions between paramagnetic centers (Bencini and Gatteschi, 1990). Strong interactions are reflected in line shapes. Weaker interactions are reflected in relaxation times.

6. What is the mobility of the paramagnetic centers?

EPR spectra are sensitive to motion on the time scale of anisotropies in resonance energies that arise from anisotropies in g values and/or electron-nuclear hyperfine interactions. This time scale typically is microseconds to nanoseconds, depending on the experiment. Thus, EPR spectra can be used as probes of local mobility and microviscosity (Berliner and Reuben, 1989).

7. Are the paramagnetic centers uniformly distributed through the sample or spatially localized?

This issue would be best addressed with an EPR imaging experiment (Eaton and Eaton, 1988a, 1990b, 1991, 1993a, 1995a, 1996b, 1999a; Eaton et al., 1991; Sueki et al., 1990).

Bulk magnetic susceptibility (GENERATION AND MEASUREMENT OF MAGNETIC FIELDS & MAGNETIC MOMENT AND MAGNETIZATION) also can be used to study systems with unpaired electrons, and can be used to determine the nature of the interactions between spins in concentrated spin systems. Rather large samples are required for bulk magnetic susceptibility measurements, whereas EPR typically is limited to rather small samples. Bulk susceptibility and EPR are complementary techniques. EPR has the great advantage that it uniquely measures unpaired electrons in low concentrations, and in fact is most informative for systems that are magnetically dilute. It is particularly powerful for identification of paramagnetic species present in the sample and characterization of the environment of the paramagnetic species.

In this unit, we seek to provide enough information so that a potential user can determine whether EPR is likely to be informative for a particular type of sample, and which type of EPR experiment would most likely be useful.

PRINCIPLES OF EPR

The fundamental principles of EPR are similar to those of nuclear magnetic resonance (NMR; Carrington and

McLachlan, 1967) and magnetic resonance imaging (MRI), which are described elsewhere in this volume (NUCLEAR MAGNETIC RESONANCE IMAGING). However, several major differences between the properties of unpaired electron spins and of nuclear spins result in substantial differences between NMR and EPR spectroscopy. First, the magnetogyric ratio of the electron is 658 times that of the proton, so for the same magnetic field the frequency for electron spin resonance is 658 times the frequency for proton resonance. Many EPR spectrometers operate in the 9 to 9.5 GHz frequency range (called "X-band"), which corresponds to resonance for an organic radical ($g \sim 2$) at a magnetic field of 3200 to 3400 gauss [G; 1 G=10^{-4} tesla (T)]. Second, electron spins couple to the electron orbital angular momentum, resulting in shorter relaxation times than are observed in NMR. Electron spin relaxation times are strongly dependent on the type of paramagnetic centers. For example, typical room temperature electron spin relaxation times range from $\sim 10^{-6}$ s for organic radicals to as short as 10^{-12} s for low-spin Fe(III). Third, short relaxation times can result in very broad lines. Even for organic radicals with relatively long relaxation times, EPR lines frequently are relatively broad due to unresolved coupling to neighboring nuclear spins. Line widths for detectable EPR signals range from fractions of a gauss to tens or hundreds of gauss. Since 1 G corresponds to about 2.8 MHz, these EPR line widths correspond to 10^6 to 10^8 Hz, which is much greater than NMR line widths. Fourth, coupling to the electron orbital angular momentum also results in larger spectral dispersion for EPR than for NMR. A room temperature spectrum of an organic radical might extend over 10 to 100 G. However, the spectrum for a transition metal ion might extend over one hundred to several thousand gauss. Finally, the larger magnetic moment of the unpaired electron also results in larger spin–spin interactions, so high-resolution spectra require lower concentrations for EPR than for NMR. These differences make the EPR measurement much more technically challenging than the NMR measurement. For example, pulsed Fourier transform NMR provides major advances in sensitivity; however pulsed Fourier transform spectroscopy has more restricted applicability in EPR because a pulse of finite width cannot excite the full spectrum for many types of paramagnetic samples (Kevan and Schwartz, 1979; Kevan and Bowman, 1990).

In a typical EPR experiment the sample tube is placed in a structure that is called a resonator. The most common type of resonator is a rectangular cavity. This name is used because microwaves from a source are used to set up a standing wave pattern. In order to set up the standing wave pattern, the resonator must have dimensions appropriate for a relatively narrow range of microwave frequencies. Although the use of a resonator enhances the signal-to-noise (S/N) ratio for the experiment, it requires that the microwave frequency be held approximately constant and that the magnetic field be swept to achieve resonance. The spectrometer detects the microwave energy reflected from the resonator. When spins are at resonance, energy is absorbed from the standing wave pattern, the reflected energy decreases, and a signal is detected. In a typical continuous wave (CW) EPR experiment, the magnetic field is swept and the change in reflected energy (the EPR signal) is recorded. To further improve S/N, the magnetic field is usually modulated (100 kHz modulation is commonly used), and the EPR signal is detected using a phase-sensitive detector at this modulation frequency. This detection scheme produces a first derivative of the EPR absorption signal. A sketch of how magnetic-field modulation results in a derivative display is found in Eaton and Eaton, (1990a, 1997a). Because taking a derivative is a form of resolution enhancement and EPR signals frequently are broad, it is advantageous for many types of samples to work directly with the first-derivative display. Thus, the traditional EPR spectrum is a plot of the first derivative of the EPR absorption as a function of magnetic field. A block diagram for a CW EPR spectrometer is given in Figure 2.

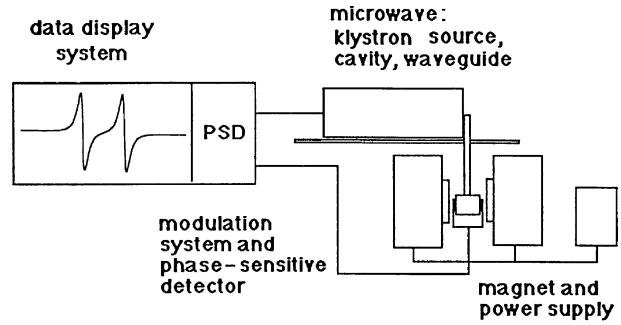


Figure 2. Block diagram of an EPR spectrometer. The fundamental modules of a CW EPR spectrometer include: the microwave system comprising source, resonator, and detector; the magnet system comprising power supply and field controller; the magnetic field modulation and phase-sensitive detection system; and the data display and manipulation system. Each of these subsystems can be largely independent of the others. In modern spectrometers there is a trend toward increasing integration of these units via interfaces to a computer. The computer is then the controller of the spectrometer and also provides the data display and manipulation. In a pulse system a pulse timing unit is added and magnetic field modulation is not used.

Types of EPR Experiments

One can list over 100 separate types of EPR spectra. One such list is in an issue of the *EPR Newsletter* (Eaton and Eaton, 1997b). For this unit we focus primarily on experiments that can be done with commercial spectrometers. For the near future, those are the experiments that are most likely to be routinely available to materials scientists. In the following paragraphs we outline some of the more common types of experiments and indicate what characteristics of the sample might indicate the desirability of a particular type of experiment. Experimental details are reserved for a later section.

1. CW Experiments at X-band with a Rectangular Resonator

This category represents the vast majority of experiments currently performed, whether for routine analysis or for research, in materials or in biomedical areas. In this experiment the microwaves are on continuously and the magnetic field is scanned to achieve resonance. This is the experiment that was described in the introductory

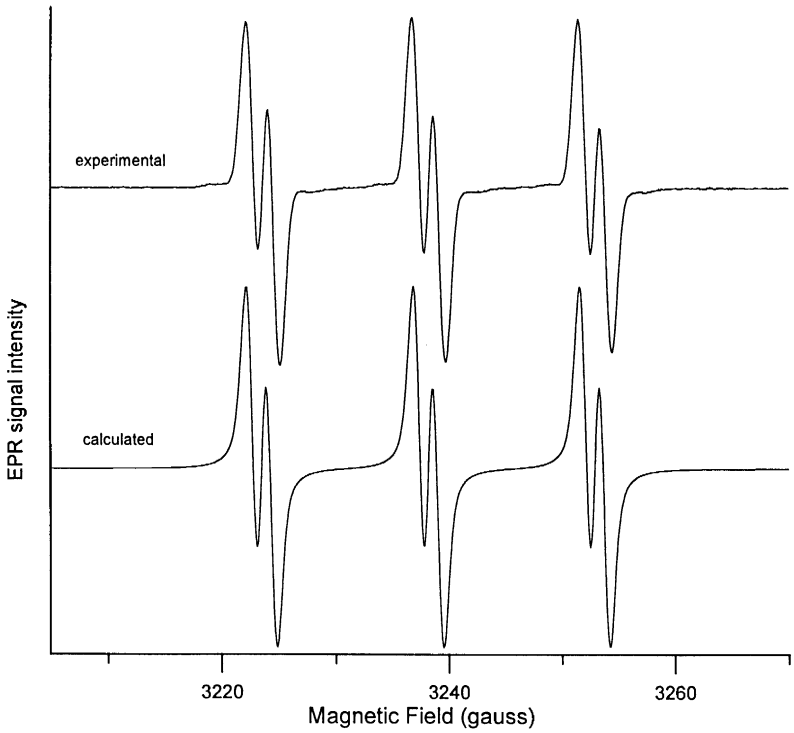


Figure 3. Room temperature CW spectrum of a nitroxyl radical obtained with a microwave frequency of 9.09 GHz. The major three-line splitting is due to coupling to the nitroxyl nitrogen ($I = 1$) and the small doublet splitting is due to interaction with a unique proton ($I = 1/2$). The calculated spectrum was obtained with hyperfine coupling to the nitrogen of 14.7 G, coupling to the proton of 1.6 G, and a Gaussian line width of 1.1 G due to additional unresolved proton hyperfine coupling. The g value (2.0056) is shifted from 2.0023 because of spin-orbit coupling involving the nitrogen heteroatom.

paragraphs and which will be the focus of much of this unit. Examples of spectra obtained at X-band in fluid solution and for an immobilized sample are shown in Figures 3 and 4. The parameters obtained by simulation of these spectra are indicated in the figure captions.

2. CW Experiments at Frequencies Other than X-band

Until relatively recently, almost all EPR was performed at approximately 9 GHz (X-band). However, experiments at lower frequencies can be advantageous in dealing with larger samples and samples composed of materials that absorb microwaves strongly at X-band, as well as for resolving nitrogen hyperfine splitting in certain Cu(II) complexes (Hyde and Froncisz, 1982).

Experiments at higher microwave frequencies are advantageous in resolving signals from species with small g-value differences (Krinichnyi, 1995; Eaton and Eaton, 1993c, 1999b), in obtaining signals from

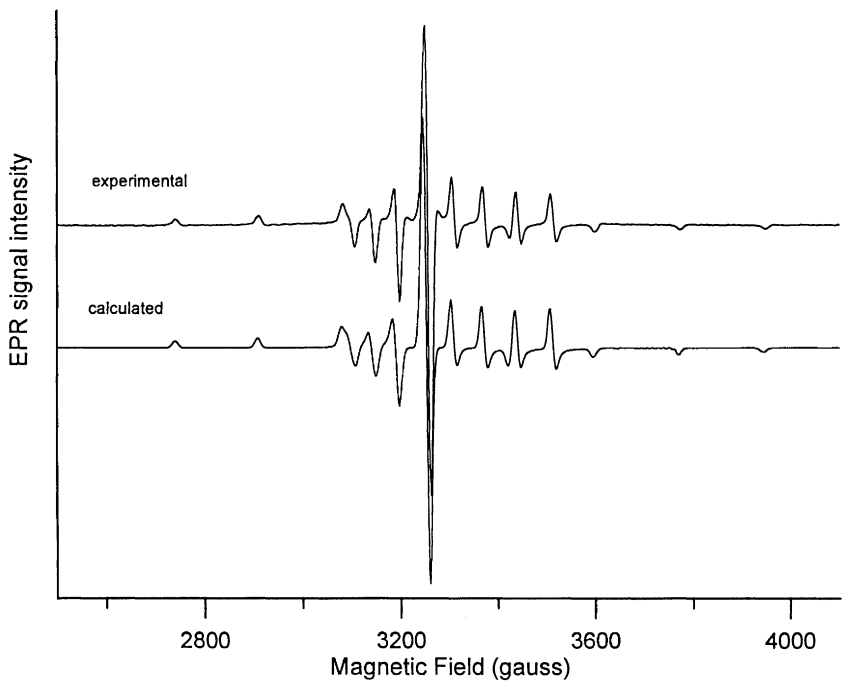


Figure 4. Room temperature CW spectrum of a vanadyl porphyrin in a 2:1 toluene:chloroform glass at 100 K, obtained with a microwave frequency of 9.200 GHz. The anisotropy of the g and hyperfine values results in different resonant conditions for the molecules as a function of orientation with respect to the external field. For this random distribution of orientations, the EPR absorption signal extends from about 2750 to 3950 G. The first derivative display emphasizes regions of the spectrum where there are more rapid changes in slope, so "peaks" in the first derivative curve occur at extrema in the powder distribution. These extrema define the values of g and A along the principal axes of the magnetic tensors. The hyperfine splitting into eight approximately equally spaced lines is due to the nuclear spin ($I = 3.5$) of the vanadium nucleus. The hyperfine splitting is greater along the normal to the porphyrin plane (z axis) than in the perpendicular plane. The calculated spectrum was obtained with $g_x = 1.984$, $g_y = 1.981$, $g_z = 1.965$, $A_x = 55 \times 10^{-4}$ cm^{-1}, $A_y = 53 \times 10^{-4}$ cm^{-1}, and $A_z = 158 \times 10^{-4}$ cm^{-1}.

paramagnetic centers with more than one unpaired electron, and in analyzing magnetic interactions in magnetically concentrated samples (Date, 1983). In principle, experiments at higher microwave frequency should have higher sensitivity than experiments at X-band, due to the larger difference in the Boltzmann populations of the electron spin states. Sensitivity varies with $(\text{frequency})^{11/4}$ if the microwave magnetic field at the sample is kept constant and the size of the resonator is scaled inversely with frequency (Rinard et al., 1999). This is a realistic projection for very small samples, as might occur in some materials research applications. To achieve the expected increase in sensitivity (Rinard et al., 1999) will require substantial engineering improvements in the sources, resonators and detectors used at higher frequency.

Computer simulations of complicated spectra as a function of microwave frequency provide a much more stringent test of the parameters than can be obtained at a single frequency.

3. Electron-Nuclear Double Resonance (ENDOR)

Due to the large line widths of many EPR signals, it is difficult to resolve couplings to nuclear spins. However, these couplings often are key to identifying the paramagnetic center that gives rise to the EPR signal. By simultaneously applying radio frequency (RF) and microwave frequency energies, one can achieve double resonance of both the electron and nuclear spins. The line widths of the signals in this experiment are much narrower than typical EPR line widths and the total number of lines in the spectrum is smaller, so it is easier to identify the nuclear spins that are interacting with the electron spin in an ENDOR spectrum (Box, 1977; Schweiger, 1982; Atherton, 1993; Piekara-Sady and Kispert, 1994; Goslar et al., 1994) than in an experiment without the RF frequency—i.e., a "normal" CW spectrum.

4. Pulsed/Fourier Transform EPR

Measurements of relaxation times via pulsed or time-domain techniques typically are much more accurate than continuous wave measurements. Pulse sequences also can be tailored to obtain detailed information about the spins. However, the much shorter relaxation times for electron spins than for nuclear spins limit pulse sequences to shorter times than can be used in pulsed NMR (Kevan and Schwartz, 1979). A technique referred to as electron spin echo envelope modulation (ESEEM) is particularly useful in characterizing interacting nuclear spins, including applications in materials (Kevan and Bowman, 1990). Interaction between inequivalent unpaired electrons can cause changes in relaxation times that depend on distance between the spins (Eaton and Eaton, 1996a).

5. Experiments with Resonators Other than the Rectangular Resonator

New lumped-circuit resonators, such as the loop-gap resonator (LGR) and others designed on this principle (Hyde and Froncisz, 1986) can be adapted to a specific experiment to optimize aspects of the measurement (Rinard et al., 1993, 1994, 1996a, 1996b). These structures are particularly important for pulsed experiments and for experiments at frequencies lower than X-band.

PRACTICAL ASPECTS OF THE CW METHOD

The selection of the microwave power and modulation amplitude for recording CW EPR signals is very important in obtaining reliable results. We therefore give a brief discussion of issues related to the selection of these parameters. Greater detail concerning the selection of parameters can be found in Eaton and Eaton (1990a, 1997a).

Microwave Power

If microwave power is absorbed by the sample at a rate that is faster than the sample can dissipate energy to the lattice (i.e., faster than the electron spin relaxation rate), the EPR signal will not be proportional to the number of spins in the sample. This condition is called saturation. The power that can be used without saturating the sample depends on the relaxation rate. To avoid saturation of spectra, much lower powers must be used for samples that have long relaxation times, such as organic radicals, than for samples with much shorter relaxation times, such as transition metals. Relaxation times can range over more than 18 orders of magnitude, from hours to 10^{-14} s, with a wide range of temperature dependence, from T^9 to temperature-independent, and a wide variety of magnetic field dependence and/or dependence on position in the EPR spectrum (Du et al., 1995). One can test for saturation by recording signal amplitude at a series of microwave powers, P, and plotting signal amplitude as a function of $\sqrt{P}$. This plot is called a power-saturation curve. The point where the signal dependence on $\sqrt{P}$ begins to deviate from linearity is called the onset of saturation. Data must be recorded at a power level below the onset of saturation in order to use integrated data to determine the spin concentration in the sample. Higher powers also cause line broadening, so powers should be kept below the onset of saturation if data analysis is based on line shape information. Above the onset of saturation, the signal intensity typically increases with increasing power and goes through a maximum. If the primary goal is to get maximum S/N at the expense of line shape or quantitation, then one can operate at the power level that gives maximum signal amplitude, but under such conditions quantitation of spins can be erroneous.

Modulation Amplitude

Provided that the modulation amplitude is less than 1/10 of the peak-to-peak line width of the first-derivative EPR signal to be recorded, increasing modulation amplitude improves S/N without distorting line shape. However, as the modulation amplitude is increased further, it causes broadening and distortion (Poole, 1967). If the key information to be obtained from the spectrum is based on the line shape, care must be taken in selecting the modulation

amplitude. Unlike the problems that occur when too high a microwave power is used to record the spectrum, increasing the modulation amplitude does not change the linear relationship between integrated signal intensity and modulation amplitude. Thus, if the primary information to be obtained from a spectrum is the integrated signal intensity (to determine spin concentration), it can be useful to increase modulation amplitude, and thereby improve S/N, at the expense of some line shape distortion. The maximum spectral amplitude, and therefore the best S/N, occurs at a peak-to-peak modulation amplitude ~2 times the peak-to-peak width of the derivative EPR spectrum.

Sensitivity

The most important issue concerning practical applications of EPR is sensitivity (Eaton and Eaton 1980, 1992; Rinard et al., 1999). In many analytical methods, one can describe sensitivity simply in terms of minimum detectable concentrations. The problem is more complicated for EPR because of the wide range of line widths and large differences in power saturation among species of interest. Vendor literature cites sensitivity as about $0.8 \times 10^{10}/\Delta B$ spins/G, where ΔB is the line width of the EPR signal. This is a statement of the minimum number of detectable spins with a S/N of 1. It is based on the assumption that spectra can be recorded at a power of 200 mW, with a modulation amplitude of 8 G, and that there is no hyperfine splitting that divides intensity of the signal into multiple resolved lines. To apply this statement to estimate sensitivity for another sample requires taking account of the actual conditions appropriate for a particular sample. Those conditions depend upon the information that one requires from the spectra. One can consider two types of conditions as described below.

Case 1. Selection of microwave power, modulation amplitude, and time constant of the signal detection system that provide undistorted line shapes. This mode of operation is crucial if one wishes to obtain information concerning mobility of the paramagnetic center or partially resolved hyperfine interactions. Typically this would also require a S/N significantly greater than 1. This case might be called "minimum number of detectable spins with desired spectral information." It would require the use of a microwave power that did not cause saturation and a modulation amplitude that did not cause line broadening.

Case 2. Selection of microwave power, modulation amplitude, and time constant of the signal detection system that provide the maximum signal amplitude, at the expense of line shape information. This mode of operation might be selected, for example, in some spin trapping experiments, where one seeks to determine whether an EPR signal is present and one can obtain adequate information for identification of the species from observed hyperfine couplings even in the presence of line shape distortions that come from power broadening or overmodulation. This case might be called "minimum number of detectable spins."

To correct the minimum number of detectable spins for microwave power, multiply $0.8 \times 10^{10}/\Delta B$ by

$$\sqrt{\frac{P_{sample}}{200}} \tag{1}$$

where P_{sample} is the power incident on the critically coupled resonator, in mW. To correct for changes in modulation amplitude (MA), multiply by MA_{sample} (in gauss)/8 (G). For samples with narrow lines and long relaxation times, the minimum number of detectable spins may be higher than $0.8 \times 10^{10}/\Delta B$ by several orders of magnitude.

Calibration

Quality assurance requires a major effort in EPR (Eaton and Eaton 1980, 1992). Because the properties of the sample affect the characteristics of the resonator, and because the microwave magnetic field varies over the sample, selecting a standard for quantitation of the number of spins requires understanding of the spin system and of the spectrometer [Poole, 1967, 1983; Wilmshurst, 1968 (see especially THERMAL ANALYSIS); Alger 1968; Czoch and Francik, 1989]. In general, the standard should be as similar as possible to the sample to be measured, in the same solvent or solid host, and in a tube that is as close as possible to the same size. The EPR tube has to be treated as a volumetric flask, and calibrated accordingly. Determining the number of spins in the sample requires subtraction of any background signal, and then integrating twice (derivative → absorption → area). If there is background slope, the double integration may not be very accurate. All of this requires that the EPR spectrum be in digital form. If the line shape is the same for the sample and the reference, then peak heights can be used instead of integrated areas. This is sometimes difficult to ensure, especially in the wings of the spectrum, which may constitute a lot of the area under the spectrum. However, for fairly widespread applications, such as spin trapping and spin labeling, reasonable estimates of spin concentration can be obtained by peak-height comparison with gravimetrically prepared standard solutions. Very noisy spectra with well defined peak positions might best be quantitated using simulated spectra that best fit the experimental spectra.

Another important calibration is of magnetic field scan. The best standard is an NMR gaussmeter, but note that the magnetic field varies across the pole face, and the gaussmeter probe has to be as close to the sample as possible—ideally inside the resonator. A few carefully-measured samples that can be prepared reproducibly have been reported in the literature (Eaton and Eaton 1980, 1992) and are good secondary standards for magnetic field calibrations.

The magnetic field control device on electromagnet-based EPR spectrometers is a Hall probe. These are inherently not very accurate devices, but EPR field control units contain circuitry to correct for their inaccuracies. Recent Bruker field controllers, for example, use a ROM containing the characteristics of the specific Hall probe with which it is matched. Such systems provide sufficient accuracy for most EPR measurements.

METHOD AUTOMATION

Early EPR spectrometers did not include computers. Many, but not all of those systems, have been retrofitted with computers for data acquisition. Historically, the operator spent full time at the EPR spectrometer console while the spectrometer was being operated. In the next generation, a computer was attached to the EPR spectrometer for data acquisition, but the spectrometer was still substantially manually operated. Even when computers were built into the spectrometer, there was a control console at which the operator sat. The newer spectrometers are operated from the computer, and have a console of electronic modules that control various functions of the magnet, microwave system, and data acquisition. The latest pulsed EPR spectrometers are able to execute long (including overnight) multipulse experiments without operator intervention. The most recent cryogenic temperature-control systems also can be programmed and can control the cryogen valve in addition to the heater. There is one research lab that has highly automated ENDOR spectrometers (Corradi et al, 1991). Except for these examples, EPR has not been automated to a significant degree.

For many samples, the sample positioning is sufficiently critical, as is the resonator tuning and coupling, that automation of sample changing for unattended operation is less feasible than for many other forms of spectroscopy, including NMR. Automation would be feasible for multiple samples that shared the same geometry, dielectric constant, and microwave loss factor, if they could be run near room temperature. The only currently available spectrometer with an automatic sample changer is one designed by Bruker for dosimetry. This system is customized to handle a particular style of alanine dosimeter and includes software to quantitate the radical signal and to prepare reports. References to computer use for acquisition and interpretation are given by Vancamp and Heiss (1981) and Kirste (1994).

DATA ANALYSIS AND INITIAL INTERPRETATION

The wide range of queries one may make of the spin system, listed elsewhere in this unit, makes the data analysis and interpretation diverse. An example of a spectrum of an organic radical in fluid solution is shown in Figure 3. Computer simulation of the spectrum gives the g value for the radical and the nuclear hyperfine splittings which can be used to identify the nuclear spins with the largest couplings to the unpaired electrons. An example of a spectrum of a vanadyl complex immobilized in 1:1 toluene:chloroform at 100 K is shown in Figure 4. The characteristic 8-line splitting patterns permit unambiguous assignment as a vanadium species ($I = 3.5$). Computer simulation of the spectrum gives the anisotropic components of the g and A values, which permits a more detailed characterization of the electronic structure of the paramagnetic center than could be obtained from the isotropic averages of these parameters observed in fluid solution.

As discussed above, under instrument calibration, a first stage in data analysis and interpretation is to determine how many spins are present, as well as the magnitude of line widths and hyperfine splittings. The importance of quantitation should not be underestimated. There are many reports in the literature of detailed analysis of a signal that accounts for only a very small fraction of the spins in the sample, the others having been overlooked. The problem is particularly acute in samples that contain both sharp and broad signals. The first-derivative display tends to emphasize the sharp signals, and it is easy to overlook broad signals unless spectra are recorded under a wide range of experimental conditions.

The software available as part of the spectrometer systems is increasingly powerful and versatile. Additional software is available, some commercially and some from individual labs (easily locatable via the EPR Society software exchange; *http://www.ierc.scs.uiuc.edu*), for simulation of more specialized spin systems. The understanding of many spin systems is now at the stage where a full simulation of the experimental line shape is a necessary step in interpreting an EPR spectrum. For $S = 1/2$ organic radicals in fluid solution, one should be able to fit spectra within experimental error if the radical has been correctly identified. However, there will always remain important problems for which the key is to understand the spin system, and for these systems simulation may push the state of the art. For example, for many high-spin Fe(III) systems, there is little information about zero-field splitting (ZFS) terms, so one does not even know which transitions should be included in the simulation.

In pulsed EPR measurements, the first step after quantitation is to determine the relaxation times, and, if relevant, the ESEEM frequencies in the echo decay by Fourier transformation of the time-domain signal. Software for these analyses and for interpretation of the results is still largely resident within individual research groups, especially for newly evolved experiments.

SAMPLE PREPARATION

Samples for study by EPR can be solid, dissolved solid, liquid, or, less frequently, gas. In the gas phase the electron spin couples to the rotational angular momentum, so EPR spectroscopy has been used primarily for very small molecules. Thus, the focus of this unit is on samples in solid or liquid phases. For most EPR resonators at X-band or lower frequency, the sample is placed in a 4-mm outside diameter (o.d.) quartz tube. Usually, synthetic fused silica is used to avoid paramagnetic metal impurities. The weaker the sample signal, the higher the quality of quartz required. Pyrex or Kimax (trade names of Corning and Kimble) have very strong EPR signals.

Often, oxygen must be removed from the sample because the sample reacts with oxygen, or because the paramagnetic O_2 broadens EPR signals by Heisenberg exchange during collisions. The oxygen broadening is a problem for high-resolution spectra of organic radicals in solution and for the study of carbonaceous materials. This problem is so severe that one can turn this around and use oxygen broadening as a measure of oxygen concentration. EPR oximetry is a powerful analytical tool (Hyde and Subczynski, 1989).

For some problems one would want to study the concentrated solid, which may not exist in any other form. Maybe one wants to examine the spin-spin interaction as a function of temperature. In this case, one would put the solid directly into the EPR tube, either as a powder or as a single crystal, which might be oriented. There are a number of caveats for these types of samples. First, one must avoid overloading the resonator. More sample is not always better. With a "standard" TE_{102} (transverse electric 102 mode cavity) rectangular resonator (which until recently was the most common EPR resonator), one cannot use more than $\sim 10^{17}$ spins (Goldberg and Crowe, 1977). If the number of spins is too high, one violates the criterion that the energy absorption on resonance is a small perturbation to the energy reflected from the resonator. This is not an uncommon problem, especially with "unknown" samples, which might contain, for example, large quantities of iron oxide.

A similar problem occurs with samples that are lossy—i.e., that absorb microwaves other than by magnetic resonance (Dalal et al., 1981). Water is a prime example. For aqueous samples, and especially biological samples, there are special flat cells to keep the lossy material in a nodal plane of the microwave field in the resonator. There are special resonators to optimize the S/N for lossy samples. Usually, these operate with a TM (transverse magnetic) mode rather than a TE mode. Lower frequency can also be advantageous for lossy samples. There are many examples of lossy materials commonly encountered in the study of materials. Carbonaceous materials often have highly conducting regions. Some semiconductors are sufficiently conducting that only small amounts of sample can be put in the resonator without reducing the Q (the resonator quality factor) so much that the spectrometer cannot operate.

These problems are obvious to the experienced operator, because general-purpose commercial EPR spectrometers provide a readout of the microwave power as a function of frequency during the "tune" mode of setting up the measurement. The power absorbed by the resonator appears as a "dip" in the power versus frequency display. The broader the dip, the lower the resonator Q. One useful definition of Q is the resonant frequency divided by the half-power bandwidth of the resonator, $Q = \nu/\Delta\nu$. A lossy or conducting sample lowers Q, which is evident to the operator as a broadening of the dip. If the sample has much effect on Q, care must be taken in quantitation of the EPR signal, since the signal amplitude is proportional to Q. Ideally, one would measure the Q for every sample if spin quantitation is a goal. No commercial spectrometer gives an accurate readout of Q, so the best approach is to make the Q effect of the samples under comparison as precisely the same as possible.

Magnetically concentrated solids can exhibit the full range of magnetic interactions, including antiferromagnetism, ferromagnetism, and ferrimagnetism, in addition to paramagnetism, many of which are amenable to study by EPR (Bencini and Gatteschi, 1990; Stevens, 1997). Because of the strength of magnetic interactions in magnetically concentrated solids, ferrimagnetic resonance and ferromagnetic resonance have become separate subfields of magnetic resonance. High-field EPR is especially valuable to studies of magnetically concentrated solids, because it is possible in some cases to achieve external magnetic fields on the order of or greater than the internal magnetic fields.

If the goal is the study of isolated spins, they need to be magnetically dilute. In practice, this means that the spin concentration in a doped solid or liquid or solid solution should be less than $\sim$1 mM ($\sim 6 \times 10^{17}$ spins per cm^3). For narrow-line spectra, such as defect centers in solids or organic radicals, either in liquid or solid solution, concentrations less than 1 mM are required to achieve minimum line width. For accurate measurement of relaxation times, concentrations have to be lower than those required for accurate measurement of line widths in CW spectra. For some materials, it is convenient to prepare a solution and cool it to obtain an immobilized sample. If this is done, it is important to select a solvent or solvent mixture that forms a glass, not crystals, when it is cooled. Crystallization excludes solute from the lattice and thereby generates regions of locally high concentrations, which can result in poor resolution of the spectra.

The discussion of lossy samples hints at part of the problem of preparing fluid solution samples. The solvent has to be nonlossy, or the sample has to be very small and carefully located. Some solvents are more lossy at given temperatures than others. Most solvents are nonlossy when frozen. Broadening by oxygen is usually not detectable in frozen solution line widths, but there still can be an impact on relaxation times. Furthermore, the paramagnetic O_2 also can yield an EPR signal in frozen solution (it is detectable in the gas phase and in the solid, but not in fluid solution), and can be mistaken for signals in the sample under study.

Care must be taken when examining powdered crystalline solids to get a true powder (random orientation) line shape. If the particles are not small enough they may preferentially orient when placed in the sample tube. At high magnetic fields, magnetically anisotropic solids can align in the magnetic field. Either of these effects can yield spectra that are representative of special crystal orientations, rather than a powder (spherical) average. The combination of these effects can dominate attempts to obtain spectra at high magnetic fields, such as at 95 GHz and above, where very small samples are needed and magnetic fields are strong enough to macroscopically align microcrystals.

Although experiments are more time consuming, very detailed information concerning the electronic structure of the paramagnetic center can be obtained by examining single crystals. In these samples, data are obtained as a function of the orientation of the crystal, which then provides a full mapping of the orientation dependence of the g-matrix and of the electron-nuclear hyperfine interaction (Byrn and Strouse, 1983; Weil et al., 1994).

SPECIMEN MODIFICATION

The issue of specimen modification is inherent in the discussion of sample preparation. EPR is appropriately described as a nondestructive technique. If the sample fits the EPR resonator, it can be recovered unchanged.

However, some of the detailed sample preparations discussed above do result in modification of the sample. Grinding to achieve a true powder average spectrum destroys the original crystal. Sometimes grinding itself introduces defects that yield EPR signals, and this could be a study in itself. Sometimes grinding exposes surfaces to reactants that convert the sample to something different. Similarly, dissolving a sample may irreversibly change it, due to reactivity or association or dissociation equilibria. Cooling or heating a single crystal sometimes causes phase transitions that shatter the crystal. Unfortunately freezing a solvent may shatter the quartz EPR tube, due to differences in coefficient of expansion, causing loss of the sample.

These potential problems notwithstanding, EPR is a nondestructive and noninvasive methodology for study of materials. Consider the alternatives to EPR imaging to determine spatial distribution of paramagnetic centers. Even using EPR to monitor the radical concentration, one would grind or etch away the surface of the sample, monitoring changes in the signal after each step of surface removal. The sample is destroyed in the process. With EPR imaging, the sample is still intact and can be used for a further study, e.g., additional irradiation or heat treatment (Sueki et al., 1995).

PROBLEMS

How to Decide Whether and How EPR Should be Used

What do you want to learn about the sample? Is EPR the right tool for the task? This section attempts to guide thinking about these questions.

Is there an EPR signal? The first thing we learn via EPR is whether the sample in a particular environment exhibits resonance at the selected frequency, magnetic field, temperature, microwave power, and time after a pulse. This is a nontrivial observation, and by itself may provide almost the full answer to a question about the material. One should be careful, however, about jumping to conclusions based on only qualitative EPR results. With care it is possible to find an EPR signal in almost any sample, but that signal may not even be relevant to the information desired about the sample. There are spins everywhere. Dirt and dust in the environment, even indoors, will yield an EPR signal, often with strong Mn(II) signals and large broad signals due to iron oxides. Quartz sample tubes and Dewars contain impurities that yield EPR signal close to $g = 2$ and slightly higher. Most often, these are "background" interferences that have to be subtracted for quantitative work.

What if the sample does not exhibit a nonbackground EPR signal? Does this mean that there are no unpaired spins in the sample? Definitely not! The experienced spectroscopist is more likely to find a signal than the novice, because of the extensive number of tradeoffs needed to optimize an EPR spectrum (sample positioning, microwave frequency, resonator coupling, magnetic field sweep, microwave power, modulation frequency and amplitude, and time constant). The sample might have relaxation times such that the EPR signal is not observable at the temperature selected (Eaton and Eaton 1990a). In general, relaxation times increase with decreasing temperature, and there will be some optimum temperature at which the relaxation time is sufficiently long that the lines are not immeasurably broad, but not so long that the line is saturated at all available microwave power levels.

Furthermore, the failure to observe an EPR signal could be a fundamental property of the spin system. If the electron spin, S, for the paramagnetic center of interest is half-integer, usually at least the transition between the $m_s = \pm 1/2$ levels will be observable at some temperature. However, for integer spin systems ($S = 1, 2, 3$) the transition may not be observable at all in the usual spectrometer configuration, even if the right frequency, field, temperature, and so forth are chosen (Abragam and Bleaney, 1970; Pilbrow, 1990). This is because such spin transitions are forbidden when the microwave magnetic field, B_1, is perpendicular to the external magnetic field, B_0, i.e., in the usual arrangement. For integer spin systems one needs B_1 parallel to B_0, and a special cavity is required. Fortunately, these are commercially available at X-band. Thus, if the question is "does this sample contain paramagnetic Ni(II)," a special parallel mode resonator is needed. An overview of EPR-detectable paramagnetic centers and corresponding parameters can be found in Eaton and Eaton (1990a).

Another reason for not observing an EPR signal, or at least the signal from the species of interest, is that spin-spin interaction resulted in a singlet-triplet separation such that only one of these states is significantly populated under the conditions of measurement. A classic example is dimeric Cu(II) complexes, which have singlet ground states, so the signal disappears as the sample is cooled to He temperatures. Similarly, in fullerene chemistry it was found that the EPR signal from C_{60}^{2-} is due to a thermally excited triplet state and could only be seen in a restricted temperature range (Trulove et al., 1995).

Even if the goal is to measure relaxation times with pulsed EPR, one would usually start with CW survey spectra to ensure the integrity of the sample. Usually, pulsed EPR is less sensitive on a molar spin basis than CW EPR when both are optimized, because of the better noise filtering that can be done in CW EPR. However, if the relaxation time is long, it may not be possible to obtain a valid slow-passage CW EPR spectrum, and alternative techniques will be superior. Sometimes rapid-passage CW EPR, or dispersion spectra instead of absorption spectra, could be the EPR methodology of choice, especially for broad signals. However, for EPR signals that are narrow relative to the bandwidth of the resonator, pulsed FT EPR may provide both better S/N and more fidelity in line shape than CW EPR. This is the case for the signal of the E' center in the γ-irradiated fused quartz sample that has been proposed as a standard sample for time-domain EPR (Eaton and Eaton, 1993b; available from Wilmad Glass).

How many species are there? In many forms of spectroscopy one can monitor isosbestic points to determine limits on the number of species present. Note that, since the common CW EPR spectrum is a derivative display, one should in general not interpret spectral positions that

are constant as a function of some variable as isosbestic points. Because of the derivative display they are isoclinic points—points of the same slope, not points of the same absorption. Integrated spectra need to be used to judge relative behavior of multiple species in the way that has become common in UV-visible electronic spectroscopy.

Relatively few paramagnetic species yield single-line EPR spectra, due to electron-nuclear coupling and electron-electron coupling. If the nuclear coupling is well resolved in a CW spectrum, it is a straightforward exercise to figure out the number of nuclear spins of each nuclear moment. However, unresolved coupling is common, and more powerful (and expensive) techniques have to be used. Electron-nuclear double resonance (ENDOR), which uses RF at the nuclear resonant frequency simultaneous with microwaves resonant at particular EPR transition, is a powerful tool for identifying nuclear couplings not resolved in CW EPR spectra (Kevan and Kispert, 1976; Box, 1977; Dorio and Freed, 1979; Schweiger, 1982; Piekara-Sady and Kispert, 1994; Goslar et al., 1994). A good introduction is given in (Atherton, 1993). Nuclear couplings too small to be observed in ENDOR commonly can be observed with spin echo EPR, where they result in electron spin echo envelope modulation (ESEEM, also abbreviated ESEM; Dikanov and Tsvetkov, 1992). This technique requires a pulsed EPR spectrometer. ESEEM is especially powerful for determining the nuclei in the environment of an electron spin, and is extensively used in metalloprotein studies.

Where are the species that yield the EPR signals? The ENDOR and ESEEM techniques provide answers about the immediate environments of the paramagnetic centers, but there may be a larger-scale question: are the paramagnetic species on the surface of the sample (Sueki et al., 1995), in the center, uniformly distributed, or localized in high-concentration regions throughout the sample? These questions also are answerable via various EPR techniques. EPR imaging can tell where in the sample the spins are located (Eaton et al., 1991). EPR imaging is analogous to NMR imaging, but is most often performed with slowly stepped magnetic field gradients and CW spectra, whereas most NMR imaging is done with pulsed-field gradients and time-domain (usually spin echo) spectra (see NUCLEAR MAGNETIC RESONANCE IMAGING). The differences in technique are due to the differences in spectral width and relaxation times. Low-frequency (L-band) spectrometers are available from JEOL and Bruker.

Uniform versus nonuniform distributions of paramagnetic centers can be ascertained by several observations. Nonuniformity is a more common problem than is recognized. Dopants (impurities) in solids usually distort the geometry of the lattice and sometimes the charge balance, so they are often not distributed in a truly random fashion. Frozen solutions (at whatever temperature) often exhibit partial precipitation and/or aggregation of the paramagnetic centers. The use of pure solvents rarely gives a good glass. For example, even flash freezing does not yield glassy water under normal laboratory conditions. A cosolvent such as glycerol is needed to form a glass. If there is even partial devitrification, it is likely that the paramagnetic species (an impurity in the solvent) will be localized at grain boundaries. Local high concentrations sometimes can be recognized by broadening of the spectra (if one knows what the "normal" spectrum should look like), or even exchange narrowing of the spectra when concentrations are very high. Time-domain (pulse) spectra are more sensitive than CW spectra to these effects, and provide a convenient measure through a phenomenon called instantaneous diffusion (Eaton and Eaton, 1991), which is revealed when the two-pulse echo decay rate depends on the pulse power. The observation of instantaneous diffusion is an indication that the spin concentration is high. If the bulk concentration is known to be low enough that instantaneous diffusion is not likely to be observed (e.g., less than $\sim$1 mM for broad spectra and less than $\sim$0.1 mM for narrow spectra), then the observation of instantaneous diffusion indicates that there are locally high concentrations of spins—i.e., that the spin distribution is microscopically nonuniform.

Multiple species can show up as multiple time constants in pulsed EPR, but there may be reasons other than multiple species for the multiple time constants. Many checks are needed, and the literature of time-domain EPR should be consulted (Standley and Vaughan, 1969; Muus and Atkins, 1972; Kevan and Schwartz, 1979; Dikanov and Tsvetkov, 1992; Kevan and Bowman, 1990).

One of the most common questions is "how long will it take to run an EPR spectrum?" The answer depends strongly on what one wants to learn from the sample, and can range from a few minutes to many weeks. Even the simple question as to whether there are any unpaired electrons present may take quite a bit of effort to answer, unless one already knows a lot about the sample. The spins may have relaxation times so long that they are difficult to observe without saturation (e.g., defect centers in solids) or so short that they cannot be observed except at very low temperature, where the relaxation times become longer [e.g., high-spin Co(II) in many environments]. At the other extreme, column fractions of a nitroxyl-spin-labeled protein can be monitored for radicals about as fast as the samples can be put in the spectrometer. This is an example of an application that could be automated. Similarly, alanine-based radiation dosimeters, if of a standard geometry, can be loaded automatically into a Bruker spectrometer designed for this application, and the spectra can be run in a few minutes. If, on the other hand, one wants to know the concentration of high-spin Co(II) in a sample, the need for quantitative sample preparation, accurate cryogenic temperature control, careful background subtraction, and skillful setting of instrument parameters leads to a rather time-consuming measurement.

Current research on EPR is usually reported in the *Journal of Magnetic Resonance*, *Applied Magnetic Resonance*, *Chemical Physics Letters*, *Journal of Chemical Physics*, and journals focusing on application areas, such as *Macromolecules*, *Journal of Non-Crystalline Solids*, *Inorganic Chemistry*, and numerous biochemical journals. Examination of the current literature will suggest applications of EPR to materials science beyond those briefly mentioned in this unit. The *Specialist Periodical Report* on ESR of the Royal Society of Chemistry is the best source of annual updates on progress in EPR.

Reports of two workshops on the future of EPR (in 1987 and 1992) and a volume celebrating the first 50 years of EPR (Eaton et al., 1998) provide a vision of the future directions of the field (Eaton and Eaton, 1988b, 1995b).

Finally, it should be pointed out that merely obtaining an EPR spectrum properly, and characterizing its dependence on the factors discussed above, is just the first step. Its interpretation in terms of the physical properties of the material of interest is the real intellectual challenge and payoff.

LITERATURE CITED

Abragam, A. and Bleaney, B. 1970. Electron Paramagnetic Resonance of Transition Ions. Oxford University Press, Oxford.

Alger, R. S. 1968. Electron Paramagnetic Resonance: Techniques and Applications. Wiley-Interscience, New York.

Atherton, N. M. 1993. Principles of Electron Spin Resonance. Prentice Hall, London.

Baden-Fuller, A. J. 1990. Microwaves: An Introduction to Microwave Theory and Techniques, 3rd ed. Pergamon Press, Oxford.

Bencini, A. and Gatteschi, D. 1990. EPR of Exchange Coupled Systems. Springer-Verlag, Berlin.

Berliner, L. J., ed., 1976. Spin Labeling: Theory and Applications. Academic Press, New York.

Berliner, L. J., ed., 1979. Spin Labeling II. Academic Press, New York.

Berliner, L. J. and Reuben, J., eds., 1989. Spin Labeling: Theory and Applications. Plenum, New York.

Box, H. C. 1977. Radiation Effects: ESR and ENDOR Analysis. Academic Press, New York.

Byrn, M. P. and Strouse, C. E. 1983. *g*-Tensor determination from single-crystal ESR data. *J. Magn. Reson*. 53:32–39.

Carrington, A. and McLachlan, A. D. 1967. Introduction to Magnetic Resonance. Harper and Row, New York.

Corradi, G., Söthe, H., Spaeth, J.-M., and Polgar, K. 1991. Electron spin resonance and electron-nuclear double resonance investigation of a new Cr^{3+} defect on an Nb site in $LiNbO_3$: Mg:Cr. *J. Phys. Condens. Matter* 3:1901–1908.

Czoch, R. and Francik, A. 1989. Instrumental Effects of Homodyne Electron Paramagnetic Resonance Spectrometers. Wiley-Halsted, New York.

Dalal, D. P., Eaton, S. S., and Eaton, G. R. 1981. The effects of lossy solvents on quantitative EPR studies. *J. Magn. Reson*. 44:415–428.

Dalton, L. R., ed., 1985. EPR and Advanced EPR Studies of Biological Systems. CRC Press, Boca Raton, Fla.

Date, M., ed., 1983. High Field Magnetism. North-Holland Publishing Co., Amsterdam.

Dikanov, S. A. and Tsvetkov, Yu. D. 1992. Electron Spin Echo Envelope Modulation (ESEEM) Spectroscopy. CRC Press, Boca Raton, Fla.

Dorio, M. M. and J. H. Freed, eds., 1979. Multiple Electron Resonance Spectroscopy. Plenum Press, New York.

Du, J.-L., Eaton, G. R., and Eaton, S. S. 1995. Temperature, orientation, and solvent dependence of electron spin lattice relaxation rates for nitroxyl radicals in glassy solvents and doped solids. *J. Magn. Reson*. A115:213–221.

Eaton, S. S. and Eaton, G. R. 1980. Signal area measurements in EPR. *Bull. Magn. Reson.* 1:130–138.

Eaton, G. R. and Eaton, S. S. 1988a. EPR imaging: progress and prospects. *Bull. Magn. Reson.* 10:22–31.

Eaton, G. R. and Eaton S. S. 1988b. Workshop on the future of EPR (ESR) instrumentation: Denver, Colorado, August 7, 1987. *Bull. Magn. Reson.* 10:3–21.

Eaton, G. R. and Eaton, S. S. 1990a. Electron paramagnetic resonance. *In* Analytical Instrumentation Handbook (G. W. Ewing, ed.) pp. 467–530. Marcel Dekker, New York.

Eaton, S. S. and Eaton, G. R. 1990b. Electron spin resonance imaging. *In* Modern Pulsed and Continuous-Wave Electron Spin Resonance (L. Kevan and M. Bowman, eds.) pp. 405–435 Wiley Interscience, New York.

Eaton, G. R., Eaton, S., and Ohno, K., eds. 1991. EPR Imaging and in Vivo EPR, CRC Press, Boca Raton, Fla.

Eaton, S. S. and Eaton, G. R. 1991. EPR imaging, *In* Electron Spin Resonance Specialist Periodical Reports (M. C. R. Symons, ed.) 12b:176–190. Royal Society of London, London.

Eaton, S. S. and Eaton, G. R. 1992. Quality assurance in EPR. *Bull. Magn. Reson.* 13:83–89.

Eaton, G. R. and Eaton, S. S. 1993a. Electron paramagnetic resonance imaging. *In* Microscopic and Spectroscopic Imaging of the Chemical State (M. Morris, ed.) pp. 395–419. Marcel Dekker, New York.

Eaton, S. S. and Eaton, G. R. 1993b. Irradiated fused quartz standard sample for time domain EPR. *J. Magn. Reson.* A102:354–356.

Eaton, S. S. and Eaton, G. R. 1993c. Applications of high magnetic fields in EPR spectroscopy. *Magn. Reson. Rev.* 16:157–181.

Eaton, G. R. and Eaton, S. S. 1995a. Introduction to EPR imaging using magnetic field gradients. *Concepts in Magnetic Resonance* 7:49–67.

Eaton, S. S. and Eaton, G. R. 1995b. The future of electron paramagnetic resonance spectroscopy. *Bull. Magn. Reson.* 16:149–192.

Eaton, S. S. and Eaton, G. R. 1996a. Electron spin relaxation in discrete molecular species. *Current Topics in Biophysics* 20: 9–14.

Eaton, S. S. and Eaton, G. R. 1996b. EPR imaging. *In* Electron Spin Resonance Specialist Periodical Reports (M. C. R. Symons, ed.) 15:169–185. Royal Society of London, London.

Eaton, S. S. and Eaton, G. R. 1997a. Electron paramagnetic resonance. *In* Analytical Instrumentation Handbook (G. W. Ewing, ed.), 2nd ed. pp. 767–862. Marcel Dekker, New York.

Eaton, S. S. and Eaton, G. R., 1997b, EPR methodologies: Ways of looking at electron spins. *EPR Newsletter* 9:15–18.

Eaton, G. R., Eaton, S. S., and Salikhov, K., eds. 1998. Foundations of Modern EPR. World Scientific Publishing, Singapore.

Eaton, G. R. and Eaton, S. S. 1999a. ESR imaging. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.), vol 2. Springer-Verlag, New York.

Eaton, S. S. and Eaton, G. R. 1999b. Magnetic fields and high frequencies in ESR spectroscopy. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.), vol 2. Springer-Verlag, New York.

Goldberg, I. B. and Crowe, H. R. 1977. Effect of cavity loading on analytical electron spin resonance spectrometry. *Anal. Chem.* 49:1353.

Goslar, J., Piekara-Sady, L. and Kispert, L. D. 1994. ENDOR data tabulation. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.). AIP Press, NY.

Hyde, J. S. and Froncisz, W. 1982. The role of microwave frequency in EPR spectroscopy of copper complexes. *Ann. Rev. Biophys. Bioeng*. 11:391–417.

Hyde, J. S. and Froncisz, W. 1986. Loop gap resonators. *In* Electron Spin Resonance Specialist Periodical Reports (M. C. R. Symons, ed.) 10:175–185. Royal Society, London.

Hyde, J. S. and Subczynski, W. K. 1989. Spin-label oximetry. *In* Spin Labeling: Theory and Applications, (L. J. Berliner, and J. Reuben, eds.) Plenum, New York.

Ikeya, M. 1993. New Applications of Electron Spin Resonance. Dating, Dosimetry, and Microscopy. World Scientific, Singapore.

Kevan, L. and Kispert, L. D. 1976. Electron Spin Double Resonance Spectroscopy. John Wiley & Sons, New York.

Kevan, L. and Bowman, M. K., eds. 1990. Modern Pulsed and Continuous-Wave Electron Spin Resonance. Wiley-Interscience, New York.

Kevan, L. and Schwartz, R. N., eds. 1979. Time Domain Electron Spin Resonance. John Wiley & Sons, New York.

Kirste, B., 1994. Computer techniques. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.). AIP Press, NY.

Krinichnyi, V. I. 1995. 2-mm Wave Band EPR Spectroscopy of Condensed Systems. CRC Press, Boca Raton, Fla.

Lancaster, G., 1967. Electron Spin Resonance in Semiconductors. Plenum, New York.

Muus, L. T. and Atkins, P. W., eds. 1972. Electron Spin Relaxation in Liquids, Plenum Press, New York.

Pake, G. E. and Estle, T. L. 1973. The Physical Principles of Electron Paramagnetic Resonance, 2nd ed. W. A. Benjamin, Reading, Mass.

Piekara-Sady, L. and Kispert, L. D. 1994. ENDOR spectroscopy. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.). AIP Press, NY.

Pilbrow, J. R., 1990. Transition Ion Electron Paramagnetic Resonance. Oxford University Press, London.

Poole, C. P., Jr., 1967. Electron Spin Resonance: A Comprehensive Treatise on Experimental Techniques, pp. 398–413. John Wiley & Sons, New York.

Poole, C. P., Jr., 1983. Electron Spin Resonance: A Comprehensive Treatise on Experimental Techniques, 2nd ed. Wiley-Interscience, New York.

Quine, R. W., Eaton, G. R., and Eaton, S. S. 1987. Pulsed EPR spectrometer. *Rev. Sci. Instrum.* 58:1709–1724.

Quine, R. W., Eaton, S. S., and Eaton, G. R. 1992. A saturation recovery electron paramagnetic resonance spectrometer. *Rev. Sci. Instrum.* 63:4251–4262.

Quine, R. W., Rinard, G. A., Ghim, B. T., Eaton, S. S. and Eaton, G. R. 1996. A 1-2 GHz pulsed and continuous wave electron paramagnetic resonance spectrometer. *Rev. Sci. Instrum.* 67:2514–2527.

Rinard, G. A., Quine, R. W., Eaton, S. S. and Eaton, G. R. 1993. Microwave coupling structures for spectroscopy. *J. Magn. Reson.* A105:134–144.

Rinard, G. A., Quine, R. W., Eaton, S. S., Eaton, G. R., and Froncisz, W. 1994. Relative benefits of overcoupled resonators vs. inherently low-Q resonators for pulsed magnetic resonance. *J. Magn. Reson.* A108:71–81.

Rinard, G. A., Quine, R. W., Ghim, B. T., Eaton, S. S., and Eaton, G. R. 1996a. Easily tunable crossed-loop (bimodal) EPR resonator. *J. Magn. Reson.* A122:50–57.

Rinard, G. A., Quine, R. W., Ghim, B. T., Eaton, S. S., and Eaton, G. R. 1996b. Dispersion and superheterodyne EPR using a bimodal resonator. *J. Magn. Reson.* A122:58–63.

Rinard, G. A., Eaton, S. S., Eaton, G. R., Poole, C. P., Jr., and Farach, H. A., 1999. Sensitivity of ESR Spectrometers: Signal, Noise, and Signal-to-Noise. *In* Handbook of Electron Spin Resonance (C. P. Poole, Jr., and H. A. Farach, eds.), vol. 2. Springer-Verlag, New York.

Rudowicz, C. Z., Yu, K. N., and Hiraoka, H., eds. 1998. Modern Applications of EPR/ESR: From Biophysics to Materials Science. Springer-Verlag, Singapore.

Schweiger, A. 1982. Electron Nuclear Double Resonance of Transition Metal Complexes with Organic Ligands, Structure and Bonding vol. 51. Springer-Verlag, New York.

Standley, K. J. and Vaughan, R. A. 1969. Electron Spin Relaxation Phenomena in Solids. Plenum Press, New York.

Stevens, K. W. H. 1997. Magnetic Ions in Crystals. Princeton University Press, Princeton, N.J.

Sueki, M., Eaton, G. R., and Eaton, S. S. 1990. Electron spin echo and CW perspectives in 3D EPR imaging. *Appl. Magn. Reson.* 1:20–28.

Sueki, M., Austin, W. R., Zhang, L., Kerwin, D. B., Leisure, R. G., Eaton, G. R., and Eaton, S. S. 1995. Determination of depth profiles of E' defects in irradiated vitreous silica by electron paramagnetic resonance imaging. *J. Appl. Phys.*77:790–794.

Swartz, H. M., Bolton, J. R., and Borg, D. C., eds. 1972. Biological Applications of Electron Spin Resonance, John Wiley & Sons, New York.

Trulove, P. C., Carlin, R. T., Eaton, G. R., and Eaton, S. S. 1995. Determination of the Singlet-Triplet Energy Separation for C_{60}^{2-} in DMSO by Electron Paramagnetic Resonance. *J. Am. Chem. Soc.* 117:6265–6272.

Vancamp, H. L. and Heiss, A. H. 1981. Computer applications in electron paramagnetic resonance. *Magn. Reson. Rev.* 7:1–40.

Weil, J. A., Bolton, J. R., and Wertz, J. E. 1994. Electron Paramagnetic Resonance: Elementary Theory and Practical Applications, John Wiley & Sons, New York.

Wilmshurst, T. H. 1968. Electron Spin Resonance Spectrometers, Plenum, N. Y., see especially ch. 4.

KEY REFERENCES

Poole, 1967. See above.

Poole, 1983. See above.

The two editions of this book provide a truly comprehensive coverage of EPR. They are especially strong in instrumentation and technique, though most early applications to materials science are cited also. This is "the bible" for EPR spectroscopists.

Eaton and Eaton, 1990a. See above.

Eaton and Eaton, 1997a. See above.

These introductory chapters in two editions of the Analytical Instrumentation Handbook provide extensive background and references. They are a good first source for a person with no background in the subject.

Weil et al., 1994. See above.

Atherton, 1993. See above.

These two very good comprehensive textbooks have been updated recently. They assume a fairly solid understanding of physical chemistry. Detailed reviews of these texts are given in:

Eaton, G.R. 1995. *J. Magn. Reson.* A113:135-136.

Detailed review of Principles of Electron Spin Resonance (Atherton, 1993).

Eaton, S.S. 1995. *J. Magn. Reson.* A113:137.

Detailed book review of Electron Paramagnetic Resonance: Theory and Practical Applications (Weil et al., 1994).

INTERNET RESOURCES

http://ierc.scs.uiuc.edu

The International EPR (ESR) Society strives to broadly disseminate information concerning EPR spectroscopy. This Web site contains a newsletter published for the Society by the staff of the Illinois EPR Research Center, University of Illinois. Supported by National Institutes of Health.

http://www.biophysics.mcw.edu

EPR center at the Medical College of Wisconsin. Supported by National Institutes of Health.

http://spin.aecom.yu.edu

EPR center at Albert Einstein College of Medicine. Supported by National Institutes of Health.

epr-list@xenon.che.ilstu.edu

E-mail service for rapid communication, run by Professor P.D. Morse, Illinois State University.

http://www.du.edu/~seaton

Web page for the authors' lab. The above Centers and the authors' lab attempt to link to each other and collectively provide many references to recent papers, notices of meetings, and background on magnetic resonance. There now exist many electronic data bases that one can search to find recent papers on EPR. However, only a small portion of the relevant papers will be found by using just "EPR" as the search term. Our experience is that at least the following set of keywords have to be used to retrieve the bulk of the papers: ESR, EPR, electron spin, electron paramag, ESEEM, ENDOR, and echo modulation. Other relevant papers will be found using keywords for the application field, since words relating to EPR may not occur in the title or abstract of the paper even if EPR is a primary tool of the study.*

APPENDIX: INSTRUMENTATION

EPR measurements require a significant investment in instrumentation. There are three fundamental parts to an EPR spectrometer: the microwave system, the magnet system, and the data acquisition and analysis system (Figure 2). Brief discussions of spectrometer design are in introductory texts (Weil et al., 1994, Eaton and Eaton, 1990, 1997). Extensive discussions of microwaves relevant to EPR are available (Poole, 1967, 1983; Baden-Fuller, 1990). Some recent papers give details on pulsed EPR design (Quine et al., 1987, 1992, 1996). The magnets used in EPR range from ~3-in. to 12-in. diameter pole faces, the larger magnets providing better magnetic field homogeneity for larger gaps. Larger magnet gaps permit a wider range of accessory instrumentation for various measurements, such as controlling the sample temperature. A 3-in. magnet weighs roughly 300 pounds, and a 12-in. magnet weighs roughly 4000 pounds. Power supplies for the magnet, and cooling water systems, which optimally are temperature-controlled-recirculating deionized water systems, also require an investment in floor space. The microwave system usually sits on a table supported on the magnet for stability.

EPR spectrometers are expensive to purchase and to operate. They require a well trained operator, and except for a few tabletop models, they take up a lot of floor space and use substantial electrical power and cooling water for the electromagnet. Newer EPR spectrometer systems are operated via a computer, so the new operator who is familiar with use of computers can learn operation much faster than in the older equipment where remembering a critical sequence of manual valves and knobs was part of doing spectroscopy. Even so, the use of a high-Q resonator results in much stronger interaction between the sample and the spectrometer than in most other analytical techniques, and a lot of judgment is needed to get useful, especially quantitatively meaningful results.

Commercial Instruments

Most materials science needs will be satisfied by commercial spectrometers. Some references are provided below to recent spectrometers built by various research labs. These provide some details about design that may be helpful to the reader who wants to go beyond the level of this article. The following brief outline of commercial instruments is intended to guide the reader to the range of spectrometers available. The largest manufacturers—Bruker Instruments EPR Division, and JEOL—market general-purpose spectrometers intended to fulfill most analytical needs. The focus is on X-band (~9 to 10 GHz) CW spectrometers, with a wide variety of resonators to provide for many types of samples. Accessories facilitate control of the sample temperature from <4 to ~700 K. Magnets commonly range from 6- to 12-in. pole face diameters. Smaller tabletop spectrometers are available from Bruker, JEOL, and Resonance Instruments. Some of these have permanent magnets and sweep coils for applications that focus on spectra near $g = 2$, but some have electromagnets permitting wide field sweeps. Bruker makes one small system optimized for quantitation of organic radicals and defect centers, such as for dosimetry.

Bruker and JEOL market pulsed, time-domain spectrometers as well as CW spectrometers. These suppliers also market spectrometers for frequencies lower than X-band, which are useful for study of lossy samples. Bruker and Resonance Technologies market high-frequency, high-field EPR spectrometers, which require superconducting magnets, not resistive electromagnets.

There are numerous vendors of accessories and supplies essential for various aspects of EPR measurements, including Oxford Instruments (cryostats), Cryo Industries (cryostats), Wilmad (sample tubes, quartz Dewars, standard samples), Oxis (spin traps), Medical Advances (loop gap resonators), Research Specialties (accessories and service), Summit Technology (small spectrometers, accessories), Scientific Software Services (data acquisition), Cambridge Isotope Labs (labeled compounds), and CPI (klystrons). Addresses and up-to-date information can be accessed via the EPR Newsletter.

SANDRA S. EATON
GARETH R. EATON
University of Denver
Denver, Colorado

CYCLOTRON RESONANCE

INTRODUCTION

Cyclotron resonance (CR) is a method for measuring the effective masses of charge carriers in solids. It is by far the most direct and accurate method for providing such information. In the simplest description, the principle of the method can be stated as follows. A particle of effective mass m^* and charge e in a DC magnetic field B executes a helical motion around B with the cyclotron frequency $\omega_c = eB/m^*$. If, at the same time, an AC electric field of frequency $\omega = \omega_c$ is applied to the system, perpendicular to B, the particle will resonantly absorb energy from the AC field. Since B and/or ω can be continuously swept through the resonance and known to a very high degree of accuracy, m^* can be directly determined with high accuracy by $m^* = eB/\omega$.

In crystalline solids, the dynamics of charge carriers (or Bloch electrons) can be most conveniently described by the use of the effective mass tensor, defined for simple nondegenerate bands as

$$(\bar{m})_{\mu\nu} = \left(\frac{1}{\hbar^2}\frac{\partial^2 E(\vec{k})}{\partial k_\mu \partial k_\nu}\right)^{-1}, \qquad (\mu, \nu = 1, 2, 3) \qquad (1)$$

where $E = E(\vec{k})$ is the energy dispersion relation near the band edge ($\vec{p} = \hbar\vec{k}$ is the crystal momentum). Hence, the primary purpose of CR in solids is to determine the components of the effective mass tensor, or the curvature of the energy surface, at the extrema of the conduction and valence bands (or at the Fermi surface in the case of a metal). In the 1950s, the first CR studies were carried out in germanium and silicon crystals (Dresselhaus et al., 1953, 1955; Lax et al., 1953), which, in conjunction with the effective-mass theory (e.g., Luttinger and Kohn, 1955; Luttinger, 1956), successfully determined the band-edge parameters for these materials with unprecedented accuracy. Since then CR has been investigated in a large number of elementary and compound materials and their alloys and heterostructures.

Quantum mechanically, the energy of a free electron in a magnetic field is quantized as $E_N = (N + 1/2)\hbar\omega_c$, where $N = 0, 1, 2, \ldots \infty$ and $\hbar\omega_c$ is called the cyclotron energy. These equally-spaced energy levels, or an infinite ladder, are the well-known Landau levels (see, e.g., Kittel, 1987). At high magnetic fields and low temperatures (T), where $k_B T < \hbar\omega_c$ (k_B is the Boltzmann constant), this magnetic quantization becomes important, and CR may then be viewed as a quantum transition between adjacent Landau levels ($\Delta N = \pm 1$).

In real solids, Landau levels are generally not equally-spaced since the energy dispersion relation, E versus $\vec{k}$, is not generally given by a simple parabolic dispersion relation $E(\vec{k}) = \hbar^2|\vec{k}|^2/2m^*$. The degree of nonparabolicity and anisotropy depends on the material, but, in general, the effective mass is energy-dependent as well as direction-dependent. Landau levels for free carriers in solids cannot be obtained analytically, but several useful approximation models exist (e.g., Luttinger, 1956; Bowers and Yafet, 1959; Pidgeon and Brown, 1966). These calculations could be complex, especially when one is concerned with degenerate bands such as the valence bands of group IV, III-V and II-VI semiconductors. However, a detailed comparison between theory and experiment on quantum CR can provide a critical test of the band theory of solids (e.g., Suzuki and Hensel, 1974; Hensel and Suzuki, 1974).

As a secondary purpose, one can also use CR to study carrier scattering phenomena in solids by examining the scattering lifetime τ (the time between collisions, also known as the collision time or the transport/momentum relaxation time), which can be found from the linewidth of CR peaks. In the classical regime, where most electrons reside in states with high Landau indices, τ is directly related to the static (or DC) conductivity in the absence of the magnetic field. The temperature dependence of τ then shows markedly different characteristics for phonon scattering, impurity scattering, and scattering from various imperfections. However, in the quantum regime, where most electrons are in the first few Landau levels, the effect of the magnetic field on scattering mechanisms is no longer negligible and τ loses its direct relationship with the DC mobility (for theory, see, e.g., Kawabata, 1967).

As the frequency of scattering events (or the density of scatterers) increases, CR linewidth increases, and, eventually, CR becomes unobservable when scattering occurs too frequently. More quantitatively, in order to observe CR, τ must be long enough to allow the electron to travel at least $1/2\pi$ of a revolution between two scattering events, i.e.,

$$\tau > \frac{T_c}{2\pi} = \frac{1}{\omega_c}$$

or

$$\omega_c\tau = \frac{eB}{m^*}\tau = \frac{e\tau}{m^*}B = \mu B > 1 \qquad (2)$$

where T_c is the period of cyclotron motion and $\mu = e\tau/m^*$ is the DC mobility of the electron. Let us examine this CR observability condition for a realistic set of parameters. If $m^* = 0.1m_0$ (where $m_0 = 9.1 \times 10^{-31}$ kg) and $B = 1$ Tesla, then $\omega_c = eB/m^* \approx 2 \times 10^{12}$ sec^{-1}. Thus one needs a microwave field with a frequency of $f_c = \omega_c/2\pi \approx 3 \times 10^{11}$ Hz $= 300$ GHz (or a wavelength of $\lambda_c = c/f_c \approx 1$ mm). Then, in order to satisfy Equation 2, one needs a minimum mobility of $\mu = 1$ m^2/V-sec $= 1 \times 10^4$ cm^2/V-sec. This value of mobility can be achieved only in a limited number of high-purity semiconductors at low temperatures, thereby posing a severe limit on the observations of microwave CR.

From the resonance condition $\omega = \omega_c$, it is obvious that if a higher magnetic field is available (see GENERATION AND MEASUREMENT OF MAGNETIC FIELDS) one can use a higher frequency (or a shorter wavelength), which should make Equation 2 easier to satisfy. Hence modern CR methods almost invariably use far-infrared (FIR) [or Terahertz (THz)] radiation instead of microwaves. Strong magnetic

fields are available either in pulsed form (up to $\sim 10^3$ T) or in steady form by superconducting magnets (up to ~20 T), water-cooled magnets (up to ~30 T), or hybrid magnets (up to ~45 T). In these cases, even at room temperature, Equation 2 may be fulfilled. Here we are only concerned with the methods of FIR-CR. The reader particularly interested in microwave CR is referred to Lax and Mavroides (1960).

Although this unit is mainly concerned with the simplest case of free carrier CR in bulk semiconductors, one can also study a wide variety of FIR magneto-optical phenomena with essentially the same techniques as CR. These phenomena ("derivatives" of CR) include: (a) spin-flip resonances, i.e., electron spin resonance and combined resonance, (b) resonances of bound carriers, i.e., internal transitions of shallow impurities and excitons, (c) polaronic coupling, i.e., resonant interactions of carriers with phonons and plasmons, and (d) 1-D and 2-D magnetoplasmon excitations. It should be also mentioned that in 2-D systems in the magnetic quantum limit, there are still unresolved issues concerning the effects of disorder and electron-electron interactions on CR (for a review, see, e.g., Petrou and McCombe, 1991; Nicholas, 1994).

It is important to note that all the early CR studies were carried out on semiconductors, not on metals. This is because of the high carrier concentrations present in metals, which preclude direct transmission spectroscopy except in the case of very thin films where thickness is less than the depth of penetration (skin depth) of the electromagnetic fields. In bulk metals, special geometries are thus required to detect CR, the most important of which is the Azbel-Kaner geometry (Azbel and Kaner, 1958). In this geometry, both the DC magnetic field $\vec{B}$ and the AC electric field $\vec{E}$ are applied parallel to the sample surface, either $\vec{B}//\vec{E}$ or $\vec{B} \perp \vec{E}$. The electrons then execute a spiral motion along $\vec{B}$, moving in and out of the skin depth, where $\vec{E}$ is present. Thus, whenever the electron enters the skin depth, it is accelerated by $\vec{E}$ and if the phase of $\vec{E}$ is the same every time the electron enters the skin depth, then the electron can resonantly absorb energy from the AC field. The condition for resonance here is $n\omega_c = \omega (n = 1, 2, 3, \ldots)$. For more details on CR in metals, see, e.g., Mavroides (1972).

Many techniques can provide information on effective masses, but none can rival CR for directness and accuracy. Effective masses can be estimated from the temperature dependence of the amplitude of the galvanomagnetic effects, i.e., the Shubnikov–de Haas and de Haas–van Alphen effects. Interband magneto-optical absorption can determine the reduced mass $\mu = (1/m_e + 1/m_h)^{-1}$ of photo-created electrons and holes. Measurements of the infrared Faraday rotation effect due to free carriers can provide information on the anisotropy of elliptical equienergy surfaces. The temperature dependence of electronic specific heat provides a good measure of the density of levels at the Fermi level, which in turn is proportional to the effective mass. Nonresonant free carrier absorption (see CARRIER LIFETIME: FREE CARRIER ABSORPTION, PHOTOCONDUCTIVITY, AND PHOTOLUMINESCENCE) can be used to estimate effective masses, but, of course, this simply represents a tail or shoulder of a CR absorption curve.

It is worth pointing out here that several different definitions of effective masses exist in the literature and care must be taken when one discusses masses. The band-edge mass, defined as in Equation 1 at band extrema (e.g., $\vec{k} = 0$ in most semiconductors), is the most important band parameter to characterize a material. The specific heat mass is directly related to the density of states at the Fermi level, and is thus also called the density-of-states mass. The cyclotron mass is defined as $m_c^* = (\hbar^2/2\pi)\partial A/\partial E$, where A is the momentum-space area enclosed by the cyclotron orbit; this definition follows naturally from calculation of ω_c in momentum space. The spectroscopic mass can be defined for any resonance peak, and is identical to the cyclotron mass when the resonance is due to free-carrier CR (also see Data Analysis and Initial Interpretation).

The basic theory and experimental methods of cyclotron resonance is presented in this unit. Basic theoretical back ground will first be presented (see Principles of the Method). A detailed description will be given of the actual experimentation procedures (see Practical Aspects of the Method). Finally, typical data analysis procedures are presented (see Data Analysis and Initial Interpretation).

PRINCIPLES OF THE METHOD

As described in the Introduction, the basic physics of CR is the interaction of electromagnetic (EM) radiation with charge carriers in a magnetic field. Here, more quantitative descriptions of this physical phenomenon will be presented, based on (1) a semiclassical model and (2) a quantum-mechanical model. In analyzing CR data, judicious combination, modification, and refinement of these basic models are necessary, depending upon the experimental conditions and the material under study, in order to obtain the maximum amount of information from a given set of data.

The most commonly used method for describing the motion of charge carriers in solids perturbed by external fields is the effective-mass approximation (EMA), developed by many workers in the early history of the quantum theory of solids. The beauty of this method lies in the ability to replace the effect of the lattice periodic potential on electron motion by a mass tensor, the elements of which are determined by the unperturbed band structure. In other words, instead of considering electrons in a lattice we may consider the motion of effective-mass particles, which obey simple equations of motion in the presence of external fields. Rigorous treatments and full justification of the EMA can be found in the early original papers (e.g., Wannier, 1937; Slater, 1949; Luttinger, 1951; Luttinger and Kohn, 1955).

Semiclassical Drude Description of CR

In many cases it is satisfactory to use the semiclassical Drude model (e.g., Aschcroft and Mermin, 1976) to describe the conductivity tensor of free carriers in a magnetic field (see MAGNETOTRANSPORT IN METALS AND ALLOYS). In this

model each electron is assumed to independently obey the equation of motion

$$\overleftrightarrow{m}^* \cdot \frac{d\vec{v}}{dt} + \overleftrightarrow{m}^* \cdot \frac{\vec{v}}{\tau} = e(\vec{E} + \vec{v} \times \vec{B}) \tag{3}$$

where $\overleftrightarrow{m}^*$ is the effective-mass tensor (see, e.g., Equation 1 for nondegenerate bands), $\vec{v}$ is the drift velocity of the electrons, τ is the scattering lifetime (which is assumed to be a constant), $\vec{E}$ is the AC electric field, and $\vec{B}$ is the DC magnetic field. The complex conductivity tensor $\overleftrightarrow{\sigma}$ is then defined by $\vec{J} = ne\vec{v} = \overleftrightarrow{\sigma} \cdot \vec{E}$, where $\vec{J}$ is the current density and n is the carrier density. Assuming that the AC field and the drift velocity have the harmonically varying form, i.e., $\vec{E}(t) = \vec{E}_0 \exp(-i\omega t)$, $\vec{v}(t) = \vec{v}_0 \exp(-i\omega t)$, one can easily solve Equation 3. In particular, for cubic materials and for $\vec{B} \parallel \hat{z}$, $\overleftrightarrow{\sigma}$ is given by

$$\overleftrightarrow{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & 0 \\ \sigma_{yx} & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{pmatrix} \tag{4a}$$

$$\sigma_{xx} = \sigma_{yy} = \sigma_0 \frac{i\omega\tau + 1}{(i\omega\tau + 1)^2 + \omega_c^2\tau^2} \tag{4b}$$

$$\sigma_{xy} = -\sigma_{yx} = \sigma_0 \frac{\omega_c\tau}{(i\omega\tau + 1)^2 + \omega_c^2\tau^2} \tag{4c}$$

$$\sigma_{zz} = \sigma_0 \frac{1}{i\omega\tau + 1} \tag{4d}$$

$$\sigma_0 = ne\mu = \frac{ne^2\tau}{m^*} \tag{4e}$$

where μ is the carrier mobility and σ_0 is the DC conductivity.

Once we know the conductivity tensor, we can evaluate the power P absorbed by the carriers from the AC field as

$$P = \langle \vec{J}(t) \cdot \vec{E}(t) \rangle = \frac{1}{2} Re(\vec{j} \cdot \vec{E}^*) \tag{5}$$

where $\langle \ldots \rangle$ represents the time average and $\vec{E}^*$ is the complex conjugate of $\vec{E}$. For an EM wave linearly polarized in the x-direction, i.e., $\vec{E} = (E_x, 0, 0)$, Equation 5 simplifies to

$$P = \frac{1}{2} Re(J_x E_x^*) = \frac{1}{2} |E_0|^2 Re(\sigma_{xx}) \tag{6}$$

Substituting part b of Equation 4 into Equation 6, we obtain

$$\begin{aligned} P(\omega) &= \frac{1}{2} |E_0|^2 \sigma_0 Re\left\{ \frac{i\omega\tau + 1}{(i\omega\tau + 1)^2 + \omega_c^2\tau^2} \right\} \\ &= \frac{1}{4} |E_0|^2 \sigma_0 \left[\frac{1}{(\omega - \omega_c)^2\tau^2 + 1} + \frac{1}{(\omega + \omega_c)^2\tau^2 + 1} \right] \end{aligned} \tag{7}$$

This is plotted in Fig. 1 for different values of the parameter $\omega_c\tau$. It is evident from this figure that the absorption peak occurs when $\omega = \omega_c$ and $\omega_c\tau > 1$.

Note that Equation 7 contains two resonances—one at $\omega = \omega_c$ and the other at $\omega = -\omega_c$. These two resonances correspond to the two opposite senses of circular polarization. It can be shown (see, e.g., Palik and Furdyna, 1970) that in the Faraday geometry $\vec{q} \parallel \vec{B} \parallel \hat{z}$, where $\vec{q}$ is the wavevector of the EM wave) a 3-D magneto-plasma can support only two propagating EM modes represented by

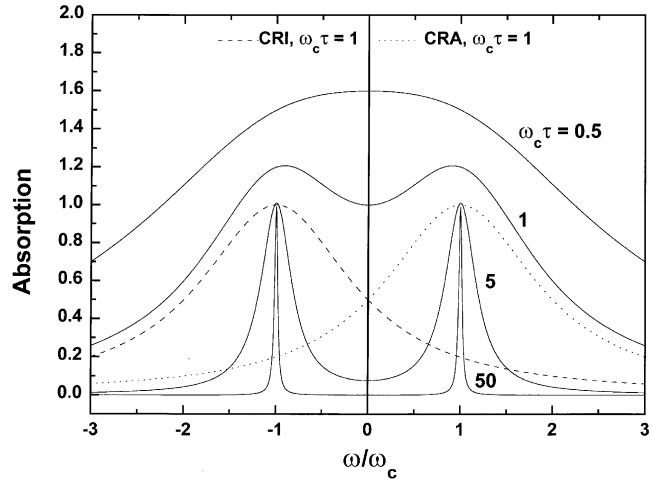


Figure 1. The CR absorption power versus ω for different values of $\omega_c\tau$. The traces are obtained from Equation 7. CR occurs at $\omega = \omega_c$ when $\omega_c\tau > 1$. The absorption is expressed in units of $|E_0|^2\sigma_0/4$.

$$\vec{E}_\pm = E_0 \frac{1}{\sqrt{2}} (\vec{e}_x \pm i\vec{e}_y) \exp[i(q_\pm z - \omega t)] \tag{8}$$

where $\vec{e}_x$ and $\vec{e}_y$ are the unit vectors in the x and y directions, respectively. The dispersion relations, $q_\pm$ versus ω, for the two modes are obtained from

$$q_\pm = \frac{\omega N_\pm}{c} \tag{9a}$$

$$\begin{cases} N_\pm^2 = \kappa_\pm = \kappa_{xx} \pm ik_{xy} & (9b) \\ \kappa_{xx} = \kappa_l + \dfrac{i}{\omega\varepsilon_0}\sigma_{xx} = \kappa_l + \dfrac{i\sigma_0}{\omega\varepsilon_0} \dfrac{i\omega\tau + 1}{(i\omega\tau + 1)^2 + \omega_c^2\tau^2} & (9c) \\ \kappa_{xy} = \dfrac{i}{\omega\varepsilon_0}\sigma_{xy} = \dfrac{i\sigma_0}{\omega\varepsilon_0} \dfrac{\omega_c\tau}{(i\omega\tau + 1)^2 + \omega_c^2\tau^2} & (9d) \end{cases}$$

where $N_\pm$ is the complex refractive index, κ_l is the relative dielectric constant of the lattice (assumed to be constant in the FIR), and κ_{xx} and κ_{xy} are components of the generalized dielectric tensor $\overleftrightarrow{\kappa} = \kappa_l \overleftrightarrow{I} + (i/\omega\varepsilon_0)\overleftrightarrow{\sigma}$, where $\overleftrightarrow{I}$ is the unit tensor and $\overleftrightarrow{\sigma}$ is the conductivity tensor (Equation 4). The positive sign corresponds to a circularly-polarized FIR field, rotating in the same sense as a negatively-charged particle, and is traditionally referred to as cyclotron-resonance-active (CRA) for electrons. Similarly, the negative sign represents the opposite sense of circular polarization, cyclotron-resonance-inactive (CRI) for electrons. The CRA mode for electrons is the CRI mode for holes, and vice versa. For linearly-polarized FIR radiation, which is an equal-weight mixture of the two modes, both terms in Equation 7 contribute to the absorption curve, as represented in Figure 1.

Quantum Mechanical Description of CR

According to the EMA, if the unperturbed energy-momentum relation for the band n, $E_n(\vec{p})$, is known, then the allowed energies E of a crystalline system perturbed by a uniform DC magnetic field $\vec{B} = \vec{\nabla} \times \vec{A}$ (where $\vec{A}$ is the vector potential) are given approximately by solving the effective Schrödinger equation

$$\hat{H}F_n(\vec{r}) = \hat{E}_n(-i\hbar\vec{\nabla} + e\vec{A})F_n(\vec{r}) = EF_n(\vec{r}) \tag{10}$$

Here the operator $\hat{E}_n(-i\hbar\vec{\nabla} + e\vec{A})$ means that we first replace the momentum $\vec{p}$ in the function $E_n(\vec{p})$ with the kinematic (or mechanical) momentum $\vec{\pi} = \vec{p} + e\vec{A}$ (see, e.g., Sakurai, 1985) and then transform it into an operator by $\vec{p} \rightarrow -i\hbar\vec{\nabla}$. The function $F_n(\vec{r})$ is a slowly varying "envelope" wavefunction; the total wavefunction $\Psi(\vec{r})$ is given by a linear combination of the Bloch functions at $\vec{p} = 0$ (or the cell-periodic function) $\psi_{n0}(\vec{r})$

$$\Psi(\vec{r}) = \sum_n F_n(\vec{r})\psi_{n0}(\vec{r}) \tag{11}$$

For simplicity let us consider a 2-D electron in a conduction-band with a parabolic and isotropic dispersion $E_n(\vec{p}) = |\vec{p}|^2/2m^* = (P_x^2 + P_y^2)/2m^*$. The Hamiltonian in Equation 10 is then simply expressed as

$$\hat{H}_0 = \frac{|\vec{\pi}|^2}{2m^*} = \frac{1}{2m^*}(\pi_x^2 + \pi_y^2) \tag{12}$$

These kinematic momentum operators obey the following commutation relations

$$[x, \pi_x] = [y, \pi_y] = i\hbar \tag{13a}$$
$$[x, \pi_y] = [y, \pi_x] = [x, y] = 0 \tag{13b}$$
$$[\pi_x, \pi_y] = -i\hbar^2/\ell^2 \tag{13c}$$

where $\ell = (\hbar/eB)^{1/2}$ is the magnetic length, which measures the spatial extent of electronic (envelope) wavefunctions in magnetic fields—the quantum mechanical counterpart of the cyclotron radius r_c. Note the non-commutability between π_x and π_y (Equation 13, line c), in contrast to p_x and p_y in the zero magnetic-field case.

We now introduce the Landau level raising and lowering operators

$$\hat{a}_\pm = \frac{\ell}{\sqrt{2}\hbar}(\pi_x \pm i\pi_y) \tag{14}$$

for which we can show that

$$[\hat{a}_-, \hat{a}_+] = 1 \tag{15a}$$
$$[\hat{H}_0, \hat{a}_\pm] = \pm\hbar\omega_c\hat{a}_\pm \tag{15b}$$

Combining Equations 13 to 15, we can see that $\hat{a}_\pm$ connects the state $|N\rangle$ and the states $|N \pm 1\rangle$ and that $\hat{H}_0 = \hbar\omega_c(\hat{a}_+\hat{a}_- + 1/2)$, and hence the eigenenergies are

$$E = \left(N + \frac{1}{2}\right)\hbar\omega_c, \quad (N = 0, 1, \ldots) \tag{16}$$

These discrete energy levels are well-known as Landau levels.

If we now apply a FIR field with CRA (CRI) polarization (Equation 8) to the system, then, in the electric dipole approximation, the Hamiltonian becomes

$$\hat{H} = \frac{1}{2m^*}|\vec{p} + e\hat{A} + e\hat{A}'_\pm|^2 \approx \frac{1}{2m^*}|\vec{\pi}|^2 + \frac{e}{m^*}\vec{\pi}\cdot\vec{A}'_\pm = \hat{H}_0 + \hat{H}' \tag{17}$$

where, from Equation 8, the vector potential for the FIR field $\vec{A}'_\pm$ is given by

$$\vec{A}'_\pm = -\frac{E_0}{\sqrt{2}i\omega}(\vec{e}_x \pm i\vec{e}_y)\exp(-i\omega t) \tag{18}$$

Thus the perturbation Hamiltonian $\hat{H}'$ is given by

$$\begin{aligned}\hat{H}' &= \frac{e}{m^*}(\pi_x A'_{\pm,x} + \pi_y A_{\pm,y}) \\ &= -\frac{eE_0}{\sqrt{2}i\omega m^*}(\pi_x \pm i\pi_y)\exp(-i\omega t) \\ &= -\frac{e\hbar E_0}{i\omega m^*\ell}\hat{a}_\pm\exp(-i\omega t)\end{aligned} \tag{19}$$

We can immediately see that this perturbation, containing $\vec{a}_\pm$, connects the state $|N\rangle$ with the states $|N \pm 1\rangle$, so that a sharp absorption occurs at $\omega = \omega_c$.

PRACTICAL ASPECTS OF THE METHOD

CR spectroscopy, or the more general FIR magneto-spectroscopy, is performed in two distinct ways—Fourier-transform magneto-spectroscopy (FTMS) and laser magneto-spectroscopy (LMS). The former is wavelength-dependent spectroscopy and the latter is magnetic-field-dependent spectroscopy. Generally speaking, these two methods are complimentary.

Narrow spectral widths and high output powers are two of the main advantages of lasers. The former makes LMS suitable for investigating spectral features that cannot be resolved by using FTMS, and the latter makes it suitable for studying absorption features in the presence of strong background absorption or reflection. It is also much easier to conduct polarization-dependent measurements by LMS than by FTMS. Moreover, LMS can be easily combined with strong pulsed magnets. Finally, with intense and short-pulse lasers such as the free-electron laser (FEL; see e.g., Brau, 1990), LMS can be extended to nonlinear and time-resolved FIR magneto-spectroscopy.

On the other hand, FTMS has some significant advantages with respect to LMS. First, currently available FIR laser sources (except for FELs) can produce only discrete wavelengths, whereas FTMS uses light sources that produce continuous spectra. Second, it is needless to say that LMS can only detect spectral features that are magnetic-field-dependent, so that it is unable to generate zero-field spectra. Third, LMS often overlooks or distorts features that have a very weak field dependence, in which case only FTMS can give unambiguous results, the

$1s \rightarrow 2p_-$ transition of shallow neutral donors being a good example (e.g., McCombe and Wagner, 1975). Finally, for studying 2-D electron systems, spectroscopy at fixed filling factors, namely at fixed magnetic fields, is sometimes crucial. In this section, after briefly reviewing FIR sources, these two modes of operation—FTMS and LMS—will be described in detail. In addition, short descriptions of two other unconventional methods—cross-modulation (or photoconductivity), in which the sample itself is used as a detector, and optically detected resonance (ODR) spectroscopy, which is a recently developed, highly sensitive method—will be provided. For the reader interested in a detailed description of FIR detectors and other FIR techniques, see, e.g., Kimmitt (1970), Stewart (1970), and Chantry (1971).

Far-Infrared Sources

The two "classic" sources of FIR radiation commonly used in FTMS are the globar and the Hg arc lamp. The globar consists of a rod of silicon carbide usually ~2 cm long and 0.5 cm in diameter. It is heated by electrical conduction; normally 5 A are passed through it, which raises its temperature to ~1500 K. The globar is bright at wavelengths between 2 and 40 μm, but beyond 40 μm its emissivity falls slowly, although it is still sufficient for spectroscopy up to ~100 μm. The mercury lamp has higher emissivity than the globar at wavelengths longer than 100 μm. It is normally termed a "high-pressure" arc, although the actual pressure is only 1 to 2 atm. (Low-pressure gaseous discharges are not useful here because they emit discrete line spectra.) It is contained in a fused quartz inner envelope. At the shorter wavelengths of the FIR, quartz is opaque, but it becomes very hot and emits thermal radiation. At the longer wavelengths, radiation from the mercury plasma is transmitted by the quartz and replaces the thermal radiation. Originally used by Rubens and von Baeyer (1911), the mercury arc lamp is still the most widely employed source in the FIR.

Three types of laser sources currently available to FIR spectroscopists are molecular-gas lasers, the FEL, and the *p*-type germanium laser. The most frequently used among these are the hundreds of laser lines available from a large number of molecular gases. The low-pressure gas consisting of HCN, H_2O, D_2O, CH_3OH, CH_3CN, etc., flows through a glass or metal tube, where population inversion is achieved either through a high-voltage discharge or by optical excitation with a CO_2 laser. Output powers range from a few μW to several hundred mW, depending on the line, gas pressure, pumping power, and whether continuous or pulsed excitation is used. The FEL, first operated in 1977, is an unusual laser source which converts the kinetic energy of free electrons to EM radiation. It is tunable in a wide range of frequencies, from millimeter to ultraviolet. An FEL consists of an electron gun, an accelerator, an optical cavity, and a periodic array of magnets called an undulator or wiggler. The wavelength of the output optical beam is determined by (1) the kinetic energy of the incident electrons, (2) the spatial period of the wiggler, and (3) the strength of the wiggler magnets, all of which are continuously tunable. With FEL's enormously high peak powers (up to ~1 GW) and short pulse widths (down to ~200 fsec), a new class of nonequilibrium phenomena is currently being explored in the FIR. The *p*-Ge laser is a new type of tunable solid-state FIR laser (for a review, see Gornik, 1991; Pidgeon, 1994). Its operation relies on the fact that streaming holes in the valence band in crossed strong electric and magnetic fields can result in an inverted hot-carrier distribution. Two different lasing processes, employing light hole–light hole and light hole–heavy hole transitions, respectively, have been identified (the former is the first realization of a CR laser). The lasing wavelength is continuously tunable, by adjusting the electric and magnetic fields. Lasing in a wide spectral range (75 to 500 μm) with powers up to almost 1 W has been reported.

Nonlinear optical effects provide tunable sources in the FIR. Various schemes exist, but the most thoroughly studied method has been Difference Frequency Mixing using the 9.6 and 10.6 μm lines from two CO_2 lasers (for a review, see, Aggarwal and Lax, 1977). InSb is usually used as the mixing crystal. Phase matching is achieved either through the use of temperature dependence of the anomalous dispersion or by using the free-carrier contribution to the refractive index in a magnetic field. As the CO_2 laser produces a large number of closely spaced lines at 9.6 and 10.6 μm, thousands of lines covering the FIR region from ~70 μm to mm can be produced. However, the efficiency of this method is very low: considerable input laser powers are necessary to obtain output powers only in the μW range. Another important method for generating FIR radiation is Optical Parametric Oscillation (OPO; see, e.g., Byer and Herbst, 1977). In this method, a birefringent crystal in an optical cavity is used to split the pump beam at frequency ω_3 into simultaneous oscillations at two other frequencies ω_1 (idler) and ω_2 (signal), where $\omega_1 + \omega_2 = \omega_3$. This is achieved either spontaneously (through vacuum fluctuation) or by feeding a beam at frequency ω_1. The advantages of OPO are high efficiency, wide tuning range, and an all solid-state system. The longest wavelength currently available from OPO is ~25 μm. More recently, the remarkable advances in high-speed opto-electronic and NIR/visible femtosecond laser technology have enabled generation and detection of ultrashort pulses of broadband FIR radiation (more frequently referred to as THz radiation or "T Rays" in this context). The technique has proven to be extremely useful for FIR spectroscopic measurements in the time domain. Many experiments have shown that ultrafast photoexcitation of semiconductors and semiconductor heterostructures can be used to generate coherent charge oscillations, which emit transient THz EM radiation. This is currently an active topic of research, and the interested reader is referred to, e.g., Nuss and Orenstein (1998) and references therein.

Fourier Transform FIR Magneto-Spectroscopy

A Fourier-transform spectrometer is essentially a Michelson type two-beam interferometer, the basic components of which are collimating optics, a fixed mirror, a beam splitter, and a movable mirror. The basic operation principle can be stated as follows. IR radiation emitted from the

light source is divided by the beam splitter into two beams with approximately the same intensity. One of the beams reaches the fixed mirror, and the other reaches the movable mirror. The two beams bounce back from the two mirrors and recombine at the beam splitter. When the movable mirror is at the zero-path-difference (ZPD) position, the output of the light intensity becomes maximum, since the two beams constructively interfere at all wavelengths. When the path difference, x, measured from the ZPD position is varied, an interference pattern as a function of x, called an interferogram, is obtained that is the FT of the spectrum of the light passing through the interferometer. Hence, by taking the inverse Fourier transform of the interferogram using a computer, one obtains the spectrum.

Two different types of FT spectrometers exist: (1) "slow-scan" (or step-scan) and (2) "fast-scan" spectrometers. In a slow-scan FT spectrometer, a stepping motor drives the movable mirror. A computer controls the step size in multiples of the fundamental step size, the dwell time at each mirror position, and the total number of steps. The product of the step size and the number of steps determines the total path difference, and hence the spectral resolution. A mechanical chopper (see Fig. 2) usually chops the FIR beam. The AC signal at this frequency from the detector is fed into a lock-in amplifier and the reference signal from the chopper into the reference input of the lock-in amplifier. The data acquisition occurs at each movable-mirror position, and thus an interferogram is constructed as the magnitude of the output versus the position of the movable mirror. Computer Fourier analysis with a Fast Fourier Transform algorithm converts the interferogram into an intensity versus frequency distribution—the spectrum. Rapid-scan FT spectrometers operate quite differently, although the basic principles are the same. The movable mirror of a rapid-scan FT machine is driven at a constant velocity. Instead of using a mechanical chopper, the constant velocity of the mirror produces a sinusoidal intensity variation with a unique frequency for each spectral element ω. The modulation frequency is given by $\vartheta = 2V\omega$, where V is the velocity of the mirror. High precision of parallel alignment between the two mirrors and the constant velocity of the moving mirror is provided *in situ* by a dynamic feedback controlling system. The signal sampling takes place at equally spaced mirror displacements, and is determined by the fringes of a He-Ne laser reference.

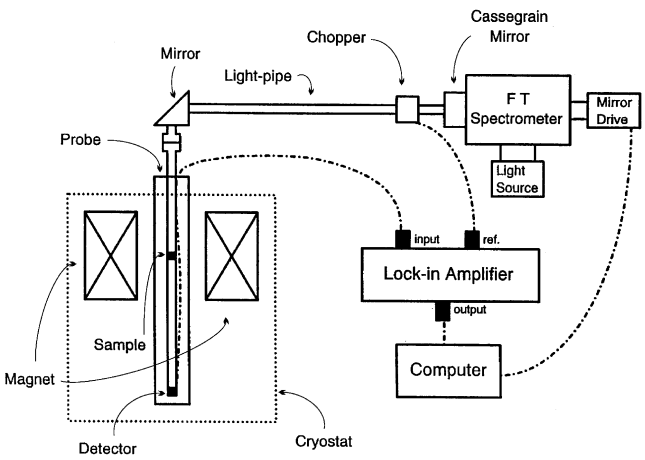


Figure 2. Schematic diagram of an experimental setup for CR studies with a (step-scan) Fourier transform spectrometer.

A slow-scan FTMS system for transmission CR studies is schematically shown in Fig. 2. FIR radiation generated by a Hg-arc lamp inside the spectrometer is coupled out by a Cassegrain output mirror and guided through a 3/4 -inch ("oversized") brass light-pipe to a 45° mirror. The beam reflected by the mirror is then directed down and passes through a white polyethylene window into a sample probe, which consists of a stainless-steel light pipe, a sample-holder/heater/temperature-sensor complex, metallic light cones, and a detector. The probe is sealed by the white polyethylene window and a stainless steel vacuum jacket, and inserted into a superconducting-magnet cryostat. The beam is focused by a condensing cone onto the sample located at the end of the light-cone at the center of the field. A black polyethylene filter is placed in front of the sample in order to filter out the high frequency part ($\geq$500 cm^{-1}) of the radiation from the light source. The FIR light transmitted through the sample is further guided by light-pipe/light-cone optics into a detector, which is placed at the bottom of the light-pipe system, far away from the center of the magnet. If a cancellation coil is available, the detector is placed at the center of the cancellation coil where $B = 0$. The sample and detector are cooled by helium exchange gas contained in the vacuum jacket of the probe.

Figures 3A and 3B show a typical interferogram and spectrum, respectively. The spectrum obtained contains not only the spectral response (transmittance in this case) of the sample but also the spectral response of combined effects of any absorption, filtering, and reflection when the light travels from the source to the detector, in addition to the output intensity spectrum of the radiation source. Therefore, in most experimental situations, spectra such as those obtained Fig. 3B are ratioed to an appropriate background spectrum taken under a different condition such as a different magnetic field, temperature, optical pumping intensity, or some other parameter that would only change the transmittance of the sample. In CR studies, spectra are usually ratioed to a zero-magnetic-field spectrum. In this way all the unwanted field-insensitive spectral structures are canceled out.

Laser FIR Magneto-Spectroscopy

LMS is generally easier to carry out than FTMS, although the experimental setup is almost identical to that of FTMS (the only difference is that the FT spectrometer is replaced by a laser in Fig. 2). This is partly because of the high power available from lasers compared with conventional radiation sources employed in FTMS, and also because mathematical treatments of the recorded data are not required. The data acquisition simply consists of monitoring an output signal from the detector that is proportional to the amount of FIR light transmitted by the sample while

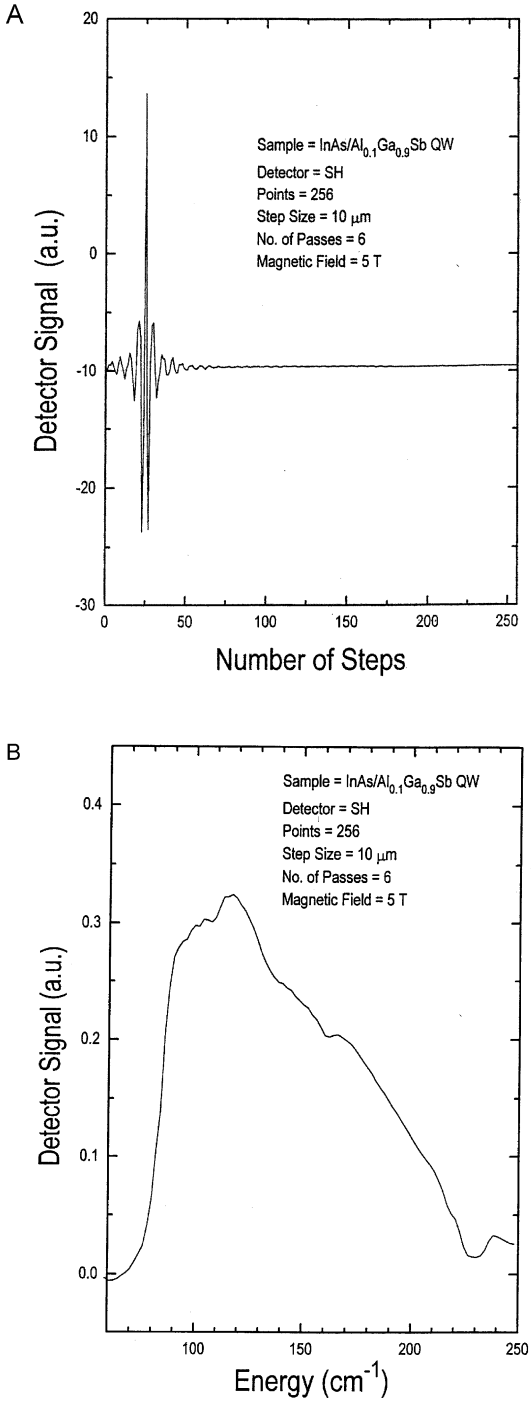


Figure 3. (**A**) An interferogram obtained with the FTMS setup shown in Fig. 2. (**B**) Spectrum obtained after Fourier transforming the interferogram in (A). This spectrum contains the output intensity spectrum of the Hg arc lamp, the spectral response of all the components between the source and the detector (filters, light-pipes, light cones, etc.), the responsivity spectrum of the detector (Ge:Ge extrinsic photoconductor), as well as the transmission spectrum of the sample.

the magnetic field is swept. The signal changes with magnetic field, decreasing resonantly and showing minima at resonant magnetic fields (see Fig. 4). Only magnetic-field-dependent features can thus be observed. If the laser output is stable while the field is being swept, no ratioing is necessary.

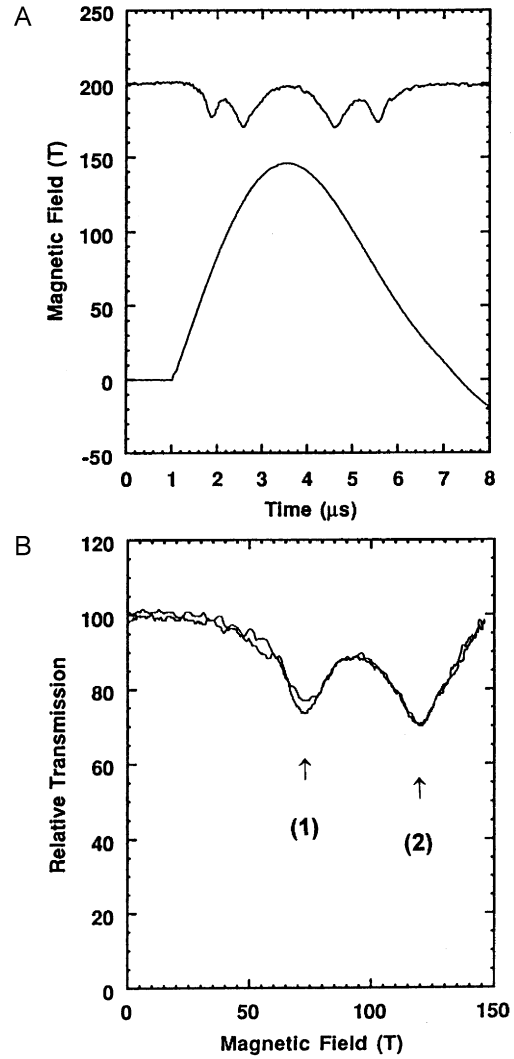


Figure 4. Example of CR with a laser in very high pulsed magnetic fields. (**A**) FIR transmission and magnetic field as functions of time. (**B**) Replot of the transmission signal as a function of magnetic field where the two traces arise from the rising and falling portions of the field strength shown in (A). Data obtained for *n*-type 3C-SiC.

A very important feature of LMS is that it can easily incorporate pulsed magnets, thus allowing observations of CR at very high magnetic fields (see Miura, 1984). Pulsed magnetic fields up to 40 to 60 Tesla with millisecond pulse durations can be routinely produced at various laboratories. Stronger magnetic fields in the megagauss (1 megagauss = 100 Tesla) range can be also generated by special destructive methods (see, e.g., Herlach, 1984) at some institutes. These strong fields have been used to explore new physics in the ultra-quantum limit (see, e.g., Herlach, 1984; Miura, 1984) as well as to observe CR in wide-gap, heavy-mass, and low-mobility materials unmeasurable by other techniques (see, e.g., Kono et al., 1993). The cyclotron energy of electrons in megagauss fields can

exceed other characteristic energies in solids such as binding energies of impurities and excitons, optical phonon energies, plasma energies, and even the fundamental band-gap in some materials, causing strong modifications in CR spectra.

An example of megagauss CR is shown in Fig. 4. Here, the recorded traces of the magnetic field pulse and the transmitted radiation from an *n*-type silicon carbide sample are plotted in Fig. 4A. The transmitted signal is then plotted as a function of magnetic field in Fig. 4B. The 36-μm line from a water-cooled H_2O vapor pulse laser and a Ge:Cu impurity photoconductive detector are used. The magnetic field is generated by a destructive single-turn coil method (see, e.g., Herlach, 1984), in which one shot of large current (~2.5 MA) from a fast capacitor bank (100 kJ, 40 kV) is passed through a thin single-turn copper coil. Although the coil is destroyed during the shot due to the EM force and the Joule heat, the sample can survive a number of such shots, so that repeated measurements can be made on the same sample. The resonance absorption is observed twice in one pulse, in the rising and falling slopes of the magnetic field, as seen in Fig. 4. It should be noted that the coincidence of the two traces (Fig. 4B) confirms a sufficiently fast response of the detector system.

Cross-Modulation (or Photoconductivity)

It is well known that the energy transferred from EM radiation to an electron system due to CR absorption induces significant heating in the system (free-carrier or Drude heating). This type of heating in turn induces a change in the conductivity tensor (e.g., an increase in $1/\tau$ or m^*), which causes (third-order) optical nonlinearity and modulation at other frequencies, allowing an observation of CR at a second frequency. Most conveniently, the DC conductivity $\sigma(\omega = 0)$ shows a pronounced change at a resonant magnetic field. This effect is known as Cross-Modulation or CR-induced photoconductivity, and has been described as a very sensitive technique to detect CR (Zeiger et al., 1959; Lax and Mavroides, 1960). The beauty of this method is that the sample acts as its own detector, so that there is no detector-related noise. The disadvantage is that the detection mechanism(s) is not well understood, so that quantitative lineshape analysis is difficult, unlike direct absorption. Either a decrease or increase in conductivity is observed, depending on a number of experimental conditions. Although many suggestions concerning the underlying mechanism(s) have been proposed, a complete understanding has been elusive.

Contrary to the situation of CR, which is a free carrier resonance, the mechanism of photoconductivity due to bound carrier resonances is much more clearly understood (see CARRIER LIFETIME: FREE CARRIER ABSORPTION, PHOTOCONDUCTIVITY, AND PHOTOLUMINESCENCE). For example, a resonant absorption occurs due to the 1*s* to 2*p* hydrogenic impurity transition in zero magnetic field in GaAs. Although the 2*p* state is below the continuum, an electron in the 2*p* state can be thermally excited into the conduction band, increasing conductivity (photo-thermal ionization). Other *p*-like excited states also can be studied in this manner, and these states evolve with increasing magnetic fields. As a result, one can map out the hydrogenic spectrum of the GaAs impurities simply by studying the photoconductivity of the sample. Since this is a null technique (i.e., there is photoresponse only at resonances), it is much more sensitive than transmission studies.

Optically-Detected Resonance Spectroscopy

Recently, a great deal of development work has centered on a new type of detection scheme, Optically-Detected Resonance (ODR) spectroscopy in the FIR. This novel technique possesses several significant advantages over conventional CR methods, stimulating considerable interest among workers in the community. With this technique, FIR resonances are detected through the change in the intensity of photoluminescence (PL; see CARRIER LIFETIME: FREE CARRIER ABSORPTION, PHOTOCONDUCTIVITY, AND PHOTOLUMINESCENCE) while the magnetic field is swept, rather than measuring FIR absorption directly. This technique, originally developed for the microwave region, was extended to the FIR by Wright et al. (1990) in studies of epitaxial GaAs, and subsequently by others. Remarkable sensitivity in comparison with conventional FIR methods has been demonstrated in studies of CR (e.g., Ahmed et al., 1992), impurity transitions (Michels et al., 1994; Kono et al., 1995), and internal transitions in excitons (Cerne et al., 1996; Salib et al., 1996). Since carriers are optically created, ODR enables detection of CR in "clean" systems with no intentional doping, with increased scattering time, and in materials for which doping is difficult. Furthermore, with ODR it is possible to select a specific PL feature in the spectrum, among various band-edge features, as the detection "channel." It is then possible to use the specificity of the FIR spectrum to obtain information about recombination mechanisms and the interactions that give rise to the various lines.

Figure 5A is a schematic of the experimental apparatus used for an ODR study (Kono et al., 1995). The sample is mounted in the Faraday geometry in a FIR lightpipe at the center of a 9-T superconducting magnet cooled to 4.2 K. PL is excited with the 632.8 nm line of a He-Ne laser via a 600-μm-diameter optical fiber. The signals are collected with a second 600-μm fiber, and analyzed with 0.25-m single-grating spectrometer/Si diode detector combination. A CO_2-pumped FIR laser is used to generate FIR radiation. The FIR laser power supply is chopped with the lock-in amplifier referenced to this chopped signal. A computer is used to simultaneously record the magnetic field values and the detected changes in the PL, and to step the monochromator to follow the center of the desired PL peak as it shifts with magnetic field. Two scans of the change in PL intensity as a function of magnetic field (ODR signal) for a FIR laser line of 118.8 μm are presented in Fig. 6. The upper trace shows a positive change (i.e., increase) in the intensity of the free exciton PL from a well-center-doped GaAs quantum well, whereas the lower trace shows a negative change (i.e., decrease) in the intensity of the bound exciton luminescence, demonstrating spectral specificity. For both scans, four different FIR resonances are clearly seen with an excellent signal-to-noise ratio (the sharp feature near 6T is the electron CR, and the other features are donor-related features).

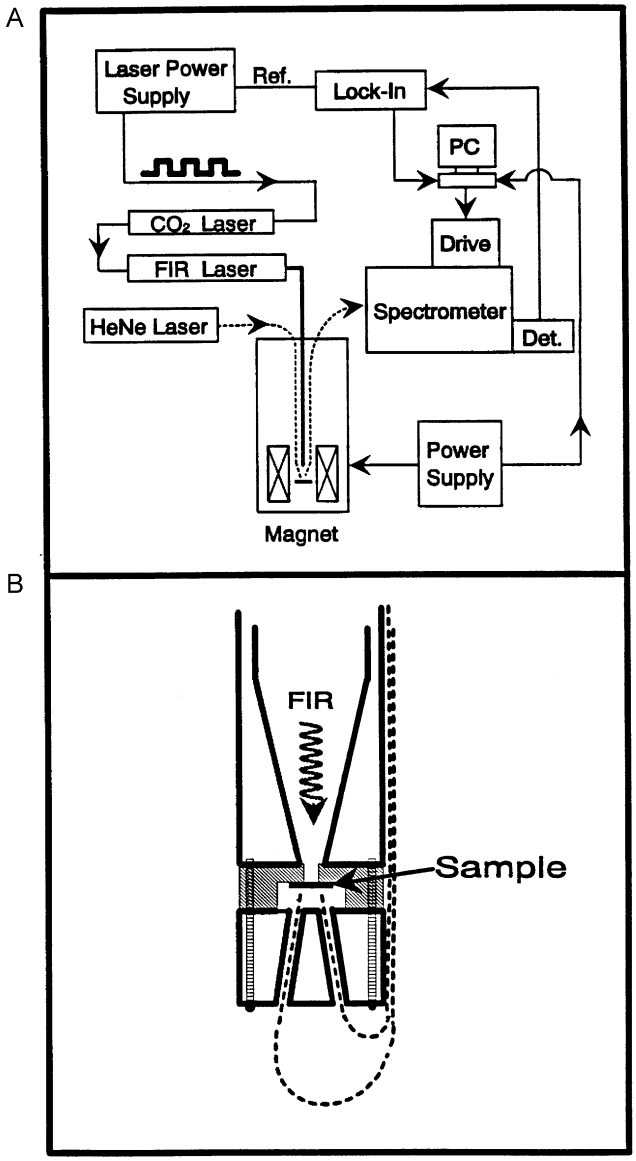


Figure 5. (**A**) Schematic diagram of an experimental setup for ODR spectroscopy. (**B**) Diagram of the FIR light cone, sample, and optical fiber arrangement used.

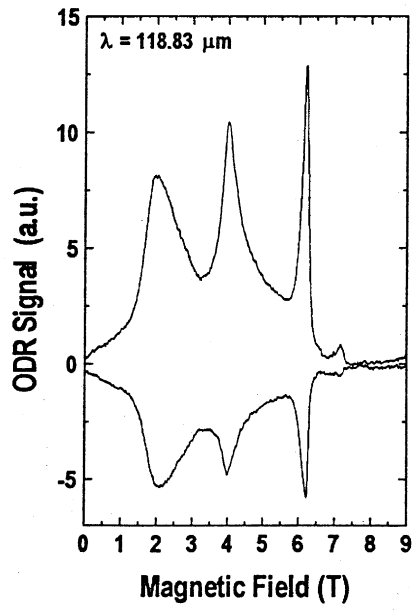


Figure 6. Two ODR spectra for a well-center-doped GaAs quantum well at a FIR wavelength of 118.8 μm. The upper trace shows a positive change in the intensity of the free exciton PL, whereas the lower trace shows a negative change in the intensity of the bound exciton luminescence, demonstrating the spectral specificity of ODR signals.

DATA ANALYSIS AND INITIAL INTERPRETATION

After minimizing noise to get as clean a spectrum as possible, and making sure that the spectrum is free from any artifacts and multiple-reflection interference effects, one can analyze resonance "positions" (i.e., magnetic fields in LMS and frequencies in FTMS). For each resonance feature, an effective mass m^* in units of m_0 (free electron mass in vacuum) can be obtained in different unit systems as follows

$$\frac{m^*}{m_0} = \frac{eB}{m_0\omega} = \frac{0.1158 \cdot B[\mathrm{T}]}{\hbar\omega[\mathrm{meV}]} = \frac{0.9337 \cdot B[\mathrm{T}]}{\tilde{\nu}[\mathrm{cm}^{-1}]} = \frac{27.99 \cdot B[\mathrm{T}]}{f[\mathrm{GHz}]} = 93.37 \cdot B[\mathrm{T}] \cdot \lambda[\mathrm{m}] \qquad (20)$$

Note that one can obtain an effective mass (a spectroscopic mass) for any resonance. For example, for the resonances (1) and (2) for SiC in Fig. 4B one can obtain $m_1 = 0.247m_0$ and $m_2 = 0.406m_0$, and for the resonances (a) – (d) for doped GaAs quantum wells in Fig. 5C one can obtain $m_a = 0.023m_0$, $m_b = 0.044m_0$, $m_c = 0.069m_0$, and $m_d = 0.079m_0$, irrespective of their origins. However, the spectroscopic mass is identical to the cyclotron mass only when the feature is due to free carrier CR; bound-carrier resonances such as the donor-related features in Fig. 6, have different spectroscopic masses at different frequencies/fields. Hence one needs to know which spectroscopic feature(s) arises from free-carrier CR. This can be found by two methods: temperature dependence and magnetic field (or frequency) dependence.

Examining the temperature dependence of a spectral feature is the easiest way to check the origin of the feature. As a rule of thumb, features associated with bound carriers increase in intensity with decreasing temperature at the expense of free carrier resonances. This is because in bulk semiconductors with doping level below the Mott condition, free carriers freeze out onto impurities, leaving no electrical conductivity at the lowest temperature. Thus, free-carrier CR grows with increasing temperature, but it broadens with the resulting increase of carrier scattering.

A more stringent test of whether a particular feature originates from CR is the response of its frequency versus magnetic field. In the case of LMS, one can take data only at discrete frequencies, but in the case of FTMS one can generate a continuous relationship between frequency and magnetic field. This is one of the advantages of FTMS over LMS as discussed earlier. Therefore, LMS is

usually performed to study, at a fixed wavelength with a higher resolution than FTMS and with circular polarizers if necessary, those spectral features whose magnetic-field-dependence is already known by FTMS. The resonance frequency versus magnetic field thus obtained for CR should show a straight line passing through the origin, if the nonparabolicity of the material is negligible and there are no localization effects (e.g., Nicholas, 1994). The slope of this straight line then provides the spectroscopic mass, which is constant and identical to the cyclotron mass and is also equal to the band-edge mass in this idealized situation. In the case of a nonparabolic band, the cyclotron mass gradually increases with magnetic field. This means that the slope versus B for CR becomes smaller (i.e., the line bends down) with increasing B.

Impurity-related lines, on the other hand, extrapolate to *finite* frequencies at zero magnetic field, corresponding to the zero-field binding energies of impurities. Different transitions have different slopes versus B, but all transitions originating from the same impurity atoms converge to an approximately common intersect at zero field. The most dominant donor-related line, the $1s$ to $2p_+$ transition, is sometimes called impurity-shifted CR (ICR). This is because its slope versus B becomes nearly equal to that of free electron CR in the high-field limit, i.e., $\hbar\omega_c \gg R_y^*$, where R_y^* is the binding energy of the impurity at zero field (McCombe and Wagner, 1975).

In many cases, multiple CR lines are observed in one spectrum (see, e.g., Dresselhaus et al., 1955; Otsuka, 1991; Petrou and McCombe, 1991; Nicolas, 1994). This generally indicates the existence of multiple types of carriers with different masses. Possible origins include: multi-valley splitting in the conduction band, light holes and heavy holes in the valence band, splitting due to resonant electron-phonon coupling, nonparabolicity-induced spin splitting, Landau-level splitting (see THEORY OF MAGNETIC PHASE TRANSITIONS), and population of two sub-bands in a quantum well. Explanation of each of these phenomena would be beyond the scope of this unit.

As discussed in the Introduction, CR linewidth is a sensitive probe for studying carrier scattering phenomena. In general, the linewidth of a CR line is related to the scattering lifetime, and the integrated absorption intensity of a CR line is related to the density of carriers participating in CR. Thus, if carrier density is constant, the product of absorption linewidth and depth is constant, even though the width and depth are not individually constant. More quantitatively, if the observed lineshape is well fitted by a Lorentzian, in the small absorption and reflection approximation it may be compared with

$$\frac{T(\omega,B)}{T(\omega,0)} = 1 - A(\omega,B) = 1 - \frac{1}{2}\frac{dne^2\tau}{c\varepsilon_0\kappa_l^{1/2}m^*}\frac{1}{1+(\omega-\omega_c)^2\tau^2} \tag{21}$$

where T is transmission, A is absorption, d is the sample thickness, c is the speed of light, and the other symbols have been defined earlier. The half-width at half maximum (HWHM) is thus equal to $1/\tau$. The peak absorption depth gives an estimate of the carrier density. For a more detailed and complete description on carrier transport studies using CR, see Otsuka (1991).

SAMPLE PREPARATION

Cyclotron resonance does not require any complicated sample preparation unless it is combined with other experimental techniques. The minimum sample size required depends on the design of the FIR optics used, i.e., how tightly the FIR beam can be focused onto the sample. Highly absorptive samples and samples with high carrier concentrations need to be polished down so that they are thin enough for transmission studies. In any case, wedging the sample substrates 2° to 3° is necessary to avoid multiple-reflection interference effects. Choosing the right sample is crucial for the success of a CR study. Samples with the highest possible carrier mobility are always preferable, if available. The DC mobility and density (see CONDUCTIVITY MEASUREMENT) can provide a rough estimate for the CR lineshape expected.

PROBLEMS

Generally speaking, the FIR (or THz) frequency regime, where CR is usually observed, is a difficult spectral range in which to carry out sophisticated spectroscopy. This range lies in the so-called "technology-gap" existing between electronics ($\leq$100 GHz) and optics ($\geq$10 THz) frequencies. The well-developed NIR/visible technology does not extend to this range; sources are dimmer and detectors are less sensitive. In addition, because of the lack of efficient non-linear crystals, there exist no amplitude or phase modulators in the FIR, except for simple mechanical choppers. Therefore, in many experiments, one deals with small signals having large background noise. In steady-state experiments with a step-scan FT spectrometer, lock-in techniques are always preferable. Modulating a property of the sample that only changes the size of the signal of interest—e.g., modulating the carrier density with tunable gate electrodes—has proven to be a very efficient way to detect small signals. The cross-modulation (or photoconductivity) technique is also frequently used to detect small signals since it is a very sensitive method, as discussed earlier.

Aside from this signal-to-noise problem inherent in the FIR, there are some additional problems that CR spectroscopists might encounter. A problem particularly important in bulk semiconductors is the carrier freeze-out effect mentioned earlier. In most semiconductors, low-temperature FIR magneto-spectra are dominated by impurity transitions. At high temperatures, free carriers are liberated from the impurities, but at the same time CR often becomes too broad to be observed because of the increased scattering rate. So one has to be careful in choosing the right temperature range to study CR. In very pure semiconductors, the only way to get any CR signal is by optical pumping. In Si and Ge, whose carrier lifetimes are very long ($\sim$msec in high-quality samples), one can create a large number of carriers sufficient for steady-state

FIR absorption spectroscopy. In direct-gap semiconductors such as GaAs, carrier lifetimes are very short ($\leq$1 nsec), so that it is nearly impossible to create enough carriers for steady-state FIR experiments, although short-pulse NIR-FIR two-color spectroscopy with an FEL is able to capture transient FIR absorption by photo-created nonequilibrium carriers. In low-dimensional semiconductor systems, so-called modulation doping is possible, where carriers can be spatially separated from their parent impurities so that they do not freeze out even at the lowest temperature.

The use of a strong magnet introduces a new class of problems. As we have seen above, in all CR studies, either in the form of FTMS or LMS, the transmission through the sample at finite magnetic field is compared with the transmission at zero magnetic field. The success of this ratioing relies on the assumption that it is only the sample that changes transmissivity with magnetic field. If anything else in the system changes some property with magnetic field, this method fails. Therefore, great care must be taken in order to make sure that no optical components have magnetic-field-dependent characteristics, that the FIR source and detector are not affected by the magnetic field, and that no component moves with magnetic field.

ACKNOWLEDGMENTS

The author would like to thank Prof. B. D. McCombe for useful discussions, comments, and suggestions. He is also grateful to Prof. N. Miura for critically reading the article, Prof. R. A. Stradling and Prof. C. R. Pidgeon for useful comments, and G. Vacca and D. C. Larrabee for proofreading the manuscript. This work was supported in part by NSF DMR-0049024 and ONR N00014-94-1-1024.

LITERATURE CITED

Aggarwal, R. L. and Lax, B. 1977. Optical mixing of CO_2 lasers in the far-infrared. *In* Nonlinear Infrared Generation—Vol. 16 of Topics in Applied Physics (Y.-R. Shen, ed.) pp. 19–80. Springer-Verlag, Berlin.

Ahmed, N., Agool, I. R., Wright, M. G., Mitchell, K., Koohian, A., Adams, S. J. A., Pidgeon, C. R., Cavenett, B. C., Stanley, C. R., and Kean, A. H. 1992. Far-infrared optically detected cyclotron resonance in GaAs layers and low-dimensional structures. *Semicond. Sci. Technol.* 7:357–363.

Aschcroft, N. W. and Mermin, N. D. 1976. Solid State Physics. Holt, Rinehart and Winston, Philadelphia.

Azbel, M. Ya. and Kaner, E. A. 1958. Cyclotron resonance in metals. *J. Phys. Chem. Solids* 6:113–135.

Bowers, R. and Yafet, Y. 1959. Magnetic susceptibility of InSb. *Phys. Rev.* 115:1165–1172.

Brau, C. A. 1990. Free-Electron Lasers. Academic Press, San Diego.

Byer, R. L. and Herbst, R. L. 1977. Parametric oscillation and mixing. *In* Nonlinear Infrared Generation—Vol. 16 of Topics in Applied Physics (Y.-R. Shen, ed.) pp. 81–137. Springer-Verlag, Berlin.

Cerne, J., Kono, J., Sherwin, M. S., Sundaram, M., Gossard, A. C., and Bauer, G. E. W. 1996. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. *Phys. Rev. Lett.* 77:1131–1134.

Chantry, G. W. 1971. Submillimeter Spectroscopy. Academic Press, New York.

Dresselhaus, G., Kip, A. F., and Kittel, C. 1953. Observation of cyclotron resonance in germanium crystals. *Phys. Rev.* 92:827.

Dresselhaus, G., Kip, A. F., and Kittel, C. 1955. Cyclotron resonance of electrons and holes in silicon and germanium crystals. *Phys. Rev.* 98:368–384.

Gornik, E. 1991. Landau emission. *In* Landau Level Spectroscopy—Vol. 27.2 of Modern Problems in Condensed Matter Sciences (G. Landwehr and E. I. Rashba, eds.) pp. 912–996. Elsevier Science, Amsterdam.

Herlach, F. ed. 1984. Strong and Ultrastrong Magnetic Fields and Their Applications—Vol. 57 of Topics in Applied Physics. Springer-Verlag, Berlin.

Hensel, J. C. and Suzuki, K. 1974. Quantum resonances in the valence bands of germanium. II. Cyclotron resonances in uniaxially stressed crystals. *Phys. Rev. B* 9:4219–4257.

Kawabata, A. 1967. Theory of cyclotron resonance linewidth. *J. Phys. Soc. Japan* 23:999–1006.

Kimmitt, M. F. 1970. Far-infrared techniques. Pion Limited, London.

Kittel, C. 1987. Quantum Theory of Solids. John Wiley & Sons, New York.

Kono, J., Miura, N., Takeyama, S., Yokoi, H., Fujimori, N., Nishibayashi, Y., Nakajima, T., Tsuji, K., and Yamanaka, M. 1993. Observation of cyclotron resonance in low-mobility semiconductors using pulsed ultra-high magnetic fields. *Physica B* 184:178–183.

Kono, J., Lee, S. T., Salib, M. S., Herold, G. S., Petrou, A., and McCombe, B. D. 1995. Optically detected far-infrared resonances in doped GaAs quantum wells. *Phys. Rev. B* 52: R8654-R8657.

Lax, B. and Mavroides, J. G. 1960. Cyclotron resonance. *In* Solid State Physics Vol. 11 (F. Seitz and D. Turnbull, eds.) pp. 261–400. Academic Press, New York.

Lax, B, Zeiger, H. J., Dexter, R. N., and Rosenblum, E. S. 1953. Directional properties of the cyclotron resonance in germanium. *Phys. Rev.* 93:1418–1420.

Luttinger, J. M. 1951. The effect of a magnetic field on electrons in a periodic potential. *Phys. Rev.* 84:814–817.

Luttinger, J. M. 1956. Quantum theory of cyclotron resonance in semiconductors: General theory. *Phys. Rev.* 102:1030–1041.

Luttinger, J. M. and Kohn, W. 1955. Motion of electrons and holes in perturbed periodic fields. *Phys. Rev.* 97:869–883.

Mavroides, J. G. 1972. Magneto-optical properties. *In* Optical Properties of Solids (F. Abeles, ed.) pp. 351–528. North-Holland, Amsterdam.

McCombe, B. D. and Wagner, R. J. 1975. Intraband magneto-optical studies of semiconductors in the far-infrared. *In* Advances in Electronics and Electron Physics (L. Marton, ed.) Vol. 37, pp. 1–78 and Vol. 38, pp. 1–53. Academic Press, New York.

Michels, J. G., Warburton, R. J., Nicholas, R. J., and Stanley, C. R. 1994. An optically detected cyclotron resonance study of bulk GaAs. *Semicond. Sci. Technol.* 9:198–206.

Miura, N. 1984. Infrared magnetooptical spectroscopy in semiconductors and magnetic materials in high pulsed magnetic fields. *In* Infrared and Millimeter Waves Vol. 12 (K. J. Button, ed.) pp. 73–143. Academic Press, New York.

Nicholas, R. J. 1994. Intraband optical properties of low-dimensional semiconductor systems. *In* Handbook on Semiconductors Vol. 2 "Optical Properties" (M. Balkanski, ed.) pp. 385–461. Elsevier, Amsterdam.

Nuss, M. C. and Orenstein, J. 1998. Terahertz time-domain spectroscopy. *In* Millimeter and Submillimeter Wave Spectroscopy of Solids (G. Grüner, ed.) pp. 7–44. Springer-Verlag, Berlin.

Otsuka, E. 1991. Cyclotron resonance. *In* Landau Level Spectroscopy—Vol. 27.1 of Modern Problems in Condensed Matter Sciences (G. Landwehr and E. I. Rashba, eds.) pp. 1–78. Elsevier Science, Amsterdam.

Palik, E. D. and Furdyna, J. K. 1970. Infrared and microwave magnetoplasma effects in semiconductors. *Rep. Prog. Phys.* 33: 1193–1322.

Petrou, A. and McCombe, B. D. 1991. Magnetospectroscopy of confined semiconductor systems. *In* Landau Level Spectroscopy—Vol. 27.2 of Modern Problems in Condensed Matter Sciences (G. Landwehr and E. I. Rashba, eds.) pp. 679–775. Elsevier Science, Amsterdam.

Pidgeon, C. R. 1994. Free carrier Landau level absorption and emission in semiconductors. *In* Handbook on Semiconductors Vol. 2 "Optical Properties" (M. Balkanski, ed.) pp. 637–678. Elsevier, Amsterdam.

Pidgeon, C. R. and Brown, R. N. 1966. Interband magneto-absorption and Faraday rotation in InSb. *Phys. Rev.* 146:575–583.

Rubens, H. and von Baeyer, O. 1911. On extremely long waves, emitted by the quartz mercury lamp. *Phil. Mag.* 21:689–695.

Sakurai, J. J. 1985. Modern Quantum Mechanics. Addison-Wesley Publishing Co., Redwood City, California.

Salib, M. S., Nickel, H. A., Herold, G. S., Petrou, A., McCombe, B. D., Chen, R., Bajaj, K. K., and Schaff, W. 1996. Observation of internal transitions of confined excitons in GaAs/AlGaAs quantum wells. *Phys. Rev. Lett.* 77:1135–1138.

Slater, J. C. 1949. Electrons in perturbed periodic lattices. *Phys. Rev.* 76:1592–1601.

Stewart, J. E. 1970. Infrared Spectroscopy. Marcel Dekker, New York.

Suzuki, K. and Hensel, J. C. 1974. Quantum resonances in the valence bands of germanium. I. Theoretical considerations. *Phys. Rev. B* 9:4184–4218.

Wannier, G. H. 1937. The structure of electronic excitation levels in insulating crystals. *Phys. Rev.* 52:191–197.

Wright, M. G., Ahmed, N., Koohian, A., Mitchell, K., Johnson, G. R., Cavenett, B. C., Pidgeon, C. R., Stanley, C. R., and Kean, A. H. 1990. Far-infrared optically detected cyclotron resonance observation of quantum effects in GaAs. *Semicond. Sci. Technol.* 5:438–441.

Zeiger, H. J., Rauch, C. J., and Behrndt, M. E. 1959. Cross modulation of D. C. resistance by microwave cyclotron resonance. *Phys. Chem. Solids* 8:496–498.

KEY REFERENCES

Dresselhaus et al.,1955. See above.

This seminal article is still a useful reference not only for CR spectroscopists but also for students beginning to study semiconductor physics. Both experimental and theoretical aspects of CR in solids as well as the band structure of these two fundamental semiconductors (Si and Ge) are described in great detail.

Landwehr, G. and Rashba, E. I. eds. 1991. Landau Level Spectroscopy—Vol. 27.1 and 27.2 of Modern Problems in Condensed Matter Sciences. Elsevier Science, Amsterdam.

These two volumes contain a number of excellent review articles on magneto-optical and magneto-transport phenomena in bulk semiconductors and low-dimensional semiconductor quantum structures.

Lax and Mavroides, 1960. See above.

This review article provides a thorough overview on early, primarily microwave, CR studies in the 1950s. A historical discussion is given also to CR of electrons and ions in ionized gases, which had been extensively investigated before CR in solids was first observed.

McCombe and Wagner, 1975. See above.

Describes a wide variety of far-infrared mangeto-optical phenomena observed in bulk semiconductors in the 1960s and 1970 s. Detailed describtions are given to basic theoretical formulations and experimental techniques as well as extensive coverage of experimental results.

J. KONO
Rice University
Houston, Texas

MÖSSBAUER SPECTROMETRY

INTRODUCTION

Mössbauer spectrometry is based on the quantum-mechanical "Mössbauer effect," which provides a nonintuitive link between nuclear and solid-state physics. Mössbauer spectrometry measures the spectrum of energies at which specific nuclei absorb γ rays. Curiously, for one nucleus to emit a γ ray and a second nucleus to absorb it with efficiency, the atoms containing the two nuclei must be bonded chemically in solids. A young R. L. Mössbauer observed this efficient γ-ray emission and absorption process in ^{191}Ir, and explained why the nuclei must be embedded in solids. Mössbauer spectrometry is now performed primarily with the nuclei ^{57}Fe, ^{119}Sn, ^{151}Eu, ^{121}Sb, and ^{161}Dy. Mössbauer spectrometry can be performed with other nuclei, but only if the experimenter can accept short radioisotope half-lives, cryogenic temperatures, and the preparation of radiation sources in hot cells.

Most applications of Mössbauer spectrometry in materials science utilize "hyperfine interactions," in which the electrons around a nucleus perturb the energies of nuclear states. Hyperfine interactions cause very small perturbations of 10^{-9} to 10^{-7} eV in the energies of Mössbauer γ rays. For comparison, the γ rays themselves have energies of 10^4 to 10^5 eV. Surprisingly, these small hyperfine perturbations of γ-ray energies can be measured easily, and with high accuracy, using a low-cost Mössbauer spectrometer.

Interpretations of Mössbauer spectra have few parallels with other methods of materials characterization. A Mössbauer spectrum looks at a material from the "inside out," where "inside" means the Mössbauer nucleus. Hyperfine interactions are sensitive to the electronic structure at the Mössbauer atom, or at its nearest neighbors. The important hyperfine interactions originate with the electron density at the nucleus, the gradient of the electric field at the nucleus, or the unpaired electron spins at the nucleus. These three hyperfine interactions are called the "isomer shift," "electric quadrupole splitting," and "hyperfine magnetic field," respectively.

The viewpoint from the nucleus is sometimes too small to address problems in the microstructure of materials.

Over the past four decades, however, there has been considerable effort to learn how the three hyperfine interactions respond to the environment around the nucleus. In general, it is found that Mössbauer spectrometry is best for identifying the electronic or magnetic structure at the Mössbauer atom itself, such as its valence, spin state, or magnetic moment. The Mössbauer effect is sensitive to the arrangements of surrounding atoms, however, because the local crystal structure will affect the electronic or magnetic structure at the nucleus. Different chemical and structural environments around the nucleus can often be assigned to specific hyperfine interactions. In such cases, measuring the fractions of nuclei with different hyperfine interactions is equivalent to measuring the fractions of the various chemical and structural environments in a material. Phase fractions and solute distributions, for example, can be determined in this way.

Other applications of the Mössbauer effect utilize its sensitivity to vibrations in solids, its timescale for scattering, or its coherence. To date these phenomena have seen little use outside the international community of a few hundred Mössbauer spectroscopists. Nevertheless, some new applications for them have recently become possible with the advent of synchrotron sources for Mössbauer spectrometry.

There have been a number of books written about the Mössbauer effect and its spectroscopies (see Key References). Most include reviews of materials research. These reviews typically demonstrate applications of the measurable quantities in Mössbauer spectrometry, and provide copious references. This unit is not a review of the field, but an instructional reference that gives the working materials scientist a basis for evaluating whether or not Mössbauer spectrometry may be useful for a research problem.

Recent research publications on Mössbauer spectrometry of materials have involved, in descending order in terms of numbers of papers: oxides, metals and alloys, organometallics, glasses, and minerals. For some problems, materials characterization by Mössbauer spectrometry is now "routine." A few representative applications to materials studies are presented. These applications were chosen in part according to the taste of the author, who makes no claim to have reviewed the literature of approximately 40,000 publications utilizing the Mössbauer effect (see Internet Resources for Mössbauer Effect Data Center Web site).

PRINCIPLES OF THE METHOD

Nuclear Excitations

Many properties of atomic nuclei and nuclear matter are well established, but these properties are generally not of importance to materials scientists. Since Mössbauer spectrometry measures transitions between states of nuclei, however, some knowledge of nuclear properties is necessary to understand the measurements.

A nucleus can undergo transitions between quantum states, much like the electrons of an atom, and doing so involves large changes in energy. For example, the first excited state of ^{57}Fe is 14.41 keV above its ground state. The Mössbauer effect is sometimes called "nuclear resonant γ-ray scattering" because it involves the emission of a γ ray from an excited nucleus, followed by the absorption of this γ ray by a second nucleus, which becomes excited. The scattering is called "resonant" because the phase and energy relationships for the γ-ray emission and absorption processes are much the same as for two coupled harmonic oscillators.

The state of a nucleus is described in part by the quantum numbers E, I, and I_z, where E is energy and I is the nuclear spin with orientation I_z along a z axis. In addition to these three internal nuclear coordinates, to understand the Mössbauer effect we also need spatial coordinates, X, for the nuclear center of mass as the nucleus moves through space or vibrates in a crystal lattice. These center-of-mass coordinates are decoupled from the internal excitations of the nucleus.

The internal coordinates of the nucleus are mutually coupled. For example, the first excited state of the nucleus ^{57}Fe has spin $I = 3/2$. For $I = 3/2$, there are four possible values of I_z, namely, $-3/2$, $-1/2$, $+1/2$, and $+3/2$. The ground state of ^{57}Fe has $I = 1/2$ and two allowed values of I_z. In the absence of hyperfine interactions to lift the energy degeneracies of spin levels, all allowed transitions between these spin levels will occur at the same energy, giving a total cross-section for nuclear absorption, σ_0, of 2.57×10^{-18} cm^2. Although σ_0 is smaller by a factor of 100 than a typical projected area of an atomic electron cloud, σ_0 is much larger than the characteristic size of the nucleus. It is also hundreds of times larger than the cross-section for scattering a 14.41-keV photon by the atomic electrons at ^{57}Fe.

The characteristic lifetime of the excited state of the ^{57}Fe nucleus, τ, is 141 ns, which is relatively long. An ensemble of independent ^{57}Fe nuclei that are excited simultaneously, by a flash of synchrotron light, for example, will decay at various times, t, with the probability per unit time of $\frac{1}{\tau}\exp(-1/\tau)$. The time uncertainty of the nuclear excited state, τ, is related to the energy uncertainty of the excited state, ΔE, through the uncertainty relationship, $\hbar \sim \Delta E\tau$. For $\tau = 141$ ns, the uncertainty relationship provides $\Delta E = 4.7 \times 10^{-9}$ eV. This is remarkably small—the energy of the nuclear excited state is extremely precise in energy. A nuclear resonant γ-ray emission or absorption has an oscillator quality factor, Q, of 3×10^{12}. The purity of phase of the γ ray is equally impressive. In terms of information technology, it is possible in principle to transmit high-quality audio recordings of all the Beethoven symphonies on a single Mössbauer γ-ray photon. The technology for modulating and demodulating this information remains problematic, however.

In the absence of significant hyperfine interactions, the energy dependence of the cross-section for Mössbauer scattering is of Lorentzian form, with a width determined by the small lifetime broadening of the excited state energy

$$\sigma_j(E) = \frac{\sigma_0 p_j}{1 + \left(\frac{E - E_j}{\Gamma/2}\right)^2} \quad (1)$$

where for ^{57}Fe, $\Gamma = \Delta E = 4.7 \times 10^{-9}$ eV, and E_j is the mean energy of the nuclear level transition ($\sim$14.41 keV). Here p_j is the fraction of nuclear absorptions that will occur with energy E_j. In the usual case where the energy levels of the different Mössbauer nuclei are inequivalent and the nuclei scatter independently, the total cross section is

$$\sigma(E) = \sum_j \sigma_j(E) \tag{2}$$

A Mössbauer spectrometry measurement is usually designed to measure the energy dependence of the total cross-section, $\sigma(E)$, which is often a sum of Lorentzian functions of natural line width Γ.

A highly monochromatic γ ray from a first nucleus is required to excite a second Mössbauer nucleus. The subsequent decay of the nuclear excitation need not occur by the reemission of a γ ray, however, and for ^{57}Fe only 10.9% of the decays occur in this way. Most of the decays occur by "internal conversion" processes, where the energy of the nuclear excited state is transferred to the atomic electrons. These electrons typically leave the atom, or rearrange their atomic states to emit an x ray. These conversion electrons or conversion x rays can themselves be used for measuring a Mössbauer spectrum. The conversion electrons offer the capability for surface analysis of a material. The surface sensitivity of conversion electron Mössbauer spectrometry can be as small as a monolayer (Faldum et al., 1994; Stahl and Kankeleit, 1997; Kruijer et al., 1997). More typically, electrons of a wide range of energies are detected, providing a depth sensitivity for conversion electron Mössbauer spectrometry of $\sim$100 nm (Gancedo et al., 1991; Williamson, 1993).

It is sometimes possible to measure coherent Mössbauer scattering. Here the total intensity, $I(E)$, from a sample is not the sum of independent intensity contributions from individual nuclei. One considers instead the total wave, $\Psi(\mathbf{r},E)$, at a detector located at $\mathbf{r}$. The total wave, $\Psi(\mathbf{r},E)$, is the sum of the scattered waves from individual nuclei, j

$$\Psi(\mathbf{r},E) = \sum_j \Psi_j(\mathbf{r},E) \tag{3}$$

Equation 3 is fundamentally different from Equation 2, since wave amplitudes rather than intensities are added. Since we add the individual Ψ_j, it is necessary to account precisely for the phases of the waves scattered by the different nuclei.

The Mössbauer Effect

Up to this point, we have assumed it possible for a second nucleus to become excited by absorbing the energy of a γ ray emitted by a first nucleus. There were a few such experiments performed before Mössbauer's discovery, but they suffered from a well recognized difficulty. As mentioned above, the energy precision of a nuclear excited state can be on the order of 10^{-8} eV. This is an extremely small energy target to hit with an incident γ ray. At room temperature, for example, vibrations of the nuclear center of mass have energies of 2.5×10^{-2} eV/atom. If there were any change in the vibrational energy of the nucleus caused by γ-ray emission, the γ ray would be far too imprecise in energy to be absorbed by the sharp resonance of a second nucleus. Such a change seems likely, since the emission of a γ ray of momentum $p_\gamma = E_\gamma/c$ requires the recoil of the emitting system with an opposite momentum (where E_γ is the γ-ray energy and c is the speed of light). A mass, m, will recoil after such a momentum transfer, and the kinetic energy in the recoil, E_{recoil}, will detract from the γ-ray energy

$$E_{\text{recoil}} = \frac{p_\gamma^2}{2m} = \left(\frac{E_\gamma^2}{2mc^2}\right) \tag{4}$$

For the recoil of a single nucleus, we use the mass of a ^{57}Fe nucleus for m in Equation 4, and find that $E_{\text{recoil}} = 1.86 \times 10^{-3}$ eV. This is again many orders of magnitude larger than the energy precision required for the γ ray to be absorbed by a second nucleus.

Rudolf Mössbauer's doctoral thesis project was to measure nuclear resonant scattering in ^{191}Ir. His approach was to use thermal Doppler broadening of the emission line to compensate for the recoil energy. A few resonant nuclear absorptions could be expected this way. To his surprise, the number of resonant absorptions was large, and was even larger when his radiation source and absorber were cooled to liquid nitrogen temperature (where the thermal Doppler broadening is smaller). Adapting a theory developed by W. E. Lamb for neutron resonant scattering (Lamb, 1939), Mössbauer interpreted his observed effect and obtained the equivalent of Equation 9, below. Mössbauer further realized that by using small mechanical motions, he could provide Doppler shifts to the γ-ray energies and tune through the nuclear resonance. He did so, and observed a spectrum without thermal Doppler broadening. In 1961, R. L. Mössbauer won the Nobel prize in physics. He was 32.

Mössbauer discovered (Mössbauer, 1958) that under appropriate conditions, the mass, m, in Equation 4 could be equal to the mass of the crystal. In such a case, the recoil energy is trivially small, the energy of the outgoing γ ray is precise to better than 10^{-9} eV, and the γ ray can be absorbed by exciting a second nucleus. The question is now how the mass, m, could be so large. The idea is that the nuclear mass is attached rigidly to the mass of the crystal. This sounds rather unrealistic, of course, and a better model is that the ^{57}Fe nucleus is attached to the crystal mass by a spring. This is the problem of a simple harmonic oscillator, or equivalently the Einstein model of a solid with Einstein frequency ω_E. The solution to this quantum mechanical problem shows that some of the nuclear recoils occur as if the nucleus is indeed attached rigidly to the crystal, but other γ-ray emissions occur by changing the state of the Einstein oscillator.

Nearly all of the energy of the emitted γ ray comes from changes in the internal coordinates of the nucleus, independently of the motion of the nuclear center of mass. The concern about the change in the nuclear center of mass coordinates arises from the conservation of the

momentum during the emission of a γ ray with momentum $p_\gamma = E_\gamma/c$. Eventually, the momentum of the γ-ray emission will be taken up by the recoil of the crystal as a whole. However, it is possible that the energy levels of a simple harmonic oscillator (comprising the Mössbauer nucleus bound to the other atoms of the crystal lattice) could be changed by the γ-ray emission. An excitation of this oscillator would depreciate the γ-ray energy by $n\hbar$ if n phonons are excited during the γ-ray emission. Since $\hbar\omega_E$ is on the order of 10^{-2} eV, any change in oscillator energy would spoil the possibility for a subsequent resonant absorption. In essence, changes in the oscillator excitation (or phonons in a periodic solid) replace the classical recoil energy of Equation 4 for spoiling the energy precision of the emitted γ ray. We need to calculate the probability that phonon excitation will not occur during γ-ray emission.

Before γ-ray emission, the wavefunction of the nuclear center of mass is $\psi_i(X)$, which can also be represented in momentum space through the Fourier transformation

$$\phi_i(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \exp\left(\frac{-ipX}{\hbar}\right) \psi_i(X')\, dX' \tag{5}$$

or

$$\psi_i(X) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \exp\left(\frac{ipX}{\hbar}\right) \phi_i(p)\, dp \tag{6}$$

The momentum space representation can handily accommodate the impulse of the γ-ray emission, to provide the final state of the nuclear center of mass, $\psi_f(X)$. Recall that the impulse is the time integral of the force, $F = dp/dt$, which equals the change in momentum. The analog to impulse in momentum space is a translation in real-space, such as $X \rightarrow X - X_0$. This corresponds to obtaining a final state by a shift in origin of an initial eigenstate. With the emission of a γ ray having momentum p_γ, we obtain the final state wave function from the initial eigenstate through a shift of origin in momentum space, $\phi_i(p) \rightarrow \phi_i(p - p_\gamma)$. We interpret the final state in real-space, $\psi_f(X)$, with Equation 6

$$\psi_f(X) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \exp\left(\frac{i(p + p_\gamma)X}{\hbar}\right) \phi_i(p) dp \tag{7}$$

Now, substituting Equation 5 into Equation 7

$$\psi_f(X) = \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} \exp\left(\frac{i(p + p_\gamma)X}{\hbar}\right) \times \int_{-\infty}^{\infty} \exp\left(\frac{-ipX'}{\hbar}\right) \psi_i(X') dX' dp \tag{8}$$

Isolating the integration over momentum, p

$$\psi_f(X) = \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} \exp\left(\frac{ip_\gamma X}{\hbar}\right) \psi_i(X') \left[\int_{-\infty}^{\infty} \exp\left(\frac{ip(X - X')}{\hbar}\right) dp \right] dX' \tag{9}$$

because this integration over p gives a Dirac delta function (times $2\pi\hbar$)

$$\psi_f(X) = \int_{-\infty}^{\infty} \exp\left(\frac{ip_\gamma X}{\hbar}\right) \psi_i(X') \delta(X - X')\, dX' \tag{10}$$

$$\psi_f(X) = \exp\left(\frac{ip_\gamma X}{\hbar}\right) \psi_i(X) \tag{11}$$

The exponential in Equation 11 is a translation of the eigenstate, $\Psi_i(X)$, in position for a fixed momentum transfer, $-p_\gamma$. It is similar to the translation in time, t, of an eigenstate with fixed energy, E, which is $\exp(-iEt/\hbar)$ or a translation in momentum for a fixed spatial translation, X_0, which is $\exp(-ipX_0/\hbar)$. (If the initial state is not an eigenstate, p_γ in Equation 11 must be replaced by an operator.)

For the nuclear center-of-mass wavefunction after γ-ray emission, we seek the amplitude of the initial state wavefunction that remains in the final state wavefunction. In Dirac notation

$$\langle i|f\rangle = \int_{-\infty}^{\infty} \psi_i^*(X) \psi_f(X)\, dX \tag{12}$$

Substituting Equation 11 into Equation 12, and using Dirac notation

$$\langle i|f\rangle = \langle i| \exp\left(\frac{ip_\gamma X}{\hbar}\right) |i\rangle \tag{13}$$

Using the convention for the γ-ray wavevector, $k_\gamma \equiv 2\pi\nu/c = E_\gamma/\hbar c$

$$\langle i|f\rangle = \langle i| \exp(ik_\gamma X) |i\rangle \tag{14}$$

The inner product $\langle i|f\rangle$ is the projection of the initial state of the nuclear center of mass on the final state after emission of the γ ray. It provides the probability that there is no change in the state of the nuclear center of mass caused by γ-ray emission. The probability of this recoilless emission, f, is the square of the matrix element of Equation 14, normalized by all possible changes of the center-of-mass eigenfunctions

$$f = \frac{|\langle i| \exp(ik_\gamma X) |i\rangle|^2}{\sum_j |\langle j| \exp(ik_\gamma X) |i\rangle|^2} \tag{15}$$

$$f = \frac{|\langle i| \exp(ik_\gamma X) |i\rangle|^2}{\sum_j \langle i| \exp(-ik_\gamma X) j\rangle \langle j| \exp(+ik_\gamma X) |i\rangle} \tag{16}$$

Using the closure relation $\Sigma_j |j\rangle\langle j| = 1$ and the normalization $\langle i|i\rangle = 1$, Equation 16 becomes

$$f = |\langle i| \exp(ik_\gamma X) |i\rangle|^2 \tag{17}$$

The quantity f is the "recoil-free-fraction." It is the probability that, after the γ ray removes momentum p_γ from the nuclear center of mass, there will be no change in the lattice state function involving the nuclear center of mass. In other words, f is the probability that a γ ray will be emitted with no energy loss to phonons. A similar factor is required for the absorption of a γ ray by a nucleus in a second crystal (e.g., the sample). The evaluation of f is straightforward for the ground state of the Einstein solid. The ground state wavefunction is

$$\psi^0_{\mathrm{CM}}(\mathrm{X}) = \left(\frac{m\omega_\mathrm{E}}{\pi\hbar}\right)^{1/4} \exp\left(-\frac{m\omega_\mathrm{E}\mathrm{X}^2}{2\hbar}\right) \qquad (18)$$

Inserting Equation 18 into Equation 17, and evaluating the integral (which is the Fourier transform of a Gaussian function)

$$f = \exp\left(-\frac{\hbar^2 k_\gamma^2}{2m\hbar\omega_\mathrm{E}}\right) = \exp\left(-\frac{E_\mathrm{R}}{\hbar\omega_\mathrm{E}}\right) = \exp\left(-k_\gamma^2\langle \mathrm{X}^2\rangle\right) \qquad (19)$$

where E_R is the recoil energy of a free ^{57}Fe nucleus, and $\langle \mathrm{X}^2\rangle$ is the mean-squared displacement of the nucleus when bound in an oscillator. It is somewhat more complicated to use a Debye model for calculating f with a distribution of phonon energies (Mössbauer, 1958). When the lattice dynamics are known, computer calculations can be used to obtain f from the full phonon spectrum of the solid, including the phonon polarizations. These more detailed calculations essentially confirm the result of Equation 19. The only nontrivial point is that low-energy phonons do not alter the result significantly. The recoil of a single nucleus does not couple effectively to long wavelength phonons, and there are few of them, so their excitation is not a problem for recoilless emission.

The condition for obtaining a significant number of "recoilless" γ-ray emissions is that the characteristic recoil energy of a free nucleus, E_R, is smaller than, or on the order of, the energy of the short wavelength phonons in the solid. These phonon energies are typically estimated from the Debye or Einstein temperatures of the solid to be a few tens of meV. Since $E_\mathrm{R} = 1.86 \times 10^{-3}$ eV for ^{57}Fe, this condition is satisfied nicely. It is not uncommon for most of the γ-ray emissions or absorptions from ^{57}Fe to be recoil-free. It is helpful that the energy of the γ ray, 14.41 keV, is relatively low. Higher-energy γ rays cause E_R to be large, as seen by the quadratic relation in Equation 4. Energies of most γ rays are far greater than 14 keV, so Mössbauer spectrometry is not practical for most nuclear transitions.

Overview of Hyperfine Interactions

Given the existence of the Mössbauer effect, the question remains as to what it can do. The answer is given in two parts: what are the phenomena that can be measured, and then what do these measurables tell us about materials? The four standard measurable quantities are the recoil-free fraction and the three hyperfine interactions: the isomer shift, the electric quadrupole splitting, and the hyperfine magnetic field. To date, the three hyperfine interactions have proved the most useful measurable quantities for the characterization of materials by Mössbauer spectrometry. This overview provides a few rules of thumb as to the types of information that can be obtained from hyperfine interactions. The section below (see More Exotic Measurable Quantities) describes quantities that are measurable, but which have seen fewer applications so far. For specific applications of hyperfine interactions for studies of materials, see Practical Aspects of the Method.

The isomer shift is the easiest hyperfine interaction to understand. It is a direct measure of electron density, albeit at the nucleus and away from the electron density responsible for chemical bonding between the Mössbauer atom and its neighbors. The isomer shift changes considerably with the valence of the Mössbauer atom in the cases of ^{57}Fe and ^{119}Sn. It is possible to use the isomer shift to estimate the fraction of Mössbauer isotope in different valence states, which may originate from different crystallographic site occupancies or from the presence of multiple phases in a sample. Valence analysis is often straightforward, and is probably the most common type of service work that Mössbauer spectroscopists provide for other materials scientists. The isomer shift has proven most useful for studies of ionic or covalently bonded materials such as oxides and minerals. Unfortunately, although the isomer shift is in principle sensitive to local atomic coordinations, it has usually not proven useful for structural characterization of materials, except when changes in valence are involved. The isomer shifts caused by most local structural distortions are generally too small to be useful.

Electric field gradients (EFG) are often correlated to isomer shifts. The existence of an EFG requires an asymmetric (i.e., noncubic) electronic environment around the nucleus, however, and this usually correlates with the local atomic structure. Again, like the isomer shift, the EFG has proven most useful for studies of oxides and minerals. Although interpretations of the EFG are not so straightforward as the isomer shift, the EFG is more capable of providing information about the local atomic coordination of the Mössbauer isotope. For ^{57}Fe, the shifts in peak positions caused by the EFG tend to be comparable to, or larger than, those caused by the isomer shift.

While isomer shifts are universal, hyperfine magnetic fields are confined to ferro-, ferri-, or antiferromagnetic materials. However, while isomer shifts tend to be small, hyperfine magnetic fields usually provide large and distinct shifts of Mössbauer peaks. Because their effects are so large and varied, hyperfine magnetic fields often permit detailed materials characterizations by Mössbauer spectrometry. For body-centered cubic (bcc) Fe alloys, it is known how most solutes in the periodic table alter the magnetic moments and hyperfine magnetic fields at neighboring Fe atoms, so it is often possible to measure the distribution of hyperfine magnetic fields and determine solute distributions about ^{57}Fe atoms. In magnetically ordered Fe oxides, the distinct hyperfine magnetic fields allow for ready identification of phase, sometimes more readily than by x-ray diffractometry.

Even in cases where fundamental interpretations of Mössbauer spectra are impossible, the identification of the local chemistry around the Mössbauer isotope is often possible by "fingerprint" comparisons with known standards. Mössbauer spectrometers tend to have similar instrument characteristics, so quantitative comparisons with published spectra are often possible. A literature search for related Mössbauer publications is usually enough to locate standard spectra for comparison. The Mössbauer Effect Data Center (see Internet Resources) is another resource that can provide this information.

Recoil-Free Fraction

An obvious quantity to measure with the Mössbauer effect is its intensity, given by Equation 19 as the recoil-free fraction, f. The recoil-free fraction is reminiscent of the Debye-Waller factor for x-ray diffraction. It is large when the lattice is stiff and ω_E is large. Like the Debye-Waller factor, f is a weighted average over all phonons in the solid. Unlike the Debye-Waller factor, however, f must be determined from measurements with only one value of wavevector k, which is of course k_γ.

It is difficult to obtain f from a single absolute measurement, since details about the sample thickness and absorption characteristics must be known accurately. Comparative studies may be possible with in situ experiments where a material undergoes a phase transition from one state to another while the macroscopic shape of the specimen is unchanged.

The usual way to determine f for a single-phase material is by measuring Mössbauer spectral areas as a function of temperature, T. Equation 19 shows that the intensity of the Mössbauer effect will decrease with $\langle X^2 \rangle$, the mean-squared displacement of the nuclear motion. The $\langle X^2 \rangle$ increases with T, so measurements of spectral intensity versus T can provide the means for determining f, and hence the Debye or Einstein temperature of the solid.

Another effect that occurs with temperature provides a measure of $\langle v^2 \rangle$, where v is the velocity of the nuclear center of mass. This effect is sometimes called the "second order Doppler shift," but it originates with special relativity. When a nucleus emits a γ ray and loses energy, its mass is reduced slightly. The phonon occupation numbers do not change, but the phonon energy is increased slightly owing to the diminished mass. This reduces the energy available to the γ-ray photon. This effect is usually of greater concern for absorption by the specimen, for which the energy shift is

$$E_{\text{therm}} = -\frac{1\langle V^2 \rangle}{2\, c^2} E_0 \qquad (20)$$

The thermal shift scales with the thermal kinetic energy in the sample, which is essentially a measure of temperature. For ^{57}Fe, $E_{\text{therm}} = -7.3 \times 10^{-4}$ mm/s K.

Isomer Shift

The peaks in a Mössbauer spectrum undergo observable shifts in energy when the Mössbauer atom is in different materials. These shifts originate from a hyperfine interaction involving the nucleus and the inner electrons of the atom. These "isomer shifts" are in proportion to the electron density at the nucleus. Two possibly unfamiliar concepts underlie the origin of the isomer shift. First, some atomic electron wavefunctions are actually present inside the nucleus. Second, the nuclear radius is different in the nuclear ground and excited states.

In solving the Schrödinger equation for radial wavefunctions of electrons around a point nucleus, it is found that for $r \to 0$ (i.e., toward the nucleus) the electron wavefunctions go as r^l, where l is the angular momentum quantum number of the electron. For s electrons (1s, 2s, 3s, 4s, etc.) with $l = 0$, the electron wavefunction is quite large at $r = 0$. It might be guessed that the wavefunctions of s electrons could make some sort of sharp wiggle so they go to zero inside the nucleus, but this would cost too much kinetic energy. The s electrons (and some relativistic p electrons) are actually present inside the nucleus. Furthermore, the electron density is essentially constant over the size of the nucleus.

The overlap of the s-electron wavefunction with the finite nucleus provides a Coulomb perturbation which lowers the nuclear energy levels. If the excited state and ground-state energy levels were lowered equally, however, the energy of the nuclear transition would be unaffected, and the emitted (or absorbed) γ ray would have the same energy. It is well known that the radius of an atom changes when an electron enters an excited state. The same type of effect occurs for nuclei—the nuclear radius is different for the nuclear ground and excited states. For ^{57}Fe, the effective radius of the nuclear excited state, R_{ex}, is smaller than the radius of the ground state, R_g, but for ^{119}Sn it is the other way around. For the overlap of a finite nucleus with a constant charge density, the total electrostatic attraction is stronger when the nucleus is smaller. This leads to a difference in energy between the nuclear excited state and ground state in the presence of a constant electron density $|\psi(0)|^2$. This shift in transition energy will usually be different for nuclei in the radiation source and nuclei in the sample, giving the following shift in position of the absorption peak in the measured spectrum

$$\Delta E_{\text{IS}} = \left[CZe^2(R_{\text{ex}}^2 - R_g^2)\right]\left[|\psi_{\text{sample}}(0)|^2 - |\psi_{\text{source}}(0)|^2\right] \qquad (21)$$

The factor C depends on the shape of the nuclear charge distribution, which need not be uniform or spherical. The sign of Equation 21 for ^{57}Fe is such that with an increasing s-electron density at the nucleus, the Mössbauer peaks will be shifted to more negative velocity. For ^{119}Sn, the difference in nuclear radii has the opposite sign. With increasing s-electron density at a ^{119}Sn nucleus, the Mössbauer peaks shift to more positive velocity.

There remains another issue in interpreting isomer shifts, however. In the case of Fe, the 3d electrons are expected to partly screen the nuclear charge from the 4s electrons. An increase in the number of 3d electrons at an ^{57}Fe atom will therefore increase this screening, reducing the s-electron density at the ^{57}Fe nucleus and causing a more positive isomer shift. The s-electron density at the

nucleus is therefore not simply proportional to the number of valence *s* electrons at the ion. The effect of this 3*d* electron screening is large for ionic compounds (Gütlich, 1975). In these compounds there is a series of trend lines for how the isomer shift depends on the 4*s* electron density, where the different trends correspond to the different number of 3*d* electrons at the ^{57}Fe atom (Walker et al., 1961). With more 3*d* electrons, the isomer shift is more positive, but also the isomer shift becomes less sensitive to the number of 4*s* electrons at the atom. Determining the valence state of Fe atoms from isomer shifts is generally a realistic type of experiment, however (see Practical Aspects of the Method).

For metals it has been more recently learned that the isomer shifts do not depend on the 3*d* electron density (Akai et al., 1986). In Fe alloys, the isomer shift corresponds nicely to the 4*s* charge transfer, in spite of changes in the 3*d* electrons at the Fe atoms. For the first factor in Equation 21, a proposed choice for ^{57}Fe is $\left[CZe^2(R_{ex}^2 - R_g^2)\right] = -0.24\ a_0^3$ mm/s (Akai et al., 1986), where a_0 is the Bohr radius of 0.529 Å.

Electric Quadrupole Splitting

The isomer shift, described in the previous section, is an electric monopole interaction. There is no static dipole moment of the nucleus. The nucleus does have an electric quadrupole moment that originates with its asymmetrical shape. The asymmetry of the nucleus depends on its spin, which differs for the ground and excited states of the nucleus. In a uniform electric field, the shape of the nuclear charge distribution has no effect on the Coulomb energy. An EFG, however, will have different interaction energies for different alignments of the electric quadrupole moment of the nucleus. An EFG generally involves a variation with position of the *x*, *y*, and *z* components of the electric field vector. In specifying an EFG, it is necessary to know, for example, how the *x* component of the electric field, $V_x \equiv \partial V/\partial x$ varies along the *y* direction, $V_y \equiv \partial^2 V/\partial y \partial x$ [here $V(x,y,z)$ is the electric potential]. The EFG involves all such partial derivatives, and is a tensor quantity. In the absence of competing hyperfine interactions, it is possible to choose freely a set of principal axes so that the off-diagonal elements of the EFG tensor are zero. By convention, we label the principal axes such that $|V_{zz}| > |V_{yy}| > |V_{xx}|$. Furthermore, because the Laplacian of the potential vanishes, $V_{xx} + V_{yy} + V_{zz} = 0$, there are only two parameters required to specify the EFG. These are chosen to be V_{zz} and an asymmetry parameter, $\eta \equiv (V_{xx} - V_{yy})/V_{zz}$.

The isotopes ^{57}Fe and ^{119}Sn have an excited-state spin of $I = 3/2$ and a ground-state spin of 1/2. The shape of the excited nucleus is that of a prolate spheroid. This prolate spheroid will be oriented with its long axis pointing along the *z* axis of the EFG when $I_z = \pm 3/2$. There is no effect from the sign of I_z, since inverting a prolate spheroid does not change its charge distribution. The $I_z = \pm 3/2$ states have a low energy compared to the $I_z = 1/2$ orientation of the excited state. In the presence of an EFG, the excited-state energy is split into two levels. Since $I_z = \pm 1/2$ for the ground state, however, the ground state

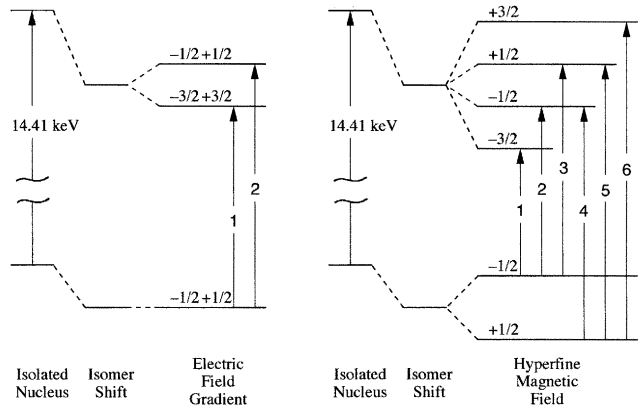


Figure 1. Energy level diagrams for ^{57}Fe in an electric field gradient (EFG; left) or hyperfine magnetic field (HMF; right). For an HMF at the sample, the numbers 1 to 6 indicate progressively more energetic transitions, which give experimental peaks at progressively more positive velocities. Sign convention is that an applied magnetic field along the direction of lattice magnetization will reduce the HMF and the magnetic splitting. The case where the nucleus is exposed simultaneously to an EFG and HMF of approximately the same energies is much more complicated than can be presented on a simple energy level diagram.

energy is not split by the EFG. With an electric quadrupole moment for the excited state defined as *Q*, for ^{57}Fe and ^{119}Sn the quadrupole splitting of energy levels is

$$\Delta E_q = \frac{\pm 1}{4} eQV_{zz}\left(1 + \frac{\eta^2}{3}\right)^{1/2} \qquad (22)$$

where often there is the additional definition $eq \equiv V_{zz}$. The energy level diagram is shown in Figure 1. By definition, $\eta < 1$, so the asymmetry factor can vary only from 1 to 1.155. For ^{57}Fe and ^{119}Sn, for which Equation 22 is valid, the asymmetry can usually be neglected, and the electric quadrupole interaction can be assumed to be a measure of V_{zz}. Unfortunately, it is not possible to determine the sign of V_{zz} easily (although this has been done by applying high magnetic fields to the sample).

The EFG is zero when the electronic environment of the Mössbauer isotope has cubic symmetry. When the electronic symmetry is reduced, a single line in the Mössbauer spectrum appears as two lines separated in energy as described by Equation 22 (as shown in Fig. 1). When the ^{57}Fe atom has a 3*d* electron structure with orbital angular momentum, V_{zz} is large. High- and low-spin Fe complexes can be identified by differences in their electric quadrupole splitting. The electric quadrupole splitting is also sensitive to the local atomic arrangements, such as ligand charge and coordination, but this sensitivity is not possible to interpret by simple calculations. The ligand field gives an enhanced effect on the EFG at the nucleus because the electronic structure at the Mössbauer atom is itself distorted by the ligand. This effect is termed "Sternheimer antishielding," and enhances the EFG from the ligands by a factor of about 7 for ^{57}Fe (Watson and Freeman, 1967).

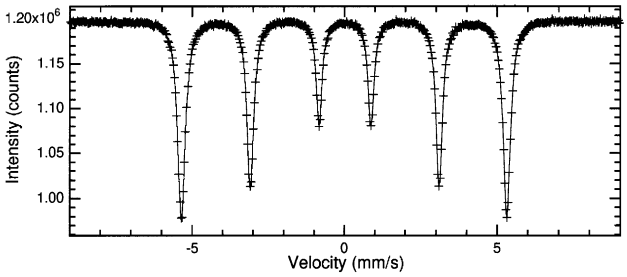


Figure 2. Mössbauer spectrum from bcc Fe. Data were acquired at 300 K in transmission geometry with a constant acceleration spectrometer (Ranger MS900). The points are the experimental data. The solid line is a fit to the data for six independent Lorentzian functions with unconstrained centers, widths, and depths. Also in the fit was a parabolic background function, which accounts for the fact that the radiation source was somewhat closer to the specimen at zero velocity than at the large positive or negative velocities. A ^{57}Co source in Rh was used, but the zero of the velocity scale is the centroid of the Fe spectrum itself. Separation between peaks 1 and 6 is 10.62 mm/s.

Hyperfine Magnetic Field Splitting

The nuclear states have spin, and therefore associated magnetic dipole moments. The spins can be oriented with different projections along a magnetic field. The energies of nuclear transitions are therefore modified when the nucleus is in a magnetic field. The energy perturbations caused by this HMF are sometimes termed the "nuclear Zeeman effect," in analogy with the more familiar splitting of energy levels of atomic electrons when there is a magnetic field at the atom.

A hyperfine magnetic field lifts all degeneracies of the spin states of the nucleus, resulting in separate transitions identifiable in a Mössbauer spectrum (see, e.g., Fig. 2). The I_z range from $-I$ to $+I$ in increments of 1, being $\{-3/2, -1/2, +1/2, +3/2\}$ for the excited state of ^{57}Fe and $\{-1/2, +1/2\}$ for the ground state. The allowed transitions between ground and excited states are set by selection rules. For the M1 magnetic dipole radiation for ^{57}Fe, six transitions are allowed: $\{(-1/2\rightarrow-3/2)\ (-1/2\rightarrow-1/2)\ (-1/2\rightarrow+1/2)\ (+1/2\rightarrow-1/2)\ (+1/2\rightarrow+1/2)\ (+1/2\rightarrow+3/2)\}$. The allowed transitions are shown in Figure 1. Notice the inversion in energy levels of the nuclear ground state.

In ferromagnetic iron metal, the magnetic field at the ^{57}Fe nucleus, the HMF, is 33.0 T at 300 K. The enormity of this HMF suggests immediately that it does not originate from the traditional mechanisms of solid-state magnetism. Furthermore, when an external magnetic field is applied to a sample of Fe metal, there is a decrease in magnetic splitting of the measured Mössbauer peaks. This latter observation shows that the HMF at the ^{57}Fe nucleus has a sign opposite to that of the lattice magnetization of Fe metal, so the HMF is given as -33.0 T.

It is easiest to understand the classical contributions to the HMF, denoted H_{mag}, H_{dip} and H_{orb}. The contribution H_{mag} is the magnetic field from the lattice magnetization, M, which is $4\pi M/3$. To this contribution we add any magnetic fields applied by the experimenter, and we subtract the demagnetization caused by the return flux. Typically, $H_{\mathrm{mag}} < +0.7$ T. The contribution H_{dip} is the classical dipole magnetic field caused by magnetic moments at atoms near the Mössbauer nucleus. In Fe metal, H_{dip} vanishes owing to cubic symmetry, but contributions of $+0.1$ T are possible when neighboring Fe atoms are replaced with nonmagnetic solutes. Finally, H_{orb} originates with any residual orbital magnetic moment from the Mössbauer atom that is not quenched when the atom is a crystal lattice. This contribution is about $+2$ T (Akai, 1986), and it may not change significantly when Fe metal is alloyed with solute atoms, for example. These classical mechanisms make only minor contributions to the HMF.

The big contribution to the HMF at a Mössbauer nucleus originates with the "Fermi contact interaction." Using the Dirac equation, Fermi and Segre discovered a new term in the Hamiltonian for the interaction of a nucleus and an atomic electron

$$\hbar_{\mathrm{FC}} = -\frac{8\pi}{3} g_{\mathrm{e}} g_{\mathrm{N}} \mu_{\mathrm{e}} \mu_{\mathrm{N}} \mathbf{I} \cdot \mathbf{S}\, \delta(\mathbf{r}) \qquad (23)$$

Here $\mathbf{I}$ and $\mathbf{S}$ are spin operators that act on the nuclear and electron wavefunctions, respectively, μ_{e} and μ_{N} are the electron and nuclear magnetons, and $(\mathbf{r})$ ensures that the electron wavefunction is sampled at the nucleus. Much like the electron gyromagnetic ratio, g_e, the nuclear gyromagnetic ratio, g_{N}, is a proportionality between the nuclear spin and the nuclear magnetic moment. Unlike the case for an electron, the nuclear ground and excited states do not have the same value of g_{N}; that of the ground state of ^{57}Fe is larger by a factor of -1.7145. The nuclear magnetic moment is $g_{\mathrm{N}}\ \mu_{\mathrm{N}} I$, so we can express the Fermi contact energy by considering this nuclear magnetic moment in an effective magnetic field, H_{eff}, defined as

$$H_{\mathrm{eff}} = \frac{8\pi}{3} g_{\mathrm{e}} \mu_{\mathrm{e}} S |\psi(0)|^2 \qquad (24)$$

where the electron spin is $\pm 1/2$, and $|\psi(0)|^2$ is the electron density at the nucleus. If two electrons of opposite spin have the same density at the nucleus, their contributions will cancel and H_{eff} will be zero. A large HMF requires an unpaired electron density at the nucleus, expressed as $|S| > 0$.

The Fermi contact interaction explains why the HMF is negative in ^{57}Fe. As described above (see Isomer Shift), only s electrons of Fe have a substantial presence at the nucleus. The largest contribution to the ^{57}Fe HMF is from $2s$ electrons, however, which are spin-paired core electrons. The reason that spin-paired core electrons can make a large contribution to the HMF is that the $2s\uparrow$ and $2s\downarrow$ wavefunctions have slightly different shapes when the Fe atom is magnetic. The magnetic moment of Fe atoms originates primarily with unpaired $3d$ electrons, so the imbalance in numbers of $3d\uparrow$ and $3d\downarrow$ electrons must affect the shapes of the paired $2s\uparrow$ and $2s\downarrow$ electrons.

These shapes of the $2s\uparrow$ and $2s\downarrow$ electron wavefunctions are altered by exchange interactions with the $3d\uparrow$ and $3d\downarrow$ electrons. The exchange interaction originates with the Pauli exclusion principle, which requires that a multielectron wavefunction be antisymmetric under the exchange

of electron coordinates. The process of antisymmetrization of a multielectron wavefunction produces an energy contribution from the Coulomb interaction between electrons called the "exchange energy," since it is the expectation value of the Coulomb energy for all pairs of electrons of like spin exchanged between their wavefunctions.

The net effect of the exchange interaction is to decrease the repulsive energy between electrons of like spin. In particular, the exchange interaction reduces the Coulomb repulsion between the $2s\uparrow$ and $3d\uparrow$ electrons, allowing the more centralized $2s\uparrow$ electrons to expand outward away from the nucleus. The same effect occurs for the $2s\downarrow$ and $3d\downarrow$ electrons, but to a lesser extent because there are fewer $3d\downarrow$ electrons than $3d\uparrow$ electrons in ferromagnetic Fe. The result is a higher density of $2s\downarrow$ than $2s\uparrow$ electrons at the ^{57}Fe nucleus. The same effect occurs for the $1s$ shell, and the net result is that the HMF at the ^{57}Fe nucleus is opposite in sign to the lattice magnetization (which is dominated by the $3d\uparrow$ electrons). The $3s$ electrons contribute to the HMF, but are at about the same mean radius as the $3d$ electrons, so their spin unbalance at the ^{57}Fe nucleus is smaller. The $4s$ electrons, on the other hand, lie outside the $3d$ shell, and exchange interactions bring a higher density of $4s\uparrow$ electrons into the ^{57}Fe nucleus, although not enough to overcome the effects of the $1s\downarrow$ and $2s\downarrow$ electrons. These $4s$ spin polarizations are sensitive to the magnetic moments at nearest neighbor atoms, however, and provide a mechanism for the ^{57}Fe atom to sense the presence of neighboring solute atoms. This is described below (see Solutes in bcc Fe Alloys).

More Exotic Measurable Quantities

Relaxation Phenomena. Hyperfine interactions have natural time windows for sampling electric or magnetic fields. This time window is the characteristic time, τ_{hf}, associated with the energy of a hyperfine splitting, $\tau_{hf} = \hbar/E_{hf}$. When a hyperfine electric or magnetic field undergoes fluctuations on the order of τ_{hf} or faster, observable distortions appear in the measured Mössbauer spectrum. The lifetime of the nuclear excited state does not play a direct role in setting the timescale for observing such relaxation phenomena. However, the lifetime of the nuclear excited state does provide a reasonable estimate of the longest characteristic time for fluctuations that can be measured by Mössbauer spectrometry.

Sensitivity to changes in valence of the Mössbauer atom between Fe(II) and Fe(III) has been used in studies of the Verwey transition in Fe_3O_4, which occurs at $\sim$120 K. Above the Verwey transition temperature the Mössbauer spectrum comprises two sextets, but when Fe_3O_4 is cooled below the Verwey transition temperature the spectrum becomes complex (Degrave et al., 1993).

Atomic diffusion is another phenomenon that can be studied by Mössbauer spectrometry (Ruebenbauer et al., 1994). As an atom jumps to a new site on a crystal lattice, the coherence of its γ-ray emission is disturbed. The shortening of the time for coherent γ-ray emission causes a broadening of the linewidths in the Mössbauer spectrum. In single crystals this broadening can be shown to occur by different amounts along different crystallographic directions, and has been used to identify the atom jump directions and mechanisms of diffusion in Fe alloys (Feldwisch et al., 1994; Vogl et al., 1994; Sepiol et al., 1996).

Perhaps the most familiar example of a relaxation effect in Mössbauer spectrometry is the superparamagnetic behavior of small particles. This phenomenon is described below (see Crystal Defects and Small Particles).

Phonons. The phonon partial density of states (DOS) has recently become measurable by Mössbauer spectrometry. Technically, nuclear resonant scattering that occurs with the creation or annihilation of a phonon is inelastic scattering, and therefore not the Mössbauer effect. However, techniques for measuring the phonon partial DOS have been developed as a capability of synchrotron radiation sources for Mössbauer scattering. The experiments are performed by detuning the incident photon energies above and below the nuclear resonance by 100 meV or so. This range of energy is far beyond the energy width of the Mössbauer resonance or any of its hyperfine interactions. However, it is in the range of typical phonon energies. The inelastic spectra so obtained are called "partial" phonon densities of states because they involve the motions of only the Mössbauer nucleus. The recent experiments (Seto et al., 1995; Sturhahn et al., 1995; Fultz et al., 1997) are performed with incoherent scattering (a Mössbauer γ ray into the sample, a conversion x ray out), and are interpreted in the same way as incoherent inelastic neutron scattering spectra (Squires, 1978). Compared to this latter, more established technique, the inelastic nuclear resonant scattering experiments have the capability of working with much smaller samples, owing to the large cross-section for nuclear resonant scattering. The vibrational spectra of monolayers of ^{57}Fe atoms at interfaces of thin films have been measured in preliminary experiments.

Coherence and Diffraction. Mössbauer scattering can be coherent, meaning that the phase of the incident wave is in a precise relationship to the phase of the scattered wave. For coherent scattering, wave amplitudes are added (Equation 3) instead of independent photon intensities (Equation 2). For the isotope ^{57}Fe, coherency occurs only in experiments where a 14.41 keV γ ray is absorbed and a 14.41 keV γ ray is reemitted through the reverse nuclear transition. The waves scattered by different coherent processes interfere with each other, either constructively or destructively. The interference between Mössbauer scattering and x-ray Rayleigh scattering undergoes a change from constructive in-phase interference above the Mössbauer resonance to destructive out-of-phase interference below. This gives rise to an asymmetry in the peaks measured in an energy spectrum, first observed by measuring a Mössbauer energy spectrum in scattering geometry (Black and Moon, 1960).

Diffraction is a specialized type of interference phenomenon. Of particular interest to the physics of Mössbauer diffraction is a suppression of internal conversion processes when diffraction is strong. With multiple transfers of energy between forward and diffracted beams, there is a nonintuitive enhancement in the rate of decay of the

nuclear excited state (Hannon and Trammell, 1969; van Bürck et al., 1978; Shvyd'ko and Smirnov, 1989), and a broadening of the characteristic linewidth. A fortunate consequence for highly perfect crystals is that with strong Bragg diffraction, a much larger fraction of the reemissions from ^{57}Fe nuclei occur by coherent 14.41 keV emission. The intensities of Mössbauer diffraction peaks therefore become stronger and easier to observe. For solving unknown structures in materials or condensed matter, however, it is difficult to interpret the intensities of diffraction peaks when there are multiple scatterings. Quantification of diffraction intensities with kinematical theory is an advantage of performing Mössbauer diffraction experiments on polycrystalline samples. Such samples also avoid the broadening of features in the Mössbauer energy spectrum that accompanies the speedup of the nuclear decay. Unfortunately, without the dynamical enhancement of coherent decay channels, kinematical diffraction experiments on small crystals suffer a serious penalty in diffraction intensity. Powder diffraction patterns have not been obtained until recently (Stephens et al., 1994), owing to the low intensities of the diffraction peaks. Mössbauer diffraction from polycrystalline alloys does offer a new capability, however, of combining the spectroscopic capabilities of hyperfine interactions to extract a diffraction pattern from a particular chemical environment of the Mössbauer isotope (Stephens and Fultz, 1997).

PRACTICAL ASPECTS OF THE METHOD

Radioisotope Sources

The vast majority of Mössbauer spectra have been measured with instrumentation as shown in Figure 3. The spectrum is obtained by counting the number of γ-ray photons that pass through a thin specimen as a function of the γ-ray energy. At energies where the Mössbauer effect is strong, a dip is observed in the γ-ray transmission. The γ-ray energy is tuned with a drive that imparts a Doppler shift, ΔE, to the γ ray in the reference frame of the sample:

$$\Delta E = \frac{v}{c} E_\gamma \qquad (25)$$

where v is the velocity of the drive. A velocity of 10 mm/s will provide an energy shift, ΔE, of 4.8×10^{-7} eV to a 14.41 keV γ ray of ^{57}Fe. Recall that the energy width of the Mössbauer resonance is 4.7×10^{-9} eV, which corresponds to 0.097 mm/s. An energy range of 10 mm/s is usually more than sufficient to tune through the full Mössbauer energy spectrum of ^{57}Fe or ^{119}Sn. It is conventional to present the energy axis of a Mössbauer spectrum in units of mm/s.

The equipment required for Mössbauer spectrometry is simple, and adequate instrumentation is often found in instructional laboratories for undergraduate physics students. In a typical coursework laboratory exercise, students learn the operation of the detector electronics and the spectrometer drive system in a few hours, and complete a measurement or two in about a week. (The understanding of the measured spectrum typically takes much longer.) Most components for the Mössbauer spectrometer in Figure 3 are standard items for x-ray detection and data acquisition. The items specialized for Mössbauer spectrometry are the electromagnetic drive and the radiation source. Abandoned electromagnetic drives and controllers are often found in university and industrial laboratories, and hardware manufactured since about 1970 by Austin Science Associates, Ranger Scientific, Wissel/Oxford Instruments, and Elscint, Ltd. are all capable of providing excellent results. Half-lives for radiation sources are: ^{57}Co, 271 days, ^{119m}Sn, 245 days, ^{151}Sm, 93 years, and ^{125}Te, 2.7 years. A new laboratory setup for ^{57}Fe or ^{119}Sn work may require the purchase of a radiation source. Suppliers include Amersham International, Dupont/NEN, and Isotope Products. It is also possible to obtain high-quality radiation sources from the former Soviet Union. Specifications for the purchase of a new Mössbauer source, besides activity level (typically 20 to 50 mCi for ^{57}Co), should include linewidth and sometimes levels of impurity radioisotopes.

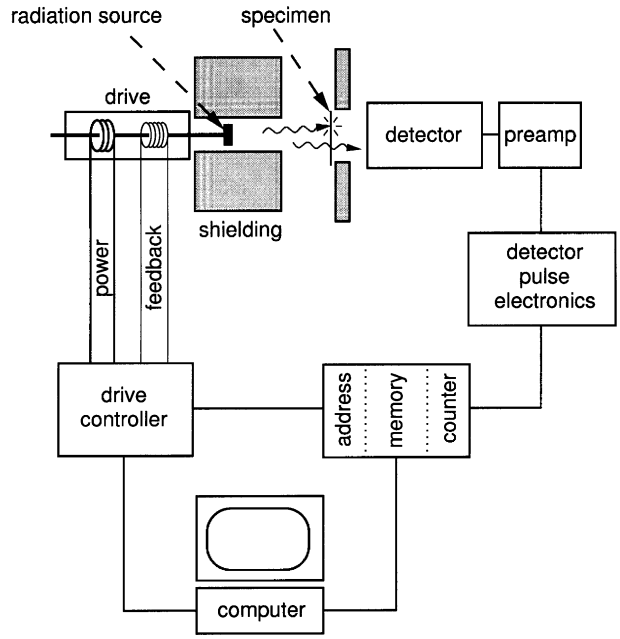


Figure 3. Transmission Mössbauer spectrometer. The radiation source sends γ rays to the right through a thin specimen into a detector. The electromagnetic drive is operated with feedback control by comparing a measured velocity signal with a desired reference waveform. The drive is cycled repetitively, usually so the velocity of the source varies linearly with time (constant acceleration mode). Counts from the detector are accumulated repetitively in short time intervals associated with memory addresses of a multichannel scaler. Each time interval corresponds to a particular velocity of the radiation source. Typical numbers are 1024 data points of 50-μs time duration and a period of 20 Hz.

The measured energy spectrum from the sample is convoluted with the energy spectrum of the radiation source. For a spectrum with sharp Lorentzian lines of natural linewidth, Γ (see Equation 1), the convolution of the source and sample Lorentzian functions provides a measured

Lorentzian function of full width at half-maximum of 0.198 mm/s. An excellent ^{57}Fe spectrum from pure Fe metal over an energy range of 10 mm/s may have linewidths of 0.23 mm/s, although instrumental linewidths of somewhat less than 0.3 mm/s are not uncommon owing to technical problems with the purity of the radiation source and vibrations of the specimen or source.

Radiation sources for ^{57}Fe Mössbauer spectrometry use the ^{57}Co radioisotope. The unstable ^{57}Co nucleus absorbs an inner-shell electron, transmuting to ^{57}Fe and emitting a 122-keV γ ray. The ^{57}Fe nucleus thus formed is in its first excited state, and decays about 141 ns later by the emission of a 14.41-keV γ ray. This second γ ray is the useful photon for Mössbauer spectrometry. While the 122-keV γ ray can be used as a clock to mark the formation of the ^{57}Fe excited state, it is generally considered a nuisance in Mössbauer spectrometry, along with emissions from other contamination radioisotopes in the radiation source. A Mössbauer radiation source is prepared by diffusing the ^{57}Co isotope into a matrix material such as Rh, so that atoms of ^{57}Co reside as dilute substitutional solutes on the fcc Rh crystal lattice. Being dilute, the ^{57}Co atoms have a neighborhood of pure Rh, and therefore all ^{57}Co atoms have the same local environment and the same nuclear energy levels. They will therefore emit γ rays of the same energy. Although radiation precautions are required for handling the source, the samples (absorbers) are not radioactive either before or after measurement in the spectrometer.

Enrichment of the Mössbauer isotope is sometimes needed when, for example, the 2.2% natural abundance of ^{57}Fe is insufficient to provide a strong spectrum. Although ^{57}Fe is not radioactive, material enriched to 95% ^{57}Fe costs approximately $15 to $30 per mg, so specimen preparation usually involves only small quantities of isotope. Biochemical experiments often require growing organisms in the presence of ^{57}Fe. This is common practice for studies on heme proteins, for example. For inorganic materials, it is sometimes possible to study dilute concentrations of Fe by isotopic enrichment. It is also common practice to use ^{57}Fe as an impurity, even when Fe is not part of the structure. Sometimes it is clear that the ^{57}Fe atom will substitute on the site of another transition metal, for example, and the local chemistry of this site can be studied with ^{57}Fe dopants. The same approach can be used with the ^{57}Co radioisotope, but this is not a common practice because it involves the preparation of radioactive materials. With ^{57}Co doping, the sample material itself serves as the radiation source, and the sample is moved with respect to a single-line absorber to acquire the Mössbauer spectrum. These "source experiments" can be performed with concentrations of ^{57}Co in the ppm range, providing a potent local probe in the material. Another advantage of source experiments is that the samples are usually so dilute in the Mössbauer isotope that there is no thickness distortion of the measured spectrum. The single-line absorber, typically sodium ferrocyanide containing 0.2 mg/cm^2 of ^{57}Fe, may itself have thickness distortion, but it is the same for all Doppler velocities. The net effect of absorber thickness is a broadening of spectral features without a distortion of intensities.

Synchrotron Sources

Since 1985 (Gerdau et al., 1985), it has become increasingly practical to perform Mössbauer spectrometry measurements with a synchrotron source of radiation, rather than a radioisotope source. This work has become more routine with the advent of Mössbauer beamlines at the European Synchrotron Radiation Facility at Grenoble, France, the Advanced Photon Source at Argonne National Laboratory, Argonne, Illinois, and the SPring-8 facility in Harima, Japan. Work at these facilities first requires success in an experiment approval process. Successful beamtime proposals will not involve experiments that can be done with radioisotope sources. Special capabilities that are offered by synchrotron radiation sources are the time structure of the incident radiation, its brightness and collimation, and the prospect of measuring energy spectra off-resonance to study phonons and other excitations in solids.

Synchrotron radiation for Mössbauer spectrometry is provided by an undulator magnet device inserted in the synchrotron storage ring. The undulator has tens of magnetic poles, positioned precisely so that the electron accelerations in each pole are arranged to add in phase. This provides a high concentration of radiation within a narrow range of angle, somewhat like Bragg diffraction from a crystal. This highly parallel radiation can be used to advantage in measurements through narrow windows, such as in diamond anvil cells. The highly parallel synchrotron radiation should permit a number of new diffraction experiments, using the Mössbauer effect for the coherent scattering mechanism.

Measurements of energy spectra are impractical with a synchrotron source, but equivalent spectroscopic information is available in the time domain. The method may be perceived as "Fourier transform Mössbauer spectrometry." A synchrotron photon, with time coherence less than 1 ns, can excite all resonant nuclei in the sample. Over the period of time for nuclear decay, 100 ns or so, the nuclei emit photon waves with energies characteristic of their hyperfine fields. Assume that there are two such hyperfine fields in the solid, providing photons of energy $E_\gamma^0 + \varepsilon_1$ and $E_\gamma^0 + \varepsilon_2$. In the forward scattering direction, the two photon waves can add in phase. The time dependence of the photon at the detector is obtained by the coherent sum as in Equation 3

$$T(t) = \exp[i(E_\gamma^0 + \varepsilon_1)t/\hbar] + \exp[i(E_\gamma^0 + \varepsilon_2)t/\hbar] \qquad (26)$$

The photon intensity at the detector, $I(t)$, has the time dependence

$$I(t) = T^*(t)T(t) = 2\{1 + \cos[\varepsilon_2 - \varepsilon_1)t/\hbar]\} \qquad (27)$$

When the energy difference between levels, $\varepsilon_2 - \varepsilon_1$, is greater than the natural linewidth, Γ, the forward scattered intensity measured at the detector will undergo a number of oscillations during the time of the nuclear decay. These "quantum beats" can be Fourier transformed to provide energy differences between hyperfine levels of the nucleus (Smirnov, 1996). It should be mentioned that forward scattering from thick samples also shows a

phenomenon of "dynamical beats," which involve energy interchanges between scattering processes.

Valence and Spin Determination

The isomer shift, with supplementary information provided by the quadrupole splitting, can be used to determine the valence and spin of ^{57}Fe and ^{119}Sn atoms. The isomer shift is proportional to the electron density at the nucleus, but this is influenced by the different σ- and π-donor acceptance strengths of surrounding ligands, their electronegativities, covalency effects, and other phenomena. It is usually best to have some independent knowledge about the electronic state of Fe or Sn in the material before attempting a valency determination. Even for unknown materials, however, valence and spin can often be determined reliably for the Mössbauer isotope.

The ^{57}Fe isomer shifts shown in Figure 4 are useful for determining the valence and spin state of Fe ions. If the ^{57}Fe isomer shift of an unknown compound is +1.2 mm/s with respect to bcc Fe, for example, it is identified as high-spin Fe(II). Low-spin Fe(II) and Fe(III) compounds show very similar isomer shifts, so it is not possible to distinguish between them on the basis of isomer shift alone. Fortunately, there are distinct differences in the electric quadrupole splittings of these electronic states. For low-spin Fe(II), the quadrupole splittings are rather small, being in the range of 0 to 0.8 mm/s. For low spin Fe(III) the electric quadrupole splittings are larger, being in the range 0.7 to 1.7 mm/s. The other oxidation states shown in Figure 4 are not so common, and tend to be of greater interest to chemists than materials scientists.

The previous analysis of valence and spin state assumed that the material is not magnetically ordered. In cases where a hyperfine magnetic field is present, identification of the chemical state of Fe is sometimes even easier. Table 1 presents a few examples of hyperfine magnetic fields and isomer shifts for common magnetic oxides and oxyhydroxides (Simmons and Leidheiser, 1976). This table is given as a convenient guide, but the hyperfine parameters may depend on crystalline quality and stoichiometry (Bowen et al., 1993).

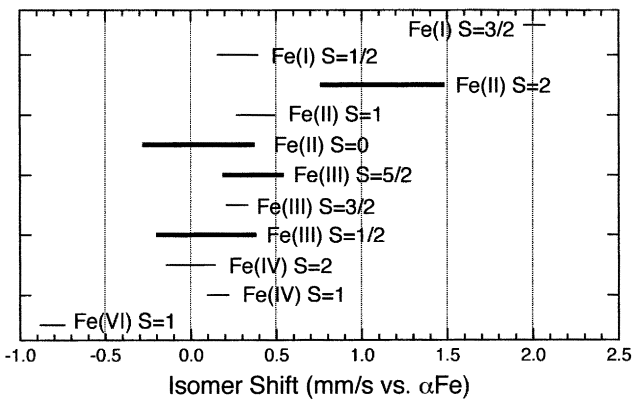


Figure 4. Ranges of isomer shifts in Fe compounds with various valences and spin states, with reference to bcc Fe at 300 K. Thicker lines are more common configurations (Greenwood and Gibb, 1971; Gütlich, 1975).

Table 1. Hyperfine Parameters of Common Oxides and Oxyhydroxides[a]

Compound (Fe Site)	HMF (T)	Q.S.	I.S. (vs. Fe)	Temp. (K)
α-FeOOH	50.0	−0.25		77
α-FeOOH	38.2	−0.25	+0.61	300
β-FeOOH	48.5	0.64	+0.38	80
β-FeOOH	0	0.62	+0.39	300
γ-FeOOH	0	0.60	+0.38	295
δ-FeOOH (large cryst.)	42.0		+0.35	295
FeO		0.8	+0.93	295
Fe_3O_4 (Fe(III), A)	49.3		+0.26	298
Fe_3O_4 (Fe(II, III), B)	46.0		+0.67	298
α-Fe_2O_3	51.8	+0.42	+0.39	296
γ-Fe_2O_3(A)	50.2		+0.18	300
γ-Fe_2O_3(B)	50.3		+0.40	300

[a]Abbreviations: HMF, hyperfine magnetic field; I.S., isomer shift; Q.S., quadrupole splitting; T, tesla.

The isomer shifts for ^{119}Sn compounds have a wider range than for ^{57}Fe compounds. Isomer shifts for compounds with Sn(IV) ions have a range from −0.5 to +1.5 mm/s versus SnO_2. For Sn(II) compounds, the range of isomer shifts is +2.2 to +4.2 versus SnO_2. Within these ranges it is possible to identify other chemical trends. In particular, for Sn compounds there is a strong correlation of isomer shift with the electronegativity of the ligands. This correlation between isomer shift and ligand electronegativity is especially reliable for Sn(IV) ions. Within a family of Sn(IV) compounds of similar coordination, the isomer shift depends on the electronegativity of the surrounding ligands as $-1.27\,\chi$ mm/s, where χ is the Pauling electronegativity. The correlation with Sn(II) is less reliable, in part because of the different coordinations found for this ion.

Finally, it should be mentioned that there have been a number of efforts to correlate the local coordination of ^{57}Fe with the electric quadrupole splitting. These correlations are often reliable within a specific class of compounds, typically showing a semiquantitative relationship between quadrupole splitting and the degree of distortion of the local atomic structure.

Phase Analysis

When more than one crystallographic phase is present in a material containing ^{57}Fe or ^{119}Sn, it is often possible to determine the phase fractions at least semiquantitatively. Usually some supplemental information is required before quantitative information can be derived. For example, most multiphase materials contain several chemical elements. Since Mössbauer spectrometry detects only the Mössbauer isotope, to determine the volume fraction of each phase, it is necessary to know its concentration of Mössbauer isotope. Quantitative phase analysis tends to be most reliable when the material is rich in the Mössbauer atom. Phase fractions in iron alloys, steels, and iron oxides can often be measured routinely by Mössbauer

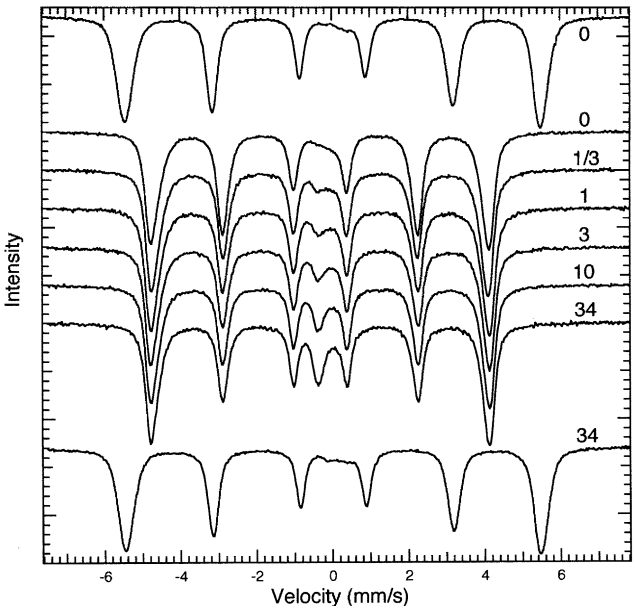


Figure 5. Mössbauer spectra of an alloy of Fe−8.9 atomic % Knee. The initial state of the material was ferromagnetic bcc phase, shown by the six-line spectrum at the top of the figure. This top spectrum was acquired at 23°C. The sample was heated in situ in the Mössbauer spectrometer to 600°C, for the numbers of hours marked on the curves, to form increasing amounts of fcc phase, evident as the single paramagnetic peak near −0.4 mm/s. This fcc phase is stable at 500°C, but not at 23°C, so the middle spectra were acquired at 500°C in interludes between heatings at 600°C for various times. At the end of the high-temperature runs, the sample temperature was again reduced to 23°C, and the final spectrum shown at the bottom of the figure showed that the fcc phase had transformed back into bcc phase. A trace of oxide is evident in all spectra as additional intensity around +0.4 mm/s at 23°C.

spectrometry (Schwartz, 1976; Simmons and Leidheiser, 1976; Cortie and Pollak, 1995; Campbell et al., 1995).

Figure 5 is an example of phase analysis of an Fe-Knee alloy, for which the interest was in determining the kinetics of fcc phase formation at 600°C (Fultz, 1982). The fcc phase, once formed, is stable at 500°C but not at room temperature. To determine the amount of fcc phase formed at 600°C it was necessary to measure Mössbauer spectra at 500°C without an intervening cooling to room temperature for spectrum acquisition. Mössbauer spectrometry is well suited for detecting small amounts of fcc phase in a bcc matrix, since the fcc phase is paramagnetic, and all its intensity appears as a single peak near the center of the spectrum. Amounts of fcc phase ("austenite") of 0.5% can be detected in iron alloys and steels, and quantitative analysis of the fcc phase fraction is straightforward.

The spectra in Figure 5 clearly show the six-line pattern of the bcc phase and the growth of the single peak at −0.4 mm/s from the fcc phase. These spectra show three other features that are common to many Mössbauer spectra. First, the spectrum at the top of Figure 5 shows a broadening of the outer lines of the sextet with respect to the inner lines (also see Fig. 2). This broadening originates with a distribution of hyperfine magnetic fields in alloys. The different numbers of Knee neighbors about the various ^{57}Fe atoms in the bcc phase cause different perturbations of the ^{57}Fe HMF. Second, the Curie temperature of bcc Fe−8.9 atomic % Knee is ∼700°C. At the Curie temperature the average lattice magnetization is zero, and the HMF is also zero. At 500°C the alloy is approaching the Curie temperature, and shows a strong reduction in its HMF as evidenced by the smaller splitting of the six line pattern with respect to the pattern at 23°C. Finally, at 500°C the entire spectrum is shifted to the left towards more negative isomer shift. This is the relativistic thermal shift of Equation 20.

To obtain the phase fractions, the fcc and bcc components of the spectrum were isolated and integrated numerically. Isolating the fcc peak was possible by digital subtraction of the initial spectrum from spectra measured after different annealing times. The fraction of the fcc spectral component then needed two correction factors to convert it into a molar phase fraction. One factor accounted for the different chemical compositions of the fcc and bcc phases (the fcc phase was enriched in Knee to about 25%). A second factor accounted for the differences in recoil-free fraction of the two phases. Fortunately, the Debye temperatures of the two phases were known, and they differed little, so the differences in recoil-free fraction were not significant. The amount of fcc phase in the alloy at 500°C was found to change from 0.5% initially to 7.5% after 34 h of heating at 600°C.

Solutes in bcc Fe Alloys

The HMF in pure bcc Fe is −33.0 T for every Fe atom, since every Fe atom has an identical chemical environment of 8 Fe first-nearest-neighbors (1nn), 6 Fe 2nn, 12 Fe 3nn, etc. In ferromagnetic alloys, however, the ^{57}Fe HMF is perturbed significantly by the presence of neighboring solute atoms. In many cases, this perturbation is about +2.5 T (a reduction in the magnitude of the HMF) for each 1nn solute atom. A compilation of some HMF perturbations for 1nn solutes and 2nn solutes is presented in Figure 6. These data were obtained by analysis of Mössbauer spectra from dilute bcc Fe-X alloys (Vincze and Campbell, 1973; Vincze and Aldred, 1974; Fultz, 1993).

In general, the HMF perturbations at ^{57}Fe nuclei from nearest-neighbor solute atoms originate from several sources, but for nonmagnetic solutes such as Si, the effects are fairly simple to understand. When the Si atom substitutes for an Fe atom in the bcc lattice, a magnetic moment of 2.2 μ_B is removed (the Fe) and replaced with a magnetic hole (the Si). The 4*s* conduction electrons redistribute their spin density around the Si atom, and this redistribution is significant at 1nn and 2nn distances. The Fermi contact interaction and H_{eff} (Equation 24) are sensitive to the 4*s* electron spin density, which has finite probability at the ^{57}Fe nucleus. Another important feature of 3*p*, 4*p*, and 5*p* solutes is that their presence does not significantly affect the magnetic moments at neighboring Fe atoms. Bulk magnetization measurements on Fe-Si and Fe-Al alloys, for example, show that the magnetic moment of the material decreases approximately in proportion to the fraction of Al or Si in the alloy. The core electron

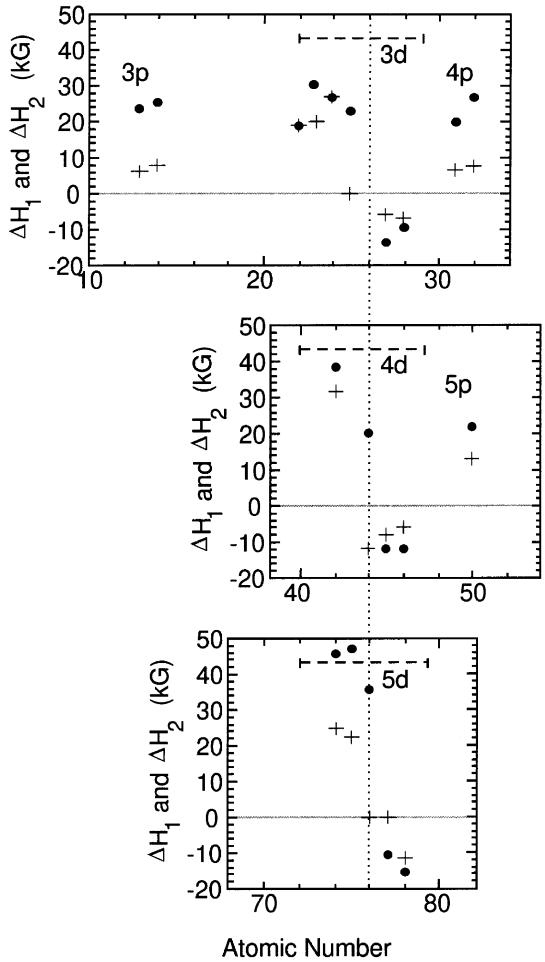


Figure 6. The hyperfine magnetic field perturbation, ΔH_1^X, at a Fe atom caused by one 1nn solute of type X, and the 2nn perturbation, ΔH_2^X, versus the atomic number of the solute. The vertical line denotes the column of Fe in the periodic table.

polarization, which involves exchange interactions between the unpaired 3*d* electrons at the ^{57}Fe atom and its inner-shell *s* electrons, is therefore not much affected by the presence of Si neighbors. The dominant effect comes from the magnetic hole at the solute atom, which causes the redistribution of 4*s* spin density. Figure 6 shows that the nonmagnetic 3*p*, 4*p*, and 5*p* elements all cause about the same HMF perturbation at neighboring ^{57}Fe atoms, as do the nonmagnetic early 3*d* transition metals.

For solutes that perturb significantly the magnetic moments at neighboring Fe atoms, especially the late transition metals, the core polarization at the ^{57}Fe atom is altered. There is an additional complicating effect from the matrix Fe atoms near the ^{57}Fe atom, whose magnetic moments are altered enough to affect the 4*s* conduction electron polarization at the ^{57}Fe (Fultz, 1993).

The HMF distribution can sometimes provide detailed information on the arrangements of solutes in nondilute bcc Fe alloys. For most solutes (that do not perturb significantly the magnetic moments at Fe atoms), the HMF at a ^{57}Fe atom depends monotonically on the number of solute atoms in its 1nn and 2nn shells. Hyperfine magnetic field perturbations can therefore be used to measure the chemical composition or the chemical short-range order in an alloy containing up to 10 atomic % solute or even more. In many cases, it is possible to distinguish among Fe atoms having different numbers of solute atoms as first neighbors, and then determine the fractions of these different first neighbor environments. This is considerably more information on chemical short-range order (SRO) than just the average number of solute neighbors, as provided by a 1nn Warren-Cowley SRO parameter, for example.

An example of chemical short range order in an Fe–26 atomic % Al alloy is presented in Figure 7. The material was cooled at a rate of 10^6 K/s from the melt by piston-anvil quenching, producing a polycrystalline ferromagnetic alloy with a nearly random distribution of Al atoms on the bcc lattice. With low-temperature annealing, the material evolved toward its equilibrium state of $D0_3$ chemical order. The Mössbauer spectra in Figure 7A change significantly as the alloy evolves chemical order. The overlap of several sets of six line patterns does confuse the physical picture, however, and further analysis requires the extraction of an HMF distribution from the experimental data. Software packages available for such work are described below (see Data Analysis). Figure 7B shows HMF distributions extracted from the three spectra of Figure 7A. At the top of Figure 7B are markers indicating the numbers of Al atoms in the 1nn shell of the ^{57}Fe nucleus associated with the HMF. With low-temperature annealing, there is a clear increase in the numbers of ^{57}Fe atoms with 0 and 4 Al neighbors, as expected when $D0_3$ order is evolving in the material. The perfectly ordered $D0_3$ structure has two chemical sites for Fe atoms, one with 0 Al neighbors and the other with 4 Al neighbors, in a 1:2 ratio. The HMF distributions were fit to a set of Gaussian functions to provide data on the chemical short range order in the alloys. These

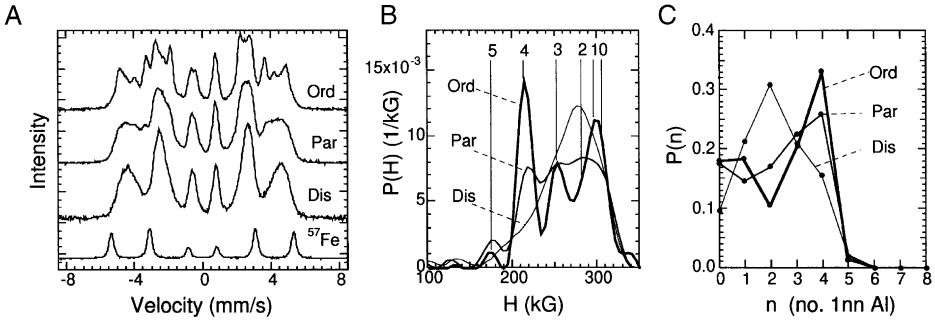


Figure 7. (**A**) Conversion electron Mössbauer spectra from a specimen of bcc ^{57}Fe and three specimens of disordered, partially ordered, and $D0_3$-ordered $^{57}Fe_3Al$. (**B**) HMF distribution of the $^{57}Fe_3Al$ specimens. Peaks in the HMF distribution are labeled with numbers indicating the different numbers of 1nn Al neighbors about the ^{57}Fe atom. (**C**) Probabilities for the ^{57}Fe atom having various numbers, *n*, of Al atoms as 1nn.

data on chemical short-range order are presented in Figure 7C.

Crystal Defects and Small Particles

Since Mössbauer spectrometry probes local environments around a nucleus, it has often been proposed that Mössbauer spectra should be sensitive to the local atomic structures at grain boundaries and defects such as dislocations and vacancies. This is in fact true, but the measured spectra are an average over all Mössbauer atoms in a sample. Unless the material is chosen carefully so that the Mössbauer atom is segregated to the defect of interest, the spectral contribution from the defects is usually overwhelmed by the contribution from Mössbauer atoms in regions of perfect crystal. The recent interest in nanocrystalline materials, however, has provided a number of new opportunities for Mössbauer spectrometry (Herr et al., 1987; Fultz et al., 1995). The number of atoms at and near grain boundaries in nanocrystals is typically 35% for bcc Fe alloys with crystallite sizes of 7 nm or so. Such a large fraction of grain boundary atoms makes it possible to identify distinct contributions from Mössbauer atoms at grain boundaries, and to identify their local electronic environment.

Mössbauer spectrometry can provide detailed information on some features of small-particle magnetism (Mørup, 1990). When a magnetically ordered material is in the form of a very small particle, it is easier for thermal energy to realign its direction of magnetization. The particle retains its magnetic order, but the change in axis of magnetization will disturb the shape of the Mössbauer spectrum if the magnetic realignment occurs on the time scale t_s, which is $\hbar$ divided by the hyperfine magnetic field energy (see Relaxation Phenomena for discussion of the time window for measuring hyperfine interactions). An activation energy is associated with this "superparamagnetic" behavior, which is the magnetocrystalline anisotropy energy times the volume of the crystallite, κV. The probability of activating a spin rotation in a small particle is the Boltzmann factor for overcoming the anisotropy energy, so the condition for observing a strong relaxation effect in the Mössbauer spectrum is

$$t_s = A \exp(-\kappa V / k_B T_b) \qquad (28)$$

The temperature, T_b, satisfying Equation 28 is known as the "blocking temperature." The prefactor of Equation 28, the attempt frequency, is not so well understood, so studies of superparamagnetic behavior often study the blocking temperature versus the volume of the particles. In practice, most clusters of small particles have a distribution of blocking temperatures, and there are often interactions between the magnetic moments at adjacent particles. These effects can produce Mössbauer spectra with a wide variety of shapes, including very broad Lorentzian lines.

At temperatures below T_b, the magnetic moments of small particles undergo small fluctuations in their alignment. These small-amplitude fluctuations can be considered as vibrations of the particle magnetization about an average orientation, which serve to reduce the HMF by a modest amount. At increasing temperatures around T_b, however, large fluctuations occur in the magnetic alignment. The result is first a severe uncertainty of the HMF distribution, leading to a very broad background in the spectrum, followed by the growth of a paramagnetic peak near zero velocity. All of these effects can be observed in the spectra shown in Figure 8. Here, the biomaterial samples

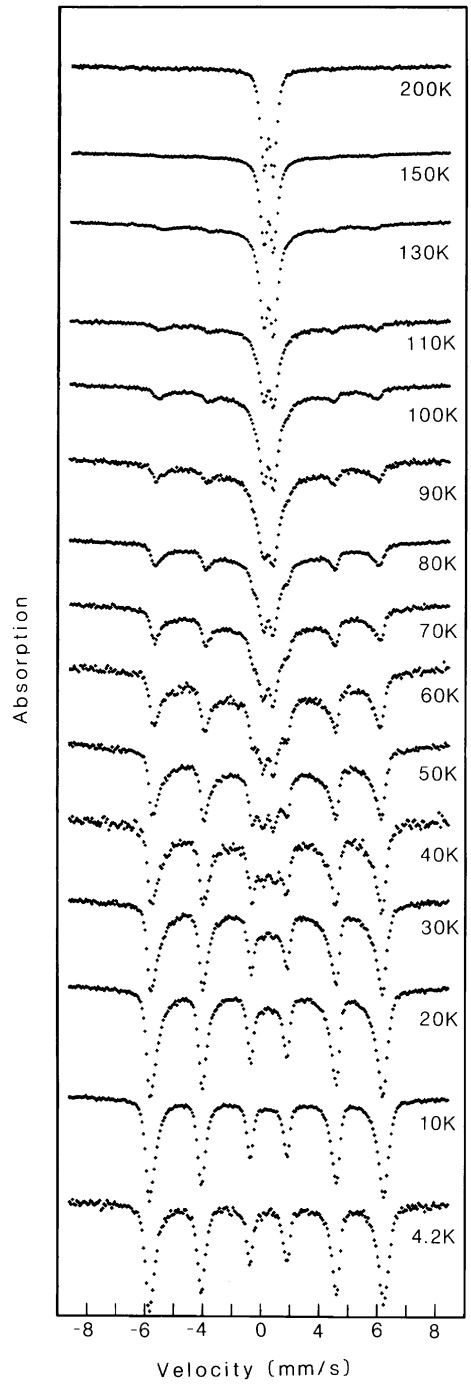


Figure 8. Mössbauer spectra from a specimen of haemosiderin, showing the effects of superparamagnetism with increasing temperature (Bell et al., 1984; Dickson, 1987).

comprised a core of haemosiderin, an iron storage compound, encapsulated within a protein shell. A clear six-line pattern is observed at 4.2 K, but the splitting of these six lines is found to decrease with temperature owing to small amplitude fluctuations in magnetic alignment. At temperatures around 40 to 70 K, a broad background appears under the measured spectrum, and a paramagnetic doublet begins to grow in intensity with increasing temperature. These effects are caused by large thermal reorientations of the magnetization. Finally, by 200 K, the thermally induced magnetic fluctuations are all of large amplitude and of characteristic times too short to permit a HMF to be detected by Mössbauer spectrometry.

DATA ANALYSIS AND INITIAL INTERPRETATION

Mössbauer spectra are often presented for publication with little or no processing. An obvious correction that can be applied to most transmission spectra is a correction for thickness distortion (see Sample Preparation). This correction is rarely performed, however, in large part because the thickness of the specimen is usually not known or is not uniform. The specimen is typically prepared to be thin, or at least this is assumed, and the spectrum is assumed to be representative of the Mössbauer absorption cross-section.

A typical goal of data analysis is to find individual hyperfine parameters, or more typically a distribution of hyperfine parameters, that characterize a measured spectrum. For example, the HMF distribution of Figure 2 should resemble a delta function centered at 330 kG. On the other hand, the HMF distribution of Figure 7B shows a number of peaks that are characteristic of different local chemical environments. Distributions of electric quadrupole splittings and isomer shifts are also useful for understanding heterogeneities in the local atomic arrangements in materials.

Several software packages are available to extract distributions of hyperfine parameters from Mössbauer spectra (Hesse and Rutbartsch, 1974; Le Car and DuBoise, 1979; Brand and Le Caër, 1988; Lagarec and Rancourt, 1997). These programs are often distributed by their authors who may be located with the Mössbauer Information eXchange (see Internet Resources). The different programs extract hyperfine distributions from experimental spectra with different numerical approaches, but all will show how successfully the hyperfine distribution can be used to regenerate the experimental spectrum.

In the presence of statistical noise, the reliability of these derived hyperfine distributions must be considered carefully. In particular, over small ranges of hyperfine parameters, the hyperfine distributions are not unique. For example, it may be unrealistic to distinguish one Lorentzian-shaped peak centered at a particular velocity from the sum of several peaks distributed within a quarter of a linewidth around this same velocity. This nonuniqueness can lead to numerical problems in extracting hyperfine distributions from experimental data. Some software packages use smoothing parameters to penalize the algorithm when it picks a candidate HMF distribution with sharp curvature. When differences in hyperfine distributions are small, there is always an issue of their uniqueness. Sometimes the data analysis cannot distinguish between different types of hyperfine distributions. For example, a spectrum that has been broadened by an EFG distribution, or even an HMF distribution, can be fit perfectly with an IS distribution. The physical origin of hyperfine distributions may not be obvious, especially when the spectrum shows little structure. Application of an external magnetic field may be helpful in identifying a weak HMF, however.

In general, distributions of all three hyperfine parameters (IS, EFG, HMF) will be present simultaneously in a measured spectrum. These parameters may be correlated; for example nuclei having the largest HMF may have the largest (or smallest) IS. Sorting out these correlations is often a research topic in itself, although the software for calculating hyperfine distributions typically allows for simple linear correlations between the distributions.

Both the EFG and the HMF use an axis of quantization for the nuclear spin. However, the direction of magnetization (for the HMF axis) generally does not coincide with the directions of the chemical bonds responsible for the EFG. The general case with comparable hyperfine interaction energies of HMFs and EFGs is quite complicated, and is well beyond the scope of this unit. Some software packages using model spin Hamiltonians are available to calculate spectra acquired under these conditions, however. In the common case when the HMF causes much larger spectral splittings than the EFG, with polycrystalline samples the usual effect of the EFG is a simple broadening of the peaks in the magnetic sextet, with no shifts in their positions.

With even modest experimental care, Mössbauer spectra can be highly reproducible from run to run. For example, the Mössbauer spectrum in Figure 2 was repeated many times over a time period of several years. Almost all of these bcc Fe spectra had data points that overlaid on top of each other, point by point, within the accuracy of the counting statistics. Because of this reproducibility, it is tempting and often appropriate to try to identify spectral features with energy width smaller than the characteristic linewidth. An underexploited technique for data analysis is "partial deconvolution" or "thinning." Since the lineshape of each nuclear transition is close to a Lorentzian function, and can be quite reproducible, it is appropriate to deconvolute a Lorentzian function from the experimental spectrum. This is formally the same as obtaining an IS distribution, but no assumptions about the origin of the hyperfine distributions are implied by this process. The net effect is to sharpen the peaks from the experimental Mössbauer spectrum, and this improvement in effective resolution can be advantageous when overlapping peaks are present in the spectra. The method does require excellent counting statistics to be reliable, however.

Finally, in spite of all the ready availability of computing resources, it is always important to look at differences in the experimental spectra themselves. Sometimes a digital subtraction of one normalized spectrum from another is an excellent way to identify changes in a material. In any

event, if no differences are detected by direct inspection of the data, changes in the hyperfine distributions obtained by computer software should not be believed. For this reason it is still necessary to show actual experimental spectra in research papers that use Mössbauer spectrometry.

SAMPLE PREPARATION

A central concern for transmission Mössbauer spectrometry is the choice and control of specimen thickness. The natural thickness of a specimen is t

$$t \equiv (f_a n_a \sigma_a)^{-1} \tag{29}$$

where f_a is the recoil-free fraction of the Mössbauer isotope in the specimen, n_a is the number of Mössbauer nuclei per cm^3, and σ_a is the cross-section in units of cm^2. The f_a can be estimated from Equation 19, for which it is useful to know that the γ-ray energies are 14.41 keV for ^{57}Fe, 23.875 keV for ^{119}Sn, and 21.64 keV for ^{151}Eu. To obtain n_a it is important to know that the natural isotopic abundance is 2.2% for ^{57}Fe, 8.6% for ^{119}Sn, and 48% for ^{151}Eu. The cross-sections for these isotopes are, in units of 10^{-19} cm^2, 25.7 for ^{57}Fe, 14.0 for ^{119}Sn, and 1.14 for ^{151}Eu. Finally, natural linewidths, Γ, of Equation 1, are 0.097 mm/s for ^{57}Fe, 0.323 mm/s for ^{119}Sn, and 0.652 mm/s for ^{151}Eu.

The observed intensity in a Mössbauer transmission spectrum appears as a dip in count rate as the Mössbauer effect removes γ rays from the transmitted beam. Since this dip in transmission increases with sample thickness, thicker samples provide better signal-to-noise ratios and shorter data acquisition times. For quantitative work, however, it is poor practice to work with samples that are the natural thickness, t, or thicker owing to an effect called "thickness distortion." In a typical constant-acceleration spectrometer, the incident beam will have uniform intensity at all velocities of the source, and the top layer of sample will absorb γ rays in proportion to its cross-section (Equation 2). On the other hand, layers deeper within the sample will be exposed to a γ-ray intensity diminished at velocities where the top layers have absorbed strongly. The effect of this "thickness distortion" is to reduce the overall sample absorption at velocities where the Mössbauer effect is strong. Broadening of the Mössbauer peaks therefore occurs as the samples become thicker. This broadening can be modeled approximately as increasing the effective linewidth of Equation 1 from the natural Γ to $\Gamma(1 + 0.135t/t)$. However, it is important to note that in the tails of the Mössbauer peaks, where the absorption is weak, there is less thickness distortion. The peak shape in the presence of thickness distortion is therefore not a true Lorentzian function. Numerical corrections for the effects of thickness distortion are sometimes possible, but are rarely performed owing to difficulties in knowing the precise sample thickness and thickness uniformity. For quantitative work the standard practice is to use samples of thickness/2 or so.

We calculate an effective thickness, t, for the case of natural Fe metal, which is widely used as a calibration standard for Mössbauer spectrometers. If there are no impurities in the Fe metal, its Mössbauer spectrum has sharp lines as shown in Figure 2. The recoil-free-fraction of bcc Fe is 0.80 at 300 K, and other quantities follow Equation 29

$$= \left(0.80 \times 0.022 \times \frac{7.86\,\mathrm{g}}{\mathrm{cm}^3} \times \frac{1\,\mathrm{mol}}{55.85\,\mathrm{g}} \times \frac{6.023 \times 10^{23}\ \mathrm{atoms}}{\mathrm{mole}} \times 25 \times 10^{-19}\,\mathrm{cm}^2 \frac{1}{4}\right)^{-1} \tag{30}$$

$$= 11 \times 10^{-4}\,\mathrm{cm} = 11\,\mu\mathrm{m} \tag{31}$$

The factor of 1/4 in Equation 30 accounts for the fact that the absorption cross-section is split into six different peaks owing to the hyperfine magnetic field in bcc Fe. The strongest of these comprises 1/4 of the total absorption. Figure 2 was acquired with a sample of natural bcc Fe 25 μm in thickness. The linewidths of the inner two peaks are 0.235 mm/s whereas the outer two are 0.291 mm/s. Although the outer two peaks are broadened by thickness distortion, effects of impurity atoms in the Fe are equally important. The widths of the inner two lines are probably a better measure of the spectrometer resolution.

LITERATURE CITED

Akai, H., Blügel, S., Zeller, R., and Dederichs, P. H., 1986. Isomer shifts and their relation to charge transfer in dilute Fe alloys. *Phys. Rev. Lett.* 56:2407–2410.

Bell, S. H., Weir, M. P., Dickson, D. P. E., Gibson, J. F., Sharp, G. A., and Peters, T. J. 1984. Mössbauer spectroscopic studies of human haemosiderin and ferritin. *Biochim. Biophys. Acta.* 787:227–236.

Black, P. J., and Moon, P. B. 1960. Resonant scattering of the 14-keV iron-57 γ-ray, and its interference with Rayleigh scattering. *Nature (London)* 188:481–484.

Bowen, L. H., De Grave, E., and Vandenberghe, R. E. 1993. Mössbauer effect studies of magnetic soils and sediments. *In* Mössbauer Spectroscopy Applied to Magnetism and Materials Science Vol. I. (G. J. Long and F. Grandjean, eds.). pp. 115–159. Plenum, New York.

Brand, R. A. and Le Car, G. 1988. Improving the validity of Mössbauer hyperfine parameter distributions: The maximum entropy formalism and its applications. *Nucl. Instr. Methods Phys. Res. B* 34:272–284.

Campbell, S. J., Kaczmarek, W. A., and Wang, G. M. 1995. Mechanochemical transformation of haematite to magnetite. *Nanostruct. Mater.* 6:35–738.

Cortie, M. B. and Pollak, H. 1995. Embrittlement and aging at 475°C in an experimental ferritic stainless-steel containing 38 wt.% chromium. *Mater. Sci. Eng. A* 199:153–163.

Degrave, E., Persoons, R. M., and Vandenberghe, R. E. 1993. Mössbauer study of the high temperature phase of co-substituted magnetite. *Phys. Rev. B* 47:5881–5893.

Dickson, D. P. E. 1987. Mössbauer Spectroscopic Studies of Magnetically Ordered Biological Materials. *Hyperfine Interact.* 33: 263–276.

Faldum, T., Meisel, W., and Gütlich, P. 1994. Oxidic and metallic Fe/Knee multilayers prepared from Langmuir-Blodgett films. *Hyperfine Interact.* 92:1263–1269.

Feldwisch, R., Sepiol, B., and Vogl, G. 1994. Elementary diffusion jump of iron atoms in intermetallic phases studied by Mössbauer spectroscopy 2. From order to disorder. *Acta Metall. Mater.* 42:3175–3181.

Fultz, B. 1982. A Mössbauer Spectrometry Study of Fe-Knee-X Alloys. Ph.D. Thesis. University of California at Berkeley.

Fultz, B. 1993. Chemical systematics of iron-57 hyperfine magnetic field distributions in iron alloys. *In* Mössbauer Spectroscopy Applied to Magnetism and Materials Science Vol. I. (G. J. Long and F. Grandjean, eds.). pp. 1–31. Plenum Press, New York.

Fultz, B., Ahn, C. C., Alp, E. E., Sturhahn, W., and Toellner, T. S. 1997. Phonons in nanocrystalline ^{57}Fe. *Phys. Rev. Lett.* 79: 937–940.

Fultz, B., Kuwano, H., and Ouyang, H. 1995. Average widths of grain boundaries in nanophase alloys synthesized by mechanical attrition. *J. Appl. Phys.* 77:3458–3466.

Gancedo, J. R., Gracia, M. and Marco, J. F. 1991. CEMS Methodology. *Hyperfine Interact.* 66:83–94.

Gerdau, E., Rüffer, R., Winkler, H., Tolksdorf, W., Klages, C. P., and Hannon, J. P. 1985. Nuclear Bragg diffraction of synchrotron radiation in yttrium iron garnet, *Phys. Rev. Lett.* 54: 835–838.

Greenwood, N. N. and Gibb, T. C. 1971. Mössbauer Spectroscopy. Chapman & Hall, London.

Gütlich, P. 1975. Mössbauer spectroscopy in chemistry. *In* Mössbauer Spectroscopy. (U. Gonser, ed.). Chapter 2. Springer-Verlag, New York.

Hannon, J. P. and Trammell, G. T. 1969. Mössbauer diffraction. II. Dynamical Theory of Mössbauer Optics. *Phys. Rev.* 186:306–325.

Herr, U., Jing, J., Birringer, R., Gonser, U., and Gleiter, H. 1987. Investigation of nanocrystalline iron materials by Mössbauer spectroscopy. *Appl. Phys. Lett.* 50:472–474.

Hesse, J. and Rutbartsch, A. 1974. Model independent evaluation of overlapped Mössbauer spectra. *J. Phys. E: Sci. Instrum.* 7: 526–532.

Kruijer, S., Keune, W., Dobler, M., and Reuther, H. 1997. Depth analysis of phase formation in Si after high-dose Fe ion-implantation by depth-selective conversion electron Mössbauer spectroscopy. *Appl. Phys. Lett.* 70:2696–2698.

Lagarec, K. and Rancourt, D. G. 1997. Extended Voigt-based analytic lineshape method for determining n-dimensional correlated hyperfine parameter distributions in Mössbauer spectroscopy. *Nucl. Inst. Methods Phys. Res. B* 128:266–280.

Lamb, W. E. Jr. 1939. Capture of neutrons by atoms in a crystal. *Phys. Rev.* 55:190–197.

Le Caër, G. and Duboise, J. M. 1979. Evaluation of hyperfine parameter distributions from overlapped Mössbauer spectra of amorphous alloys. *J. Phys. E: Sci. Instrum.* 12:1083–1090.

Mørup, S. 1990. Mössbauer effect in small particles. *Hyperfine Interact.* 60:959–974.

Mössbauer, R. L. 1958. Kernresonanzfluoreszenz von Gammastrahlung in Ir^{191}. *Z. Phys.* 151:124.

Ruebenbauer, K., Mullen, J. G., Nienhaus, G. U., and Shupp, G. 1994. Simple model of the diffusive scattering law in glass-forming liquids. *Phys. Rev. B* 49:15607–15614.

Schwartz, L. H. 1976. Ferrous alloy phase transformations. *In* Applications of Mössbauer Spectroscopy, Vol. 1. pp. 37–81. (R. L. Cohen, ed.). Academic Press, New York.

Seto, M., Yoda, Y., Kikuta, S., Zhang, X. W., and Ando, M. 1995. Observation of nuclear resonant scattering accompanied by phonon excitation using synchrotron radiation. *Phys. Rev. Lett.* 74:3828–3831.

Sepiol, B., Meyer, A., Vogl, G., Ruffer, R., Chumakov, A. I., and Baron, A. Q. R. 1996. Time domain study of Fe-57 diffusion using nuclear forward scattering of synchrotron radiation. *Phys. Rev. Lett.* 76:3220–3223.

Shvyd'ko, Yu. V., and Smirnov, G. V. 1989. Experimental study of time and frequency properties of collective nuclear excitations in a single crystal. *J. Phys. Condens. Matter* 1:10563–10584.

Simmons, G. W. and Leidheiser, Jr., H. 1976. Corrosion and interfacial reactions. *In* Applications of Mössbauer Spectroscopy, Vol. 1. pp. 92–93. (R. L. Cohen, ed.). Academic Press, New York.

Squires, G. L. 1978. Introduction to the Theory of Thermal Neutron Scattering. p. 54. Dover Publications, New York.

Stahl, B. and Kankeleit, E. 1997. A high-luminosity UHV orange type magnetic spectrometer developed for depth-selective Mössbauer spectroscopy. *Nucl. Instr. Meth. Phys. Res. B* 122: 149–161.

Stephens, T. A. and Fultz, B. 1997. Chemical environment selectivity in Mössbauer diffraction from $^{57}Fe_3Al$. *Phys. Rev. Lett.* 78:366–369.

Stephens, T. A., Keune, W., and Fultz, B. 1994. Mössbauer effect diffraction from polycrystalline ^{57}Fe. *Hyperfine Interact.* 92: 1095–1100.

Sturhahn, W., Toellner, T. S., Alp, E. E., Zhang, X., Ando, M., Yoda, Y., Kikuta, S., Seto, M., Kimball, C. W., and Dabrowski, B. 1995. Phonon density-of-states measured by inelastic nuclear resonant scattering. *Phys. Rev. Lett.* 74:3832–3835.

Smirnov, G. V. 1996. Nuclear resonant scattering of synchrotron radiation. *Hyperfine Interact.* 97/98:551–588.

van Bürck, U., Smirnov, G. V., Mössbauer, R. L., Parak, F., and Semioschkina, N. A. 1978. Suppression of nuclear inelastic channels in nuclear resonance and electronic scattering of γ-quanta for different hyperfine transtions in perfect ^{57}Fe single crystals. *J. Phys. C Solid State Phys.* 11:2305–2321.

Vincze, I. and Aldred, A. T. 1974. Mössbauer measurements in iron-base alloys with nontransition elements. *Phys. Rev. B* 9:3845–3853.

Vincze, I. and Campbell, I. A. 1973. Mössbauer measurements in iron based alloys with transition metals. *J. Phys. F Metal Phys.* 3:647–663.

Vogl, G. and Sepiol, B. 1994. Elementary diffusion jump of iron atoms in intermetallic phases studied by Mössbauer spectroscopy 1. Fe-Al close to equiatomic stoichiometry. *Acta Metall. Mater.* 42:3175–3181.

Walker, L. R., Wertheim, G. K., and Jaccarino, V. 1961. Interpretation of the Fe^{57} isomer shift. *Phys. Rev. Lett.* 6:98–101.

Watson, R. E. and Freeman, A. J. 1967. Hartree-Fock theory of electric and magnetic hyperfine interactions in atoms and magnetic compounds. *In* Hyperfine Interactions. (A. J. Freeman and R. B. Frankel, eds.) Chapter 2. Academic Press, New York.

Williamson, D. L. 1993. Microstructure and tribology of carbon, nitrogen, and oxygen implanted ferrous materials. *Nucl. Instr. Methods Phys. Res. B* 76:262–267.

KEY REFERENCES

Bancroft, G. M. 1973. Mössbauer Spectroscopy: An Introduction for Inorganic Chemists and Geochemists. John Wiley & Sons, New York.

Belozerski, G. N. 1993. Mössbauer Studies of Surface Layers. Elsevier/North Holland, Amsterdam.

Cohen, R. L. (ed.). 1980. Applications of Mössbauer Spectroscopy, Vols. 1, 2. Academic Press, New York.

These 1980 volumes by Cohen contain review articles on the applications of Mössbauer spectrometry to a wide range of materials and phenomena, with some exposition of the principles involved.

Cranshaw, T. E., Dale, B. W., Longworth, G. O., and Johnson, C. E. 1985. Mössbauer Spectroscopy and its Applications. Cambridge University Press, Cambridge.

Dickson, D. P. E. and Berry, F. J. (eds.). 1986. Mössbauer Spectroscopy. Cambridge University Press, Cambridge.

Frauenfelder, H. 1962. The Mössbauer Effect: A Review with a Collection of Reprints. W. A. Benjamin, New York.

This book by Frauenfelder was written in the early days of Mössbauer spectrometry, but contains a fine exposition of principles. More importantly, it contains reprints of the papers that first reported the phenomena that are the basis for much of Mössbauer spectrometry. It includes an English translation of one of Mössbauer's first papers.

Gibb, T. C. 1976. Principles of Mössbauer Spectroscopy. Chapman and Hall, London.

Gonser, U. (ed.). 1975. Mössbauer Spectroscopy. Springer-Verlag, New York.

Gonser, U. (ed.). 1986. Microscopic Methods in Metals, Topics in Current Physics, 40, Springer-Verlag, Berlin.

Gruverman, I. J. (ed.). 1976. Mössbauer Effect Methodology, Vols. 1-10. Plenum Press, New York.

Gütlich, P., Link, R., and Trautwein, A. (eds.). 1978. Mössbauer Spectroscopy and Transition Metal Chemistry. Springer-Verlag, Berlin.

Long, G. J. and Grandjean, F. (eds.). 1984. Mössbauer Spectroscopy Applied to Inorganic Chemistry, Vols. 1-3. Plenum Press, New York.

Long, G. J. and Grandjean, F. (eds.). 1996. Mössbauer Spectroscopy Applied to Magnetism and Materials Science, Vols. 1 and 2. Plenum Press, New York.

These 1996 volumes by Long and Grandjean contain review articles on different classes of materials, and on different techniques used in Mössbauer spectrometry.

Long, G. J. and Stevens, J. G. (eds.). 1986. Industrial Applications of the Mössbauer Effect. Plenum, New York.

May, L. (ed.). 1971. An Introduction to Mössbauer Spectroscopy. Plenum Press, New York.

Mitra, S. (ed.). 1992. Applied Mössbauer Spectroscopy: Theory and Practice for Geochemists and Archaeologists. Pergamon Press, Elmsford, New York.

Thosar, B. V. and Iyengar, P. K. (eds.). 1983. Advances in Mössbauer Spectroscopy, Studies in Physical and Theoretical Chemistry 25. Elsevier/North Holland, Amsterdam.

Wertheim, G. 1964. Mössbauer Effect: Principles and Applications. Academic Press, New York.

INTERNET RESOURCES

http://www.kfki.hu/~mixhp/

The Mössbauer Information eXchange, MIX, is a project of the KFKI Research Institute for Particle and Nuclear Physics, Budapest, Hungary. It is primarily for scientists, students, and manufacturers involved in Mössbauer spectroscopy and other nuclear solid-state methods.

http://www.unca.edu/medc

medc@unca.edu

The Mössbauer Effect Data Center (University of North Carolina; J.G.. Stevens, Director) maintains a library of most publications involving the Mössbauer effect, including hard-to-access publications. Computerized databases and database search services are available to find papers on specific materials.

BRENT FULTZ
California Institute of Technology
Pasadena, California

X-RAY TECHNIQUES

INTRODUCTION

X-ray scattering and spectroscopy methods can provide a wealth of information concerning the physical and electronic structure of crystalline and noncrystalline materials in a variety of external conditions and environments. X-ray powder diffraction, for example, is generally the first, and perhaps the most widely used, probe of crystal structure. Over the last nine decades, especially with the introduction of x-ray synchrotron sources during the last 25 years, x-ray techniques have expanded well beyond their initial role in structure determination.

This chapter explores the wide variety of applications of x-ray scattering and spectroscopy techniques to the study of materials. Materials, in this sense, include not only bulk condensed matter systems, but liquids and surfaces as well. The term "scattering," is generally used to include x-ray measurements on noncrystalline systems, such as glasses and liquids, or poorly crystallized materials, such as polymers, as well as x-ray diffraction from crystalline solids. Information concerning long-range, short-range, and chemical ordering as well as the existence, distribution and characterization of various defects is accessible through these kinds of measurements. Spectroscopy techniques generally make use of the energy dependence of the scattering or absorption cross-section to study chemical short-range order, identify the presence and location of chemical species, and probe the electronic structure through inelastic excitations.

X-ray scattering and spectroscopy provide information complementary to several other techniques found in this volume. Perhaps most closely related are electron and neutron scattering methods described in Chapters 11 and 13, respectively. The utility of all three methods in investigations of atomic scale structure arises from the close match of the wavelength of these probes to typical interatomic distances (a few Ångstroms). Of the three methods, the electron scattering interaction with matter is the strongest, so this technique is most appropriate to the study of surfaces or very thin samples. In addition, samples must be studied in an ultrahigh-vacuum environment. The relatively weak absorption of neutrons by most isotopes allows investigations of bulk samples. The principal neutron scattering interaction involves the nuclei of constituent elements and the magnetic moment of the outer electrons. Indeed, the cross-section for scattering from magnetic electrons is of the same order as scattering from the nuclei, so that this technique is of great utility in studying magnetic structures of magnetic materials. Neutron energies are typically on the order of a few meV to tens of meV, the same energy scale as many important elementary excitations in solids. Therefore inelastic neutron scattering has become a critical probe of elementary excitations including phonon and magnon dispersion in solids. The principal x-ray scattering interaction in materials involves the atomic electrons, but is significantly weaker than the electron-electron scattering cross-section. X-ray absorption by solids is relatively strong at typical energies found in laboratory apparatus (8 keV), penetrating a few tens of microns into a sample. At x-ray synchrotron sources, x-ray beams of energies above 100 keV are available. At these high energies, the absorption of x-rays approaches small values comparable to neutron scattering. The energy dependence of the x-ray absorption cross-section is punctuated by absorption edges or resonances that are associated with interatomic processes of constituent elements. These resonances can be used to great advantage to investigate element-specific properties of the material and are the basis of many of the spectroscopic techniques described in this chapter.

Facilities for x-ray scattering and spectroscopy range from turn-key powder diffraction instruments for laboratory use to very sophisticated beam lines at x-ray synchrotron sources around the world. The growth of these synchrotron facilities in recent years, with corresponding increases in energy and scattering angle resolution, energy tunability, polarization tunability, and raw intensity has spurred tremendous advances in this field. For example, the technique of x-ray absorption spectroscopy (XAS) requires a source of continuous radiation that can be tuned over approximately 1 keV through the absorption edge of interest. Although possible using the bremstrahlung radiation from conventional x-ray tubes, full exploitation of this technique requires the high intensity, collimated beam of synchrotron radiation. As another example, the x-ray magnetic scattering cross-section is quite small (but finite) compared to the charge scattering cross-section. The high flux of x-rays available at synchrotron sources compared to neutron fluxes from reactors, however, allows x-ray magnetic scattering to effectively complement neutron magnetic scattering in several instances. Again, largely due to high x-ray fluxes and strong collimation, energy resolution in the meV range are achievable at current third generation x-ray sources, paving the way for new x-ray inelastic scattering studies of elementary excitations. Third generation sources include the Advanced Photon Source (APS) at the Argonne National Laboratory and the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory in the United States, the European Synchrotron Radiation Facility (ESRF) in Europe, and Spring-8 in Japan.

ALAN I. GOLDMAN

X-RAY POWDER DIFFRACTION

INTRODUCTION

X-ray powder diffraction is used to determine the atomic structure of crystalline materials without the need for large ($\sim$100-μm) single crystals. "Powder" can be a misnomer; the technique is applicable to polycrystalline phases

such as cast solids or films grown on a substrate. X-ray powder diffraction can be useful in a wide variety of situations. Below we list a number of questions that can be effectively addressed by this technique. This is an attempt to illustrate the versatility of x-ray powder diffraction and not, by any means, a complete list. Six experiments (corresponding to numbers 1 through 6 below), described later as concrete examples (see Practical Aspects of the Model), constitute an assortment of problems of varying difficulty and involvement that we have come across over the course of several years.

1. The positions and integrated intensities of a set of peaks in an x-ray powder diffraction pattern can be compared to a database of known materials in order to identify the contents of the sample and to determine the presence or absence of any particular phase.
2. A mixture of two or more crystalline phases can be easily and accurately analyzed in terms of its phase fractions, whether or not the crystal structures of all phases are known. This is called quantitative phase analysis, and it is particularly valuable if some or all of the phases are chemically identical and hence cannot be distinguished while in solution.
3. The crystal structure of a new or unknown material can be determined when a similar material with a known structure exists. Depending on the degree of similarity between the new and the old structure, this can be fairly straightforward.
4. The crystal structure of a new or unknown material can be solved *ab initio* even if no information about the material other than its stoichiometry is known. This case is significantly more difficult than the previous one, and it requires both high-resolution data and a significant investment of time and effort on the part of the investigator.
5. Phase transitions and solid-state reactions can be investigated in near real time by recording x-ray powder diffraction patterns as a function of time, pressure, and/or temperature.
6. Subtle structural details such as lattice vacancies of an otherwise known structure can be extracted. This also usually requires high-resolution data and a very high sample quality.

The variation of peak positions with sample orientation can be used to deduce information about the internal strain of a sample. This technique is not covered in this unit, and the interested reader is directed to references such as Noyan and Cohen (1987). Another related technique, not covered here, is texture analysis, the determination of the distribution of orientations in a polycrystalline sample.

X-ray powder diffraction is an old technique, in use for most of this century. The capabilities of the technique have recently grown for two main reasons: (1) development of x-ray sources and optics (e.g., synchrotron radiation, Göbel mirrors) and (2) the increasing power of computers and software for analysis of powder data. This unit discusses the fundamental principles of the technique, including the expression for the intensity of each diffraction peak, gives a brief overview of experimental techniques, describes several illustrative examples, gives a description of the procedures to interpret and analyze diffraction data, and discusses weaknesses and sources of possible errors.

Competitive and Related Techniques

Alternative Methods

1. Single-crystal x-ray diffraction requires single crystals of appropriate size (10 to 100 μm). Solution of single-crystal structures is usually more automated with appropriate software than powder structures. Single-crystal techniques can generally solve more complicated structures than powder techniques, whereas powder diffraction can determine the constituents of a mixture of crystalline solid phases.
2. Neutron powder diffraction generally requires larger samples than x rays. It is more sensitive to light atoms (especially hydrogen) than x rays are. Deuterated samples are often required if the specimen has a significant amount of hydrogen (see NEUTRON POWDER DIFFRACTION). Neutrons are sensitive to the magnetic structure (see MAGNETIC NEUTRON SCATTERING). Measurements must be performed at a special facility (reactor or spallation source).
3. Single-crystal neutron diffraction requires single-crystal samples of millimeter size, is more sensitive to light atoms (especially hydrogen) than are x rays, and often requires deuterated samples if the specimen has a significant amount of hydrogen. As in the powder case, it is sensitive to the magnetic structure and must be performed at a special facility (reactor or spallation source).
4. Electron diffraction can give lattice information for samples that have a strain distribution too great to be indexed by the x-ray powder technique. It provides spatial resolution for inhomogeneous samples and can view individual grains but requires relatively sophisticated equipment (electron microscope; see LOW-ENERGY ELECTRON DIFFRACTION).

PRINCIPLES OF THE METHOD

If an x ray strikes an atom, it will be weakly scattered in all directions. If it encounters a periodic array of atoms, the waves scattered by each atom will reinforce in certain directions and cancel in others. Geometrically, one may imagine that a crystal is made up of families of lattice planes and that the scattering from a given family of planes will only be strong if the x rays reflected by each plane arrive at the detector in phase. This leads to a relationship between the x-ray wavelength λ, the spacing d between lattice planes, and the angle of incidence θ known as Bragg's law, $\lambda = 2d\sin\theta$. Note that the angle of deviation of the x ray is 2θ from its initial direction. This is fairly restrictive for a single crystal: for a given λ, even

if the detector is set at the correct 2θ for a given d spacing within the crystal, there will be no diffracted intensity unless the crystal is properly aligned to both the incident beam and the detector. The essence of the powder diffraction technique is to illuminate a large number of crystallites, so that a substantial number of them are in the correct orientation to diffract x rays into the detector.

The geometry of diffraction in a single grain is described through the reciprocal lattice. The fundamental concepts are in any of the Key References or in introductory solid-state physics texts such as Kittel (1996) or Ashcroft and Mermin (1976); also see SYMMETRY IN CRYSTALLOGRAPHY. If the crystal lattice is defined by three vectors **a**, **b**, and **c**, there are three reciprocal lattice vectors defined as

$$\mathbf{a}^* = \frac{2\pi \mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}} \tag{1}$$

and cyclic permutations thereof for $\mathbf{b}^*$ and $\mathbf{c}^*$ (in the chemistry literature the factor 2π is usually eliminated in this definition). These vectors define the reciprocal lattice, with the significance that any three integers (hkl) define a family of lattice planes with spacing $d = 2\pi/|h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*|$, so that the diffraction vector $\mathbf{K} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ satisfies $|\mathbf{K}| = 2\pi/d = 4\pi\sin\theta/\lambda$ (caution: most chemists and some physicists define $|\mathbf{K}| = 1/d = 2\sin\theta/\lambda$).

The intensity of the diffracted beam is governed by the unit cell structure factor, defined as

$$F_{hkl} = \sum_j e^{i\mathbf{K}\cdot\mathbf{R}_j} f_j e^{-2W} \tag{2}$$

where R_j is the position of the jth atom in the unit cell, the summation is taken over all atoms in the unit cell, and f_j is the atomic scattering factor, tabulated in, e.g., the International Tables for Crystallography (Brown et al., 1992), and is equal to the number of atomic electrons at $2\theta = 0$, decreasing as a smooth function of $\sin\theta/\lambda$ (there are "anomalous" corrections to this amplitude if the x-ray energy is close to a transition in a target atom). The Debye-Waller factor $2W$ is given by $2W = K^2 u_{\mathrm{rms}}^2/3$, where u_{rms} is the (three-dimensional) root-mean-square deviation of the atom from its lattice position due to thermal and zero-point fluctuations. Experimental results are often quoted as the thermal parameter B, defined as $8\pi^2 u_{\mathrm{rms}}^2/3$, so that the Debye-Waller factor is given by $2W = 2B\sin^2\theta/\lambda^2$. Note that thermal fluctuations of the atoms about their average position weaken the diffraction lines but do not broaden them. As long as the diffraction is sufficiently weak (kinematic limit, assumed valid for most powders), the diffracted intensity is proportional to the square of the structure factor. In powder diffraction, it is always useful to bear in mind that the positions of the observed peaks indicate the geometry of the lattice, both its dimensions and any internal symmetries, whereas the intensities are governed by the arrangement of atoms within the unit cell.

In a powder experiment, various factors act to spread the intensity over a finite range of the diffraction angle, and so it is useful to consider the integrated intensity (power) of a given peak over the diffraction angle,

$$P_{hkl} = \left[\frac{P_0 \lambda^3 R_e^2 l}{16\pi R}\frac{V_s}{V^2}\right] M_{hkl} F_{hkl}^2 \left(\frac{1+\cos^2 2\theta}{\sin 2\theta \, \sin\theta}\right) \tag{3}$$

where P_0 is the power density of the incident beam, $R_e = 2.82$ fm is the classical electron radius, l and R are the width of the receiving slit and the distance between it and the sample, and V_s and V are the effective illuminated volume of the sample and the volume of one unit cell. The term M_{hkl} is the multiplicity of the hkl peak, e.g., 8 for a cubic hhh and 6 for a cubic $h00$, and $(1+\cos^2 2\theta)/(\sin\theta \sin 2\theta)$ is called the Lorentz polarization factor. The numerator takes the given form only for unpolarized incident radiation and in the absence of any other polarization-sensitive optical elements; it must be adapted, e.g., for a synchrotron-radiation source or a diffracted-beam monochromator. There are considerable experimental difficulties with measuring the absolute intensity either of the incident or the diffracted beam, and so the terms in the square brackets are usually lumped into a normalization factor, and one considers the relative intensity of different diffraction peaks, or more generally, the spectrum of intensity vs. scattering angle.

An inherent limitation on the amount of information that can be derived from a powder diffraction spectrum arises from possible overlap of Bragg peaks with different (hkl), so that their intensities cannot be determined independently. This overlap may be exact, as in the coincidence of cubic (511) and (333) reflections, or it may allow a partial degree of separation, as in the case of two peaks whose positions differ by a fraction of their widths. Because powder diffraction peaks generally become broader and more closely spaced at higher angles, peak overlap is a factor in almost every powder diffraction experiment. Perhaps the most important advance in powder diffraction during the last 30 years is the development of whole-pattern (Rietveld) fitting techniques for dealing with partially overlapped peaks, as discussed below.

Diffraction peaks acquire a nonzero width from three main factors: instrumental resolution (not discussed here), finite grain size, and random strains. If the grains have a linear dimension L, then the full-width at half-maximum (FWHM) in 2θ, expressed in radians, of the diffraction line is estimated by the well-known Scherrer equation, $\mathrm{FWHM}_{2\theta} = 0.89\lambda/L\cos\theta$. This is a reflection of the fact that the length L of a crystal made up of atoms with period d can only be determined to within d. One should not take the numerical value too seriously, as it depends on the precise shape of the crystallites and dispersion of the crystallite size distribution. If the crystallites have a needle or plate morphology, the size can be different for different families of lattice planes. On the other hand, if the crystallites are subject to a random distribution of lattice fractional strains having a FWHM of e_{FWHM}, the FWHM of the diffraction line will be $2\tan\theta e_{\mathrm{FWHM}}$. It is sometimes asserted that size broadening produces a Lorentzian and strain a Gaussian lineshape, but there is no fundamental reason for this to be true, and counterexamples are frequently observed. If the sample peak width

exceeds the instrumental resolution, or can be corrected for that effect, it can be informative to make a plot (called a Williamson-Hall plot) of FWHM$\cos\theta$ vs. $\sin\theta$. If the data points fall on a smooth curve, the intercept will give the particle size and the limiting slope the strain distribution. Indeed, the curve will be a straight line if both effects give a Lorentzian lineshape, because the shape of any peak would be the convolution of the size and strain contributions. If the points in a Williamson-Hall plot are scattered, it may give useful information (or at least a warning) of anisotropic size or strain broadening. More elaborate techniques for the deconvolution of size and strain effects from experimental data are described in the literature (Klug and Alexander, 1974; Balzar and Ledbetter, 1993).

There are two major problems in using powder diffraction measurements to determine the atomic structure of a material. First, as noted above, peaks overlap, so that the measured intensity cannot be uniquely assigned to the correct Miller indices (hkl). Second, even if the intensities were perfectly separated so that the magnitudes of the structure factors were known, one could not Fourier transform the measured structure factors to learn the atomic positions because their phases are not known.

PRACTICAL ASPECTS OF THE METHOD

The requirements to obtain a useful powder diffraction data set are conceptually straightforward: allow a beam of x rays to impinge on the sample and record the diffracted intensity as a function of angle. Practical realizations are governed by the desire to optimize various aspects of the measurement, such as the intensity, resolution, and discrimination against undesired effects (e.g., background from sample fluorescence).

Most laboratory powder x-ray diffractometers use a sealed x-ray tube with a target of copper, molybdenum, or some other metal. About half of the x rays from such a tube are in the characteristic K_α line ($\lambda = 1.54$Å for Cu, $\lambda = 0.70$Å for Mo), and the remainder are in other lines and in a continuous bremsstrahlung spectrum. Rotating-anode x-ray sources can be approximately ten times brighter than fixed targets, with an attendant cost in complexity and reliability. In either case, one can either use the x rays emitted by the anode directly (so that diffraction of the continuous component of the spectrum contributes a smooth background under the diffraction peaks) or select the line radiation by a crystal monochromator (using diffraction to pass only the correct wavelength) or by an energy-sensitive detector. The K_α line is actually a doublet (1.54051 and 1.54433 Å for Cu), which can create the added complication of split peaks unless one uses a monochromator of sufficient resolving power to pass only one component. Synchrotron radiation sources are finding increasing application for powder diffraction, due to their high intensity, intrinsically good collimation ($\sim 0.01^\circ$ in the vertical direction) of x-ray beams, and tunability over a continuous spectrum and the proliferation of user facilities throughout the world.

There are a large number of detectors suitable for powder x-ray diffraction. Perhaps the simplest is photographic film, which allows the collection of an entire diffractogram at one time and, with proper procedures, can be used to obtain quantitative intensities with a dynamic range up to 100:1 (Klug and Alexander, 1974). An updated form of photographic film is the x-ray imaging plate, developed for medical radiography, which is read out electronically (Miyahara et al., 1986; Ito and Amemiya, 1991). The simplest electronic detector, the Geiger counter, is no longer widely used because of its rather long dead time, which limits the maximum count rate. The gas-filled proportional counter offers higher count rates and some degree of x-ray energy resolution. The most widely used x-ray detector today is the scintillation counter, in which x rays are converted into visible light, typically in a thallium-doped NaI crystal, and then into electronic pulses by a photomultiplier tube. Various semiconductor detectors [Si:Li, positive-intrinsic-negative (PIN)] offer energy resolutions of 100 to 300 eV, sufficient to distinguish fluorescence from different elements and from the diffracted x rays, although their count rate capability is generally lower than that of scintillation counters. There are various forms of electronic position-sensitive detectors. Gas-filled proportional detectors can have a spatial resolution of a small fraction of a millimeter and are available as straight-line detectors, limited to several degrees of 2θ by parallax, or as curved detectors covering an angular range as large as 120°. They can operate at a count rate up to 10^5 Hz over the entire detector, but one must bear in mind that the count rate in one individual peak would be significantly less. Also, not all position-sensitive detectors are able to discriminate against x-ray fluorescence from the sample, although there is one elegant design using Kr gas and x rays just exceeding the Kr K edge that addresses this problem (Smith, 1991). Charge-coupled devices (CCDs) are two-dimensional detectors that integrate the total energy deposited into each pixel and therefore may have a larger dynamic range and/or a faster time response (Clarke and Rowe, 1991).

Some of the most important configurations for x-ray powder diffraction instruments are illustrated in Figure 1. The simple Debye-Scherrer camera in (A) records a wide range of angles on curved photographic film but suffers from limited resolution. Modern incarnations include instruments using curved position-sensitive detectors and imaging plates and are in use at several synchrotron sources. It generally requires a thin rod-shaped sample either poured as a powder into a capillary or mixed with an appropriate binder and rolled into the desired shape. The Bragg-Brentano diffractometer illustrated in (B) utilizes parafocusing from a flat sample to increase the resolution available from a diverging x-ray beam; in this exaggerated sketch, the distribution of Bragg angles is 3°, despite the fact that the sample subtends an angle of 15° from source or detector. The addition of a diffracted-beam monochromator illustrated in (C) produces a marked improvement in performance by eliminating x-ray fluorescence from the sample. For high-pressure cells, with limited access for the x-ray beam, the energy-dispersive diffraction approach illustrated in (D) can be an attractive

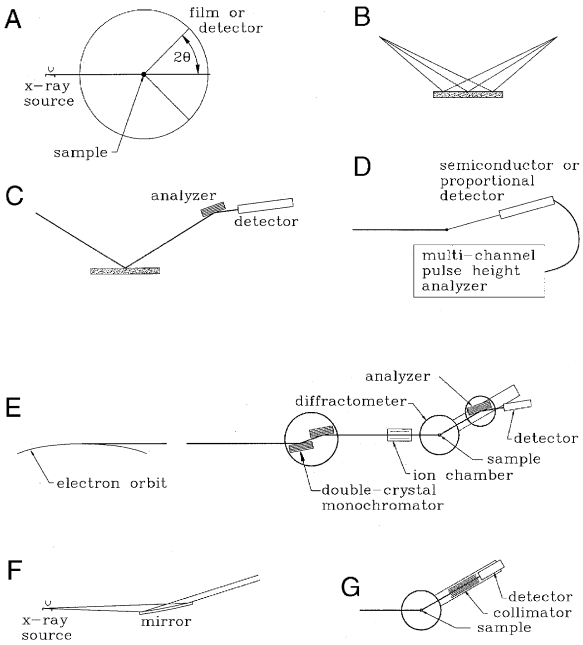


Figure 1. Schematic illustration of experimental setups for powder diffraction measurements.

solution. A polychromatic beam is scattered through a fixed angle, and the energy spectrum of the diffracted x rays is converted to d spacings for interpretation.

A typical synchrotron powder beamline such as X7A or X3B1 at the National Synchrotron Light Source (Brookhaven National Laboratory) is illustrated in (E). One particular wavelength is selected by the dual-crystal monochromator, and the x rays are diffracted again from an analyzer crystal mounted on the 2θ arm. Monochromator and analyzer are typically semiconductor crystals with rocking-curve widths of a few thousandths of a degree. Besides its intrinsically high resolution, this configuration offers the strong advantage that it truly measures the angle through which a parallel beam of radiation is scattered and so is relatively insensitive to parallax or sample transparency errors. The advantage of a parallel incident beam is available on laboratory sources by the use of a curved multilayer (Göbel) mirror (F), currently commercialized by Bruker Analytical X-ray Systems. One can also measure the diffracted angle, free from parallax, by use of a parallel-blade collimator as shown in (G). This generally gives higher intensity than an analyzer crystal at the cost of coarser angular resolution (0.1° to 0.03°), but this is often not a disadvantage because diffraction peaks from real samples are almost always broader than the analyzer rocking curve. However, the analyzer crystal also discriminates against fluorescence.

There are a number of sources of background in a powder diffraction experiment. X rays can be scattered from the sample by a number of mechanisms other than Bragg diffraction: Compton scattering, thermal diffuse scattering, scattering by defects in the crystal lattice, multiple scattering in the sample, and scattering by noncrystalline components of the sample, by the sample holder, or even by air in the x-ray beam path. The x rays scattered by all of these mechanisms will have an energy that is identical to or generally indistinguishable from the primary beam, and so it is not possible to eliminate these effects by using energy-sensitive detectors. Another important source of background is x-ray fluorescence, the emission of x rays by atoms in the target that have been ionized by the primary x-ray beam. X-ray fluorescence is always of a longer wavelength (lower energy) than the primary beam and so can be eliminated by the use of a diffracting crystal between sample and detector or an energy-sensitive detector (with the precaution that the fluorescence radiation does not saturate it) or controlled by appropriate choice of the incident wavelength.

It is not usually possible to predict the shape of this background, and so, once appropriate measures are taken to reduce it as much as practical, it is empirically separated from the relatively sharp powder diffraction peaks of interest. In data analysis such as the Rietveld method, one may take it as a piecewise linear or spline function between specified points or parameterize and adjust it to produce the best fit. There is some controversy about how to treat the statistical errors in a refinement with an adjustable background.

DATA ANALYSIS AND INTERPRETATION

A very important technique for analysis of powder diffraction data is the whole-pattern fitting method proposed by Rietveld (1969). It is based on the following properties of x-ray (and neutron) powder diffraction data: a powder diffraction pattern usually comprises a large number of peaks, many of which overlap, often very seriously, making the separate direct measurement of their integrated intensities difficult or impossible. However, it is possible to describe the shape of all Bragg peaks in the pattern by a small (compared to the number of peaks) number of profile parameters. This allows the least-squares refinement of an atomic model combined with an appropriate peak shape function, i.e., a simulated powder pattern, directly against the measured powder pattern. This may be contrasted to the single-crystal case, where the atomic structure is refined against a list of extracted integrated intensities. The Rietveld method is an extremely powerful tool for the structural analysis of virtually all types of crystalline materials not available as single crystals.

The parameters refined in the Rietveld method fall into two classes: those that describe the shape and position of the Bragg peaks in the pattern (profile parameters) and those that describe the underlying atomic model (atomic or structural parameters). The former include the lattice parameters and those describing the shape and width of the Bragg peaks. In x-ray powder diffraction, a widely used peak shape function is the pseudo-Voigt function (Thompson et al., 1987), a fast-computing approximation to a convolution of a Gaussian and a Lorentzian (Voigt function). It uses only five parameters (usually called U, V, W, X, and Y) to describe the shape of all peaks in the powder pattern. In particular, the peak widths are a smooth function of the scattering angle 2θ. Additional profile parameters are often used to describe the peak

asymmetry at low angles due to the intersection of the curved Debye-Scherrer cone of radiation with a straight receiving slit and corrections for preferred orientation. The structural parameters include the positions, types, and occupancies of the atoms in the structural model and isotropic or anisotropic thermal parameters (Debye-Waller factors). The power of the Rietveld method lies in simultaneous refinement of both profile and atomic parameters, thereby maximizing the amount of information obtained from the powder data.

The analysis of x-ray powder diffraction data can be divided into a number of separate steps. While some of these steps rely on the correct completion of the previous one(s), they generally constitute independent tasks to be completed by the experimenter. Depending on the issues to be addressed by any particular experiment, one, several, or all of these tasks will be encountered, usually in the order in which they are described here.

Crystal Lattice and Space Group Determination from X-Ray Powder Data

Obtaining the lattice parameters (unit cell) of an unknown material is always the first step on the path to a structure solution. While predicting the scattering angles of a set of diffraction peaks given the unit cell is trivial, the inverse task is not, because the powder method reduces the three-dimensional atomic distribution in real space to a one-dimensional diffraction pattern in momentum space. In all but the simplest cases, a computer is required and a variety of programs are available for the task. They generally fall into two categories: those that attempt to find a suitable unit cell by trial and error (also often called the "semiexhaustive" method) and those that use an analytical approach based on the geometry of the reciprocal lattice. Descriptions of these approaches can be found below [see Key References, particularly Bish and Post (1989, pp. 188–196) and Klug and Alexander (1974, Section 6)]. Some widely used auto-indexing programs are TREOR (Werner et al., 1985), ITO (Visser, 1969), and DICVOL (Boultif and Louër, 1991). All of these programs take either a series of peak positions or corresponding d spacings as input, and their accuracy is crucial for this step to succeed. The programs may tolerate reasonably small random errors in the peak position; however, even a small systematic error will prevent them from finding the correct (or any) solution. Hence, the data set used to extract peak positions must be checked carefully for systematic errors such as the diffractometer zero point, for example, by looking at pairs of Bragg peaks whose values of $\sin\theta$ are related by integer multiples.

Once the unit cell is known, the space group can be determined from systematic absences of Bragg peaks, i.e., Bragg peaks allowed by the unit cell but not observed in the actual spectrum. For example, the observed spectrum of a face-centered-cubic (fcc) material contains only Bragg peaks (hkl) for which h, k, and l are either all odd or all even. Tables listing possible systematic absences for each crystal symmetry class and correlating them with space groups can be found in the International Tables for Crystallography (Vos and Buerger, 1996).

Extracting Integrated Intensities of Bragg Peaks

The extraction of accurate integrated intensities of Bragg peaks is a prerequisite for several other steps in x-ray powder diffraction data analysis. However, the fraction of peaks in the observed spectrum that can be individually fitted is usually very small. The problem of peak overlap is partially overcome in the Rietveld method by describing the shapes of all Bragg peaks observed in the x-ray powder pattern by a small number of profile parameters. It is then possible to perform a refinement of the powder pattern without any atomic model using (and refining) only the lattice and profile parameters to describe the position and shape of all Bragg peaks and letting the intensity of each peak vary freely; two variations are referred to as the Pawley (1981) and LeBail (1988) methods. While this does not solve the problem of exact peak overlap [e.g., of the cubic (511) and (333) peaks], it is very powerful in separately determining the intensities of clustered, partially overlapping peaks. The integrated intensities of individual peaks that can be determined in this manner usually constitute a significant fraction of the total number of allowed Bragg peaks, and they can then be used as input to search for candidate atom positions. Many Rietveld programs (see The Final Rietveld Refinement, below) feature an intensity extraction mode that can be used for this task. The resulting fit can also serve as an upper bound for the achievable goodness of fit in the final Rietveld refinement. EXTRA (Altomare et al., 1995) is another intensity extraction routine that interfaces directly with SIRPOW.92 (see Search for Candidate Atom Positions, below), a program that utilizes direct methods. The program EXPO consists of both the EXTRA and SIRPOW.92 modules.

Search for Candidate Atom Positions

When a suitable starting model for a crystal structure is not known and cannot easily be guessed, it must be determined from the x-ray powder data before any Rietveld refinements can be performed. The Fourier transform of the distribution of electrons in a crystal is the x-ray scattering amplitude. However, only the scattering intensity can be measured, and hence all phase information is lost. Nevertheless, the Fourier transform of the intensities can be useful in finding some of the atoms in the unit cell. A plot of the Fourier transform of the measured scattering intensities is called a Patterson map, and its peaks correspond to translation vectors between pairs of atoms in the unit cell. Obviously, the strongest peak will be at the origin, and depending on the crystal structure, it may be anywhere from simple to impossible to deduce atom positions from a Patterson map. Cases where there is one (or few) relatively heavy atom(s) in the unit cell are most favorable for this approach. Similarly, if the positions of most of the atoms in the unit cell (or at least of the heavy atoms) are already known, it may be reasonable to guess that the phases are dominated by the known part of the atomic structure. Then the differences between the measured intensities and those calculated from the known atoms together with the phase information calculated from the known atoms can be used to obtain a difference Fourier map that, if successful, will indicate the positions of the

remaining atoms. The ability to calculate and plot Patterson and Fourier maps is included in common Rietveld packages (see below).

Another approach to finding atom candidate positions is the use of direct methods originally developed (and widely used) for the solution of crystal structures from single-crystal x-ray data. A mathematical description is beyond the scope of this unit; the reader is referred to the literature (e.g., Giacovazzo, 1992, pp. 335–365). In direct methods, an attempt is made to derive the phase of the structure factor directly from the observed amplitudes through mathematical relationships. This is feasible because the electron density function is positive everywhere and consists of discrete atoms. These two properties, "positivity" and "atomicity," are then used to establish likely relationships between the phases of certain groups of Bragg peaks. If the intensities of enough such groups of Bragg peaks have been measured accurately, this method yields the positions of some or all the atoms in the unit cell. The program SIRPOW.92 (Altomare et al., 1994) is an adaptation of direct methods to x-ray powder diffraction data and has been used to solve a number of organic and organometallic crystal structures from such data.

We emphasize that these procedures are not straightforward, and even for experienced researchers, success is far from guaranteed. However, the chance of success increases appreciably with the quality of the data, in particular with the resolution of the x-ray powder pattern.

The Final Rietveld Refinement

Once a suitable starting model is found, the Rietveld method allows the simultaneous refinement of structural parameters such as atomic positions, site occupancies, isotropic or anisotropic Debye-Waller factors, along with lattice and profile parameters, against the observed x-ray powder diffraction pattern. Since the refinement is usually performed using a least-squares algorithm [chi-square (χ^2) minimization], the result will be a local minimum close to the set of starting parameters. It is the responsibility of the experimenter to confirm that this is the global minimum, i.e., that the model is in fact correct. There are criteria based on the goodness-of-fit of the final refinement that, if they are not fulfilled, almost certainly indicate a wrong solution. However, such criteria are not sufficient in establishing the correctness of a model. For example, if an inspection of the final fit shows that a relatively small amount of total residual error is noticeably concentrated in a few peaks rather than being distributed evenly over the entire spectrum, that may be an indication that the model is incorrect. More details on this subject are given in the International Tables for Crystallography (Prince and Spiegelmann, 1992) and in Young (1993) and Bish and Post (1989).

A number of frequently updated Rietveld refinement program packages are available that can be run on various platforms, most of which are distant descendants of the original program by Rietveld (1969). Packages commonly in use today and freely distributed for noncommercial use by their authors include—among several—GSAS (Larson and Von Dreele, 1994) and FULLPROF (Rodriguez-Carvajal, 1990). Notably, both include the latest descriptions of the x-ray powder diffraction peak shape functions (Thompson et al., 1987), analytical descriptions of the peak shape asymmetry at low angles (Finger et al., 1994), and an elementary description for anisotropic strain broadening for different crystal symmetries (Stephens, 1999). Both programs contain comprehensive documentation, which is no substitute for the key references discussing the Rietveld method (Young, 1993; Bish and Post, 1989).

Estimated Standard Deviations in Rietveld Refinements

The assignment of estimated standard deviations (ESDs) of the atomic parameters obtained from Rietveld refinements is a delicate matter. In general, the ESDs calculated by the Rietveld program are measures of the precision (statistical variation between equivalent experiments) rather than the accuracy (discrepancy from the correct value) of any given parameter. The latter cannot, in principle, be determined experimentally, since the "truly" correct model describing experimental data remains unknown. However, it is the accuracy that the experimenter is interested in, and it is desired to make the best possible estimate of it, based both on statistical considerations and the experience and judgment of the experimenter.

The Rietveld program considers each data point of the measured powder pattern to be an independent measurement of the Bragg peak(s) contributing to it. This implicit assumption holds only if the difference between the refined profile and the experimental data is from counting statistics alone (i.e., if $\chi^2 = 1$ and the points of the difference curve form a random, uncorrelated distribution with center 0 and variance N). However, this is not true for real experiments in which the dominant contributions to the difference curve are an imperfect description of the peak shapes and/or an imperfect atomic model. Consequently, accepting the precision of any refined parameter as a measure of its accuracy cannot be justified and would usually be unreasonably optimistic.

A number of corrections to the Rietveld ESDs have been proposed, and even though all are empirical in nature and there is no statistical justification for these procedures, they can be very valuable in order to obtain estimates. If one or more extra parameters are introduced into a model, the fit to the data will invariably improve. The F statistic can be used to determine the likelihood that this improvement represents additional information (i.e., is statistically significant) and is not purely by chance (Prince and Spiegelmann, 1992). This can also be used to determine how far any given parameter must be shifted away from its ideal value in order to make a statistically significant difference in the goodness of fit. This can be used as an estimate of the accuracy of that parameter that is independent of its precision as calculated by the Rietveld program. There exists no rigid rule on whether precision or accuracy should be quoted when presenting results obtained from x-ray powder diffraction; it certainly depends on the motivation and outcome of the experiment. However, it is very important to distinguish between the two. We note that

the International Union of Crystallography has sponsored a Rietveld refinement round robin, during which identical samples of ZrO_2 were measured and Rietveld refined by 51 participants (Hill and Cranswick, 1994), giving an empirical indication of the achievable accuracy from the technique.

Quantitative Phase Analysis

Quantitative phase analysis refers to the important technique of determining the amount of various crystalline phases in a mixture. It can be performed in two ways, based on the intensities of selected peaks or on a multiphase Rietveld refinement.

In the first case, suppose that the sample is a flat plate that is optically thick to x rays. In a mixture of two phases with mass absorption coefficients μ_1^* and μ_2^* (cm^2/g), the intensity of a given peak from phase 1 will be reduced from its value for a pure sample of phase 1 by the ratio

$$\frac{I_1(x_1)}{I_1(\text{pure})} = \frac{x_1\mu_1^*}{x_1\mu_1^* + (1-x_1)\mu_2^*} \tag{4}$$

where x_1 is the weight fraction of phase 1. Note that the intensity is affected both by dilution and by the change of the sample's absorption constant. When the absorption coefficients are not known, they can be determined from experimental measurements, such as by "spiking" the mixture with an additional amount of one of its constituents. Details are given in Klug and Alexander (1974) and Bish and Post (1989).

If the atomic structures of the constituents are known, a multiphase Rietveld refinement (see above) directly yields scale factors s_j for each phase. The weight fraction w_j of the jth phase can then be calculated by

$$w_j = \frac{s_j Z_j m_j V_j}{\sum_i s_i Z_i m_i V_i} \tag{5}$$

where Z_j is the number of formula units per unit cell, m_j is the mass of a formula unit, and V_j is the unit cell volume.

In either case, the expressions above are given under the assumption that the powders are sufficiently fine, i.e., the product of the linear absorption coefficient μ (cm^{-1}, equal to $\mu^*\rho$) and the linear particle (not sample) size D is small ($\mu D < 0.01$). If not, one must make allowance for microabsorption, as described by Brindley (1945).

It is worth noting that the sensitivity of an x-ray powder diffraction experiment to weak peaks originating from trace phases is governed by the signal-to-background ratio. Hence, in order to obtain maximum sensitivity, it is desirable to reduce the relative background as much as possible.

SAMPLE PREPARATION

The preparation of samples to avoid unwanted artifacts is an important consideration in powder diffraction experiments. One issue is that preferred orientation (texture) should be avoided or controlled. The grains of a sample may tend to align, especially if they have a needle or platelike morphology, so that reflections in certain directions are enhanced relative to others. Various measures such as mixing the powder with a binder or an inert material chosen to randomize the grains or pouring the sample sideways into the flat plate sample holder are in common use. [More details are given in, e.g., Klug and Alexander (1974), Bish and Post (1989), and Jenkins and Snyder (1996).] It is also possible to correct experimental data if one can model the distribution of crystallite orientations; this option is available in most common Rietveld programs (see The Final Rietveld Refinement, above).

A related issue is that there must be a sufficient number of individual grains participating in the diffraction measurement to ensure a valid statistical sample. It may be necessary to grind and sieve the sample, especially in the case of strongly absorbing materials. However, grinding can introduce strain broadening into the pattern, and some experimentation is usually necessary to find the optimum means of preparing a sample. A useful test of whether a specimen in a diffractometer is sufficiently powdered is to scan the sample angle θ over several degrees while leaving 2θ fixed at the value of a strong Bragg peak, in steps of perhaps 0.01°; fluctuations of more than a few percent indicate trouble. It is good practice to rock (flat plate) or twirl (capillary) the sample during data collection to increase the number of observed grains.

PROBLEMS

X rays are absorbed by any material through which they pass, and so the effective volume is generally not equal to the actual volume of the sample. Absorption constants are typically tabulated as the mass absorption coefficient μ/ρ (cm^2/g), which makes it easy to work out the attenuation length in a sample as a function of its composition. A plot of the absorption lengths (in micrometers) of some "typical" samples vs. x-ray energy/wavelength is given in Figure 2. To determine the relative intensities of different

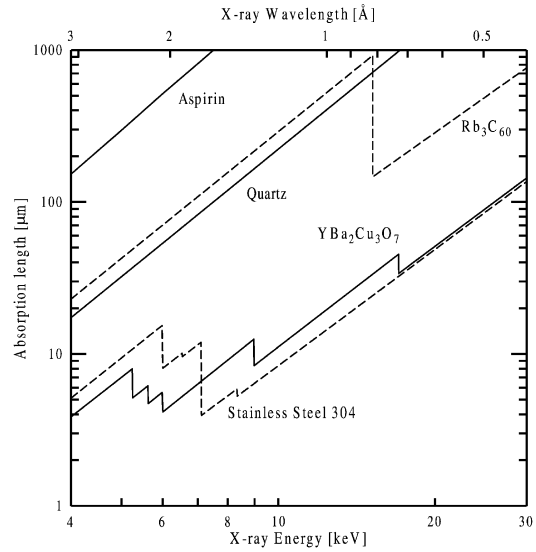


Figure 2. Calculated absorption lengths for several materials as a function of x-ray wavelength or energy.

peaks, it is important to either arrange for the effective volume to be constant or be able to make a quantitative correction. If the sample is a flat plate in symmetrical geometry, as illustrated in Figure 1B, the effective volume is independent of diffraction angle as long as the beam does not spill over the edge of the sample. If the sample is a cylinder and the absorption constant is known, the angle-dependent attenuation constant is tabulated, e.g., in the International Tables for Crystallography (Maslen, 1992). It is also important to ensure that the slits and other optical components of the diffractometer are properly aligned so that the variation of illuminated sample volume with 2θ is well controlled and understood by the experimenter.

Inasmuch as a powder diffraction experiment consists of "simply" measuring the diffracted intensity as a function of angle, one can classify sources of systematic error as arising from measurement of the intensity or the angle or from the underlying assumptions. Errors of intensity can arise from detector saturation, drift or fluctuations of the strength of the x-ray source, and the statistical fluctuations intrinsic to counting the random arrival of photons.

Error of the angle can arise from mechanical faults or instability of the instrument or from displacement of the sample from the axis of the diffractometer (parallax). A subtle form of parallax can occur for flat-plate samples that are moderately transparent to x rays, because the origin of the diffracted radiation is located below the surface of the sample by a distance that depends on the diffraction angle. Another effect that can give rise to an apparent shift of diffraction peaks is the asymmetry caused by the intersection of the curved Debye-Scherrer cone of radiation with a straight receiving slit. This geometric effect has been discussed by several authors (Finger et al., 1994, and references therein), and it can produce a significant shift in the apparent position of low-angle peaks if fitted with a model lineshape that does not properly account for it. A potential source of angle errors in very high precision work is the refraction of x rays; to an x ray, most of the electrons of a solid appear free, and if their number density is n, the index of refraction is given by

$$1 - \left(\frac{e^2}{mc^2}\right)\frac{n\lambda^2}{2\pi} \qquad (6)$$

so that the observed Bragg angle is slightly shifted from its value inside the sample.

The interpretation of powder intensities is based on a number of assumptions that may or may not correspond to experimental reality in any given case.

The integrated intensity is proportional to the square of the structure factor only if the diffracted radiation is sufficiently weak that it does not remove a significant fraction of the incident beam. If the individual grains of the powder sample are too large, strong reflections will be effectively weakened by this "extinction" effect. The basic phenomenon is described in any standard crystallography text, and a treatment specific to powder diffraction is given by Sabine (1993). This can be an issue for grains larger than $\sim$1 μm, which is well below the size that can be easily guaranteed by passing the sample through a sieve. Consequently, it is not easy to assure *a priori* that extinction is not a problem. One might be suspicious of an extinction effect if a refinement shows that the strongest peaks are weaker than predicted by a model that is otherwise satisfactory.

To satisfy the basic premise of powder diffraction, there must be enough grains in the effective volume to produce a statistical sample. This may be a particularly serious issue in highly absorbing materials, for which only a small number of grains at the surface of the sample participate. One can test for this possibility by measuring the intensity of a strong reflection at constant 2θ as a function of sample rotation; if the intensity fluctuates by more than a few percent, it is likely that there is an insufficient number of grains in the sample. One can partly overcome this problem by moving the sample during the measurement to increase the number of grains sampled. A capillary sample can be spun around its axis, while a flat plate can be rocked by a few degrees about the dividing position or rotated about its normal (or both) to achieve this aim.

If the grains of a powder sample are not randomly oriented, it will distort the powder diffraction pattern by making the peaks in certain directions stronger or weaker than they would be in the ideal case. This arises most frequently if the crystallites have a needle- or platelike morphology. [Sample preparation procedures to control this effect are described in, e.g., Klug and Alexander (1974), Bish and Post (1989), and Jenkins and Snyder (1996).] It is also possible to correct experimental data if one can model the distribution of crystallite orientations.

There are a number of standard samples that are useful for verifying and monitoring instrument performance and as internal standards. The National Institute of Standards and Technology sells several materials, such as SRM 640 (Silicon Powder), SRM 660 (LaB_6, which has exceedingly sharp lines), SRM 674a (a set of five metal oxides of various absorption lengths, useful for quantitative analysis standards), and SRM 1976 (corundum Al_2O_3 plate, with somewhat sharper peaks than the same compound in SRM 674a and certified relative intensities). Silver behenate powder has been proposed as a useful standard with a very large lattice spacing of $c = 58.4$ Å (Blanton et al., 1995).

EXAMPLES

Comparison Against a Database of Known Materials

An ongoing investigation of novel carbon materials yielded a series of samples with new and potentially interesting properties: (1) the strongest signal in a mass spectrometer was shown to correspond to the equivalent of an integer number of carbon atoms, (2) electron diffraction indicated that the samples were at least partially crystalline, and (3) while the material was predicted to consist entirely of carbon, the procedures for synthesis and purification had involved an organic Li compound. The top half of Figure 3 shows an x-ray powder diffraction pattern of one such sample recorded at $\lambda = 0.7$ Å at beamline X3B1 of the National Synchrotron Light Source. The peak positions can be indexed (Werner et al., 1985) to a monoclinic unit cell: $a = 8.36$ Å, $b = 4.97$ Å, $c = 6.19$ Å, and $\beta = 114.7°$.

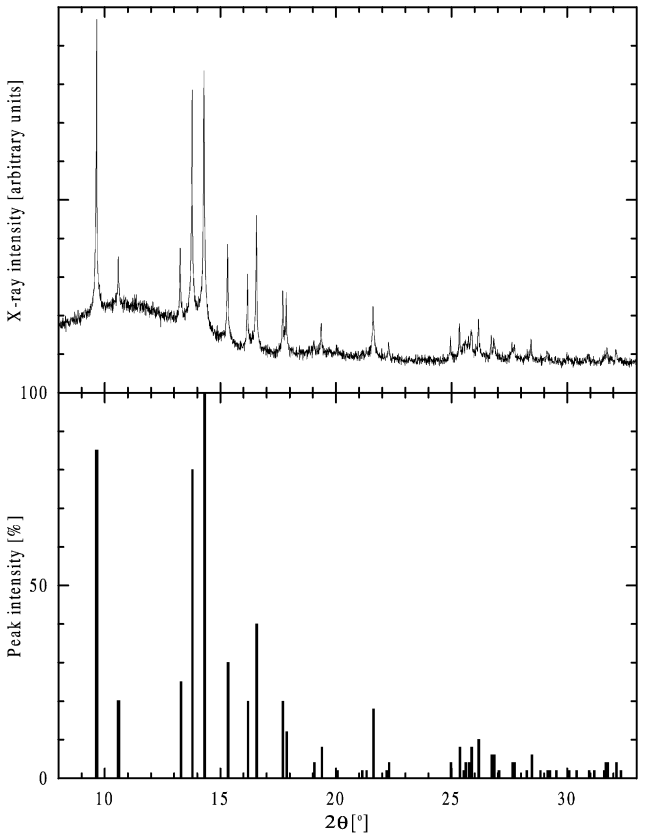


Figure 3. X-ray powder diffraction pattern of "unknown" sample (top) and relative peak intensities of Li_2O_3 as listed in the PDF database (bottom), both at $\lambda = 0.7$ Å.

The Powder Diffraction File (PDF) database provided by the International Center for Diffraction Data (ICDD, 1998) allows for searching based on several search criteria, including types of atoms, positions of Bragg peaks, or unit cell parameters. Note that it is not necessary to know the unit cell of a compound to search the database against its diffraction pattern, nor is it necessary to know the crystal structures (or even the unit cells) of all the compounds included in the database.

Searches for materials with an appropriate position of their strong diffraction peaks containing C only, or containing some or all of C, N, O, or H, did not result in any candidate materials. However, including Li in the list of possible atoms matched the measured pattern with the one in the database for synthetic Li_2CO_3 (zabuyelite). The peak intensities listed in the database, converted to 2θ for the appropriate wavelength, are shown in the bottom half of Figure 3, leading to the conclusion that the only crystalline fraction of the candidate sample is Li_2CO_3, not a new carbon material.

This type of experiment is fairly simple, and both data collection and analysis can be performed quickly. Average sample quality is sufficient, and usually so are relatively low-resolution data from an x-ray tube. In fact, this test is routinely performed before any attempt to solve the structure of an unknown and presumably new material.

Quantitative Phase Analysis

The qualitative and quantitative detection of small traces of polymorphs plays a major role in pharmacology, since many relevant molecules form two or more different crystal structures, and phase purity rather than chemical purity is required for scientific, regulatory, and patent-related legal reasons. Mixtures of the α and γ phases of the anti-inflammatory drug indomethacin ($C_{19}H_{16}ClNO_4$) provide a good example for the detection limit of x-ray powder diffraction experiments in this case (Dinnebier et al., 1996).

Mixtures of 10, 1, 0.1, 0.01, and 0 wt% α phase (with the balance γ phase) were investigated. Using its five strongest Bragg peaks, 0.1% α phase can easily be detected, and the fact that the "pure" γ phase from which the mixtures were prepared contains traces of α became evident, making the quantification of the detection limits below 0.1% difficult. Figure 4 shows the strong (120) peak of the α phase and the weak (010) peak (0.15% intensity of its strongest peak) of the γ phase for three different concentrations of the α phase.

While these experiments and the accompanying data analysis are not difficult to perform per se, it is probably necessary to have access to a synchrotron radiation source with its high resolution to obtain detection limits comparable to or even close to the values mentioned. The data of this example were collected on beamline X3B1 of the National Synchrotron Light Source.

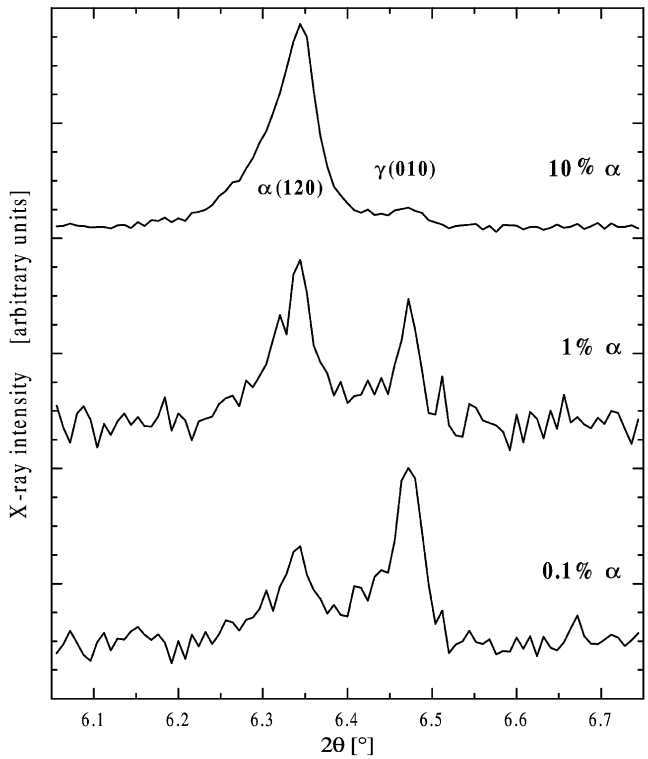


Figure 4. Strong (120) peak of the α phase and weak (010) peak of the γ phase of the drug indomethacin for different concentrations of α phase.

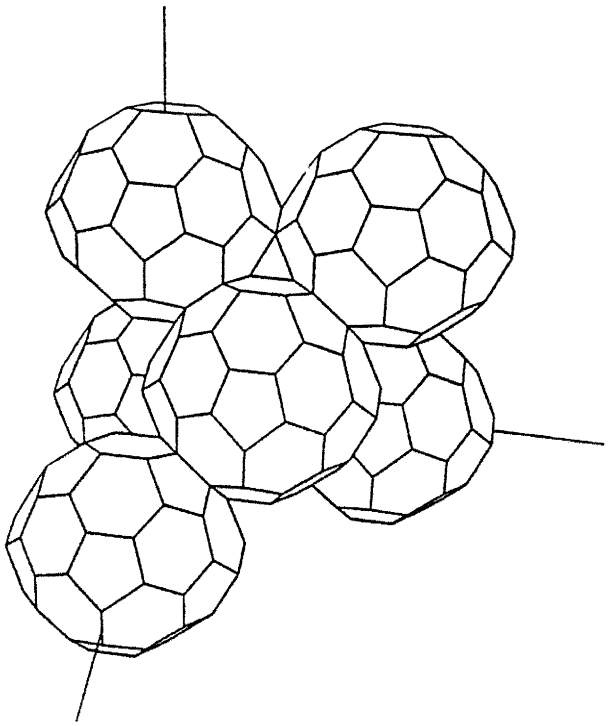


Figure 5 Atomic clusters R(Al-Li-Cu) exhibiting nearly icosahedral symmetry.

Structure Determination in a Known Analog

The icosahedral quasicrystal phase *i*(Al-Li-Cu) is close to a crystalline phase, *R*(Al-Li-Cu), which has a very similar local structure, making the crystal structure of the latter an interesting possible basis for the structure of the former. The crystal symmetry [body-centered-cubic (bcc)], lattice parameter ($a = 13.89$ Å), space group ($Im\bar{3}$), and composition ($Al_{5.6}Li_{2.9}Cu$) suggest that the crystal structure of *R*(Al-Li-Cu) is similar to that of $Mg_{32}(Zn, Al)_{49}$ (Bergman et al., 1952, 1957). However, using the atomic parameters of the latter compound does not give a satisfactory fit to the data of the former. On the other hand, a Rietveld analysis (see Data Analysis and Interpretation, above) using those parameters as starting values converges and gives a stable refinement of occupancies and atom positions (Guryan et al., 1988).

Such an experiment can be done with a standard laboratory x-ray source. This example was solved using Mo K_α radiation from a rotating-anode x-ray generator. Figure 5 shows the nearly icosahedral symmetry of the atomic clusters of which this material is comprised.

Ab Initio Structure Determination

The Kolbe-Schmitt reaction, which proceeds from sodium phenolate, C_6H_5ONa, is the starting point for the synthesis of many pigments, fertilizers, and pharmaceuticals such as aspirin. However, replacing Na with heavier alkalis thwarts the reaction, a fact known since 1874 yet not understood to date. This provides motivation for the solution of the crystal structure of potassium phenolate, C_6H_5OK (Dinnebier et al., 1997).

A high-resolution x-ray powder diffraction pattern of a sample of C_6H_5OK was measured at room temperature, $\lambda = 1.15$ Å, and $2\theta = 5°-65°$ with step size 0.005° for a total counting time of 20 hr on beamline X3B1 of the National Synchrotron Light Source. Low-angle diffraction peaks had a FWHM of 0.013°. Following are the steps that were necessary to obtain the structure solution of this compound.

The powder pattern was indexed (Visser, 1969) to an orthorhombic unit cell, $a = 14.10$ Å, $b = 17.91$ Å, and $c = 7.16$ Å. Observed systematic absences of Bragg peaks, e.g., of all ($h0l$) peaks with h odd, were used to determine the space group, $Pna2_1$ (Vos and Buerger, 1996), and the number of formula units per unit cell ($Z = 12$) was determined from geometrical considerations.

Next, using the LeBail technique (Le Bail et al., 1988), i.e., fitting the powder pattern to lattice and profile parameters without any structural model, ~300 Bragg peak intensities were extracted. Using these as input for the direct-methods program SIRPOW.92 (Altomare et al., 1994), it was possible to deduce the positions of potassium and some candidate oxygen atoms but none of the carbon atoms. In the next stage, the positions of the centers of the phenyl rings were searched by replacing them with diffuse pseudoatoms having the same number of electrons but a high temperature factor. After Rietveld refining this model (see Data Analysis and Interpretation, above) one pseudoatom at a time was back substituted for the corresponding phenyl ring, and the orientation of each phenyl ring was found via a grid search of all of its possible orientations. Upon completion of this procedure, the coordinates of all nonhydrogen atoms had been obtained.

The final Rietveld refinement (see Data Analysis and Interpretation) was performed using the program GSAS (Larson and Von Dreele, 1994), the hydrogen atoms were included, and the model remained stable when all structural parameters were refined simultaneously. Figure 6 shows a plot of the final refinement of C_6H_5OK, and Figure 7 shows the crystal structure obtained.

This example illustrates the power of high-resolution x-ray powder diffraction for the solution of rather complicated crystal structures. There are 24 nonhydrogen atoms in the asymmetric unit, which until recently would have been impossible to solve without the availability of a single crystal. At the same time, it is worth emphasizing that such an *ab initio* structure solution from x-ray powder diffraction requires both a sample of extremely high quality and an instrument with very high resolution. By comparison, it was not possible to obtain the crystal structure of C_6H_5OLi from a data set with approximately 3 times fewer independently observed peaks due to poorer crystallinity of the sample. Even with a sample and data of sufficient quality, solving a structure *ab initio* from a powder sample (unlike the single-crystal case) requires experience and a significant investment of time and effort on the part of the researcher.

Time-Resolved X-Ray Powder Diffraction

Zeolites are widely used in a number of applications, for example, as catalysts and for gas separations. Their

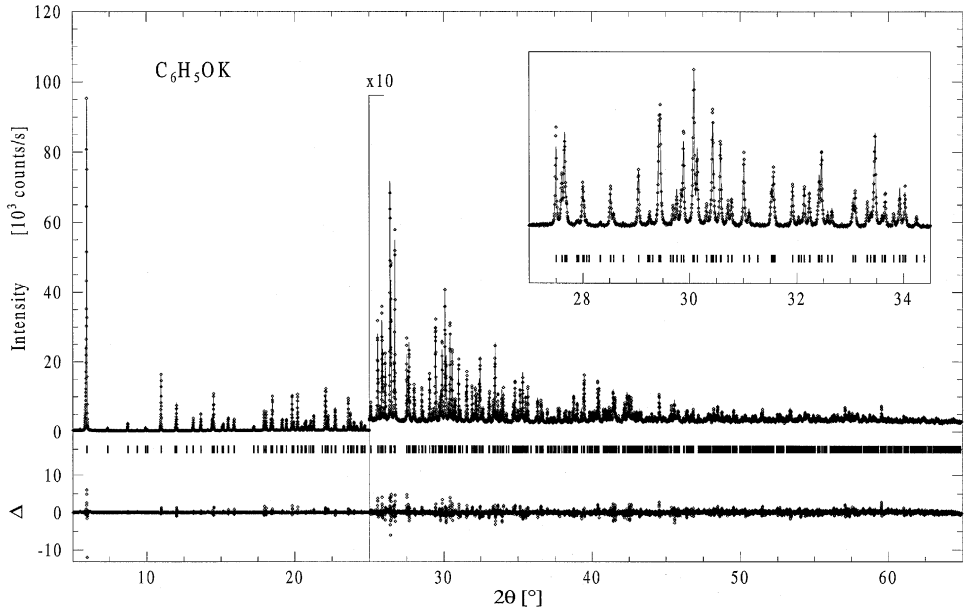


Figure 6. Rietveld refinement of potassium phenolate, C_6H_5OK. Diamonds denote x-ray powder diffraction data; solid line denotes atomic model. Difference curve is given below.

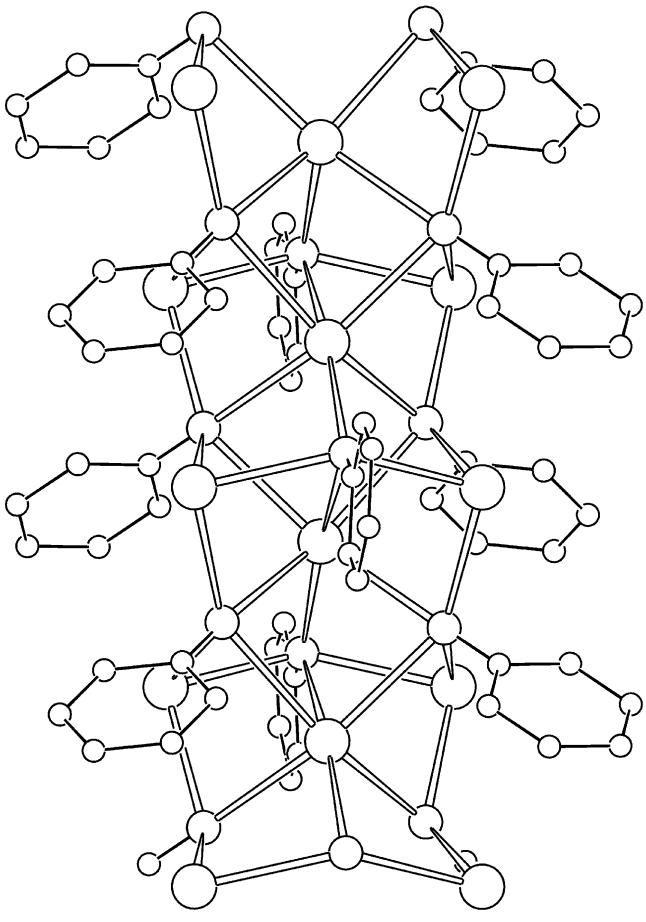


Figure 7. Crystal structure of potassium phenolate, C_6H_5OK, solved *ab initio* from x-ray powder diffraction data.

structural properties, particularly the location of the extra-framework cations, are important in understanding their role in these processes.

Samples of CsY zeolite dehydrated at 300° and at 500°C show differences in their diffraction patterns (Fig. 8). Rietveld analysis (see Data Analysis and Interpretation, above) shows that the changes occur in the extra-framework cations. The basic framework contains large so-called supercages and smaller sodalite cages, and in dehydration at 300°C, almost all of the Cs cations occupy the supercages while the Na cations occupy the sodalite cages. However, after dehydration at 500°C, a fraction of both Cs and Na ions will populate each other's cage position, resulting in mixed occupancy for both sites (Poshni et al., 1997; Norby et al., 1998).

To observe this transition while it occurs, the subtle changes in the CsY zeolite structure were followed by *in situ* dehydration under vacuum. The sample, loosely packed in a 0.7-mm quartz capillary, was ramped to a final temperature of 500°C over a period of 8 hr. Figure 9 shows the evolution of the diffraction pattern as a function of time during the dehydration process (Poshni et al., 1997). The data were collected at beamline X7B of the National Synchrotron Light Source with a translating image plate data collection stage.

To investigate phase transitions and solid-state reactions in near real time by recording x-ray powder diffraction patterns as a function of time, pressure, and/or temperature, a number of conditions must be met. For data collection, a position-sensitive detector or a translating image plate stage is required. Also, since there can be no collimator slits or analyzer crystal between the sample and the detector/image plate, such an experiment is optimally carried out at a synchrotron radiation source. These have an extremely low vertical divergence, allowing for a reasonably good angular resolution in this detector

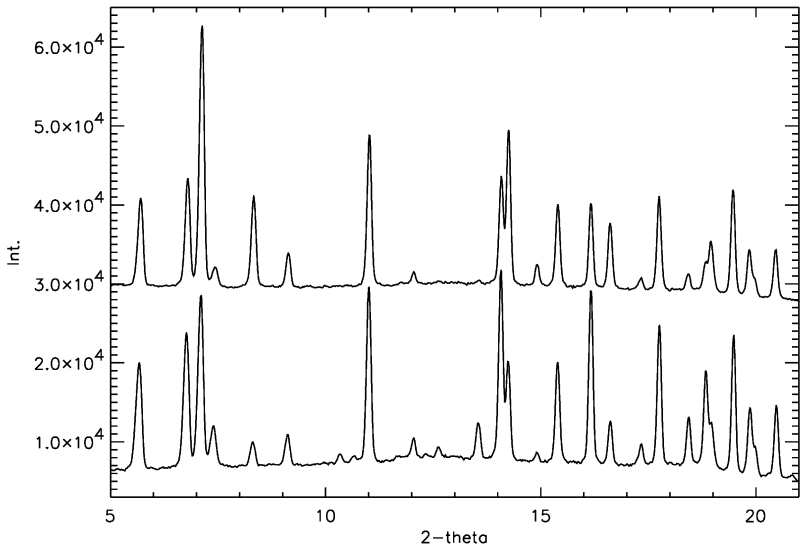


Figure 8. X-ray powder diffraction patterns of CsY zeolite dehydrated at 500°C (top) and 300°C (bottom).

geometry. The sample environment must be designed with care in order to provide, e.g., the desired control of pressure, temperature, and chemical environment compatible with appropriate x-ray access and low background. Examples include a cryostat with a Be can, a sample heater, a diamond-anvil pressure apparatus, and a setup that allows a chemical reaction to take place inside the capillary.

Determination of Subtle Structural Details

The superconductor Rb_3C_{60} has $T_c \approx 30$ K. The fullerenes form an fcc lattice, and of the three Rb^+ cations, one occupies the large octahedral (O) site at ($\frac{1}{2}$, 0, 0) and the remaining two occupy the small tetrahedral (T) sites at ($\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$). The size mismatch between the smaller Rb^+ ions and the larger octahedral site led to the suggestion that the Rb^+ cations in the large octahedral site could be displaced from the site center, supported by some nuclear magnetic resonance (NMR) and extended x-ray absorption fine structure (EXAFS) experiments (see XAFS Spectroscopy). If true, such a displacement would have a significant impact on the electronic and superconductin properties. By comparison, the Rb−C distances of the Rb^+ cation in the tetrahedral site are consistent with ionic radii.

In principle, a single x-ray pattern cannot distinguish between static displacement of an atom from its site center (as proposed for this system) and dynamic thermal fluctuation about a site (as described by a Debye-Waller factor). Hence, to address this issue, an x-ray powder diffraction study of Rb_3C_{60} at various temperatures was carried out at beamline X3B1 of the National Synchrotron Light Source (Bendele et al., 1998).

At each temperature the data were Rietveld refined (see Data Analysis and Interpretation, above) against two competing (and otherwise identical) structural models: (1) The octahedral Rb^+ cation is fixed at ($\frac{1}{2}$, 0, 0) and its isotropic

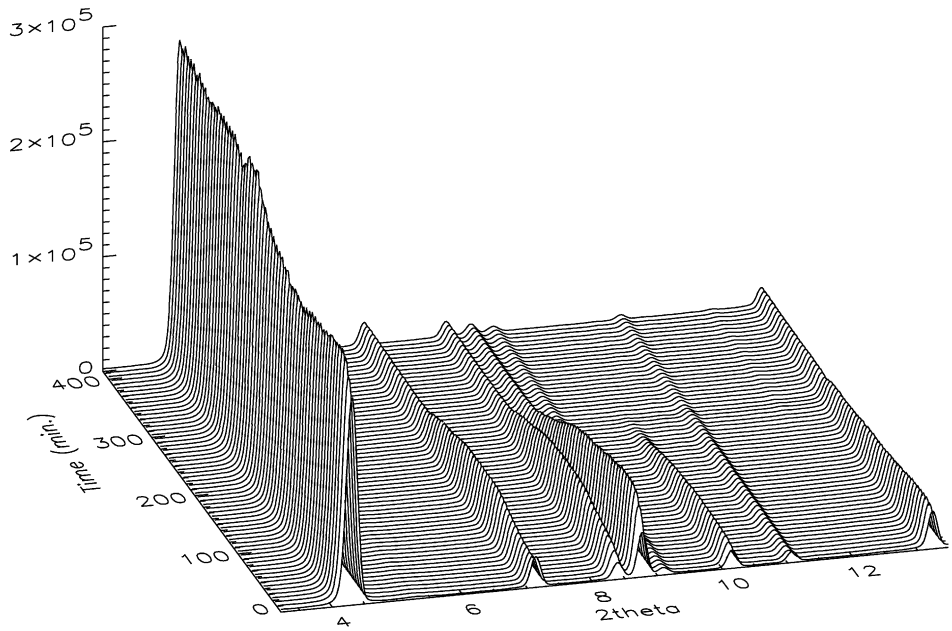


Figure 9. Evolution of the diffraction pattern during the *in situ* dehydration of CsY zeolite.

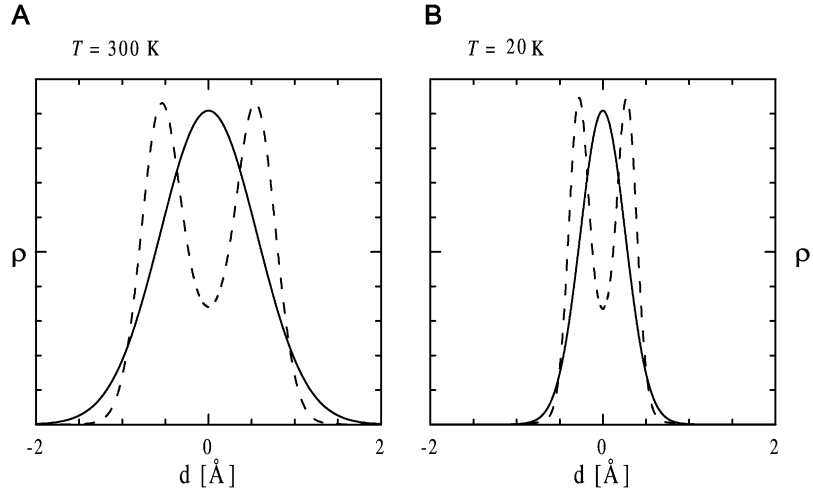


Figure 10. Electron distributions obtained from x-ray powder diffraction data in the octahedral site of Rb_3C_{60} for $T = 300$ K (**A**) and $T = 20$ K (**B**).

Debye-Waller factor B_o is refined and (2) the octahedral Rb^+ cation is shifted an increasing distance ε away from ($\frac{1}{2}$, 0, 0) until the refined B_o becomes comparable to the temperature factor of the tetrahedral ion, B_t. In each case, both models result in an identical goodness of fit, and the direction of the assumed displacement ε has no effect whatsoever. Figure 10 shows the electron distributions in the octahedral site for both models (in each case convoluted with the appropriate Debye-Waller factor). While the width of this distribution (static or dynamic) is 1.32 Å at room temperature, it has decreased to 0.56 Å upon cooling the sample to $T = 20$ K.

If the octahedral Rb^+ cations were truly displaced away from their site center (either at all temperatures or below some transition temperature T_{tr}), the amount of that displacement would act as an order parameter, i.e., increase or saturate with lower temperature and not decrease monotonically to zero as the data show it does. Hence, any off-center displacement of the octahedral Rb^+ cations can be excluded from this x-ray powder diffraction experiment. Consequently, other possible causes for the features seen in NMR and EXAFS experiments that had given rise to this suggestion must be researched. For such an experiment to be valid and successful, it must be possible to put very tight limits on the accuracy of the measured structural parameters. This requires that two conditions are met. First, the sample must have the highest possible quality, in terms of both purity and crystallinity, and the data must have both high resolution and very good counting statistic. Second, the accuracy for each measured parameter must be judged carefully, which is not trivial in x-ray powder diffraction. Note, in particular, the discussion under Estimated Standard Deviations in Rietveld Refinements, above.

LITERATURE CITED

Altomare, A., Burla, M. C., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., and Polidori, G. 1995. Extra: A program for extracting structure-factor amplitudes from powder diffraction data. *J. Appl. Crystallogr.* 28:842–846. See Internet Resources.

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Burla, M. C., Polidori, G., and Camalli, M. 1994. SIRPOW.92—a program for automatic solution of crystal structures by direct methods optimized for powder data. *J. Appl. Crystallogr.* 27:435–436. See Internet Resources.

Ashcroft, N. W. and Mermin, N. D. 1976. Solid State Physics. Holt, Rinehart and Winston, New York.

Balzar, D. and Ledbetter, H. 1993. Voigt-function modeling in Fourier analysis of size- and strain-broadened X-ray diffraction peaks. *J. Appl. Crystrallogr.* 26:97–103.

Bendele, G. M., Stephens, P. W., and Fischer, J. E. 1998. Octahedral cations in Rb_3C_{60}: Reconciliation of conflicting evidence from different probes. *Europhys. Lett.* 41:553–558.

Bergman, G., Waugh, J. L. T., and Pauling, L. 1952. Crystal structure of the intermetallic compound $Mg_{32}(Al,Zn)_{49}$ and related phases. *Nature* 169:1057.

Bergman, G., Waugh, J. L. T., and Pauling, L. 1957. The crystal structure of the metallic phase $Mg_{32}(Al, Zn)_{49}$. *Acta Crystallogr.* 10:254–257.

Bish, D. L. and Post, J. E. (eds.). 1989. Modern Powder Diffraction. Mineralogical Society of America, Washington, D.C.

Blanton, T. N., Huang, T. C., Toraya, H., Hubbard, C. R., Robie, S. B., Louër, D., Göbel, H. E., Will, G., Gilles, R., and Raftery T. 1995. JCPDS—International Centre for Diffraction Data round robin study of silver behenate. A possible low-angle X-ray diffraction calibration standard. *Powder Diffraction* 10:91–95.

Boultif, A. and Louër, D. 1991. Indexing of powder diffraction patterns for low symmetry lattices by the successive dichotomy method. *J. Appl. Crystallogr.* 24:987–993.

Brindley, G. W. 1945. The effect of grain or particle size on x-ray reflections from mixed powders and alloys, considered in relation to the quantitative determination of crystalline substances by x-ray methods. *Philos. Mag.* 36:347–369.

Brown, P. J., Fox, A. G., Maslen, E. N., O'Keefe, M. A., and Willis, B. T. M. 1992. Intensity of diffracted intensities. *In* International Tables for Crystallography, Vol. C (A. J. C. Wilson, ed.) pp. 476–516. Kluwer Academic Publishers, Dordrecht.

Clarke, R. and Rowe, W. P. 1991. Real-time X-ray studies using CCDs. *Synchrotron Radiation News* 4(3):24–28.

Dinnebier, R. E., Pink, M., Sieler, J., and Stephens, P. W. 1997. Novel alkali metal coordination in phenoxides: Powder diffraction results on C_6H_5OM [M = K, Rb, Cs]. *Inorg. Chem.* 36:3398–3401.

Dinnebier, R. E., Stephens, P. W., Byrn, S., Andronis, V., and Zografi, G. 1996. Detection limit for polymorphs in drugs using powder diffraction (J. B. Hastings, ed.). p. B47. *In* BNL-NSLS 1995 Activity Report. National Synchrotron Light Source, Upton, N.Y.

Finger, L. W., Cox, D. E., and Jephcoat, A. P. 1994. A correction for powder diffraction peak asymmetry due to axial divergence. *J. Appl. Crystallogr.* 27:892–900.

Giacovazzo, C. (ed.) 1992. Fundamentals of Crystallography. Oxford University Press, Oxford.

Guryan, C. A., Stephens, P. W., Goldman, A. I., and Gayle, F. W. 1988. Structure of icosahedral clusters in cubic $Al_{5.6}Li_{2.9}Cu$. *Phys. Rev. B* 37:8495–8498.

Hill, R. J. and Cranswick, L. M. D. 1994. International Union of Crystallography Commission on Powder Diffraction Rietveld Refinement Round Robin. II. Analysis of monoclinic ZrO_2. *J. Appl. Crystallogr.* 27:802–844.

International Center for Diffraction Data (ICDD). 1998. PDF-2 Powder Diffraction File Database. ICCD, Newtown Square, Pa. See Internet Resources.

Ito, M. and Amemiya, Y. 1991. X-ray energy dependence and uniformity of an imaging plate detector. *Nucl. Instrum. Methods Phys. Res.* A310:369–372.

Jenkins, R. and Snyder, R. L. 1996. Introduction to X-ray Powder Diffractometry. John Wiley & Sons, New York.

Kittel, C. 1996. Introduction to Solid State Physics. John Wiley & Sons, New York.

Klug, H. P. and Alexander, L. E. 1974. X-ray Diffraction Procedures for Polycrystalline and Amorphous Materials. John Wiley & Sons, New York.

Larson, A. C. and Von Dreele, R. B. 1994. GSAS: General Structure Analysis System. Los Alamos National Laboratory, publication LAUR 86–748. See Internet Resources.

Le Bail, A., Duryo, H., and Fourquet, J. L. 1988. Ab-initio structure determination of $LiSbWO_6$ by X-ray powder diffraction. *Mater. Res. Bull.* 23:447–452.

Maslen, E. N. 1992. X-ray absorption. *In* International Tables for Crystallography, Vol. C (A. J. C. Wilson, ed.). pp. 520–529. Kluwer Academic Publishers, Dordrecht.

Miyahara, J., Takahashi, K., Amemiya, Y., Kamiya, N., and Satow, Y. 1986. A new type of X-ray area detector utilizing laser stimulated luminescence. *Nucl. Instrum. Methods Phys. Res.* A246:572–578.

Norby, P., Poshni, F. I., Gualtieri, A. F., Hanson, J. C., and Grey, C. P. 1998. Cation migration in zeolites: An in-situ powder diffraction and MAS NMR study of the structure of zeolite Cs(Na)-Y during dehydration. *J. Phys. Chem. B* 102:839–856.

Noyan, I. C. and Cohen, J. B. 1987. Residual Stress. Springer- Verlag, New York.

Pawley, G. S. 1981. Unit-cell refinement from powder diffraction scans. *J. Appl. Crystallogr.* 14:357–361.

Poshni, F. I., Ciraolo, M. F., Grey, C. P., Gualtieri, A. F., Norby, P., and Hanson, J. C. 1997. An in-situ X-ray powder diffraction study of the dehydration of zeolite CsY (J. B. Hastings, ed.). p. B84. *In* BNL-NSLS 1996 Activity Report. National Synchrotron Light Source, Upton, N.Y.

Prince, E. and Spiegelmann, C. H. 1992. Statistical significance tests. *In* International Tables for Crystallography, Vol. C (A. J. C. Wilson, ed.). pp. 618–621. Kluwer Academic Publishers, Dordrecht.

Rietveld, H. M. 1969. A profile refinement method for nuclear and magnetic structures. *J. Appl. Crystallogr.* 2:65–71.

Rodriguez-Carvajal, R. 1990. FULLPROF: A program for Rietveld refinement and pattern matching analysis. *In* Abstracts of the Satellite Meeting on Powder Diffraction of the XV Congress of the IUCr, p. 127. IUCr, Toulouse, France. See Internet Resources.

Sabine, T. M. 1993. The flow of radiation in a polycrystalline material. *In* The Rietveld Method (R. A. Young, ed.). pp. 55–61. Oxford University Press, Oxford.

Smith, G. W. 1991. X-ray imaging with gas proportional detectors. *Synchrotron Radiation News* 4(3):24–30.

Stephens, P. W. 1999. Phenomenological model of anisotropic peak broadening in powder diffraction. *J. Appl. Crystallogr.* 32:281–289.

Thompson, P., Cox, D. E., and Hastings. J. B. 1987. Rietveld refinement of Debye-Scherrer synchrotron X-ray data from Al_2O_3. *J. Appl. Crystallogr.* 20:79–83.

Visser, J. W. 1969. A fully automatic program for finding the unit cell from powder data. *J. Appl. Crystallogr.* 2:89–95.

Vos, A. and Buerger, M. J. 1996. Space-group determination and diffraction symbols. *In* International Tables for Crystallography, Vol. A (T. Hahn, ed.) pp. 39–48. Kluwer Academic Publishers, Dordrecht.

Werner, P.-E., Eriksson, L., and Westdahl, M. 1985. TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Crystallogr.* 18:367–370.

Young, R. A. (ed.) 1993. The Rietveld Method. Oxford University Press, Oxford.

KEY REFERENCES

Azaroff, L. V. 1968. Elements of X-ray Crystallography. McGraw-Hill, New York.

Gives comprehensive discussions of x-ray crystallography albeit not particularly specialized to powder diffraction.

Bish and Post, 1989. See above.

An extremely useful reference for details on the Rietveld method.

Giacovazzo, 1992. See above.

Another comprehensive text on x-ray crystallography, not particularly specialized to powder diffraction.

International Tables for Crystallography (3 vols.). Kluwer Academic Publishers, Dordrecht.

An invaluable resource for all aspects of crystallography. Volume A treats crystallographic symmetry in direct space, and it includes tables of all crystallographic plane, space, and point groups. Volume B contains accounts of numerous aspects of reciprocal space, such as the structure-factor formalism, and Volume C contains mathematical, physical, and chemical information needed for experimental studies in structural crystallography.

Jenkins and Snyder, 1996. See above.

Offers a modern approach to techniques, especially of phase identification and quantification; however, there is very little discussion of synchrotron radiation or of structure solution.

Klug and Alexander, 1974. See above.

The classic text about many aspects of the technique, particularly the analysis of mixtures, although several of the experimental techniques described are rather dated.

Langford, J. I. and Louër, D. 1996. Powder diffraction. *Rep. Prog. Phys.* 59:131–234.

Offers another approach to techniques that includes a thorough discussion of synchrotron radiation and of structure solution.

Young, 1993. See above.

Another valuable reference for details on the Rietveld refinements.

INTERNET RESOURCES

http://www.ba.cnr.it/IRMEC/SirWare.html

http://ww.iccd.com

http://www.ccp14.ac.uk

An invaluable resource for crystallographic computing that contains virtually all freely available software for power (and also single-crystal) diffraction for academia, including every program mentioned in this unit.

ftp://ftp.lanl.gov/public/gsas

ftp://charybde.saclay.cea.fr/pub/divers/fullp/

PETER W. STEPHENS
State University of New York
Stony Brook, New York

GOETZ M. BENDELE
Los Alamos National Laboratory
Los Alamos, New Mexico, and
State University of New York
Stony Brook, New York

SINGLE-CRYSTAL X-RAY STRUCTURE DETERMINATION

INTRODUCTION

X-ray techniques provide one of the most powerful, if not the most powerful, methods for deducing the crystal and molecular structure of a crystalline solid. In general, it is a technique that provides a large number of observations for every parameter to be determined and results obtained thereby can usually be considered very reliable. Significant advances have taken place, especially over the last twenty years, to make the collection of data and the subsequent steps to carry out the rest of the structure determination amenable to the relative novice in the field. This has led to the adoption of single-crystal diffraction methods as a standard analytical tool. It is no accident that this has paralleled the development of the digital computer—modern crystallography is strongly dependent on the use of the digital computer. Typically, a small single crystal (a few tenths of a millimeter in dimension) of the material to be investigated is placed on a diffractometer and data are collected under computer-controlled operation. A moderate-sized unit cell (10 to 20 Å on edge) can yield 5000 or so diffraction maxima, which might require a few day's data collection time on a scintillation counter diffractometer or only hours of data collection time on the newer charge-coupled device (CCD) area detectors.

From these data the fractional coordinates describing the positions of the atoms within the cell, some characteristics of their thermal motion, the cell dimensions, the crystal system, the space group symmetry, the number of formula units per cell, and a calculated density can be obtained. This "solving of the structure" may only take a few additional hours or a day or two if no undue complications occur.

We will concentrate here on the use of X-ray diffraction techniques as would be necessary for the normal crystal structure investigation. A discussion of the basic principles of single-crystal X-ray diffraction will be presented first, followed by a discussion of the practical applications of these techniques. In the course of this discussion, a few of the more widely used methods of overcoming the lack of measured phases, i.e., the "phase problem" in crystallography, will be presented. These methods are usually sufficient to handle most routine crystal structure determinations. Space will not permit discussion of other more specialized techniques that are available and can be used when standard methods fail. The reader may wish to refer to other books devoted to this subject (e.g., Ladd and Palmer, 1985; Stout and Jensen, 1989; Lipscomb and Jacobson, 1990) for further information on these other techniques and for a more in-depth discussion of diffraction methods.

Competitive and Related Techniques

Single-crystal X-ray diffraction is certainly not the only technique available to the investigator for the determination of the structure of materials. Many other techniques can provide information that is often complementary to that obtained from single-crystal X-ray experiments. Some of these other techniques are briefly described below.

Crystalline powders also diffract X rays. In fact, it is convenient to view powder diffraction as single-crystal diffraction integrated over angular coordinates, thus retaining only sin θ dependence (assuming the powder contains crystallites in random orientation). Diffraction maxima with essentially the same sin θ values combine in the powder diffraction pattern. This results in fewer discrete observations and makes the unit cell determination and structure solution more difficult; this is especially true in lower symmetry cases. The fewer number of data points above background translates into larger standard deviations for the determined atomic parameters. Since the angular information is lost, obtaining an initial model is often quite difficult if nothing is known about the structure. Therefore, the majority of quantitative powder diffraction investigations are done in situations where an initial model can be obtained from a related structure; the calculated powder pattern is then fit to the observed one by adjusting the atomic parameters, the unit cell parameters, and the parameters that describe the peak shape of a single diffraction maximum. (This is usually termed Rietveld refinement; see X-RAY POWDER DIFFRACTION and NEUTRON POWDER DIFFRACTION) If an appropriate single crystal of the material could be obtained, single-crystal diffraction would be the preferred approach. Powder diffraction sometimes provides an alternate approach when only very small crystals can be obtained.

Neutrons also undergo diffraction when they pass through crystalline materials. In fact, the theoretical descriptions of X-ray and neutron diffraction are very similar, primarily differing in the atomic scattering factor used (see NEUTRON POWDER DIFFRACTION and SINGLE-CRYSTAL

NEUTRON DIFFRACTION). The principal interaction of X rays with matter is electronic, whereas with neutrons two interactions predominate—one is nuclear and the other is magnetic (Majkrzak et al., 1990). The interaction of the neutron with the nucleus varies with the isotope and typically does not show the larger variation seen with X rays, where the scattering is proportional to the atomic number. Therefore neutrons can be useful in distinguishing between atoms of neighboring atomic number or light atoms, especially hydrogen in the presence of heavy atoms. The neutron scattering factors for hydrogen and deuterium are also markedly different, which opens up a number of interesting possible structural studies, especially of organic and biological compounds. As noted above, neutrons are also sensitive to magnetic structure (see MAGNETIC NEUTRON SCATTERING). In fact, neutron scattering data constitute the most significant body of experimental evidence regarding long-range magnetic ordering in solids. In practice, with the neutron sources that are presently available, it is necessary to use a considerably larger sample for a neutron diffraction investigation than is necessary for an X-ray investigation. Single-crystal X-ray studies are usually carried out initially, followed by the single-crystal neutron investigation to obtain complementary information where sample preparation is not a problem. (It might be noted that for powder studies neutron diffraction can have some advantages over X-ray diffraction in terms of profile pattern fitting due to the Gaussian character of the neutron diffraction peaks and the enhanced intensities that can be obtained at higher sin θ values; see NEUTRON POWDER DIFFRACTION).

Electron diffraction can also provide useful structural information on materials. The mathematical description appropriate for the scattering of an electron by the potential of an atom has many similarities to the description used in discussing X-ray diffraction. Electrons, however, are scattered much more efficiently than either X rays or neutrons, so much so that electrons can penetrate only a few atomic layers in a solid and are likely to undergo multiple scattering events in the process. Hence, electron diffraction is most often applied to the study of surfaces (low-energy electron diffraction; LEED), in electron-microscopic studies of microcrystals, or in gas-phase diffraction studies.

X-ray absorption spectroscopy (EXAFS) (Heald and Tranquada, 1990) provides an additional technique that can be used to obtain structural information (see XAFS SPECTROSCOPY). At typical X-ray energies, absorption is a much more likely event than the scattering process that we associate with single-crystal diffraction. When an X-ray photon is absorbed by an electron in an atom, the electron is emitted with an energy of the X-ray photon minus the electron-binding energy. Thus some minimum photon energy is necessary, and the X-ray absorption spectrum shows sharp increases in absorption as the various electron-binding energies are crossed. These sharp steps are usually termed X-ray edges and have characteristic values for each element. Accurate measurements have revealed that there is structure near the edge. This fine structure is explained by modulations in the final state of the photoelectron that are caused by backscattering from the surrounding atoms. The advent of synchrotron X-ray sources has accelerated the use of EXAFS since they provide the higher intensities that are necessary to observe the fine structure. The local nature of EXAFS makes it particularly sensitive to small perturbations within the unit cell and can thus complement the information obtained in diffraction experiments. With EXAFS it is possible to focus in on a particular atom type, and changes can be monitored without determining the entire structure. It can be routinely performed on a wide variety of heavier elements. When long-range order does not exist, diffraction methods such as X-ray diffraction have less utility. EXAFS yields an averaged radial distribution function pertinent to the particular atom and can be applied to amorphous as well as crystalline materials. EXAFS can be considered as being complementary to single-crystal diffraction.

Solid-state nuclear magnetic resonance (Hendrichs and Hewitt, 1980) and Mössbauer spectroscopy (Berry, 1990; see MOSSBAUER SPECTROMETRY) can sometimes provide additional useful structural information, especially for non-single-crystal materials.

PRINCIPLES OF X-RAY CRYSTALLOGRAPHY

When a material is placed in an X-ray beam, the periodically varying electric field of the X-ray accelerates the electrons into periodic motion; each in turn emits an electromagnetic wave with a frequency essentially identical to that of the incident wave and with a definite phase relation to the source that we will take as identical to that of the incident wave. To the approximation usually employed, all electrons are assumed to scatter with the same amplitude and phase and to do so independently of all other electrons. (The energy of the X ray, approximately 20 keV for Mo Kα, is usually large in comparison with the binding energy of the electron, except for the inner electrons of the heavier atoms. It should be noted, however, that the anomalous behavior of first shell, or first- and second-shell electrons in heavy elements, can provide a useful technique for phase determination in some cases and has proven to be especially valuable in protein crystallography.)

Interference can occur from electrons occupying different spatial positions, providing the spatial difference is comparable to the wavelength being employed. Interference effects can occur within an atom, as is shown in Figure 1, where the variation of scattering with angle for

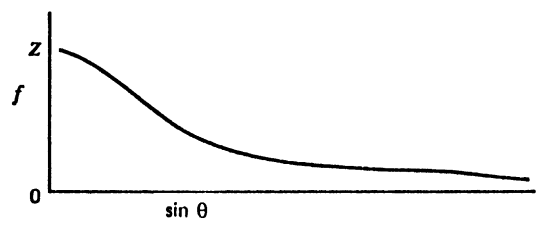


Figure 1. Decrease of amplitude of scattering from an atom of atomic number Z as a function of scattering angle θ. The function f is the atomic scattering factor. All atoms show similar shaped functions, the rate of decrease being somewhat less for atoms of larger Z; their inner electrons are closer to the nucleus.

a typical atom is plotted. These interference effects cause f, the atom's amplitude of scattering, to decrease as the scattering angle increases. Interference effects can also occur between atoms, and it is the latter that permits one to deduce the positions of the atoms in the unit cell, i.e., the crystal structure.

It is mathematically convenient to use a complex exponential description for the X-ray wave,

$$E = E_0 \exp[2\pi i(\nu t - x/\lambda + \delta)] \tag{1}$$

where x is the distance, λ is the wavelength, ν is the frequency, t is the time, and δ is a phase factor for the wave at $x = 0$ and $t = 0$. Anticipating that we will be referring to intensities (i.e., EE^*, E^* being the complex conjugate of E) and only be concerned with effects due to different spatial positions, we will use the simpler form

$$E = E_0 \exp(-2\pi i x/\lambda) \tag{2}$$

For two electrons displaced in one dimension the intensity of scattering would be given by

$$\begin{aligned} I &= EE^* = |E_1 + E_2|^2 \\ &= |E_{0_1} \exp(-2\pi i x_1/\lambda) + E_{0_2} \exp(-2\pi i x_2/\lambda)|^2 \end{aligned} \tag{3}$$

Furthermore, since E_{0_1} would be expected to equal E_{0_2},

$$\begin{aligned} I &= |E_0[\exp(-2\pi i x_1/\lambda) + \exp(-2\pi i x_2/\lambda]|^2 \\ &= E_0^2|[\{1 + \exp[-2\pi i(x_2 - x_1)/\lambda]\}\exp(-2\pi i x_1/\lambda)]|^2 \\ &= E_0^2|\{1 + \exp[-2\pi i(x_2 - x_1)/\lambda]\}|^2 \end{aligned} \tag{4}$$

Generalizing to three dimensions, consider a wave (Fig. 2) incident upon an electron at the origin and a second electron at $\mathbf{r}$. Then the difference in distance would be given by $\mathbf{r} \cdot (\mathbf{s}_0 - \mathbf{s})$, where $\mathbf{s}_0$ and $\mathbf{s}$ represent directions of the incident and scattered rays and $\mathbf{r}$ is the distance between electrons

$$I = E_0^2|1 + \exp[2\pi i \mathbf{r} \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda]|^2 \tag{5}$$

or for n electrons

$$I = E_0^2\left|\sum_{j=0}^{n-1} \exp[2\pi i \mathbf{r}_j \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda]\right|^2 \tag{6}$$

Since we will only require relative intensities, we define

$$F = \sum_{j=0}^{n-1} \exp[2\pi i \mathbf{r}_j \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda] \tag{7}$$

where F is termed the structure factor and

$$I \propto |F|^2 \tag{8}$$

The structure factor expression can be further simplified by replacing the summation by an integral,

$$F = \int_V \rho(\mathbf{r}) \exp[2\pi i \mathbf{r} \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda] dV \tag{9}$$

where $\rho(\mathbf{r})$ is the electron density. If the assumption is made that the electron density is approximated by the sum of atom electron densities, then

$$F = \sum_{j=0}^{N} f_j \exp[2\pi i \mathbf{r}_j \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda)] \tag{10}$$

where f_j is the atomic scattering factor, i.e., the scattering that would be produced by an isolated atom of that atomic number, convoluted with effects due to thermal motion.

To simplify this equation further, we now specify that the sample is a single crystal, and $\mathbf{r}_j = (x_j + p)\mathbf{a}_1 + (y_j + m)\mathbf{a}_2 + (z_j + n)\mathbf{a}_3$. The vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ describe the repeating unit (the unit cell), x_j, y_j, z_j are the fractional coordinates of atom j in the cell, and p, m, and n are integers. Since $\mathbf{r}_j \bullet (\mathbf{s} - \mathbf{s}_0)/\lambda$ must be dimensionless, the $(\mathbf{s} - \mathbf{s}_0)/\lambda$ quantity must be of dimension reciprocal length and can be represented by

$$(\mathbf{s} - \mathbf{s}_0)/\lambda = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3 \tag{11}$$

where $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ are termed the reciprocal cell vectors. They are defined such that $\mathbf{a}_i \cdot \mathbf{b}_i = 1$ and $\mathbf{a}_i \cdot \mathbf{b}_j = 0$. (It should be noted that another common convention is to use $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ to represent unit dimensions in direct space and $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ to represent the reciprocal space quantities.)

By substituting into the structure factor expression and assuming at least a few hundred cells in each direction, it can be shown (Lipscomb and Jacobson, 1990) that the restriction of p, m, and n to integer values yields nonzero diffraction maxima only when h, k, and l are integer. Hence the structure factor for single-crystal diffraction is usually written as

$$F_{hkl} = s\sum_{j=1}^{N} f_j \exp[2\pi i(hx_j + ky_j + lz_j)] \tag{12}$$

where the sum is now over only the atoms in the repeating unit and the other terms (those involving p, m, and n) contribute to a constant multiplying factor, s.

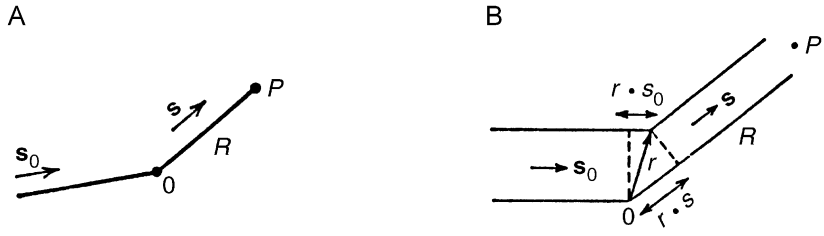


Figure 2. (**A**) Unit vectors $\mathbf{s}_0$ and $\mathbf{s}$ represent directions of the incident and scattered waves, respectively. (**B**) Scattering from a point at the origin and a second point at $\mathbf{r}$. The point of observation P is very far away compared with the distance $\mathbf{r}$.

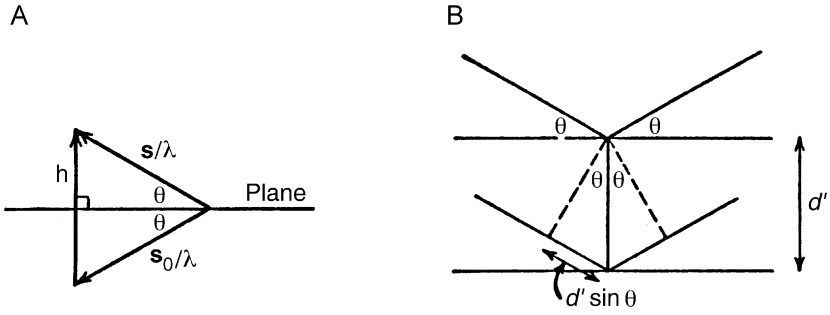


Figure 3. (**A**) Graphical representation of $\mathbf{s}/\lambda - \mathbf{s}_0/\lambda = \mathbf{h}$. (**B**) Path difference between successive planes a distance d' apart is $2d''$ sin θ, which yields a maximum when equal to $n\lambda$.

We now define a reciprocal lattice vector $\mathbf{h}$ as $\mathbf{h} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3$.
Also from Equation 11,

$$\mathbf{h}\lambda = \mathbf{s} - \mathbf{s}_0 \tag{13}$$

and this is termed the Laue equation.

The diffraction pattern from a single crystal therefore can be viewed as yielding a reciprocal lattice pattern (h, k, l all integer) that is weighted by $|F_{hkl}|^2$.

It can be readily shown that $\mathbf{h}$ is a vector normal to a plane passing through the unit cell with intercepts $\mathbf{a}_1/h$, $\mathbf{a}_2/k$, and $\mathbf{a}_3/l$. The integers h, k, and l are termed crystallographic indices (Miller indices if they are relatively prime numbers; see SYMMETRY IN CRYSTALLOGRAPHY). If 2θ is defined as the angle between $\mathbf{s}$ and $\mathbf{s}_0$, then the scalar equivalent of the Laue equation is

$$\lambda = 2d\sin\theta \tag{14}$$

i.e., Bragg's law, where $|\mathbf{h}| = 1/d$, the distance between planes. (If Miller indices are used, a factor of n, the order of the diffraction, is included on the left side of Equation 14.) A convenient graphical representation of the Laue equation is given in Figure 3 and alternately in what is termed the Ewald construction in Figure 4. These are

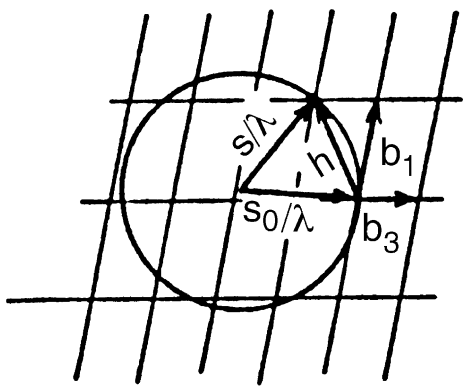


Figure 4. Ewald construction. Assume the reciprocal lattice net represents the $h0l$ zone of a monoclinic crystal. Then the $\mathbf{a}_2$ axis [010] is perpendicular to the plane of the paper. As the crystal is rotated about this axis, various points of the reciprocal lattice cross the circle of reflection; as they do so, the equation $\mathbf{s}/\lambda - \mathbf{s}_0/\lambda = \mathbf{h}$ is satisfied, and a diffracted beam occurs. Diffraction by the [10 1] plane is illustrated.

especially useful when discussing diffractometers and other devices for the collection of X-ray data. Diffraction maxima will only occur when a reciprocal lattice point intersects the sphere of diffraction.

The scattering from an isolated atom represents a combination of the scattering of an atom at rest convoluted with the effects due to thermal motion. These can be separated, and for an atom undergoing isotropic motion, a thermal factor expression

$$T_j = \exp(-B_j \sin^2\theta/\lambda^2) \tag{15}$$

can be used, where $B_j = 8\pi^2\mu^2$, μ^2 being the mean-square amplitude of vibration.

For anisotropic motion, we use

$$T_j = \exp[-2\pi^2(U_{11}h^2\mathbf{b}_1^2 + U_{22}k^2\mathbf{b}_2^2 + U_{33}l^2\mathbf{b}_3^2 + 2U_{12}hk\mathbf{b}_1 \bullet \mathbf{b}_2 + 2U_{13}hl\mathbf{b}_1 \bullet \mathbf{b}_3 + 2U_{23}kl\mathbf{b}_2 \bullet \mathbf{b}_3)] \tag{16a}$$

or the analogous expression

$$T_j = \exp[-(\beta_{11}h^2 + \beta_{22}k^2 + \beta_{33}l^2 + 2\beta_{12}hk + 2\beta_{13}hl + 2\beta_{23}kl)] \tag{16b}$$

The U's are thermal parameters expressed in terms of mean-square amplitudes in angstroms, while the β's are the associated quantities without units.
The structure factor is then written as

$$F_{\mathrm{hkl}} = s\sum_{j=1}^{N} f_j \exp[2\pi i(hx_j + ky_j + lz_j)]T_j \tag{17}$$

As we have seen, one form of the structure factor expression (Equation 9) involves $\rho(\mathbf{r})$, the electron density function, in a Fourier series. Therefore, it would be expected that there exists an inverse Fourier series in which the electron density is expressed in terms of structure factor quantities. Indeed, the electron density function for the cell can be written as

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} F_{hkl}\exp[-2\pi i(hx + ky + lz)] \tag{18}$$

In this expression V is the volume of the cell and the triple summation is in theory from $-\infty$ to ∞ and in practice over all the structure factors.

This series can be written in a number of equivalent ways. The structure factor can be expressed in terms of the real and imaginary components A and B as

$$F_{hkl} = A_{hkl} + iB_{hkl} \tag{19}$$

$$= s\sum_j f_j \cos 2\pi(hx_j + ky_j + lz_j)T_j + is\sum_j f_j \sin 2\pi(hx_j + ky_j + lz_j)T_j \tag{20}$$

and

$$\rho(xyz) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}[A_{hkl}\cos 2\pi(hx + ky + lz) + B_{hkl}\sin 2\pi(hx + ky + lz)] \tag{21}$$

Another form for the structure factor is

$$F_{hkl} = |F_{hkl}|\exp(2\pi i\alpha_{hkl}) \tag{22}$$

Then

$$\alpha_{hkl} = \tan^{-1}(B_{hkl}/A_{hkl}) \tag{23}$$

and

$$|F_{hkl}| = (A_{hkl}^2 + B_{hkl}^2)^{1/2} \tag{24}$$

The quantity α_{hkl} is termed the crystallographic phase angle. It could also be noted that, using $\bar{h}$ to designate $-h$

$$A_{\bar{h}\bar{k}\bar{l}} = A_{hkl} \tag{25}$$

and

$$B_{\bar{h}\bar{k}\bar{l}} = -B_{hkl} \tag{26}$$

if the atomic scattering factor f_j is real. Also,

$$|F_{hkl}|^2 = |F_{\bar{h}\bar{k}\bar{l}}|^2 \tag{27}$$

Equation 27 is termed Friedel's law and as noted holds as long as f_j is real.

Equation 18 implies that the electron density for the cell can be obtained from the structure factors; a knowledge of the electron density would enable us to deduce the location and types of atoms making up the cell. The magnitude of the structure factor can be obtained from the intensities (Equation 8); however, the phase is not measured, and this gives rise to what is termed the phase problem in crystallography. Considerable effort has been devoted over the last half century to developing reliable methods to deduce these phases. One of the most widely used of the current methods is that developed in large part by J. Karle and H. Hauptman (Hauptman and Karle, 1953; Karle and Hauptman, 1956; Karle and Karle, 1966), which uses a statistical and probability approach to extract phases directly from the magnitudes of the structure factors. We will give a brief discussion of the background of such "direct methods" as well as a discussion of the heavy-atom method, another commonly employed approach. It should be noted, however, that a number of other techniques could be used if these methods fail to yield a good trial model; these will not be discussed here due to space limitations.

Symmetry in Crystals

Symmetry plays an important role in the determination of crystal structures; only those atomic positional and thermal parameters that are symmetry independent need to be determined. In fact, if the proper symmetry is not recognized and attempts are made to refine parameters that are symmetry related, correlation effects occur and unreliable refinement and even divergence can result.

What restrictions are there on the symmetry elements and their combinations that can be present in a crystal? (We will confine our discussion to crystals as generated by a single tiling pattern and exclude cases such as the quasi-crystal.) Consider a lattice vector **a** operated on by a rotation axis perpendicular to the plane of the paper, as shown in Figure 5. Rotation by α or $-\alpha$ will move the head of the vector to two other lattice points, and these can be connected by a vector **b**. Since **b** must be parallel to **a**, one can relate their lengths by $b{=}ma$, where m is an integer, or by $b = 2a\cos\alpha$. Thus $m = 2\cos\alpha$ and

$$\alpha = \cos^{-1}(m/2) \tag{28}$$

Obviously the only allowed values for m are $0, \pm 1, \pm 2$, which yield 0, 60°, 90°, 120°, 180°, 240°, 270°, and 300° as possible values for the rotation angle α. This in turn implies rotation axes of order 1, 2, 3, 4, or 6. No other orders are compatible with the repeating character of the lattice. A lattice can also have inversion symmetry, as would be implied by Friedel's law (Equation 27). Coupling the inversion operation with a rotation axis produces the rotatory inversion axes: 1, 2, 3, 4, and 6. (These operations can be viewed as a rotation followed by an inversion. The 1 is just the inversion operation and 2 the mirror operation.)

Various combinations of these symmetry elements can now be made. In doing so, 32 different point groups are produced (see SYMMETRY IN CRYSTALLOGRAPHY). Since in a lattice description the repeating unit is represented by a point and the lattice has inversion symmetry, those point

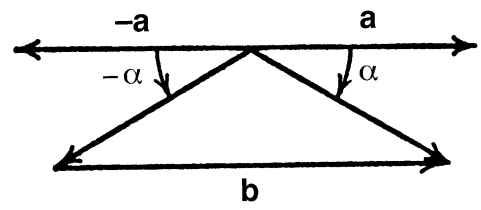


Figure 5. Rotation axis $A(\alpha)$ at a lattice point. Note that the vector **b** is parallel to **a**.

Table 1. Fourteen Bravais Lattices

Symmetry at a Lattice Point	Name	Lattice Points Per Unit Cell	Elementary Cell Relations
$C_i - 1$	Triclinic	1	
$C_{2h} - 2/2m$	Primitive monoclinic	1	$\alpha_1 = \alpha_3 = 90^\circ$
	End-centered monoclinic C	2	
D_{2h}	Primitive orthorhombic	1	$\alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$
	End-centered C, A, B	2	
	Face-centered F	4	
	Body-centered I	2	
$D_{4h} - 4/m\,mm$	Primitive tetragonal (equivalent to end-centered)	1	$\alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$ $a_1 = a_2$
	Body-centered (equivalent to face-centered) I	2	
$D_{3d} - 3m$	Trigonal	1	$\alpha_1 = \alpha_2 = \alpha_3$ $a_1 = a_2 = a_3$ (rhombohedral axes, for hexagonal description, see below)
$D_{6h} - 6/m\,mm$	Heagonal	1	$\alpha_1 = \alpha_2 = 90^\circ$ $\alpha_3 = 120^\circ$ $a_1 = a_2$
$O_h - m3m$	Primitive cubic	1	$\alpha_1 = \alpha_2 = \alpha_3 = 90^\circ$
	Face-centered cubic F	4	$a_1 = a_2 = a_3$
	Body-centered cubic I	2	

groups that differ only by an inversion center can be grouped together. This produces the seven crystal systems given in Table 1. For a number of crystal systems, in addition to the primitive cell (one lattice point per cell), there are also centered cell possibilities. (Such cells are not allowed if redrawing the centered cell to obtain a primitive one can be done without loss of symmetry consistent with the crystal system. Also, if redrawing or renaming axes converts one type to another, only one of these is included.) The notation used to describe centered cells is as follows: A—centered on the $\mathbf{a}_2$, $\mathbf{a}_3$ face; B—centered on the $\mathbf{a}_1$, $\mathbf{a}_3$ face; C—centered on the $\mathbf{a}_1$, $\mathbf{a}_2$ face; I—body centered; and F—face centered.

As noted in the table, trigonal cells can be reported using a primitive rhombohedral cell or as a hexagonal cell. If the latter, then a triply primitive R-centered cell is used. R centering has lattice points at (1/3, 2/3, 2/3) and (2/3, 1/3, 1/3) as well as at (0, 0, 0).

The next task is to convert these 32 point groups into the related space groups. This can be accomplished by generating not only space groups having these point group symmetries but also groups in which the point group operation is replaced by an "isomorphous" operation that has a translational element associated with it. However, this must be done in such a way as to preserve the multiplicative properties of the group.

The symmetry operation isomorphous to the rotation axis is the screw axis. They are designated by N_m, where the rotation is through an angle $2\pi/N$ and the translation is m/N, in terms of fractions of the cell, and is parallel to the rotation axis. Thus 2_1 is a 180° rotation coupled with a 1/2 cell translation, 4_1 is a 90° rotation coupled with 1/4 cell translation, and 6_2 represents a 60° rotation coupled with a 1/3 cell translation. Note that $2_1 \cdot 2_1$ yields a cell translation along the rotation axis direction and hence produces a result isomorphous to the identity operation; performing a 6_2 six times yields a translation of two cell lengths.

Mirror planes can be coupled with a translation and converted to glide planes. The following notation is used: a glide, translates 1/2 in $\mathbf{a}_1$; b glide, translates 1/2 in $\mathbf{a}_2$; c glide, translates 1/2 in $\mathbf{a}_3$; n glide, translates 1/2 along the face diagonal; and d glide, translates 1/4 along the face or body diagonal.

Let us digress for a moment to continue to discuss notation in general. When referring to space groups, the Hermann-Mauguin notation is the one most commonly used (see SYMMETRY IN CRYSTALLOGRAPHY). The symbol given first indicates the cell type, followed by the symmetry in one or more directions as dictated by the crystal system.

For the triclinic system, the only possible space groups are $P1$ and $P\bar{1}$.

For the monoclinic system, the cell type, P or C, is followed by the symmetry in the b direction. (The b direction is typically taken as the unique direction for the monoclinic system, although one can occasionally encounter c-unique monoclinic descriptions in the older literature.) If a rotation axis and the normal to a mirror plane are in the same direction, then a slash (/) is used to designate that the two operations are in the same direction. For the point group C_{2h}, the isomorphous space groups in the monoclinic system include $P2/m$, $P2_1/m$, $P2/c$, $P2_1/c$, $C2/m$ and $C2/c$.

In the orthorhombic system, the lattice type is followed by the symmetry along the **a**, **b**, and **c** directions. The space groups $P222$, $Pca2$, $P2_12_12_1$, $Pmmm$, $Pbca$, $Ama2$,

*Fdd*2, and *Cmcm* are all examples of orthorhombic space groups.

The notation for the higher symmetry crystal systems follows in the same manner. The symmetry along the symmetry axis of order greater than 2 is given first followed by symmetry in the other directions as dictated by the point group. For tetragonal and hexagonal systems, the symmetry along the **c** direction is given, followed by that for **a** and then for the **ab** diagonal, when isomorphous with a *D*-type point group. For cubic space groups the order is symmetry along the cell edge, followed by that along the body diagonal, followed by the symmetry along the face diagonal, where appropriate.

A space group operation can be represented by

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \mathbf{R}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \mathbf{t} \tag{29}$$

where $\mathbf{R}$ is a 3×3 matrix corresponding to the point group operation and $\mathbf{t}$ is a column vector containing the three components of the associated translation. As noted above, group properties require that operations in the space group combine in the same fashion as those in the isomorphous point group; this in turn usually dictates the positions in the cell of the various symmetry elements making up the space group. Consider, for example, the space group $P2_1/c$ in the monoclinic system. This commonly occurring space group is derived from the C_{2h} point group. Since in the point group a twofold rotation followed by the reflection is equivalent to the inversion, in the space group the 2_1 operation followed by the *c* glide must be equivalent to the inversion. The usual convention is to place an inversion at the origin of the cell. For this to be so, the screw axis has to be placed at $z = 1/4$ and the *c* glide at $y = 1/4$. Therefore, in $P2_1/c$, the following relations hold:

$$x,y,z \xrightarrow{i} -x,-y,-z \tag{30a}$$

$$x,y,z \xrightarrow{2_1} -x,1/2+y,1/2-z \tag{30b}$$

$$x,y,z \xrightarrow{c} x,1/2-y,1/2+z \tag{30c}$$

and these four equivalent positions are termed the general positions in the space group. In many space groups it is also possible for atoms to reside in "special positions," locations in the cell where a point group operation leaves the position invariant. Obviously these can only occur for symmetry elements that contain no translational components. In $P2_1/c$, the only special position possible is associated with the inversion. The number of equivalent positions in this case is two, namely, 0, 0, 0 and 0, 1/2, 1/2 or any of the translationally related pairs.

By replacing any point group operation by any allowable isomorphous space operation, the 32 point groups yield 230 space groups. A complete listing of all the space groups with associated cell diagrams showing the placement of the symmetry elements and a listing of the general and special position coordinates can be found in Volume A, International Tables for Crystallography, 1983.

Examination of the symmetry of the diffraction pattern (the Laue symmetry) permits the determination of the crystal system. To further limit the possible space groups within the crystal system, one needs to determine those classes of reflections that are systematically absent (extinct).

Consider the *c*-glide plane as present in $P2_1/c$. For every atom at (x, y, z), there is another at $(x, 1/2-y, 1/2+z)$. The structure factor can then be written as

$$F_{hkl} = \sum_{j=1}^{N/2} f_j \left\{ \exp[2\pi i(hx_j + ky_j + lz_j)] + \exp\left[2\pi i\left(hx_j + k\left(\frac{1}{2} - y_j\right) + l\left(\frac{1}{2} - z_j\right)\right)\right]\right\} \tag{31}$$

For those reflections with $k = 0$,

$$F_{h0l} = \sum_{j=1}^{N/2} f_j \exp[2\pi i(hx_j + lz_j)][1 + (-1)^l] \tag{32}$$

Thus, for $h0l$ reflections, all intensities with l odd would be absent. Different symmetry elements give different patterns of missing reflections as long as the symmetry element has a translational component. Point group elements show no systematic missing reflections.

Table 2 lists the extinction conditions caused by various symmetry elements. In many cases an examination of the extinctions will uniquely determine the space group. In others, two possibilities may remain that differ only by the presence or absence of a center of symmetry. Statistical tests (Wilson, 1944) have been devised to detect the presence of a center of symmetry; one of the most reliable is that developed by Howells et al. (1950). It should be cautioned, however, that while such tests are usually reliable, one should not interpret their results as being 100% certain.

Further discussion of space group symmetry can be found in SYMMETRY IN CRYSTALLOGRAPHY as well as in the literature (International Tables for Crystallography, 1983; Ladd and Palmer, 1985; Stout and Jensen, 1989; Lipscomb and Jacobson, 1990).

Crystal Structure Refinement

After data have been collected and the space group determined, a first approximation to the crystal structure must then be found. (Methods to do so are described in the next section.) Such a model must then be refined to obtain a final model that best agrees with the experimental data. Although a Fourier series approach could be used (using electron density map or difference electron density map calculations), this approach has the disadvantage that the structural model could be affected by systematic errors due to series termination effects, related to truncation of data at the observation limit. Consequently, a least-squares refinement method is employed instead. Changes in the positional and thermal parameters are calculated that minimize the difference between $|\mathbf{F}_{hkl}^{\text{obs}}|$ and $|\mathbf{F}_{hkl}^{\text{calc}}|$, or alternatively $\mathbf{I}_{hkl}^{\text{obs}}$ and $\mathbf{I}_{hkl}^{\text{calc}}$.

Table 2. Extinctions Caused by Symmetry Elements[a]

Class of Reflection	Condition for Nonextinction (n, an integer)	Interpretation of Extinction	Symbol for Symmetry Element
hkl	$h+k+l=2n$	Body-centered lattice	I
	$h+k=2n$	C-centered lattice	C
	$h+l=2n$	B-centered lattice	B
	$k+l=2n$	A-centered lattice	A
	$h+k,h+l$	Face-centered lattice	F
	$h+l=2n$		
	$-h+k+l=3n$	Rhombohedral lattice indexed on hexagonal lattice	R
	$h+k+l=3n$	Hexagonal lattice indexed on rhombohedral lattice	H
Okl	$k=2n$	Glide plane $\perp a$, translation $b/2$	b
	$l=2n$	Glide plane $\perp a$, translation $c/2$	c
	$k+l=2n$	Glide plane $\perp a$, translation $(b+c)/2$	n
	$k+l=4n$	Glide plane $\perp a$, translation $(b+c)/4$	d
$h0l$	$h=2n$	Glide plane $\perp b$, translation $a/2$	a
	$l=2n$	Glide plane $\perp b$, translation $c/2$	c
	$h+l=2n$	Glide plane $\perp b$, translation $(a+c)/2$	n
	$h+k=4n$	Glide plane $\perp a$, translation $(a+b)/4$	d
hhl	$l=2n$	Glide plane $\perp (a-b)$, translation $c/2$	c
	$2h+l=2n$	Glide plane $\perp (a-b)$, translation $(a+b+c)/2$	n
	$2h+l=2n$	Glide plane $\perp (a-b)$, translation $(a+b+c)/4$	d
$h00$[b]	$h=2n$	Screw axis $\parallel a$, translation $a/2$	2_1, 4_2
	$h=4n$	Screw axis $\parallel a$, translation $a/4$	4_1, 4_3
$00l$	$l=2n$	Screw axis $\parallel c$, translation $c/2$	2_1, 4_2, 6_3
	$l=4n$	Screw axis $\parallel c$, translation $c/4$	4_1, 4_3
	$i=3n$	Screw axis $\parallel c$, translation $c/3$	3_1, 3_2, 6_2, 6_4
	$l=6n$	Screw axis $\parallel c$, translation $c/6$	6_1, 6_5
$hh0$	$h=2n$	Screw axis $\parallel (a+b)$, translation $(a+b)/2$	2_1

[a] Adapted from Buerger, 1942.
[b] Similarly for k and b translations.

It is common practice to refer to one portion of the data set as observed data and the other as unobserved. The terminology is a carry-over from the early days of crystallography when the data were measured from photographic film by visual matching to a standard scale. Intensities could be measured only to some lower limit, hence the "observed" and "unobserved" categories. Although measurements are now made with digital counters (scintillation detectors or area detectors), the distinction between these categories is still made; often $I > 3\sigma(I)$ or a similar condition involving $|F|$ is used for designating observed reflections. Such a distinction is still valuable when refining on $|F|$ (see below) and when discussing R factors.

Crystallographers typically refer to a residual index (R factor) as a gauge for assessing the validity of a structure. It is defined as

$$R = \frac{\sum_{\mathbf{h}} \| F_{\mathbf{h}}^{\text{obs}}| - |F_{\mathbf{h}}^{\text{obs}} \|}{\sum_{\mathbf{h}} |F_{\mathbf{h}}^{\text{obs}}|} \tag{33}$$

Atoms placed in incorrect positions usually produce $R > 0.40$. If positional parameters are correct, R's in the 0.20 to 0.30 range are common. Structures with refined isotropic thermal parameters typically give $R \sim 0.15$, and least-squares refinement of the anisotropic thermal parameters will reduce this to $\sim$0.05 or less for a well-behaved structure.

Positional and thermal parameters occur in trigonometric and exponential functions, respectively. To refine such nonlinear functions, a Taylor series expansion is used. If f is a nonlinear function of parameters $p_1, p_2, \ldots, p_n$, then f can be written as a Taylor series

$$f = f^0 + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial p_i}\right)_{p_j} \Delta p_i + \text{higher order terms} \tag{34}$$

where f^0 represents the value of the function evaluated with the initial parameters and the Δp_i's are the shifts in these parameters. In practice, the higher order terms are neglected; if Δ_j for observation j is defined as

$$\Delta_j = f_j - f_j^0 - \sum_{i=1}^{n} \left(\frac{\partial f}{\partial p_i}\right)_{p_j} \Delta p_i \tag{35}$$

then $\sum_j w_j \Delta_j^2$ can be minimized to obtain the set of best shifts of the parameters in a least squares sense. These in turn can be used to obtain a new set of f^0s, and the process repeated until all the shifts are smaller than some prescribed minimum. The weights w_j are chosen such as to be reciprocally related to the square of the standard deviation associated with that observation.

In the X-ray case, as noted above, f can be I_{hkl} or $|F_{hkl}|$. If the latter, then

$$\begin{aligned} |F_{hkl}^c| &= F_{hkl}^c \exp(-2\pi i \alpha_{hkl}) \\ &= A_{hkl}^c \cos\alpha_{hkl} + B_{hkl}^c \sin\alpha_{hkl} \end{aligned} \tag{36}$$

In this case, only observed reflections should be used to calculate parameter shifts since the least-squares method assumes a Gaussian distribution of errors in the observations. (For net intensities that are negative, i.e., where the total intensity is somewhat less than the background measurement, $|F^0_{hkl}|$ is not defined. After refinement, structure factors can still be calculated for these unobserved reflections.)

The set of least-squares equations can be obtained from these Δ_j equations as follows for m observations. Express

$$\begin{pmatrix} \frac{\partial f_1}{\partial p_1} & \frac{\partial f_1}{\partial p_2} & \frac{\partial f_1}{\partial p_3} & \cdots & \frac{\partial f_1}{\partial p_n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial p_1} & \frac{\partial f_m}{\partial p_2} & \frac{\partial f_m}{\partial p_3} & \cdots & \frac{\partial f_m}{\partial p_n} \end{pmatrix} \begin{pmatrix} \Delta p_1 \\ \Delta p_2 \\ \vdots \\ \Delta p_n \end{pmatrix} = \begin{pmatrix} f_1 - f_1^0 \\ f_2 - f_2^0 \\ \vdots \\ f_m - f_m^0 \end{pmatrix} \quad (37)$$

It is convenient to represent these matrices by $\mathbf{MP} = \mathbf{Y}$. Then multiplying both sides by the transpose of $\mathbf{M}$, $\mathbf{M^T}$, gives $\mathbf{M^T MP} = \mathbf{M^T Y}$. Let $\mathbf{A} = \mathbf{M^T M}$ and $\mathbf{B} = \mathbf{M^T Y}$. Since $\mathbf{A}$ is now a square matrix, $\mathbf{P} = \mathbf{A}^{-1}\mathbf{B}$. Least-squares methods not only give a final set of parameters but also provide

$$s^2(p_j) = a^{jj}\left(\frac{\sum_i w_i \Delta j^2}{m - n}\right) \quad (38)$$

where s is the standard deviation associated with the parameter p_j, a^{ij} is the jth diagonal matrix element of the inverse matrix, m is the number of observations, and n is the number of parameters. The square root of the quantity in parentheses is sometimes referred to as the standard deviation of an observation of unit weight, or the goodness-of-fit parameter. If the weights are correct, i.e., if the errors in the data are strictly random and correctly estimated, and if the crystal structure is properly being modeled, then the value of this quantity should approximately equal unity.

PRACTICAL ASPECTS OF THE METHOD

In practice, the application of single-crystal X-ray diffraction methods for the determination of crystal and molecular structure can be subdivided into a number of tasks. The investigator must: (i) select and mount a suitable crystal; (ii) collect the diffracted intensities produced by the crystal; (iii) correct the data for various experimental effects; (iv) obtain an initial approximate model of the structure; (v) refine this model to obtain the best fit between the observed intensities and their calculated counterparts; and (vi) calculate derived results (e.g., bond distances and angles and stereographic drawings including thermal ellipsoid plots) from the atomic parameters associated with the final model.

For further discussion of the selection and mounting of a crystal see Sample Preparation, below; for further details on methods of data collection, see Data Collection Techniques.

The fiber or capillary holding the crystal is placed in a pin on a goniometer head that has x, y, z translations. (Preliminary investigation using film techniques—precession or Weissenberg—could be carried out at this stage, if desired, but we will not discuss such investigations here.) The goniometer head is then placed on the diffractometer and adjusted using the instrument's microscope so that the center of the crystal remains fixed during axial rotations.

The next task is to determine the cell dimensions, the orientation of the cell relative to the laboratory coordinate system, and the likely Bravais lattice and crystal system. This process is usually termed indexing.

To accomplish this, an initial survey of the low to mid angle (θ) regions of reciprocal space is carried out. This can be done via a rotation photograph using a Polaroid cassette—the user then inputs the x, y positions of the diffraction maxima and the diffractometer searches in the rotation angle—or by initiating a search procedure that varies a variety of the angles on a four-circle diffractometer. Any peaks found are centered. Peak profiles give an indication of crystal quality—one would expect Gaussian-like peaks from a crystal of good quality. Sometimes successful structure determinations can be carried out on crystals that give split peaks or have a significant side peak, but it is best to work with crystals giving well-shaped peaks if possible. With an area detector, reflections can be collected from a few frames at different rotation angles. An indexing program is then employed that uses this information to determine the most probable cell and crystal system. (Although these indexing programs are usually quite reliable, the user must remember that the information is being obtained from a reasonably small number of reflections and some caution is advisable. The user should verify that all or almost all of the reflections were satisfactorily assigned indices by the program.)

Computer programs (Jacobson, 1976, 1997) typically produce a reduced cell as a result of indexing. The reduced cell is one that minimizes the lengths of $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ that describe the repeating unit (and hence give angles as close to 90° as possible) (Niggli, 1928; Lawton and Jacobson, 1965; Santoro and Mighell, 1970). It also orders the axes according to a convention such as $|\mathbf{a}_1| \leq |\mathbf{a}_2| \leq |\mathbf{a}_3|$. This is followed by a Bravais lattice determination in which characteristic length and angle relationships are used to predict the presence of centered cells and the axes are rearranged to be consistent with the likely crystal system. The next step would then be to measure intensities of some sets of reflections that would be predicted to be equal by the Laue group symmetry (to obtain additional experimental verification as to the actual Laue group of the crystal or a possible subgroup).

Once the likely Laue symmetry has been determined, the data collection can begin. This author usually prefers to collect more than just a unique data set; the subsequent data averaging gives a good estimate of data quality. Data are typically collected to a 2θ maximum from 50 to 60° with Mo K_α radiation, the maximum chosen depending on the crystal quality and the amount of thermal motion associated with the atoms. (Greater thermal motion means weaker average intensities as 2θ increases.)

Once data collection has finished, the data will need to be corrected for geometric effects such as the Lorentz effect (the velocity with which the reciprocal lattice point moves

through the sphere of reflection) and the polarization effect; the X-ray beam becomes partially polarized in the course of the diffraction process. Usually data will also have to be corrected for absorption since as the beam traverses the crystal, absorption occurs, as given by the relation $e^{-\mu t}$, where μ is the absorption coefficient and t is the path length. "Analytical" absorption correction can be made if the orientations of crystal faces and their distances from the center of the crystal are carefully measured. Often empirical absorption corrections are employed instead. For data collected on a four-circle diffractometer, a series of psi scans can be used for this purpose. These are scans at different ϕ angle positions for reflections close to $\chi = 90°$. (For a reflection at $\chi = 90°$, diffraction is ϕ independent and any variation in observed diffraction can be attributable to effects due to absorption.) For data collected on an area detector, computer programs determine the parameters associated with an analytic representation of the absorption surface using the observed differences found between the intensities of symmetry-related reflections. Corrections can also be made for a decrease in scattering caused by crystal decomposition effects, if necessary. This composite set of corrections is often referred to as data reduction.

Following data reduction, the data can be averaged and then the probable space group or groups found by an examination of systematic absences in the data. Statistical tests, such as the Howells, Philips, and Rogers tests for the possible presence of a center of symmetry, can be carried out if more than one space group is indicated.

Next the investigator typically seeks an initial trial model using a direct method program, or if a heavy-atom approach is appropriate, the analysis of peaks between symmetry-related atoms found in the Patterson calculation might be used as discussed below. An initial electron density map would then be calculated. The bond distances and bond angles that are associated with these atom position possibilities and the residual index often indicate whether a valid solution has been found. If so, the inclusion of additional atoms found on electron density or difference electron density maps and subsequent refinement usually leads to a final structural model. Final bond distances, angles, and their standard deviations can be produced along with a plot of the atom positions (e.g., an ORTEP drawing); see Data Analysis and Initial Interpretations, Derived Results.

If a valid initial model is not obtained, the number of phase set trials might be changed if using a direct method program. Different Patterson peaks could be tried. Other space group or crystal system possibilities could be investigated. Other structure solution methods beyond the scope of this chapter might have to be employed to deduce an appropriate trial model. Fortunately, if the data are of good quality, most structures can be solved using the standard computer software available with commercial systems.

Data Collection Techniques

Film has long been used to record X-ray diffraction intensities. However, film has a number of disadvantages; it is relatively insensitive, especially for the shorter X-ray wavelengths, and does not have a large dynamic range. Gas-filled proportional counters can be used to detect X rays. These detectors can attain high efficiencies for longer wavelength radiations; with the use of a center wire of high resistance, they can be used to obtain positional information as well. The curved one-dimensional versions are used for powder diffraction studies, and the two-dimensional counterparts form the basis for some of the area detectors used in biochemical applications. They operate at reasonably high voltages (800 to 1700 V), and the passage of an X-ray photon causes ionization and an avalanche of electrons to occur that migrate to the anode. The number of electrons is proportional to the number of ion pairs, which is approximately proportional to the energy of the incident X-ray photon.

The scintillation counter is composed of a NaI crystal that has been activated by the addition of some Tl. An absorbed X-ray produces a photoelectron and one or more Auger electrons; the latter energize the Tl sites, which in turn emit visible photons. The crystal is placed against a photomultiplier tube that converts the light into a pulse of electrons. The number of color centers energized are approximately dependent on the initial X-ray energy.

Scintillation detectors are typically used on four-circle diffractometers. The four-circle diffractometer is one of the most popular of the counter diffractometers. As the name suggests, this diffractometer has four shafts that can be moved usually in an independent fashion. Such a diffractometer is shown schematically in Figure 6. The 2θ axis is used to move the detector parallel to the basal plane, while the remaining three axes position the crystal (Hamilton, 1974). The ω axis rotates around an axis that is perpendicular to the basal plane, the χ axis rotates around an axis that is perpendicular to the face of the circle, and the ϕ axis rotates around the goniometer head mount. To collect data with such an instrument, these angles must be positioned such that the Laue equation, or its scalar equivalent, Bragg's law, is satisfied (see DYNAMICAL DIFFRACTION).

Electronic area detectors have become more routinely available in the last few years and represent a major new advance in detector technology. Two such new detectors are the CCD detector and the imaging plate detector. In a typical CCD system, the X-ray photons impinge on a phosphor such as gadolinium oxysulfide. The light

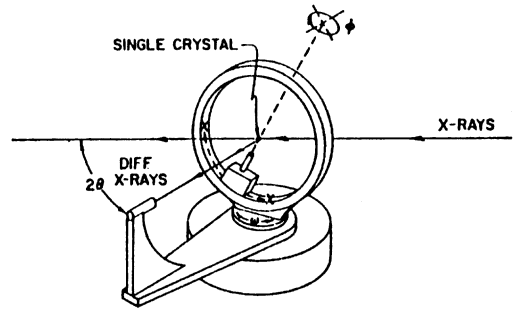


Figure 6. Four-circle diffractometer.

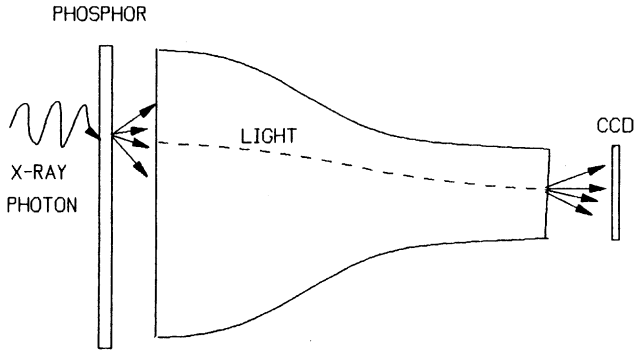


Figure 7. A CCD area detector. An X-ray photon arriving from the left is converted to light photons at the phosphor. Fiber optics conduct these to the CCD chip on the right.

produced is conducted to a CCD chip via fiber optics, as shown in Figure 7, yielding about eight electrons per X-ray photon. The typical system is maintained at $-50°C$ and has a dynamic range of 1 to $> 10^5$ photons per pixel. Readout times are typically 2 to 9 s depending on the noise suppression desired.

In a typical imaging plate system, the X-ray photons impinge on a phosphor (e.g., europium-doped barium halide) and produce color centers in the imaging plate. Following the exposure, these color centers are read by a laser and the light emitted is converted to an electronic signal using a photomultiplier. The laser action also erases the image so that the plate can be reused. In practice, usually one plate is being exposed while another is being read. The readout time is somewhat longer than with a CCD, typically 4 min or so per frame. However, the imaging plate has excellent dynamic range, 1 to $>10^6$ photons per pixel, and a considerably larger active area. It operates at room temperature.

As noted above, imaging plate systems usually provide a larger detector surface and slightly greater dynamic range than does the CCD detector, but frames from the latter system can be read much more rapidly than with imaging plate systems. Crystallographers doing small-molecule studies often tend to prefer CCD detectors, while imaging plate systems tend to be found in those laboratories doing structural studies of large biological molecules. Both types of area detectors permit the crystallographer to collect data in a manner almost independent of cell size. With the older scintillation counter system, a doubling of the cell size translates into a doubling of the time needed to collect the data set (since the number of reflections would be doubled). On the other hand, with an area detector, as long as the detector does not have to be moved further away from the source to increase the separation between reflections, the time for a data collection is essentially independent of cell size. Furthermore, if the crystal is rotated around one axis (e.g., the vertical axis) via a series of small oscillations and the associated frame data are stored to memory, most of the accessible data within a theta sphere would be collected. The remaining "blind region" could be collected by offsetting the crystal and performing a second limited rotation. Many commercial instruments use a four-circle diffractometer to provide maximum rotational capability. Due to the high sensitivity of the detector, a full data collection can be readily carried out with one of these systems in a fraction of a day for moderate-size unit cells.

METHOD AUTOMATION

Single-crystal X-ray methods involve the collection and processing of thousands to tens of thousands of X-ray intensities. Therefore, modern instrumentation comes with appropriate automation and related computer software to carry out this process in an efficient manner. Often, user input is confined to the selection of one of a few options at various stages in the data-collection process. The same is true of subsequent data processing, structure solution, and refinement.

DATA ANALYSIS AND INITIAL INTERPRETATION

Obtaining the Initial Model

There are a number of approaches that can be utilized to obtain an initial model of the structure appropriate for input into the refinement process. Many are specialized and tend to be used to "solve" particular kinds of structures, e.g., the use of isomorphous replacement methods for protein structure determination. Here we will primarily confine ourselves to two approaches: direct methods and the heavy-atom method.

Direct methods employ probability relations relying on the magnitudes of the structure factors and a few known signs or phases to determine enough other phases to reveal the structure from an electron density map calculation. The possibility that phase information could be extracted from the structure factor magnitudes has intrigued investigators throughout the history of crystallography. Karle and Hauptman (1956), building on the earlier work of Harker and Kasper (1948) and Sayre (1952), developed equations that have made these methods practical. A much more complete discussion of direct methods can be found in the Appendix.

One of the most important of the equations developed for the direct methods approach is the sigma-two equation:

$$E_{\mathbf{h}} \approx \langle E_{\mathbf{h}} E_{\mathbf{h}-\mathbf{k}} \rangle_{\mathbf{k}} \tag{39}$$

where E_{h} is derived from F_{h} and $\langle\rangle_{\mathbf{k}}$ represents the average over the $\mathbf{k}$ contributors. In practice, only those E's whose sign or phase has been determined are used in the average on the right. Since $|E_{\mathbf{h}}|$ is known, the equation is used to predict the most probable sign or place of the reflection.

The probability of $|E_{\mathbf{h}}|$ being positive is given by

$$P_+(\mathbf{h}) = \frac{1}{2} + \frac{1}{2}\tanh \sigma_3 \sigma_2^{-3/2} |E_{\mathbf{h}}| \sum_{\mathbf{k}} E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}} \tag{40}$$

For the non-centrosymmetric case

$$\tan\alpha_{\mathbf{h}} = \frac{\sum_{\mathbf{k}} |E_{\mathbf{h}}\| E_{\mathbf{h}-\mathbf{h}}| \sin(\alpha_{\mathbf{k}} + \alpha_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |E_{\mathbf{k}}\| E_{\mathbf{h}-\mathbf{k}}| \cos(\alpha_{\mathbf{k}} + \alpha_{\mathbf{h}-\mathbf{k}})} \tag{41}$$

Direct method computer programs use these equations and other related ones as described in the Appendix to attempt to obtain an approximate set of signs (or phases) directly from the magnitudes of the structure factors. These are then used to calculate an electron density map, which could reveal the approximate positions of most of the atoms in the structure. Further details on two of the more commonly used direct methods programs can be found in the Appendix.

Direct methods can be expected to be the most reliable when applied to structures in which most of the atoms are approximately equal in atomic number. In structures where one or a few atoms are present that have appreciably larger atomic numbers than the rest, the heavy-atom method is typically the approach of choice.

The basic assumption of the heavy-atom method is that these atoms alone provide good enough initial phasing of an electron density map that other atoms can be found and included, and the process repeated, until all atoms are located. See the Appendix for a more complete theoretical treatment of the heavy-atom method. The initial positions for the heavy atoms can come from direct methods or from an analysis of the Patterson function (see Appendix).

Derived Results

As a result of a crystal structure determination, one usually wishes to obtain, e.g., bond distances, bond angles, and least-squares planes, along with associated standard deviations. These can be determined from the least-squares results in a straightforward fashion.

Consider the bond distance between two atoms A and B. In vector notation, using $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ to represent the unit cell vectors.

$$\begin{aligned} d_{AB} &= |\mathbf{r}_B - \mathbf{r}_A| = [(\mathbf{r}_B - \mathbf{r}_A) \bullet (\mathbf{r}_B - \mathbf{r}_A)]^{1/2} \\ &= [(x_B - x_A)^2 a_1^2 + (y_B - y_A)^2 a_2^2 \\ &\quad + (z_B - z_A)^2 a_3^2 + 2(x_B - x_A)(y_B - y_A)\mathbf{a}_1 \bullet \mathbf{a}_2 \\ &\quad + 2(y_B - y_A)(z_B - z_A)\mathbf{a}_2 \bullet \mathbf{a}_3 \\ &\quad + 2(x_B - x_A)(z_B - z_A)\mathbf{a}_1 \bullet \mathbf{a}^3]^{1/2} \end{aligned} \tag{42}$$

For atoms *A*, *B*, and *C*, the bond angle A–B–C can be obtained from

$$\text{Angle} = \cos^{-1}\left\{\frac{(\mathbf{r}_A - \mathbf{r}_B) \bullet (\mathbf{r}_C - \mathbf{r}_B)}{d_{AB} \bullet d_{BC}}\right\} \tag{43}$$

Furthermore, the standard deviation in the bond distances can be obtained from

$$\sigma^2(d) = \sum_j \left(\frac{\partial d}{\partial p_j}\right)\sigma^2(p_j) \tag{44}$$

if the parameters are uncorrelated. The parameters in this case are the fractional coordinates (x_A, y_A, z_A, x_B, y_B, z_B) and their standard deviations ($\sigma(x_A)$, $\sigma(y_A)$, $\sigma(z_A)$, $\sigma(x_B)$, $\sigma(y_B)$, $\sigma(z_B)$) and the cell dimensions and their standard deviations. [Usually the major contributors to $\sigma(d)$ are from the errors in the fractional coordinates.] A similar expression can be written for the standard deviation in the bond angle. One can also extend these arguments to e.g., torsion angles and least-squares planes.

As noted above, anisotropic thermal parameters are usually obtained as part of a crystal structure refinement. One should keep in mind that *R* factors will tend to be lower on introduction of anisotropic temperature factors, but this could merely be an artifact due to the introduction of more parameters (five for each atom). One of the most popular ways of viewing the structure, including the thermal ellipsoids, is via plots using ORTEP [a Fortran thermal ellipsoid plot program for crystal structure illustration (Johnson, 1965)]. ORTEP plots provide a visual way of helping the investigator decide if these parameters are physically meaningful. Atoms would be expected to show less thermal motion along bonds and more motion perpendicular to them. Atoms, especially those in similar environments, would be expected to exhibit similar thermal ellipsoids. Very long or very short principal moments in the thermal ellipsoids could be due to disordering effects, poor corrections for effects such as absorption, or an incorrect choice of space groups, to mention just some of the more common causes.

Once data collection has finished, the data will need to be corrected for geometric effects such as the Lorentz effect (the velocity with which the reciprocal lattice point moves through the sphere of reflection) and the polarization effect; the X-ray beam becomes partially polarized in the course of the diffraction process. Usually data will also have to be corrected for absorption since as the beam traverses the crystal, absorption occurs, as given by the relation $e^{-\mu t}$, where μ is the absorption coefficient and t is the path length. "Analytical" absorption correction can be made if the orientations of crystal faces and their distances from the center of the crystal are carefully measured. Often empirical absorption corrections are employed instead. For data collected on a four-circle diffractometer, a series of psi scans can be used for this purpose. These are scans at different ϕ angle positions for reflections close to $\chi = 90^\circ$. (For a reflection at $\chi = 90^\circ$, diffraction is ϕ independent and any variation in observed diffraction can be attributable to effects due to absorption.) For data collected on an area detector, computer programs determine the parameters associated with an analytic representation of the absorption surface using the observed differences found between symmetry-related reflections. Corrections can also be made for a decrease in scattering caused by crystal decomposition effects, if necessary. This composite set of corrections is often referred to as data reduction.

Following data reduction, the data can be averaged, if desired, and then the probable space group or groups found by an examination of systematic absences in the data. Statistical tests such as the Howells, Philips, and Rogers tests can be carried out to check for the possible presence of a center of symmetry; if more than one space group is indicated.

Next the investigator typically seeks an initial trial model using a direct method program, or if a heavy-atom approach is appropriate, the analysis of peaks between symmetry-related atoms found in the Patterson approach might be used. An initial electron density map would then be calculated. The bond distances and bond angles that are associated with these atom position possibilities and sometimes the residual index often indicate whether a valid solution has been found. If so, the inclusion of additional atoms found on electron density or difference electron density maps and subsequent refinement usually leads to a final structural model. Final bond distances, angles, and their standard deviations can be produced along with a plot of the atom positions (e.g., an ORTEP drawing).

If a valid initial model is not obtained, the number of phase set trials might be changed if using a direct method program. Different Patterson peaks could be tried. Other space group or crystal system possibilities could be investigated. Other structure solution methods beyond the scope of this chapter might have to be employed to deduce an appropriate trial model. Fortunately, if the data are of good quality, most structures can be solved using the standard computer software available with commercial systems.

Example

The compound $(C_5H_5)(CO)_2$MnDBT, where DBT is dibenzothiophene, was synthesized in R. J. Angelici's group as part of their study of prototype hydrodesulfurization catalysts (Reynolds et al., 1999). It readily formed large, brown, parallelepiped-shaped crystals. A crystal of approximate dimensions $0.6 \times 0.6 \times 0.5$ mm was selected and mounted on a glass fiber. (This crystal was larger than what would normally be used, but it did not cleave readily and the absorption coefficient for this material is quite small, $\mu = 9.34\,\text{cm}^{-1}$. A larger than normal collimator was used.) It was then placed on a Bruker P4 diffractometer, a four-circle diffractometer equipped with a scintillation counter. Molybdenum K_α radiation was used as a source from a sealed tube target and a graphite monochromator ($\lambda = 0.71069$ Å).

A set of 42 reflections was found by a random-search procedure in the range $10 \leq 2\theta \leq 25°$. These were carefully centered, indexed, and found to fit a monoclinic cell with dimensions $a = 13.076(2)$, $b = 10.309(1)$, $c = 23.484(3)$ Å, and $\beta = 92.93(1)°$. Based on atomic volume estimates and a formula weight of 360.31 g/mol, a calculated density of 1.514 g/cm^3 was obtained with eight molecules per cell. Systematic absences of $h0l$: $h + l \neq 2n$ and $0k0$: $k \neq 2n$ indicated space group $P2_1/n$, with two molecules per asymmetric unit, which was later confirmed by successful refinement.

Data were collected at room temperature (23°C) using an ω-scan technique for reflections with $2\theta \leq 50°$. A hemisphere of data was collected ($+h, \pm k, \pm l$), yielding 12,361 reflections. Symmetry-related reflections were averaged (5556 reflections) and 4141 had $F > 3\sigma(F)$ and were considered observed. The data were corrected for Lorentz and polarization effects and were corrected for absorption using an empirical method based on psi scans. The transmission factors ranged from 0.44 to 0.57.

Table 3. Fractional Coordinates (Nonhydrogen Atoms) for One Molecule of $Mn(SC_{12}H_9)(CO)_2(C_5H_5)$

Atom	X	Y	Z	B_{eq}[a]
Mn1	.0356(1)	.0387(1)	.14887(6)	3.18(4)
S1	.0970(2)	−.1032(2)	.2155(1)	3.38(6)
C1	.4723(8)	.2495(9)	.2839(4)	3.2(3)
C2	.4098(8)	.1431(9)	.2962(4)	3.5(3)
C3	.4538(10)	.0169(10)	.2953(4)	4.7(3)
C4	.5588(11)	.0055(12)	.2818(5)	5.0(4)
C5	.6168(10)	.1102(11)	.2692(5)	4.7(3)
C6	.5742(9)	.2367(11)	.2701(4)	4.1(3)
C7	.7902(8)	.1841(10)	.8059(4)	3.4(3)
C8	.3040(8)	.1813(9)	.3090(4)	3.4(3)
C9	.2202(9)	.1044(11)	.3220(5)	4.5(3)
C10	.1274(10)	.1599(13)	.3303(5)	5.0(3)
C11	.6169(10)	.2017(14)	.8285(5)	5.5(4)
C12	.6989(8)	.1259(11)	.8158(5)	4.4(3)
C13	.0334(10)	.1004(11)	.0620(4)	4.6(3)
C14	.9012(10)	.0103(13)	.9320(5)	5.1(4l)
C15	.9462(10)	.1140(11)	.9141(4)	4.5(3)
C16	.0649(10)	.0685(11)	.9097(4)	4.3(3)
C17	.9348(10)	.0674(12)	.0753(4)	4.6(3)
C18	.9439(10)	.1038(11)	.1963(5)	4.4(3)
C19	.1292(10)	.1583(11)	.1670(4)	4.3(3)
O1	.8838(7)	.1452(9)	.2255(4)	6.6(3)
O2	.1901(7)	.2394(8)	.1754(4)	6.0(3)

[a] $B_{eq} = (8\pi^2/3)\sum_{i=1}^{3}\sum_{j=1}^{3} u_{ij} a_i^* a_j^* a_i \bullet a_j$.

The SHELX direct method procedure was used and yielded positions for the heavier atoms and possible positions for many of the carbon atoms. An electron density map was calculated using the phases from the heavier atoms; this map yielded further information on the lighter atom positions, which were then added. Hydrogen atoms were placed in calculated positions and all nonhydrogen atoms were refined anisotropically. The final cycle of full-matrix least-squares refinement converged with an unweighted residual $R = \sum \| F_0| - |F_c \| / \sum |F_0|$ of 3.3% and a weighted residual $R_w = [\sum w \ (|F_0| - |F_c|)^2 / \sum w F_0^2]^{1/2}$, where $w = 1/\sigma^2(F)$, of 3.9%. The maximum peak in the final electron density difference map was 0.26 e/Å^3. Atomic scattering factors were taken from Cromer and Weber (1974), including corrections for anomalous dispersion. Fractional atomic coordinates for one of the molecules in the asymmetric unit are given in Table 3, and their anisotropic temperature factors in Table 4.

The two independent molecules in the asymmetric unit have essentially identical geometries within standard deviations. A few selected bond distances and angles are given in Table 5 and Table 6, respectively. A thermal ellipsoid plot of one of the molecules is shown in Figure 8.

SAMPLE PREPARATION

The typical investigation starts obviously with the selection of a crystal. This is done by examining the material under a low- or moderate-power microscope. Samples that display sharp, well-defined faces are more likely to be single crystals. Crystals with average dimensions of a

Table 4. Anisotropic Thermal Parameters (Non-Hydrogen Atoms)* for One Molecule of $Mn(SC_{12}H_9)(CO)_2(C_5H_5)$

Atom	U_{11}	U_{22}	U_{33}	U_{23}	U_{13}	U_{12}
Mn1	5.11(10)	3.41(8)	3.56(7)	−.07(6)	.38(6)	.33(7)
S1	5.5(2)	3.5(1)	3.8(1)	−.1(1)	−.1(1)	.4(1)
C1	5.6(7)	4.1(5)	2.6(4)	.0(4)	−.3(4)	−.2(5)
C2	6.1(7)	3.5(5)	3.6(5)	−.6(4)	−.6(5)	−.1(5)
C3	8.8(9)	3.8(6)	5.1(6)	−.6(5)	−1.3(6)	.6(6)
C4	7.8(9)	5.8(7)	5.4(6)	−1.5(6)	.0(6)	1.5(7)
C5	6.6(8)	5.7(7)	5.7(6)	−1.8(6)	.0(6)	.7(6)
C6	5.2(7)	6.1(7)	4.2(5)	−.7(5)	.4(5)	−.3(6)
C7	4.6(6)	4.1(5)	4.1(5)	.1(4)	−.6(5)	.6(5)
C8	5.6(6)	3.6(5)	3.7(5)	−.1(4)	−.5(5)	−.6(5)
C9	6.5(8)	5.4(7)	5.2(6)	−.1(5)	−.5(6)	−1.8(6)
C10	5.6(8)	6.5(8)	6.9(7)	−.1(6)	.4(6)	−.7(7)
C11	5.9(8)	7.9(9)	6.8(7)	−.6(7)	−.4(6)	.4(7)
C12	4.6(7)	5.9(7)	6.2(7)	.4(6)	−.1(5)	−.6(6)
C13	7.9(9)	6.1(7)	3.6(5)	.7(5)	.4(5)	−.4(7)
C14	7.3(8)	8.0(9)	4.3(6)	−.8(6)	.9(6)	.0(7)
C15	7.5(8)	6.2(7)	3.4(50)	−1.1(5)	−.3(5)	.5(7)
C16	7.3(8)	5.3(6)	3.6(5)	−.5(5)	.0(5)	−.2(6)
C17	7.3(8)	6.1(7)	3.9(5)	.3(5)	−.4(5)	.2(6)
C18	6.4(8)	4.4(6)	5.7(6)	−.2(5)	−.9(6)	.2(6)
C19	6.7(8)	4.5(6)	5.2(6)	.4(5)	.8(6)	1.1(6)
O1	8.0(7)	9.1(7)	8.4(6)	−2.1(6)	3.1(5)	3.0(6)
O2	8.1(7)	5.1(5)	9.6(7)	−.6(5)	.2(5)	−1.7(5)

*U values are scaled by 100.

few tenths of a millimeter are often appropriate—somewhat larger for an organic material. (An approximately spherical crystal with diameter $2/\mu$, where μ is the absorption coefficient, would be optimum.) The longest dimension should not exceed the diameter of the X-ray beam, i.e., it is important that the crystal be completely bathed by the beam if observed intensities are to be compared to their calculated counterparts.

The crystal can be mounted on the end of a thin glass fiber or glass capillary if stable to the atmosphere using any of a variety of glues, epoxy, or even fingernail polish. If unstable, it can be sealed inside a thin-walled glass capillary. If the investigation is to be carried out at low temperature and the sample is unstable in air, it can be immersed in an oil drop and then quickly frozen; the oil should contain light elements and be selected such as to form a glassy solid when cooled.

Table 5. Selected Bond Distances (Å) for One Molecule of $Mn(SC_{12}H_9)(CO)_2(C_5H_5)$

Mn1–S1	2.255(1)	C4–C5	1.381(5)
Mn1–C13	2.116(3)	C5–C6	1.387(4)
Mn1–C14	2.139(3)	C7–C8	1.397(4)
Mn1–C15	2.151(3)	C7–C12	1.376(4)
Mn1–C16	2.148(3)	C8–C9	1.391(4)
Mn1–C17	2.123(3)	C9–C10	1.376(5)
Mn1–C18	1.764(3)	C10–C11	1.405(5)
Mn1–C19	1.766(3)	C11–C12	1.377(5)
S1–C1	1.770(3)	C13–C14	1.424(5)
S1–C7	1.769(3)	C13–C17	1.381(5)
C1–C2	1.398(4)	C14–C15	1.412(5)
C1–C6	1.380(4)	C15–C16	1.393(4)
C2–C3	1.405(4)	C16–C17	1.422(4)
C2–C8	1.459(4)	C18–O1	1.166(4)
C3–C4	1.365(5)	C19–O2	1.164(4)

SPECIMEN MODIFICATION

In most instances, little if any specimen modification occurs, especially when area detectors are employed for data collection, thereby reducing exposure times. In some cases, the X-ray beam can cause some decomposition of the crystal. Decomposition can often be reduced by collecting data at low temperatures.

PROBLEMS

Single-crystal X-ray diffraction has an advantage over many other methods of structure determination in that there are many more observations (the X-ray intensities) than parameters (atomic coordinates and thermal parameters). The residual index (R-value) therefore usually serves as a good guide to the general reliability of the solution. Thermal ellipsoids from ORTEP should also be examined, as well as the agreement between distances involving similar atoms, taking into account their standard deviations. Ideally, such distance differences should be less than 3 sigma, but may be somewhat greater if packing or

Table 6. Selected Bond Angles (deg) for One Molecule of $Mn(SC_{12}H_9)(CO)_2(C_5H_5)$

S1– Mn1 –C18	92.5(1)	C1–C2 –C8	112.3(2)
S1– Mn1 –C19	93.3(1)	S1–C7 –C8	112.2(2)
C18– Mn1 –C19	92.8(1)	C2 –C8 –C7	112.3(2)
Mn1– S1 –C1	113.5(1)	Mn1 –C18 –O1	177.6(3)
Mn1– S1 –C7	112.7(1)	Mn1 –C19 –O2	175.0(3)
S1–C1 –C2	112.1(2)		

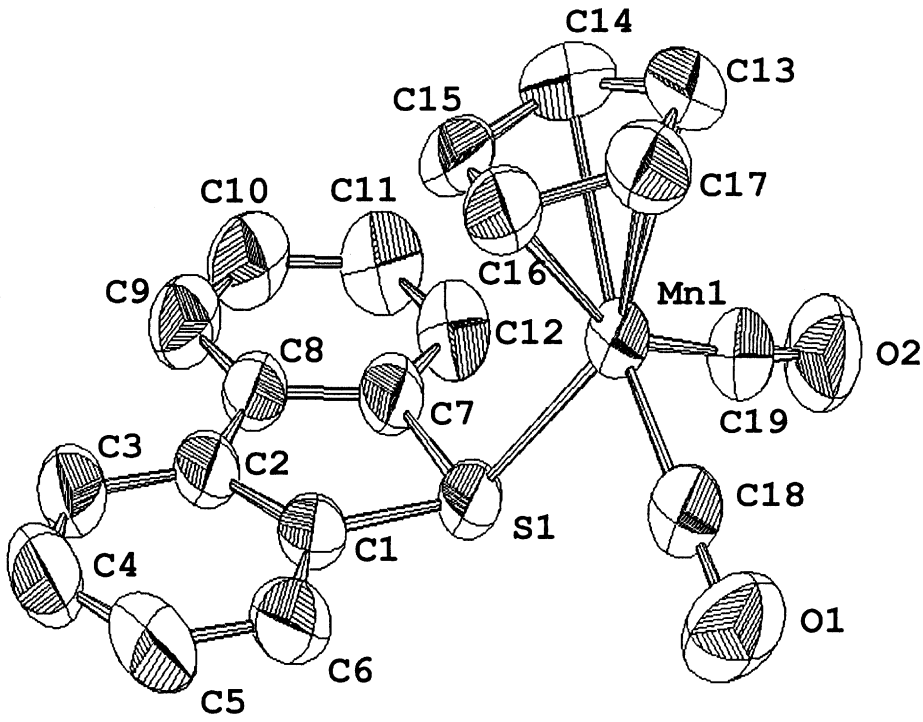


Figure 8. ORTEP of $Mn(SC_{12}Ha)(CO)_2(C_5H_5)$.

systematic error effects are present. If thermal ellipsoids are abnormal, this can be an indication of a problem with some aspect of the structure determination. This might be due to a poor correction for absorption of X rays as they pass through the crystal, a disordering of atoms in the structure, an extinction coefficient, or an incorrect choice of space group.

Approximations are inherent in any absorption correction, and therefore it is best to limit such effects by choosing a smaller crystal or using a different wavelength where the linear absorption coefficient would be less. The optimum thickness for a crystal is discussed above (see Sample Preparation).

If an atom is disordered over a couple of sites in the unit cell, an appropriate average of the electron density is observed and an elongation of the thermal ellipsoid may be observed if these sites are only slightly displaced. Sometimes it is possible to include multiple occupancies in the refinement to reasonably model such behavior with a constraint on the sum of the occupancies.

If a crystal is very perfect, extinction effects may occur, i.e., a scattered X-ray photon is scattered a second time. If this occurs in the same "mosaic block," it is termed primary extinction; if it occurs in another "mosaic block," it is termed secondary extinction. The effect is most noticeable in the largest intensities and usually manifests itself in a situation in which ten or so of the largest intensities are found to have calculated values that exceed the observed ones. Mathematical techniques exist to approximately correct for such extinction effects, and most crystallographic programs include such options.

Problems can be encountered if some dimension of the crystal exceeds the beam diameter. The crystal must be completely bathed in a uniform beam of X rays for all orientations of the crystal, if accurate intensities are to be predicted for any model.

The crystal should not decompose significantly (less than about 10%) in the X-ray beam. Some decomposition effects can be accounted for as long as they are limited.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, under contract W-7405-Eng-82.

LITERATURE CITED

Berry, F. J. 1990. Mössbauer Spectroscopy. *In* Physical Methods of Chemistry. Determination of Structural Features of Crystalline and Amorphous Solids, Vol. V, 2nd ed. (B. W. Rossiter and J. F. Hamilton, eds.). pp. 273–343. John Wiley & Sons, New York.

Buerger, M. J. 1942. X-Ray Crystallography. John Wiley & Sons, New York.

Buerger, M. J. 1959. Vector Space. John Wiley & Sons, New York.

Cochran, W. and Woolfson, M. M. 1955. The theory of sign relations between structure factors. *Acta Crystallogr.* 8:1–12.

Cromer, D. T. and Weber, J. T. 1974. International Tables for X-Ray Crystallography, Vol. IV. Kynoch Press, Birmingham, England, Tables 2.2 A, 2.3.1.

Germain, G., Main, P., and Woolfson, M. M. 1970. On the application of phase relationships to complex structures. II. Getting a good start. *Acta Crystallogr.* B26:274–285.

Hamilton, W. C. 1974. Angle settings for four-circle diffractometers. *In* International Tables for X-ray Crystallography, Vol. IV (J. A. Ibers and W. C. Hamilton, eds.). pp. 273–284. Kynoch Press, Birmingham, England.

Harker, D. 1936. The application of the three-dimensional Patterson method and the crystal structures of provstite, $Ag_3A_5S_3$ and pyrargyrite, Ag_3SbS_3. *J. Chem. Phys.* 4:381–390.

Harker, D. and Kasper, J. S. 1948. Phase of Fourier coefficients directly from crystal diffraction data. *Acta Crystallogr.* 1:70–75.

Hauptman, H. and Karle, J. 1953. Solution of the Phase Problem I. Centrosymmetric Crystal. American Crystallographic Association, Monograph No. 3. Polycrystal Book Service, Pittsburgh, Pa.

Heald, S. M. and Tranquada, J. M. 1990. X-Ray Absorption Spectroscopy: EXAFS and XANES. *In* Physical Methods of Chemistry. Determination of Structural Features of Crystalline and Amorphous Solids, Vol. V, 2nd ed. (B. W. Rossiter and J. F. Hamilton, eds.). pp. 189–272. John Wiley & Sons, New York.

Hendrichs, P. M. and Hewitt, J. M. 1980. Solid-State Nuclear Magnetic Resonance. *In* Physical Methods of Chemistry. Determination of Structural Features of Crystalline and Amorphous Solids, Vol. V, 2nd ed. (B. W. Rossiter and J. F. Hamilton, eds.). pp. 345–432. John Wiley & Sons, New York.

Howells, E. R., Phillips, D. C., and Rogers, D. 1950. The probability distribution of x-ray intensities. II. Experimental investigation and the x-ray detection of centres of symmetry. *Acta Crystallogr.* 3:210–214.

International Tables for Crystallography. 1983. Vol. A: Space-Group Symmetry. D. Reidel Publishing Company, Dordrecht, Holland.

Jacobson, R. A. 1976. A single-crystal automatic indexing procedure. *J. Appl. Crystallogr.* 9:115–118.

Jacobson, R. A. 1997. A cosine transform approach to indexing. *Z. Kristallogr.* 212:99–102.

Johnson, C. K. 1965. ORTEP: A Fortran Thermal-Ellipsoid Plot Program for the Crystal Structure Illustrations. Report ORNL-3794. Oak Ridge National Laboratory, Oak Ridge, TN.

Karle, J. and Hauptman, H. 1956. A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, $3P_12$,$3P_22$. *Acta Crystallogr.* 9:635–651.

Karle, J. and Karle, I. L. 1966. The symbolic addition procedure for phase determination for centrosymmetric and noncentrosymmetric crystals. *Acta Crystallogr.* 21:849–859.

Ladd, M. F. C. and Palmer, R. A. 1980. Theory and Practice of Direct Methods in Crystallography. Plenum, New York.

Ladd, M. F. C. and Palmer, R. A. 1985. Structure Determination by X-Ray Crystallography. Plenum Press, New York.

Lawton, S. L. and Jacobson, R. A. 1965. The Reduced Cell and Its Crystallographic Applications. USAEC Report IS-1141. Ames Laboratory, Ames, IA.

Lipscomb, W. N. and Jacobson, R. A. 1990. X-Ray Crystal Structure Analysis. *In* Physical Methods of Chemistry. Determination of Structural Features of Crystalline and Amorphous Solids, Vol. V, 2nd ed. (B. W. Rossiter and J. F. Hamilton, eds.). pp. 3–121. John Wiley & Sons, New York.

Majkrzak, C. F., Lehmann, M. S., and Cox, D. E. 1990. The Application of Neutron Diffraction Techniques to Structural Studies, Physical Methods of Chemistry. Determination of Structural Features of Crystalline and Amorphous Solids, Vol. V, 2nd ed. (B. W. Rossiter and J.F. Hamilton, eds.). pp. 123–187.

Niggli, P. 1928. Handbuch der Experimentalphysike 7 Part 1. Akademische Verlagsgesellschaft, Leipzig.

Patterson, A. L. 1935. A direct method for the determination of components of interatomic distances in crystals. Z. Kristallogr. A90:517–542; Tabulated data for the seventeen phase groups. *Z. Kristallogr.* A90:543–554.

Reynolds, M. A., Logsdon, B. C., Thomas, L. M., Jacobson, R. A., and Angelici, R. J. 1999. Transition metal complexes of Cr, Mo, W and Mn containing h^1(S)-2,5-dimethylthiophene, benzothiophene, and dibenzothiophene ligands. *Organometallics* 18: 4075–4081.

Richardson, Jr., J. W. and Jacobson, R. A. 1987. Computer-aided analysis of multi-solution Patterson superpositions. *In* Patterson and Pattersons (J. P. Glusker, B. K. Patterson, and M. Rossi, eds.). p. 311. Oxford University Press, Oxford.

Santoro, A. and Mighell, A. D. 1970. Determination of reduced cells. *Acta Crystallogr.* A26:124–127.

Sayre, D. 1952. The squaring method: A new method for phase determination. *Acta Crystallogr.* 5:60–65.

Sheldrick, G. M. 1990. Phase annealing in SHELX-90: Direct methods for larger structures. *Acta Crystallogr.* A46:467–473.

Stout, G. H. and Jensen, L. H. 1989. X-Ray Crystal Structure Determination. John Wiley & Sons, New York.

Wilson, A. J. C. 1944. The probability distribution of x-ray intensities. *Acta Crystallogr.* 2:318–321.

Woolfson, M. M. and Germain, G. 1970. On the application of phase relationships to complex structures. *Acta Crystallogr.* B24:91–96.

KEY REFERENCES

Lipscomb and Jacobson, 1990. See above.

Provides a much more detailed description of the kinematic theory of X-ray scattering and its application to single-crystal structure determination.

Ladd and Palmer, 1985. See above.

Another good text that introduces the methods of crystallography, yet does so without being too mathematical; also provides a number of problems following each chapter to facilitate subject comprehension.

Stout and Jensen, 1989. See above.

A good general text on X-ray diffraction and one that emphasizes many of the practical aspects of the technique.

APPENDIX

Direct Methods

There were a number of early attempts to use statistical approaches to develop relations that could yield the phases (or signs for the centrosymmetric case) of some of the reflections. Harker and Kasper (1948), for example, developed a set of inequalities based on the use of Cauchy's inequality. A different approach was suggested by Sayre (1952). Although later superceded by the work of Karle and Hauptman (1956), this work is still of some interest since it affords some physical insight into these methods.

Sayre (1952) noted that ρ^2 would be expected to look much like ρ, especially if the structure contains atoms that are all comparable in size. (Ideally, ρ should be nonnegative; it should contain peaks where atoms are and should equal zero elsewhere, and hence its square should

have the same characteristics with modified peak shape.) One can express ρ^2 as a Fourier series,

$$\rho^2(xyz) = \frac{1}{V}\sum_h \sum_k \sum_l G_{hkl} \exp[-2\pi i(hx + ky + lz)] \quad (45)$$

where the G's are the Fourier transform coefficients.

Since ρ^2 would be made up of "squared" atoms, G_{hkl} can be expressed as

$$G_{hkl} = \sum_{j=1}^{n} g_j \exp[2\pi i(hx_j + ky_j + lz_j)] \quad (46)$$

or

$$G_{hkl} \approx \frac{g}{f} F_{hkl} \quad (47)$$

if all atoms are assumed to be approximately equal. Therefore

$$\rho^2(xyz) = \frac{1}{V}\sum_h \sum_k \sum_l \frac{g}{f} F_{hkl} \exp[2\pi i(hx + ky + lz)] \quad (48)$$

but also

$$\begin{aligned}\rho^2(xyz) &= \frac{1}{V^2}\left(\sum_h \sum_k \sum_l F_{hkl} \exp[2\pi i(hx + ky + lz)]\right) \\ &\quad \times \left(\sum_H \sum_K \sum_L \frac{g}{f} F_{HKL} \exp[2\pi i(Hx + Ky + Lz)]\right) \\ &= \frac{1}{V^2}\sum_h \sum_k \sum_l \sum_H \sum_K \sum_L F_{hkl} \exp\{2\pi i[(h + H)x \\ &\quad + (k + K)y + (l + L)z]\} \end{aligned} \quad (49)$$

Comparing the two expressions for ρ^2 and letting $\mathbf{h} = \mathbf{k} + \mathbf{H}$ yield

$$\frac{g}{f} F_{\mathbf{h}} = \frac{1}{V}\sum_k F_{\mathbf{h}-\mathbf{k}} F_{\mathbf{k}} \quad (50)$$

Sayre's relation therefore implies that the structure factors are interrelated. The above relation would suggest that in a centrosymmetric structure, if $F_{\mathbf{k}}$ and $F_{\mathbf{h}-\mathbf{k}}$ are both large in magnitude and of known sign, the sign of $F_{\mathbf{h}}$ would likely be given by the product $F_{\mathbf{k}}F_{\mathbf{h}-\mathbf{k}}$ since this term will tend to predominate in the sum. Sayre's relation was primarily of academic interest since in practice few F's with known signs have large magnitudes, and the possibilities for obtaining new phases or signs are limited.

Karle and Hauptman (1956), using a different statistical approach, developed a similar equation along with a number of other useful statistical-based relations. They did so using E_{hkl}'s that are related to the structure factor by defining

$$|E_{hkl}|^2 = |F_{hkl}|^2/\varepsilon \sum f_j^2 \quad (51)$$

where f_j is the atomic scattering factor corrected for thermal effects and ε is a factor to account for degeneracy and is usually unity except for special projections.

For the equal atom case, a simple derivation can be given (Karle and Karle, 1966). Then

$$|E_{\mathbf{h}}|^2 = f^2 \left| \sum_j \exp(2\pi i \mathbf{h} \bullet \mathbf{r}_j) \right|^2 / Nf^2 \quad (52)$$

or

$$E_{\mathbf{h}} = \frac{1}{N^{1/2}} \sum_{j=1}^{N} \exp(2\pi i \mathbf{h} \bullet \mathbf{r}_j) \quad (53)$$

Now consider

$$\begin{aligned} E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}} &= \frac{1}{N}\left[\sum_j \exp(2\pi i \mathbf{k} \bullet \mathbf{r}_j)\right] \\ &\quad \times \left[\sum_j \exp(2\pi i(\mathbf{h} - \mathbf{k}) \bullet \mathbf{r}_j)\right] \\ &= \frac{1}{N}\left\{\sum_j \exp(2\pi i \mathbf{h} \bullet \mathbf{r}_j)\right. \\ &\quad \left. + \sum_i \sum_j \exp[2\pi i \mathbf{k} \bullet (\mathbf{r}_j - \mathbf{r}_i)][\exp(2\pi i \mathbf{h} \bullet \mathbf{r}_i)]\right\} \end{aligned} \quad (54)$$

If an average is now taken over a reasonably large number of such pairs, keeping h constant, then

$$\begin{aligned} \langle E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}\rangle &\approx \frac{1}{N}\sum \exp(2\pi i \mathbf{h} \bullet \mathbf{r}_i) \\ &\approx \frac{1}{N^{1/2}} E_{\mathbf{h}} \end{aligned} \quad (55)$$

(The second term in Equation 54 will tend to average to zero.) Since the magnitudes are measured and our concern is with the phase, we write

$$E_{\mathbf{h}} \propto \langle E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}\rangle_{\mathbf{k}} \quad (56)$$

Karle and Hauptman showed, using a more sophisticated mathematical approach, that for the unequal atom ease, a similar relation will hold. Equation 56 is often referred to as the sigma-two ($\sum_2$) relation.

For the non-centrosymmetric case, this is more conveniently written (Karle and Hauptman, 1956) as

$$\tan\alpha_{\mathbf{h}} = \frac{\sum_{\mathbf{k}} |E_{\mathbf{h}}\| E_{\mathbf{h}-\mathbf{h}}| \sin(\alpha_{\mathbf{k}} + \alpha_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |E_{\mathbf{k}}\| E_{\mathbf{h}-\mathbf{k}}| \cos(\alpha_{\mathbf{k}} + \alpha_{\mathbf{h}-\mathbf{k}})} \quad (57)$$

Karle and Hauptman argued that the relation should hold for an average using a limited set of data if the $|E|$'s are large, and Cochran and Woolfson (1955) obtained an approximate expression for $|E_{\mathrm{h}}|$ being positive from the $\sum_2$ relation,

$$P_+(\mathbf{h}) = \frac{1}{2} + \frac{1}{2}\tanh \sigma_3 \sigma_2^{-3/2} |E_{\mathbf{h}}| \sum_{\mathbf{k}} E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}} \quad (58)$$

where, Z_j being the atomic number of atom j

$$\sigma_n = \sum_j Z_j^n \tag{59}$$

The strategy then is to start with a few reflections whose phases are known and extend these using the $\sum_2$ relation accepting only those indications that have high probability (typically >95%) of being correct. The process is then repeated until no significant further changes in phases occur.

However, to employ the above strategy, phases of a few of the largest $|E|$'s must be known. How does one obtain an appropriate starting set? If working in a crystal system of orthorhombic symmetry or lower, three reflections can be used to specify the origin. Consider a centrosymmetric space group. The unit cell has centers of symmetry not only at (0, 0, 0) but also halfway along the cell in any direction. Therefore the cell can be displaced by 1/2 in any direction and an equally valid structural solution obtained (same set of intensities) using the tabulated equivalent positions of the space group. If in the structure factor expression all x_j were replaced by $1/2+x_j$, then

$$E_{hkl}^{\text{new}} = (-1)^h E_{hkl}^{\text{old}} \tag{60}$$

Therefore the signs of the structure factors for h odd would change while the signs for the even would not. Thus specifying a sign for an h odd reflection amounts to specifying an origin in the x direction. A similar argument can be made for the y and z directions as long as the reflection choices are independent. Such choices will be independent if they obey

$$\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 \neq (e,e,e) \tag{61a}$$

$$\mathbf{h}_1 + \mathbf{h}_2 \neq (e,e,e), \cdots \tag{61b}$$

$$\mathbf{h}_1 \neq (e,e,e), \cdots \tag{61c}$$

where e indicates an even value of the index.

Another useful approach to the derivation of direct method equations is through the use of what are termed structure semivariants and structure invariants.

Structure semivariants are those reflections or linear combination of reflections that are independent of the choice of permissible origins, such as has just been described above.

Structure invariants are certain linear combinations of the phases whose values are determined by the structure alone and are independent of the choice of origin. We will consider here two subgroups, the three-phase (triplet) structure invariant and the four-phase (quartet) structure invariant.

A three-phase structure invariant is a set of three reciprocal vectors $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ that satisfy the relation

$$\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0 \tag{62}$$

Let

$$A = \frac{\sigma_3}{\sigma_2^{3/2}} (|E_{\mathbf{h}_1} \parallel E_{\mathbf{h}_2} \parallel E_{\mathbf{h}_3}|) \tag{63}$$

and, in terms of phases φ

$$\varphi_3 = \varphi_{\mathbf{h}_1} + \varphi_{\mathbf{h}_2} + \varphi_{\mathbf{h}_3} \tag{64}$$

then φ_3 is found to be distributed about zero with a variance that is dependent on A—the larger its value, the narrower the distribution. It is obviously then another version of the $\sum_2$ relation.

A four-phase structure invariant is similarly four reciprocal vectors $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$, and $\mathbf{h}_4$ that satisfy the relation

$$\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 + \mathbf{h}_4 = 0 \tag{65}$$

Let

$$B = \frac{\sigma_3}{\sigma_2^{3/2}} (|E_{\mathbf{h}_1} \parallel E_{\mathbf{h}_2} \parallel E_{\mathbf{h}_3} \parallel E_{\mathbf{h}_4}|) \tag{66}$$

and

$$\varphi_4 = \varphi_{\mathbf{h}_1} + \varphi_{\mathbf{h}_2} + \varphi_{\mathbf{h}_3} + \varphi_{\mathbf{h}_4} \tag{67}$$

If furthermore three additional magnitudes are known, $|E_{\mathbf{h}_1+\mathbf{h}_2}|, |E_{\mathbf{h}_2+\mathbf{h}_3}|$, and $|E_{\mathbf{h}_3+\mathbf{h}_1}|$, then, in favorable cases, a more reliable estimate of φ_4 may be obtained, and furthermore, the estimate may lie anywhere in the interval 0 to π. If all seven of the above $|E|$'s are large, then it is likely that $\varphi_4 = 0$. However, it can also be shown that for the case where $|E_{\mathbf{h}_1}|, |E_{\mathbf{h}_2}|, |E_{\mathbf{h}_3}|$, and $|E_{\mathbf{h}_4}|$ are large but $|E_{\mathbf{h}_1+\mathbf{h}_2}|, |E_{\mathbf{h}_2+\mathbf{h}_3}|$ and $|E_{\mathbf{h}_3+\mathbf{h}_1}|$ are small, the most probable value of φ_4 is π. The latter is sometimes referred to as the negative-quartet relation. (For more details see Ladd and Palmer, 1980.)

Normally additional reflections will be needed to obtain a sufficient starting set. Various computer programs adopt their own specialized procedures to accomplish this end. Some of the more common approaches include random choices for signs or phases for a few hundred reflections followed by refinement and extension; alternately, phases for a smaller set of reflections, chosen to maximize their interaction with other reflections, are systematically varied followed by refinement and extension (for an example of the latter, see Woolfson and Germain, 1970; Germain et al., 1970). Various figures of merit have also been devised by the programmers to test the validity of the phase sets so obtained.

Direct method approaches readily lend themselves to the development of automatic techniques. Over the years, they have been extended and refined to make them generally applicable, and this has led to their widespread use for deducing an initial model.

Computer programs such as the SHELX direct method programs (Sheldrick, 1990) use both the $\sum_2$ (three-phase structure invariant) and the negative four-phase quartet to determine phases. Starting phases are produced by random-number generation techniques and then refined using

$$\text{New } \varphi_h = \text{phase of } [\alpha - \eta] \tag{68}$$

where α is defined by

$$\alpha = 2|E_{\mathbf{h}}|E_{\mathbf{k}}E_{\mathbf{h}-\mathbf{k}}/N^{1/2} \tag{69}$$

and η is defined by

$$\eta = g|E_{\mathbf{h}}|E_{\mathbf{k}}E_{\mathbf{l}}E_{\mathbf{h}-\mathbf{k}-\mathbf{l}}/N \tag{70}$$

N is the number of atoms (for an equal atom case), and g is a constant to compensate for the smaller absolute value of η compared to α. In the calculation of η, only those quartets are used for which all three cross terms have been measured and are weak. An E map is then calculated for the refined phase set giving the best figure of merit.

Heavy-Atom Method

Direct methods can be expected to be the most reliable when applied to structures in which most of the atoms are approximately equal in atomic number. In structures where one or a few atoms are present that have appreciably larger atomic numbers than the rest, the heavy-atom method is typically the approach of choice.

Assume a structure contains one atom in the asymmetric unit that is "heavier" than the rest. Also assume for the moment that the position of this atom (r_{H}) is known. The structure factor can be written as

$$F_{hkl} = f_H \exp(2\pi i\mathbf{h} \bullet \mathbf{r}_H) + \sum_{j=2}^{N} f_j \exp(2i\mathbf{h} \bullet \mathbf{r}_j) \tag{71}$$

Alternately the structure factor can be written as

$$F_{hkl} = F_{hld}^{\mathrm{heavy}} + F_{hkl}^{\mathrm{other}} \tag{72}$$

where "heavy" denotes the contributors to the structure factor from the heavy atom(s) and "other" represents the contribution to the structure factor from the remaining atoms in the structure. The latter contains many small contributions, which can be expected to partially cancel one another in most reflections. Therefore, if the "other" contribution does not contain too large a number of these smaller contributors, one would expect the sign (or the phase) of the observed structure factor to be approximately that calculated from F_{hkl}^{heavy}, although the agreement in terms of magnitudes would likely be poor. Thus the approach is to calculate F_{hkl}^{heavy} and transfer its sign or phase to the $|F_{hkl}^{\mathrm{obs}}|$, unless $|F_{hkl}^{\mathrm{heavy}}|$ is quite small. Use these phased $|F_{hkl}^{\mathrm{obs}}|$ to calculate an electron density map and examine this map to attempt to find additional positions. (As a general rule of thumb, atoms often appear with $\sim 1/3$ of their true height if not included in the structure factor calculation.) Those atoms that appear, especially if they are in chemically reasonable positions, are then added to the structure factor calculation, giving improved phases, and the process is repeated until all the atoms have been found. For this approach to be successful, it is usually necessary for $\sum Z_{\mathrm{H}}^2 \geq \sum Z_{\mathrm{L}}^2$.

How does one go about finding the position of the heavy atom or atoms, assuming the statistics indicate that the heavy-atom approach could be successful? One possibility is to use direct methods as discussed earlier. Although direct methods may not provide the full structure, the largest peaks on the calculated map may well correspond to the positions of the heavier atoms.

An alternate approach is to deduce the position of the heavy atom(s) through the use of a Patterson function (Patterson, 1935). The Patterson function is an autocorrelation function of the electron density function:

$$P(\mathbf{u}) = \int \rho(\mathbf{r}) \bullet \rho(\mathbf{r} + \mathbf{u}) d\tau \tag{73}$$

If the Fourier series expressions for ρ are substituted into this equation, it can be readily shown that

$$P(\mathbf{u}) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}|^2 \cos 2\pi\mathbf{h} \bullet \mathbf{u} \tag{74}$$

The Patterson function therefore can be directly calculated from the intensities. From a physical point of view, a peak would be expected in the Patterson function anytime two atoms are separated by a vector displacement $\mathbf{u}$. Since a vector could be drawn from A to B or B to A, a centrosymmetric function would be expected, consistent with a cosine function. Moreover, the heights of the peaks in the Patterson function should be proportional to the products of the atomic electron densities, i.e., Z_{iZj}. If a structure contained nickel and oxygen atoms, the heights of the Ni-Ni, Ni-O, and O-O vectors would be expected to be in the ratio 784:224:64 for peaks representing single interactions.

Most materials crystallize in unit cells that have symmetry higher than triclinic. This symmetry can provide an additional very useful tool in the analysis of the Patterson function. All Patterson functions have the same symmetry as the Laue group associated with the related crystal system. Often peaks corresponding to vectors between symmetry-related atoms occur in special planes or lines. This is probably easiest seen with an example. Consider the monoclinic space group $P2_1/c$. The Patterson function would have $2/m$ symmetry—the symmetry of the monoclinic Laue group. Moreover, because of the general equivalent positions in this space group, namely, $(x, y, z)(x, 1/2 - y, 1/2 + z)(\bar{x}, 1/2 + y, 1/2 - z)(\bar{x}, \bar{y}, \bar{z})$, peaks between symmetry-related atoms would be found at $(0, 1/2 - 2y, 1/2)(0, 1/2 + 2y, 1/2)$ $(-2x, 1/2, 1/2 - 2z)$ $(2x, 1/2, 1/2 + 2z)$ $(-2x, 2y, -2z)$ $(2x, 2y, 2z)$ $(-2x, 2y, -2z)$ and $(2x, -2y, 2z)$ as deduced by obtaining all the differences between these equivalent positions. Such vectors are often termed Harker vectors (Harker, 1936; Buerger, 1959). In the case of the first four peaks, two differences give the same values and yield a double peak on the Patterson function. This set of Patterson peaks can be used to determine the position of a heavy atom. First we should again note that the peaks corresponding to vectors between heavy atoms should be larger than the other types. For this space group, one can select those peaks with $u = 0$ and $w = 1/2$ and, from their v coordinate, determine y possibilities. In a similar fashion, by selecting those large peaks with $\nu = 1/2$ (remember peaks in both of these categories

must be double), an x and z pair can be determined. These results can be combined together to predict the $2x$, $2y$, $2z$ type peak position, which can then be tested to see if the combination is a valid one. The same process can be carried out if more than one heavy atom is present per asymmetric unit—the analysis just becomes somewhat more complicated. [There is another category of methods termed Patterson superposition methods, which are designed to systematically break down the multiple images of the structure present in the Patterson function to obtain essentially a single image of the structure. Because of space limitations, they will not be discussed here. An interested reader should consult Richardson and Jacobson (1987) and Lipscomb and Jacobson (1990) and references therein for further details.] It may also be appropriate here to remind the reader that the intensities are invariant if the structure is acted upon by any symmetry element in the Laue group or is displaced by half-cell displacements. Two solutions differing only in this fashion are equivalent.

ROBERT A. JACOBSON
Iowa State University
Ames, Iowa

XAFS SPECTROSCOPY

INTRODUCTION

X rays, like other forms of electromagnetic radiation, are both absorbed and scattered when they encounter matter. X-ray scattering and diffraction are widely utilized structural methods employed in thousands of laboratories around the world. X-ray diffraction techniques are undeniably among the most important analysis tools in nearly every physical and biological science. As this unit will show, x-ray absorption measurements are achieving a similar range of application and utility. Absorption methods are often complementary to diffraction methods in terms of the area of application and the information obtained.

The basic utility of x-ray absorption arises from the fact that each element has characteristic absorption energies, usually referred to as absorption edges. These occur when the x rays exceed the energy necessary to ionize a particular atomic level. Since this is a new channel for absorption, the absorption coefficient shows a sharp rise. Some examples are shown in Figure 1A for elements in a high-temperature superconductor. Note that the spectra for each element can be separately obtained and have distinctly different structure. Often the spectra are divided into two regions. The region right at the edge is often called the XANES (x-ray absorption near-edge structure), and the region starting ~20 or 30 eV past its edge is referred to as EXAFS (extended x-ray absorption fine structure). The isolated EXAFS structure is shown in Figure 1B. Recent theories have begun to treat these in a more unified manner, and the trend is to refer to the entire spectrum as the XAFS (x-ray absorption fine structure) and the technique in general as XAS (x-ray absorption spectroscopy). Historically the principal quantum number of the initial state atomic level is labeled by the letters K, L, M, N, ... for $n = 1, 2, 3, 4, \ldots$, and the angular momentum state of the level is denoted as subscripts $1, 2, 3, 4, \ldots$ for the s, $p_{1/2}$, $p_{3/2}$, $d_{3/2}, \ldots$ levels. The most common edges used for XAFS are the K edge or $1s$ initial state (the subscript 1 is omitted since it is the only possibility) and the L_3 edge or $2p_{3/2}$ initial state.

The utility of XAS is demonstrated in Figure 1. The different elements have different fine structure because they have different local atomic environments. In this unit it will be shown how these spectra can be analyzed to obtain detailed information about each atom's environment. This includes the types of neighboring atoms, their distances, the disorder in these distances, and the type of bonding. The near-edge region is more sensitive to chemical effects and can often be used to determine the formal valence of the absorbing atom as well as its site symmetry.

In the simplest picture, the spectra are a result of quantum interference of the photoelectron generated by the absorption process as it is scattered from the neighboring atoms. This interference pattern is, of course, related to the local arrangement of atoms causing the scattering. As the incoming x-ray energy is changed, the energy of the photoelectron also varies along with its corresponding

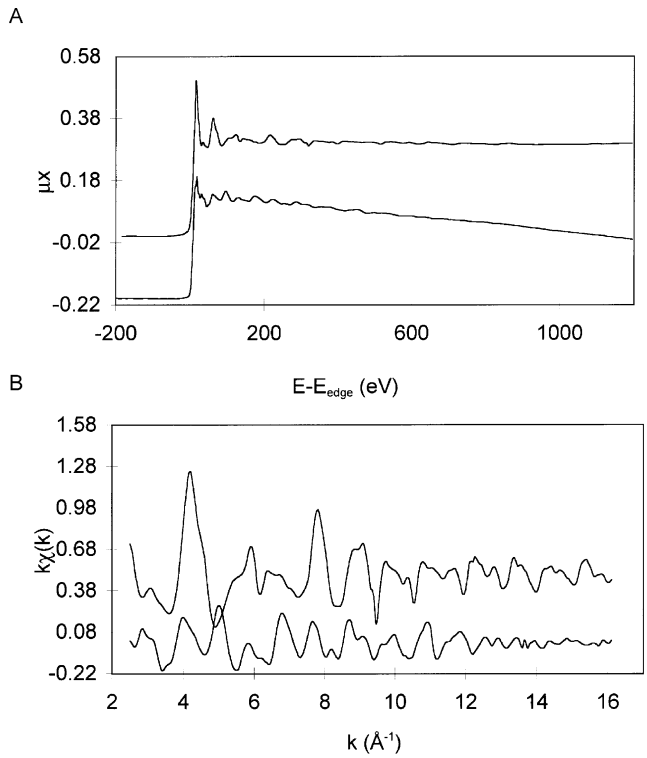


Figure 1. Examples of x-ray absorption data from the high-temperature superconductor material $YBa_2Cu_3O_7$. (**A**) X-ray absorption. (**B**) Normalized extended fine structure vs. wave vector extracted from the spectra in (A). The top spectra in both plots are for the Y K edge at 17,038 eV, and the bottom spectra are for the Cu K edge at 8,980 eV. Both edges were obtained at liquid nitrogen temperature.

wavelength. Therefore, the interference from the different surrounding atoms goes in and out of phase, giving the oscillatory behavior seen in Figure 1B. Each type of atom also has a characteristic backscattering amplitude and phase shift variation with energy. This allows different atom types to be distinguished by the energy dependence of the phase and amplitude of the different oscillatory components of the spectrum. This simple picture will be expanded below (see Principles of the Method) and will form the basis for the detailed analysis procedures used to extract the structural parameters. It is important to emphasize that the oscillations originate from a local process that does not depend on long-range order. XAFS will be observed any time an atom has at least one well-defined neighbor and, in addition to well-ordered crystalline or polycrystalline materials, has been observed in molecular gases, liquids, and amorphous materials.

The widespread use of x-ray absorption methods is intimately connected to the development of synchrotron radiation sources. The measurements require a degree of tunability and intensity that is difficult to obtain with conventional laboratory sources. Because of this need and of advances in the theoretical understanding that came at about the same time as the major synchrotron sources were first developed, the modern application of absorption methods for materials analysis is only ~25 years old. However, it has a long history as a physical phenomenon to be understood, with extensive work beginning in the 1930s. For a review of the early work see Azaroff and Pease (1974) and Stumm von Bordwehr (1989).

This unit will concentrate on the modern application of XAS at synchrotron sources. Conducting experiments at remote facilities is a process with its own style that may not be familiar to new practitioners. Some issues related to working at synchrotrons will be discussed (see Practical Aspects of the Method).

Comparison to Other Techniques

There are a wide variety of structural techniques; they can be broadly classified as direct and indirect methods. Direct methods give signals that directly reflect the structure of the material. Diffraction measurements can, in principle, be directly inverted to give the atomic positions. Of course, the difficulty in measuring both the phase and amplitude of the diffraction signals usually precludes such an inversion. However, the diffraction pattern still directly reflects the underlying symmetry of the crystal lattice, and unit cell symmetry and size can be simply determined. More detailed modeling or determination of phase information is required to place the atoms accurately within the unit cell. Because there is such a direct relation between structure and the diffraction pattern, diffraction techniques can often provide unsurpassed precision in the atomic distances.

Indirect methods are sensitive to structure but typically require modeling or comparison to standards to extract the structural parameters. This does not mean they are not extremely useful structural methods, only that the signal does not have a direct and obvious relation to the structure. Examples of such methods include Mössbauer spectroscopy (MOSSBAUER SPECTROMETRY), nuclear magnetic resonance (NMR; NUCLEAR MAGNETIC RESONANCE IMAGING), electron paramagnetic resonance (EPR; ELECTRON PARAMAGNETIC RESONANCE SPECTROSCOPY), and Raman spectroscopy (RAMAN SPECTROSCOPY OF SOLIDS). The XANES part of the x-ray absorption signal can be fit into this category. As will be shown, strong multiple-scattering and chemical effects make direct extraction of the structural parameters difficult, and analysis usually proceeds by comparison of the spectra with calculated or measured models. On the other hand, as is the case with many of the other indirect methods, the complicating factors can often be used to advantage to learn more about the chemical bonding.

The indirect methods usually share the characteristic that they are atomic probes. That is, they are foremost measures of the state of an individual atom. The structure affects the signal since it affects the atomic environment, but long range order is usually not required. Also, atomic concentration is not important in determining the form of the signal, only its detectability. Therefore, exquisite sensitivity to low concentrations is often possible. This is also true for x-ray absorption. Methods that allow detection of structural information to concentrations for which diffraction techniques would be useless will be described under Detection Methods, below.

In some respects, the EXAFS part of the absorption signal shares the characteristics of both direct and indirect structural methods. The absorbing atom acts as both the source and detector of the propagating photoelectrons. The measured signal contains both phase and amplitude information, and, as in direct methods, can be inverted to obtain structural information. However, the photoelectrons interact strongly with the surrounding atoms, which modifies both their amplitude and phase. Therefore, a direct inversion (Fourier transform) of the spectrum gives only a qualitative view of the local structure. This qualitative view can be informative, but modeling is usually required to extract detailed structural parameters.

PRINCIPLES OF THE METHOD

Single-Scattering Picture

When an x ray is absorbed, most of the time an electron is knocked out of the atom, which results in a photoelectron with energy $E = E_x - E_b$, where E_x is the x-ray energy and E_b is the binding energy of the electron. The x-ray edge occurs when $E_x = E_b$. The photoelectron propagates as a spherical wave with a wave vector given by

$$k = \frac{\sqrt{2m_e E}}{\hbar} \quad (1)$$

As shown in Figure 2, this wave can be scattered back to the central atom and interfere with the original absorption. In a classical picture this can seem like a strange concept, but this process must be treated quantum mechanically, and the absorption cross-section is governed by Fermi's golden rule:

$$\sigma_K = 4\pi^2 \alpha \hbar\omega \sum_j |\langle \Psi_f | \mathbf{e} \cdot \mathbf{r} | \Psi_K \rangle|^2 \times \delta(E_f + E_K - \hbar\omega) \quad (2)$$

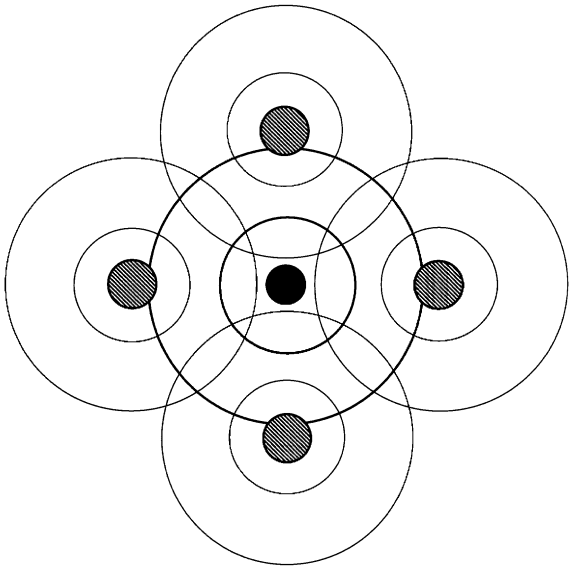


Figure 2. Two-dimensional schematic of the EXAFS scattering process. The photoelectron originates from the central (black) atom and is scattered by the four neighbors. The rings represent the maxima of the photoelectron direct and scattered waves.

where **e** is the electric field vector, **r** is the radial distance vector, α is the fine structure constant, ω is the angular frequency of the photon, and Ψ_K and Ψ_f are the initial state (e.g., the 1*s* state in the *K* shell) and the final state, respectively. The sum is over all possible final states of the photoelectron.

It is necessary to consider the overlap of the initial core state with a final state consisting of the excited atom and the outgoing and backscattered photoelectron. Since the photoelectron has a well-defined wavelength, the backscattered wave will have a well-defined phase relative to the origin (deep core electrons characteristic of x-ray edges are highly localized at the center of the absorbing atom). This phase will vary with energy, resulting in a characteristic oscillation of the absorption probability as the two waves go in and out of phase. Using this simple idea and considering only single scattering events gives the following simple expression (Sayers et al., 1971; Ashley and Doniach, 1975; Lee and Pendry, 1975):

$$\chi(k) = \frac{\mu(k) - \mu_0(k)}{\mu_0(k)} = \sum_j \frac{N_j}{kR_j^2} A_j(k) p_j(\mathbf{e}) \sin[2kR_j + \Psi_j(k)] \qquad (3)$$

where $\chi(k)$ is the normalized oscillatory part of the absorption determined by subtracting the smooth part of the absorption, μ_0, from the total absorption, μ, and normalizing by μ_0. The sum is over the neighboring shells of atoms each of which consists of N_j atoms at a similar distance R_j. The R_j^2 factor accounts for the fall-off of the spherical photoelectron wave with distance.

The sine term accounts for the interference. The total path length of the photoelectron as it travels to a shell and back is $2kR_j$. The factor $p_j(\mathbf{e})$ accounts for the polarization of the incoming x-rays. The remaining factors, $A_j(k)$ and $\Psi_j(k)$, are overall amplitude and phase factors. It is these two factors that must be calibrated by theory or experiment. The amplitude factor $A(k)$ can be further broken down to give

$$A_j(k) = S_0^2 F_j(k) Q_j(k) e^{-2R_j/\lambda} \qquad (4)$$

The amplitude reduction factor S_0^2 results from multielectron excitations in the central atom (the simple picture assumes only a single photoelectron is excited). This factor depends on the central atom type and typically has nearly k independent values from 0.7 to 0.9. The magnitude of the complex backscattering amplitude factor $f_j(k, \pi)$, $F_j(k)$ depends on the type of scattering atom. Often its k -dependence can be used to determine the type of backscattering atom. The factor $Q_j(k)$ accounts for any thermal or structural disorder in the jth shell of atoms. It will be discussed in more detail later. The final exponential factor is a mean free path factor that accounts for inelastic scattering and core hole lifetime effects.

The phase term can be broken down to give

$$\Psi_j(k) = 2\phi_c + \theta_j(k) + \varphi_j(k) \qquad (5)$$

where $\theta_j(k)$ is the phase of the backscattering factor $f_j(k, \pi)$, ϕ_c is the p-wave phase shift caused by the potential of the central atom, and $\varphi_j(k)$ is a phase factor related to the disorder of the jth shell.

In the past, the phase and amplitude terms were generally calibrated using model compounds that had known structures and chemical environments similar to those of the substance being investigated. The problem with this approach is that suitable models were often difficult to come by, since a major justification for applying XAFS is the lack of other structural information. In recent years the theory has advanced to the state where it is often more accurate to make a theoretical determination of the phase and amplitude factors. Using theory also allows for proper consideration of multiple scattering complications to the simple single scattering picture presented so far. This will be discussed below (see Data Analysis and Initial Interpretation). Figures 3, 4, and 5 show some calculated

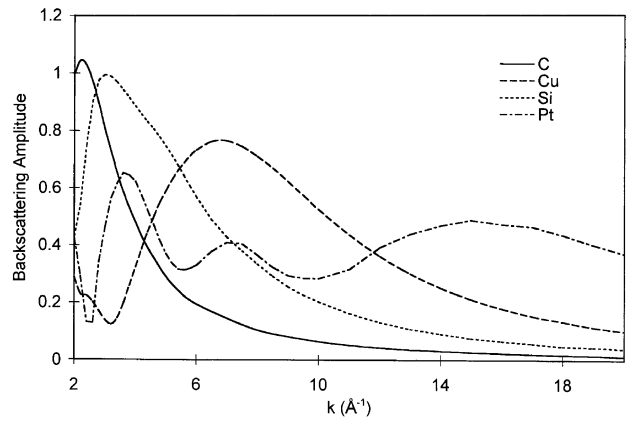


Figure 3. Theoretical backscattering amplitude versus photoelectron wave vector k for some representative backscattering atoms. Calculated using FEFF 5.0 (Zabinsky et al., 1995).

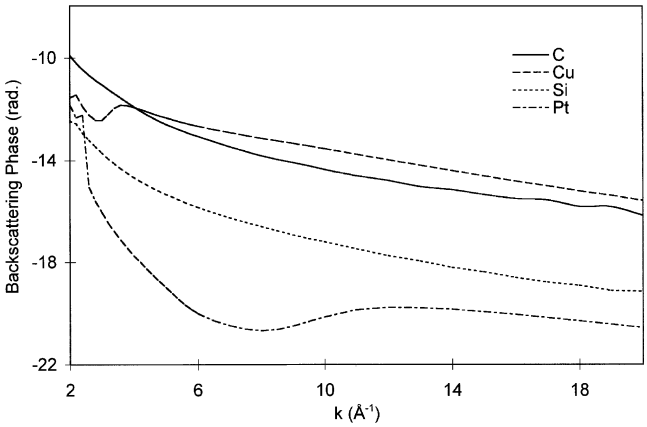


Figure 4. Theoretical backscattering phase shift versus photoelectron wave vector k for some representative backscattering atoms. Calculated using FEFF 5.0 (Zabinsky et al., 1995).

amplitudes and phases, illustrating two main points. First, there are significant differences in the backscattering from different elements. Therefore, it is possible to distinguish between different types of atoms in a shell. For neighboring atoms such as Fe and Co, the differences are generally too small to distinguish them reliably, but if the atomic numbers of atoms in a shell differ by more than ~10%, then their contributions can be separated. The figures nicely illustrate this. Platinum is quite different from the other elements and is easily distinguished. Silicon and carbon have similar backscattering amplitudes that could be difficult to distinguish if the additional amplitude variations from disorder terms are not known. However, the backscattering phase is about a factor of π different, which means their oscillations will be out of phase and easily distinguished. Similarly, copper and carbon have almost the same phase but a distinctly different amplitude.

The second point is that the phase shifts are fairly linear with k. This means that the sinusoidal oscillations in Equation 3 will maintain their periodic character. The additional phase term will primarily result in a constant frequency shift. Each shell of atoms will still have a characteristic frequency, and a Fourier transform of the spectrum can be used to separate them. This is shown in Figure 6, which is a Fourier transform of the Cu data in Figure 1B. The magnitude of the Fourier transform is related to the partial radial distribution function of the absorbing atoms. The peaks are shifted to lower R by the phase shift terms, and the amplitude of the peaks depends on the atomic type, number of atoms in a shell, and disorder of the shell. This is a complex structure and there are no well-defined peaks that are well separated from the others. Methods that can be used to extract the structural information are described below (see Data Analysis and Initial Intepretation). This is high-quality data and all the structure shown is reproducible and real. However, above ~4 Å the structure is a composite of such a large number of single- and multiple-scattering contributions that it is not possible to analyze. This is typical of most structures, which become too complex to analyze at distances above 4 to 5 Å. It is also important to note that because of the additional phase and amplitude terms, the transform does not give the true radial distribution. It should only be used as an intermediate step in the analysis and as a method for making a simple qualitative examination of the data. It is very useful for assessing the noise in the data and for pointing out artifacts that can result in nonphysical peaks.

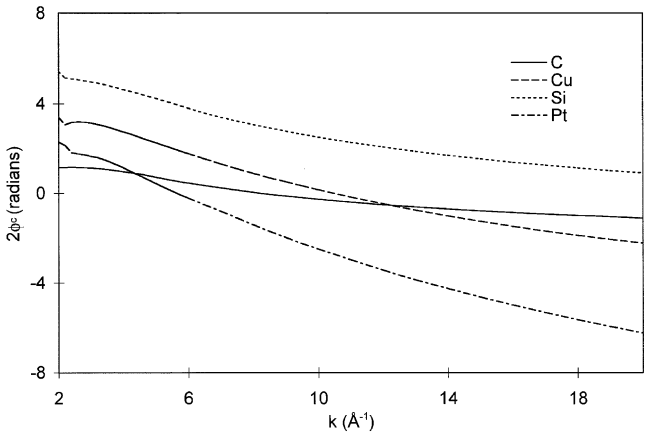


Figure 5. Theoretical central atom phase shift versus photoelectron wave vector k for some representative absorbing atoms. Calculated using FEFF 5.0 (Zabinsky et al., 1995).

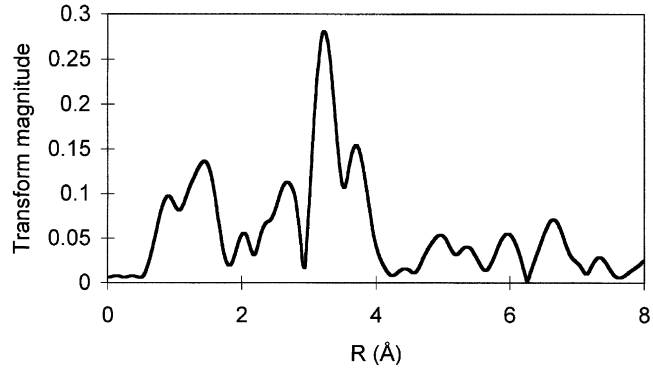


Figure 6. Magnitude of the k-weighted Fourier transform of the Cu spectrum in Figure 1B. The k range from 2.5 to 16 Å^{-1} was used.

EXTENSIONS TO THE SIMPLE PICTURE

Disorder

The x-ray absorption process is energetic and takes place in a typical time scale of 10^{-15} sec. This is much faster than the typical vibrational periods of atoms and means that the x-ray absorption takes a snapshot of the atomic configuration. Even for an ideal shell of atoms all at a single distance, thermal vibrations will result in a distribution about the average R_j. In addition, for complex or disordered structures there can be a structural contribution to the distribution in R_j. This results in the disorder factors $Q_j(k)$ and $\varphi_j(k)$ given above. To a good approximation these factors are given by the Fourier transform of the real-space distribution $P_j(r)$:

$$Q_j(k)\exp[i\varphi_j(k)] = \int dr\, P_j(r)\exp[i2k(r-R_j)] \qquad (6)$$

The simplest case is for a well-ordered shell at low temperature where the harmonic approximation for the vibrations applies. Then the distribution is Gaussian, and the transform is also Gaussian. This results in $\varphi_j(k) = 0$ and a Debye-Waller-like term for $Q : Q_j(k) = \exp(-2k^2\sigma_j^2)$, where σ_j^2 is the mean-square width of the distribution. This simple approximation is often valid and provides a good starting point for analysis. However, it is a fairly extreme approximation and can break down. If $\sigma^2 >\sim 0.01^2$, more exact methods should be used that include the contribution to the phase (Boyce et al., 1981; Crozier et al., 1988; Stern et al., 1992). One commonly used method included in many analysis packages is the cumulant expansion (Bunker, 1983).

The changes in the Debye-Waller term with temperature can often be determined relatively accurately. It can be quite sensitive to temperature, and low-temperature measurements will often produce larger signals and higher data quality.

Although the EXAFS disorder term is often called a Debye-Waller factor, this is not strictly correct. The EXAFS Debye-Waller factor is different from that determined by x-ray diffraction or Mössbauer spectroscopy (Beni and Plantzman, 1976). Those techniques measure the mean-square deviation from the ideal lattice position. The EXAFS term is sensitive to the mean-square deviation in the bond length. Long-wavelength vibrational modes in which neighboring atoms move together make relatively little contribution to the EXAFS Debye-Waller term but can dominate the conventional Debye-Waller factor. Thus, the temperature dependence of the EXAFS Debye-Waller term is a sensitive meaure of the bond strength since it directly measures the relative vibrational motion of the bonded atoms.

L Edges

The preceding discussion is limited to the K edge since that has a simple 1s initial state. For an L or higher edge the situation can be more complicated. For the L_2 or L_3 edge, the initial state has 2p symmetry and the final state can have either s or d symmetry. It is also possible to have mixed final states where an outgoing d wave is scattered back as an s state or vice versa. The result is three separate contributions to the function $\chi(k)$ that can be denoted χ_{00}, χ_{22}, and χ_{20}, where the subscripts refer to the angular momentum state of the outgoing and backscattered photoelectron wave. Each of these can have different phase and amplitude functions. The total EXAFS can then be expressed as follows (Lee, 1976; Heald and Stern, 1977):

$$\chi = \frac{M_{21}^2\chi_{22} + M_{01}^2\chi_{00} + 2M_{01}M_{21}\chi_{20}}{M_{21}^2 + M_{01}^2/2} \tag{7}$$

The matrix element terms (e.g., M_{01}) refer to radial dipole matrix elements between the core wave function with $l = 1$ and the final states with $l = 0$ and 2. For the K edge there is only one possibility for M that cancels out. The L_1 edge, which has a 2s initial state, has a similar cancellation and can be treated the same as the K edge.

This would seriously complicate the analysis of $L_{2,3}$ edges, but for two fortuitous simplifications. The most important is that M_{01} is only $\sim$0.2 of M_{21}. Therefore, the M_{01}^2 terms can be ignored. The cross-term can still be $\sim$40% of the main term. Fortunately, for an unoriented polycrystalline sample or a material with at least cubic symmetry, the angular average of this term is zero. The cross-term must be accounted for when the sample has an orientation. This most commonly occurs for surface XAFS experiments, which often employ L edges and which have an intrinsic asymmetry.

Polarization

At synchrotron sources the radiation is highly polarized. The most common form of polarization is linear polarization, which will be considered here (it is also possible to have circular polarization). For linear polarization, the photoelectron is preferentially emitted along the electric field direction. At synchrotrons the electric field is normally oriented horizontally in a direction perpendicular to the beam direction. For K and L_1 shells the polarization factor in Equation 3 is $p_j(\mathbf{e}) = 3\langle\cos^2\theta\rangle$, where

$$\langle\cos^2\theta\rangle = \frac{1}{N_j}\sum_i \cos^2\theta_i \tag{8}$$

The sum is over the i individual atoms in shell j. If the material has cubic or higher symmety or is randomly oriented, then $\langle\cos^2\theta\rangle = \frac{1}{3}$, and $p_j(\mathbf{e}) = 1$. A common situation is a sample with uniaxial symmetry such as a layered material or a surface experiment. If the symmetry in the plane is 3-fold, then the signal does not depend on the orientation of $\mathbf{e}$ within the plane. The signal only depends on the orientation of the electric field vector with respect to the plane, Θ, and $\chi(\Theta) = 2\chi_0\cos^2\Theta/3 + \chi_{90}\sin^2\Theta/3$, where χ_0 and χ_{90} are the signals with the field parallel and perpendicular to the plane, respectively.

For $L_{2,3}$ edges the situation is more complicated (Heald and Stern, 1977). There are separate polarization factors for the three contributions:

$$\begin{aligned} p_{22}^j &= \frac{1}{2}(1 + 3\langle\cos^2\theta\rangle_j) \\ p_{00}^j &= \frac{1}{2} \\ p_{02}^j &= \frac{1}{2}(1 - 3\langle\cos^2\theta\rangle_j) \end{aligned} \tag{9}$$

As mentioned previously, the 00 case can be ignored. For the unoriented case $\langle\cos^2\theta\rangle_j = 1/3$, giving $p_{22} = 1$ and $p_{02} = 0$. This is the reason the cross-term can often be ignored.

Multiple Scattering

The most important extension to the simple single-scattering picture discussed so far is the inclusion of multiple-scattering paths. For example, considering the first shell of atoms, the photoelectron can scatter from one first-shell atom to a second first-shell atom before being scattered back to the origin. Since in an ordered structure there

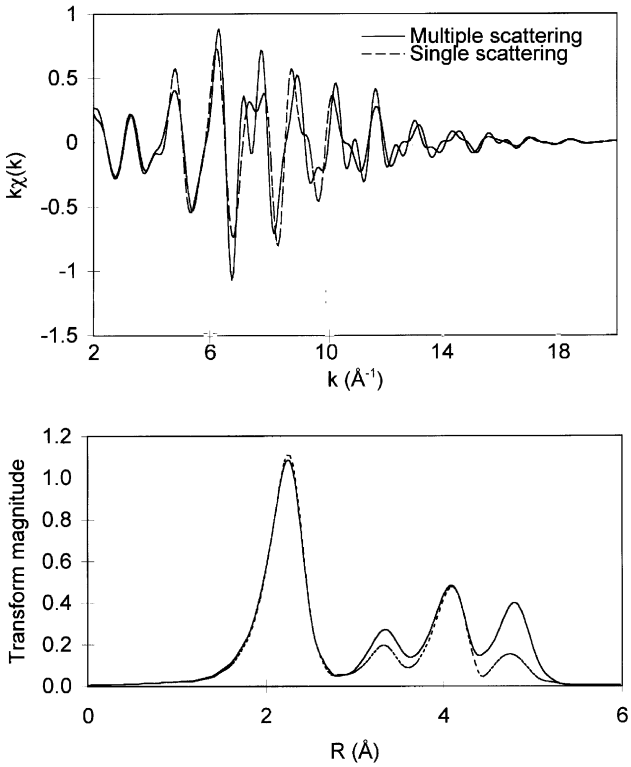


Figure 7. Comparison of the theoretical multiple- and single-scattering spectra for Cu metal at 80 K. The calculation only included the first four shells of atoms, which show up as four distinct peaks in the transform.

are typically 4 to 12 first-shell atoms, there are a large number of equivalent paths of this type, and the contribution can potentially be comparable to the single-scattering signal from the first shell. Two facts rescue the single-scattering picture. First, it is obvious that the multiple-scattering path is longer than the single-scattering path. Therefore, the multiple-scattering contribution would not contaminate the first-shell signal. It can, however, complicate the analysis of the higher shells. The second important point is that the scattering of the photoelectron is maximum when the scattering angle is either 0° (forward scattering) or 180° (backscattering). For the example given, there are two scattering events at intermediate angles and the contribution to the signal is reduced.

The single-scattering picture is still valid for the first shell and provides a reasonably accurate representation of the next one or two shells. For best results, however, modern analysis practice dictates that multiple scattering should be taken into account whenever shells past the first are analyzed, and this will be discussed further in the analysis section.

The enhancement of the photoelectron wave in the forward direction is often referred to as the focusing affect. It results in strong enhancement of shells of atoms that are directly in line with inner shells. An example is a face-centered-cubic (fcc) material such as copper. The fourth-shell atoms are directly in line with the first. This gives a strong enhancement of the fourth-shell signal and significantly changes the phase of the signal relative to a single-scattering calculation. Figure 7 compares the $\chi(k)$ and the Fourier transforms of calculated spectra for copper in the single-scattering and multiple-scattering cases. Sometimes this focusing effect can be used to advantage in the study of lattice distortions. Because the focusing effect is strongly peaked in the forward direction, small deviations from perfect colinearity are amplified in the signal, allowing such distortions to be resolved even when the change in atomic position of the atoms is too small to resolve directly (Rechav et al., 1994). The focusing effect has even allowed the detection of hydrogen atoms, which normally have negligible backscattering (Lengeler, 1986).

XANES

The primary difference in the near-edge region is the dominance of multiple-scattering contributions (Bianconi, 1988). At low energies the scattering is more isotropic, and many more multiple-scattering paths must be accounted for, including paths that have negligible amplitude in the extended region. This complicates both the quantitative analysis and calculation of the near-edge region. However, the same multiple scattering makes the near-edge region much more sensitive to the symmetry of the neighboring atoms. Often near-edge features can be used as indicators of such things as tetrahedral versus octahedral bonding. Since the bond symmetry can be intimately related to formal valence, the near edge can be a sensitive indicator of valence in certain systems. A classic example is the near edge of Cr, where, as shown in Figure 8, the chromate ion has a very strong pre-edge feature that makes it easy to distinguish from other types. This near-edge feature is much larger than any EXAFS signal and can be used to estimate the chromate-to-total-Cr ratio in many cases for which the EXAFS would be too noisy for reliable analysis.

While great strides have been made in the calculation of near-edge spectra, the level of accuracy often is much less than for the extended fine structure. In addition to the multiple scattering approach discussed, attempts have been made to relate the near-edge structure to the projected density of final states obtained from a band-structure-type calculation. In this approach, for the *K* edge,

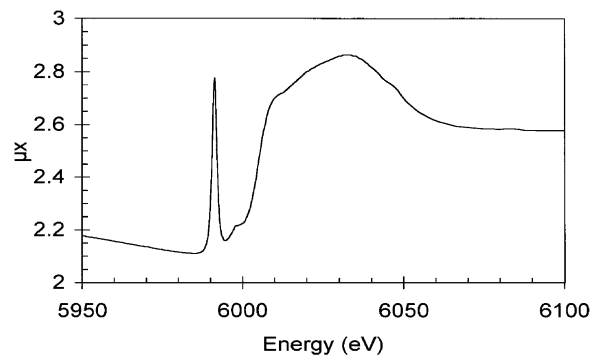


Figure 8. Near-edge region for the Cr *K* edge in a 0.5 M solution of $KCrO_4$.

the p density of states would be calculated. Often there is quite good qualitative agreement with near-edge features that allows them to be correlated to certain electronic states. This approach has some merit since one approach to band structure calculation is to consider all possible scattering paths for an electron in the solid, a quite similar calculation to that for the multiple scattering of the photoelectron. However, the band structure calculations do not account for the ionized core hole created when the photoelectron is excited. If local screening is not strong, the core hole can significantly affect the local potential and the resulting local density of states. While the band structure approach can often be informative, it cannot be taken as a definitive identification of near-edge features without supporting evidence.

It is important to point out that the near edge has the same polarization dependence as the rest of the XAFS. Thus, the polarization dependence of the XANES features can be used to associate them with different bonds. For example, in the K edge of a planar system a feature associated with the in-plane bonding will have a $\cos^2\Theta$ dependence as the polarization vector is rotated out of the plane by Θ. For surface studies this can be a powerful method of determining the orientation of molecules on a surface (Stohr, 1988).

PRACTICAL ASPECTS OF THE METHOD

Detection Methods

The simplest method for determining the absorption coefficient μ is to measure the attenuation of the beam as it passes through the sample:

$$I_t = I_0 e^{-} \quad \text{or} \quad \mu x = \ln(I_0/I_t) \tag{10}$$

Here, I_0 is measured with a partially transmitting detector, typically a gas ionization chamber with the gas chosen to absorb a fraction of the beam. It can be shown statistically that the optimum fraction of absorption for the I_0 detector is $f = 1/(1 + e^{\mu x/2})$, or $\sim 10\%$ to 30% for typical samples (Rose and Shapiro, 1948). The choice of the optimum sample thickness depends on several factors and will be discussed in more detail later. There are two methods for measuring the transmission: the standard scanning method and the dispersive method.

The scanning method uses a monochromatic beam that is scanned in energy as the transmission is monitored. The energy scanning can be either in a step-by-step mode, where the monochromator is halted at each point for a fixed time, or the so-called quick XAFS method, where the monochromator is continuously scanned and data are collected "on the fly" (Frahm, 1988). The step-scanning mode has some overhead in starting and stopping the monochromator but does not require as high a stability of the monochromator during scanning.

The dispersive method uses a bent crystal monochromator to focus a range of energies onto the sample (Matsushita and Phizackerley, 1981; Flank et al., 1982). After the sample, the focused beam is allowed to diverge. The constituent energies have different angles of divergence and eventually separate enough to allow their individual detection on a linear detector. The method has no moving parts, and spectra can be collected in a millisecond time frame. The lack of moving apparatus also means that a very stable energy calibration can be achieved, and the method is very good for looking at the time response of small near-edge changes. With existing dispersive setups, the total energy range is limited, often to a smaller range than would be desired for a full EXAFS measurement. Also, it should be noted that, in principle, the dispersive and quick XAFS methods have the same statistical noise for a given acquisition time, assuming they both use the same beam divergence from the source. For N data points, the dispersive method uses $1/N$ of the flux in each data bin for the full acquisition time, while the quick XAFS method uses the full flux for $1/N$ of the time.

The transmission method is the preferred choice for concentrated samples that can be prepared in a uniform layer of the appropriate thickness. As the element of interest becomes more dilute, the fluorescence technique begins to be preferred. In transmission the signal-to-noise ratio applies to the total signal, and for the XAFS of a dilute component the signal-to-noise ratio will be degraded by a factor related to the diluteness. With the large number of x rays (10^{10} to 10^{13} photons/s) available at synchrotron sources, if the signal-to-noise ratio was dominated by statistical errors, then the transmission method would be possible for very dilute samples. However, systematic errors, beam fluctuation, nonuniform samples, and amplifier noise generally limit the total signal-to-noise ratio to $\sim 10^4$. Therefore, if the contribution of an element to the total absorption is only a few percent, the XAFS signal, which is only a few percent of that element's absorption, will be difficult to extract from the transmission signal.

In the fluorescence method, the fluorescence photons emitted by the element of interest are detected (Jacklevic et al., 1977). The probability of fluorescence is essentially constant with energy, and the fluorescence intensity is, therefore, directly proportional to the absorption of a specific atomic species. The advantage is that each element has a characteristic fluorescence energy, and in principle the fluorescence from the element under study can be measured separately from the total absorption. For an ideal detector the signal-to-noise ratio would then be independent of the concentration, and would depend only on the number of photons collected. Of course, the number of fluorescence photons that can be collected does depend on concentration, and there would be practical limits. In actuality the real limits are determined by the efficiency of the detectors. It is difficult to achieve a high degree of background discrimination along with a high efficiency of collection. Figure 9 shows a schematic of the energy spectrum from a fluorescence sample. The main fluorescence line is $\sim 10\%$ lower than the edge energy. The main background peak is from the elastic scattered incident photons. There are also Compton-scattered photons, which are energy-shifted slightly below the elastic peak. The width and shape of the Compton scattering peak is determined by the acceptance angle of the detector. In multicomponent samples there can also be fluorescence lines from other sample components.

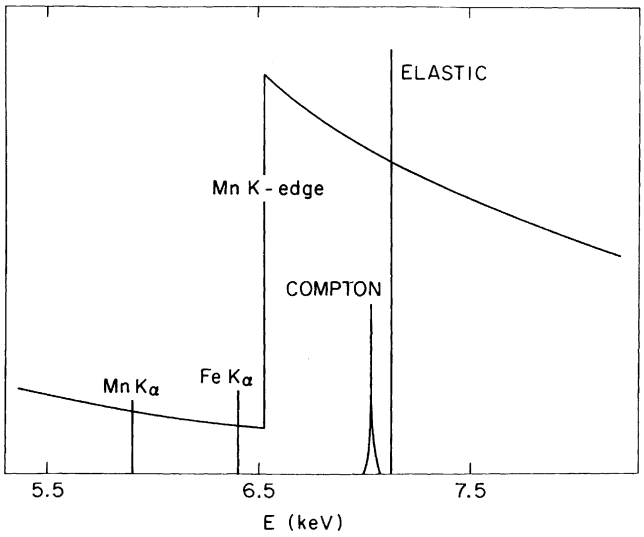


Figure 9. Schematic of the fluorescence spectrum from an Fe-containing sample with the incident energy set at the Fe K edge. The details of the Compton peak depend on the solid angle and location of the detector. This does not include complications such as fluorescence lines from other components in the sample. Also shown is the Mn K-edge absorption as an example of how filters can be used to reduce the Compton and elastic background.

To obtain a reasonable fluorescence spectrum, it is typically necessary to collect at least 10^6 photons in each energy channel. This would give a fractional statistical error of 10^{-3} for the total signal and 1% to 10% in the EXAFS. If the fluorescence signal N_f is accompanied by a background N_b, which cannot be separated by the detector, then the signal-to-noise ratio is equivalent to a signal $N_f/(1+A)$, where $A = N_b/N_f$. Therefore, a detector that is inefficient at separating the background can substantially increase the counts required.

There are three basic approaches to fluorescence detection: solid-state detectors, filters and slits (Stern and Heald, 1979), and crystal spectrometers. High-resolution Ge or Si detectors can provide energy resolution sufficient to separate the fluorescence peak. The problem is that their total counting rate (signal plus all sources of background) is limited to somewhere in the range of 2×10^4 to 5×10^5 counts per second depending on the type of detector and electronics used. Since the fluorescence from a dilute sample can be a small fraction of the total, collecting enough photons for good statistics can be time consuming. One solution has been the development of multichannel detectors with up to 100 individual detector elements (Pullia et al., 1997).

The other two approaches try to get around the count rate limitations of solid-state detectors by using other means of energy resolution. Then a detector such as an ion chamber can be used that has essentially no count rate limitations (saturation occurs at signal levels much greater than from a typical sample). The principle behind filters is shown in Figure 9. It is often possible to find a material whose absorption edge is between the fluorescence line and the elastic background peak. This will selectively attenuate the background. There is one problem with this approach: the filter will reemit the absorbed photons as its own fluorescence. Many of these can be prevented from reaching the detector by using appropriate slits, but no filter slit system can be perfect. The selectivity of a crystal spectrometer is nearly ideal since Bragg reflection is used to select only the fluorescence line. However, the extreme angular selectivity of Bragg reflection means that it is difficult to collect a large solid angle, and many fluorescent photons are wasted. Thus, the filter with a slit detector trades off energy selectivity to collect large solid angles, while the crystal spectrometer achieves excellent resolution at the expense of solid angles.

Which approach is best will depend on the experimental conditions. For samples with moderate diluteness (cases where the background ratio $A <\sim 20$), the filter slit is an excellent choice, providing simple setup and alignment, ease of changing the fluorescence energy, and good collection efficiency. For more dilute samples or samples with multiple fluorescence lines that cannot be filtered out, the standard method is to use a multielement detector. These can often collect a substantial solid angle and with the right control software can be fairly easy to set up. As sources become more intense, the count rate limitations of solid-state detectors become more serious. This has spurred renewed interest in developing either faster detectors or those with more elements (Gauthier et al., 1996; Pullia et al., 1997) and more efficient crystal detectors. The potential user would be advised to discuss with facility managers the performance of the detectors they have available. Since x-ray cross-sections are well known, it is usually possible to estimate the relative merits of the different detection methods for a particular experiment.

The emission of a fluorescence photon is only one possible outcome of an absorption event. Another, which is dominant for light elements, is the emission of Auger electrons (see AUGER ELECTRON SPECTROSCOPY). The electron yield detection method takes advantage of this. There are several schemes for detecting the electron emission (Stohr, 1988), but the total electron yield method is by far the most widely used. This involves measuring the total electron current emitted by the sample, usually by using a positively biased collecting electrode. It is also possible to measure the sample drain current, which in effect uses the sample chamber as the collecting electrode. The electron yield method can be performed in a vacuum, which makes it an attractive choice for surface XAFS experiments carried out in ultra-high-vacuum (UHV) chambers. The yield signal is composed of the excited photoelectron, directly emitted Auger electrons, and numerous secondary electrons excited by the primary Auger or photoelectrons. Since the electron mean free path in solids is very short, all of these electron signals originate from the surface region (typically within 10 to 100 nm of the surface). This effectively discriminates against absorption occurring deep within a sample and allows for monolayer sensitivity in UHV surface experiments.

For samples that do not need UHV conditions, it is also possible to use a gas such as He or H_2 to enhance the electron signal (Kordesch and Hoffman, 1984). In this case, the emitted electrons ionize the gas atoms or molecules, which are then detected as in an ionization chamber. It typically

requires ~30 eV to ionize a gas atom, and the electrons can have energies up to several keV. Thus many gas ions can be produced for each absorption event, giving an effective amplification of the signal current. The total yield method is often applied to concentrated samples that are difficult to prepare in a thin uniform manner. It is especially useful for avoiding thickness effects in samples with strong white lines, as will be discussed below (see Data Analysis and Initial Interpretation).

Synchrotron Facilities

Most XAFS experiments are carried out at synchrotron beamlines over which the user typically has little direct control. To carry out an effective experiment, it is important to understand some of the common beamline characteristics. These must be measured or obtained from the beamline operator.

Energy Resolution

Bragg reflection by silicon crystals is the nearly universal choice for monochromatizing the x rays at energies >2 keV. These crystals are nearly always perfect and, if thermal or mounting strains are avoided, have a very well defined energy resolution. For Si(111), the resolution is $\Delta E/E = 1.3 \times 10^{-4}$. Higher-order reflections such as Si(220) and Si(311) can have even better resolution (5.6×10^{-5} and 2.7×10^{-5}, respectively). This is the resolution for an ideally collimated beam. All beamlines have some beam divergence, $\Delta\theta$. Taking the derivative of the Bragg condition, this will add an additional contribution $\Delta E = \Delta\theta \cot(\theta_b)$ where θ_b is the monochromator crystal Bragg angle. Typically, these two contributions can be added in quadrature to obtain the total energy spread in the beam. The beam divergence is determined by the source and slit sizes and can also be affected by any focusing optics that precede the monochromator.

At lower energies, other types of monochromators will likely be employed. These are commonly based on gratings but can also use more exotic crystals such as quartz, InSb, beryl, or YB_{66}. In these cases, it is generally necessary to obtain the resolution information from the beamline operator.

Harmonic Content

As the later discussion on thickness effects will make clear, it is extremely important to understand the effects of harmonics on the experiment planned. These are most serious for transmission experiments but can be a problem for the other methods as well. Both crystals and gratings can diffract energies that are multiples of the fundamental desired energy. These generally need to be kept to a low level for accurate measurements. Often the beamline mirrors can be used to reduce harmonics if their high-energy cutoff is between the fundamental and harmonic energies. When this is not possible (or for unfocused beamlines), detuning of crystal monochromators is another common method. Nearly all scanning monochromators use two or more crystal reflections to allow energy selection while avoiding wide swings in direction of the output beam. Typically the harmonic reflection curves are much narrower than for the fundamental. This means that by setting the diffraction angles of the two crystals to slightly different values the intensity of the harmonic reflections can be dramatically reduced while only losing 20% to 50% of the fundamental intensity. Nearly all XAFS double-crystal monochromators allow for this possibility.

Detuning only works well for ideal crystals and for higher energies. It is generally good practice to determine the harmonic content of the beam if it is not already known. The simplest method is to insert a foil whose edge is near the harmonic energy. For Si(111) the second harmonic is forbidden, so this would involve choosing a material whose edge is ~3 times the desired energy. When the monochromator is scanned through the desired energy range, the harmonic content can be accurately estimated from the edge step of this test sample. Other methods include using an energy-analyzing detector such as a solid-state detector or Bragg spectrometer.

Energy Calibration

All mechanical systems can have some error, and it is important to verify the energy calibration of the beamline. For a single point this can be done using a standard foil containing the element planned for study. It is also important that the energy scale throughout the scan is correct. Most monochromators rely on gear-driven rotation stages, which can have some nonlinearity. This can result in nonlinearity in the energy scale. To deal with this, many modern monochromators employ accurate angle encoders. If an energy scale error is suspected, a good first check is to measure the energy difference between the edges of two adjacent materials and compare with published values. Unfortunately it is sometimes difficult to determine exactly which feature on the edge corresponds to the published value. Another approach is the use of Bragg reflection calibrators, available at some facilities.

Monochromator Glitches

An annoying feature of perfect crystal monochromators is the presence of sharp features in the output intensities, often referred to as "glitches." These are due to the excitation of multiple-diffraction conditions at certain energies and angles. Generally, these are sharp dips in the output intensity due to a secondary reflection removing energy from the primary. For ideal detectors and samples these should cancel out, but for XAFS measurements cancellation the 10^{-4} level is needed to make them undetectable. This is difficult to achieve. It is especially difficult for unfocused beamlines since the intensity reduction may only affect a part of the output beam, and then any sample nonuniformity will result in noncancellation. A sharp glitch that affects only one or two points in the spectrum can generally be handled in the data analysis. Problematic glitches extend over a broader energy range. These can be minimized only by operating the detectors at optimum linearity and making samples as uniform as possible. Usually a particular set of crystals will have only a few problematic glitches, which should be known to the beamline operator. The presence of glitches makes it essential

that the I_0 spectra be recorded separately. Then it can be assured that weak features in the spectra are not spurious by checking for correlation with glitches.

METHOD AUTOMATION

Most XAFS facilities are highly automated. At synchrotron facilities the experiments are typically located inside radiation enclosures to which access is forbidden when the beam is on. Therefore, it is important to automate sample alignment. Usually, motorized sample stages are provided by the facility, but for large or unusual sample cells these may not be satisfactory. It is important to work with the facility operators to make sure that sample alignment needs are provided for. Otherwise tedious cycles of entering the enclosure, changing the sample alignment, interlocking the radiation enclosure, and measuring the change in the signal may be necessary.

Since beamlines are complex and expensive, it is usually necessary that the installed control system be used. This may or may not be flexible enough to accommodate special needs. While standard experiments will be supported, it is difficult to anticipate all the possible experimental permutations. Again, it is incumbent upon the users to communicate early with the facility operators regarding special needs. This point has already been made several times but, to avoid disappointment, cannot be emphasized enough.

DATA ANALYSIS AND INITIAL INTERPRETATION

XAFS data analysis is a complex topic that cannot be completely covered here. There are several useful reviews in the literature (Teo, 1986; Sayers and Bunker, 1988; Heald and Tranquada, 1990). There are also a number of software packages available (see Internet Resources listing below), which include their own documentation. The discussion here will concentrate on general principles and on the types of preliminary analysis that can be conducted at the beamline facility to assess the quality of the data.

There are four main steps in the analysis: normalization, background removal, Fourier transformation and filtering, and data fitting. The proper normalization is given in Equation 3. In practice, it is simpler to normalize the measured absorption to the edge step. For sample thickness x,

$$\chi_{\exp}(k) = \frac{\mu(k)x - \mu_0(k)x}{\Delta\mu(0)x} \qquad (11)$$

where $\Delta\mu(0)x$ is the measured edge step. This can be determined by fitting smooth functions (often linear) to regions above and below the edge. This step normalization is convenient and accurate when it is consistently applied to experimental data being compared. To compare with theory, the energy dependence of $\mu_0(k)$ in the denominator must be included. This can be determined from tabulated absorption coefficients (McMaster et al., 1969; see Internet Resources). For data taken in fluorescence or electron yield, it is also necessary to compensate for the energy dependence of the detectors. For a gas-filled ionization chamber used to monitor I_0, the absorption coefficient of the gas decreases with energy. This means the I_0 signal would decrease with energy even if the flux were constant. Again, tabulated coefficients can be used. The fluorescence detector energy dependence is not important since it is detecting the constant-energy fluorescence signal. This correction is not needed for transmission since the log of the ratio is analyzed. For this case, the detector energy dependence becomes an additive factor, which is removed in the normalization and background removal process.

Background subtraction is probably the most important step. The usual method is to fit a smooth function such as a cubic spline through the data. The fitting parameters must be chosen to only fit the background while leaving the oscillations untouched. This procedure has been automated to a degree in some software packages (Cook and Sayers, 1981; Newville et al., 1993), but care is needed. Low-Z backscatterers are often problematic since their oscillation amplitude is very large at low k and damps out rapidly. It is then difficult to define a smooth background function that works over the entire k region in all cases. Often careful attention is needed in choosing the low-Z termination point. The background subtraction stage is also useful in pointing out artifacts in the data. The background curve should be reasonably smooth. Any unusual oscillation or structure should be investigated. One natural cause of background structure is multielectron excitations. These can be ignored in many cases but may be important for heavier atoms.

Once the background-subtracted $\chi(k)$ is obtained, its Fourier transform can be taken. As shown in Figure 6, this makes visible the different frequency components in the spectra. It also allows a judgment of the success of the background subtraction. Poor background removal will result in strong structure at low r. However, not all structure at low r is spurious. For example, the rapid decay of low-Z backscatterers can result in a low-r tail on the transform peak. To properly judge the source of the low-r structure, it is useful to transform theoretical calculations of similar cases for comparison.

Before Fourier transforming, the data is usually multiplied by a factor k^w, W = 1, 2, or 3. This is done to sharpen the transform peaks. Generally the transform peaks will be sharpest if the oscillation amplitude is uniform over the data window being transformed. Often, W = 1 is used to compensate for the $1/k$ factor seen in Equation 3. For low-Z atoms and systems with large disorder, higher values of W can be used to compensate for the additional k-dependent fall-off. High values of W also emphasize the higher-k part of the spectrum where the heavier backscatterers have the largest contribution. Thus, comparing transforms with different k weighting is a simple qualitative method of determining which peaks are dominated by low- or high-Z atoms. In addition to k weighting, other weighting can be applied prior to transforming. The data should only be transformed over the range for which there is significant signal. This can be accomplished by truncating $\chi(k)$ with a rectangular window prior to transforming. A sharp rectangular window can induce truncation ripples

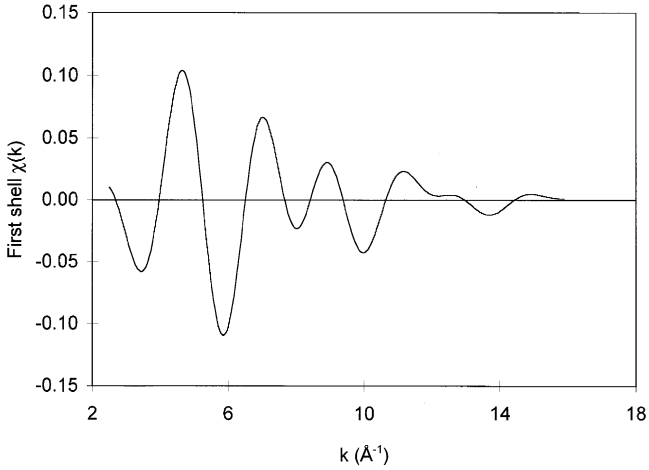


Figure 10. Inverse Fourier transform of the region 0.8 to 2 Å for the Cu data in Figure 6. This region contains contributions from three different Cu-O distances in the first coordination sphere.

in the transform. To deal with this, most analysis packages offer the option of various window functions that give a more gradual truncation. There are many types of windows, all of which work more or less equivalently.

The final stage of analysis is to fit the data either in k or r space using theoretical or experimentally derived functions. If k-space fitting is done, the data are usually filtered to extract only the shells being fit. For example, if the first and second shells are being fit in k space, an r-space window is applied to the r-space data, which includes only the first- and second-shell contributions, and then the data are transformed back to k space. This filtered spectra only includes the first and second shell contributions, and it can be fit without including all the other shells. An example of filtered data is shown in Figure 10. An important concept in fitting filtered data is the number of independent data points (Stern, 1993). This is determined by information theory to be $N_1 = 2\Delta k \Delta r/\pi + 2$, where Δk and Δr are the k- and r-space ranges of the data used for analysis. If fitting is done in r space, Δr would be the range over which the data are fit, and Δk is the k range used for the transform. The number of independent parameters N_I is the number that are needed to completely describe the filtered data. Therefore, any fitting procedure should use fewer than N_I parameters.

The relative merits of different types of fitting are beyond the scope of this unit, and fitting is generally not carried out at the beamline while data is being taken. It is important, however, to carry the analysis through the Fourier transform stage to assess the data quality. As mentioned, systematic errors often show up as strange backgrounds or unphysical transform peaks, and random statistical errors will show up as an overall noise level in the transform.

The most widely used theoretical program is FEFF (Zabinsky et al., 1995). It is designed to be relatively portable and user friendly and includes multiple scattering. It is highly desirable for new users to obtain FEFF or an equivalent package (Westre et al., 1995) and to calculate some example spectra for the expected structures.

The most common causes of incorrect data can all be classified under the general heading of thickness effects or thickness-dependent distortions. A good experimental check is to measure two different thicknesses of the same sample: if thickness effects are present, the normalized XAFS amplitude for the thicker sample will be reduced. Thickness effects are caused by leakage radiation, such as pinholes in the sample, radiation that leaks around the sample, or harmonic content in the beam. All of these are essentially unaffected by the sample absorption. As the x-ray energy passes through the edge, the transmission of the primary radiation is greatly reduced while the leakage is unchanged. Thus, the percentage of leakage radiation in the transmitted radiation increases in passing through the edge, and the edge step is apparently reduced. The greater the absorption, the greater is this effect. This means the peaks in the absorption are reduced in amplitude even more than the edge if leakage is present. In the normalized spectrum such peaks will be reduced. The distortion is larger for thick samples, which have the greatest primary beam attenuation. It can be shown statistically that the optimum thickness for transmission samples in the absence of leakage is about $\mu x = 2.5$. In practice, a $\mu x \sim 1.5$ is generally a good compromise between obtaining a good signal and avoiding thickness effects.

An analogous effect occurs in fluorescence detection. It is generally referred to as self-absorption. As the absorption increases at the edge, the penetration of the x rays into the sample decreases, and fewer atoms are available for fluorescing. Again, peaks in the absorption are reduced. Obviously this is a problem only if the absorption change is significant. For very dilute samples the absorption change is too small for a noticeable effect. Self-absorption is the reason that fluorescence detection should not be applied to concentrated samples. In this context "concentrated" refers to samples for which the absorption step is $>\sim$10% of the total absorption. When sample conditions preclude transmission measurements on a concentrated sample, electron yield detection is a better choice than fluorescence detection. For electron yield detection, the signal originates within the near-surface region where the x-ray penetration is essentially unaffected by the absorption changes in the sample. Thus, in electron yield detection, self-absorption can almost always be ignored.

SAMPLE PREPARATION

XAFS samples should be as uniform as possible. This can be a significant challenge. Typical absorption lengths for concentrated samples are from 3 to 30 μm. Foils or liquid solutions are a simple approach but cannot be universally applied. It can be difficult to grind and disperse solid particles into uniform layers of appropriate thickness. For example, to achieve a 10-μm layer of a powdered sample, the particle size should be of order 2 to 3 μm. This is difficult to achieve for many materials. Some common sample preparation methods include rubbing the powder into low-absorption adhesive tape or combining the powder with a low-absorption material such as graphite or BN prior

to pressing into pellets. The tape method has been especially successful, probably because the rubbing process removes some of the larger particles leaving behind a smaller average particle size adhered to the tape. Nevertheless, if less than about four layers of tape are required to produce a reasonable sample thickness, then the samples are likely to be unacceptably nonuniform. XAFS signals can be significantly enhanced by cooling samples to liquid nitrogen temperatures. This can also be a constraint on sample preparation since tape-prepared samples can crack, exposing leakage paths, upon cooling. This can be avoided by using Kapton-based tape, although the adhesive on such tapes can have significant absorption.

Materials with layered or nonsymmetric structures can result in anisotropic particles. These will tend to be oriented when using the tape or pellet method and result in an aligned sample. Since all synchrotron radiation sources produce a highly polarized beam, the resulting data will show an orientation dependence. This can be accounted for in the analysis if the degree of orientation is known, but this can be difficult to determine. For layered materials, one solution is to orient the sample at the "magic" angle (54.7°) relative to the polarization vector. This will result in a signal equivalent to an unoriented sample for partially oriented samples.

Sample thickness is less of a concern for fluorescence or electron yield samples, but it is still important that the samples are uniform. For electron yield, the sample surface should be clean and free of thick oxide layers or other impurities. Sample charging can also be a factor. This can be a major problem for insulating samples in a vacuum. It is less important when a He-filled detector is used. Also, the x-ray beam is less charging than typical electron beam techniques, and fairly low conductivity samples can still be successfully measured.

SPECIMEN MODIFICATION

In general, x rays are less damaging than particle probes such as the electrons used in electron microscopy. However, samples can be modified by radiation damage. This can be physical damage resulting in actual structural changes (usually disordering) or chemical changes. This is especially true for biological materials such as metalloproteins. Metals and semiconductors are generally radiation resistant. Many insulating materials will eventually suffer some damage, but often on timescales that are long compared to the measurement. Since it is difficult to predict which materials will be damaged, for potentially sensitive materials, it is wise to use other characterization techniques to verify sample integrity after exposure or to look for an exposure dependence in the x-ray measurements.

PROBLEMS

The two most common causes of incorrect data are sample nonuniformity and thickness effects, which have already been discussed. Other problems can occur that are exacerbated by the need to run measurements at remote facilities under the time constraints of a fixed schedule. The critical role played by beam harmonics has already been discussed. Eliminating harmonics requires the correct setting of beamline crystals or mirrors. It is easy to do this incorrectly at a complex beamline. Another area of concern is the correct operation of detectors. The high fluxes at synchrotrons mean that saturation or deadtime effects are often important. For ion-chamber detectors it is important to run the instrument at the proper voltage and to use the appropriate gases for linear response. Similarly, saturation must be avoided in the subsequent amplifier chain. Single-photon detectors, if used, are often run at count rates for which deadtime corrections are necessary. For fluorescence experiments this means that the incoming or total count rate should be monitored as well as the fluorescence line(s) of interest.

Guidance on many of these issues can be obtained from the beamline operators for typical running conditions. However, few experiments are truly typical in all respects, and separate tests may need to be made. In contrast to laboratory experiments, the time constraints of synchrotron experiments make it tempting to skip some simple tests needed to verify that undistorted data is being collected. This should be avoided if at all possible.

LITERATURE CITED

Ashley, C. A. and Doniach, S. 1975. Theory of extended x-ray absorption edge fine structure (EXAFS) in crystalline solids. *Phys. Rev. B* 11:1279–1288.

Azaroff, L. V. and Pease, D. M. 1974. X-ray absorption spectra. *In* X-ray Spectroscopy (L. V. Azaroff, ed.). McGraw-Hill, New York.

Beni, G. and Platzman, P. M. 1976. Temperature and polarization dependence of extended x-ray absorption fine-structure spectra. *Phys. Rev. B* 14:1514–1518.

Bianconi, A. 1988. XANES spectroscopy. *In* X-ray Absorption (D. C. Koningsberger and R. Prins, eds.). pp. 573–662. John Wiley & Sons, New York.

Boyce, J. B., Hayes, T. M., and Mikkelsen, J. C. 1981. Extended-x-ray-absorption-fine-structure of mobile-ion density in superionic AgI, CuI, CuBr, and CuCl. *Phys. Rev. B* 23:2876–2896.

Bunker, G. 1983. Application of the ratio method of EXAFS analysis to disordered systems. *Nucl. Instrum. Methods* 207:437–444.

Cook, J. W. and Sayers, D. E. 1981. Criteria for automatic x-ray absorption fine structure background removal. *J. Appl. Physiol.* 52:5024–5031.

Crozier, E. D., Rehr, J. J., and Ingalls, R. 1988. Amorphous and liquid systems. *In* X-ray Absorption (D. C. Koningsberger and R. Prins, eds.). pp. 373–442. John Wiley & Sons, New York.

Flank, A. M., Fontaine, A., Jucha, A., Lemonnier, M., and Williams, C. 1982. Extended x-ray absorption fine structure in dispersive mode. *J. Physique* 43:L-315-L-319.

Frahm, R. 1988. Quick scanning EXAFS: First experiments. *Nucl. Instrum. Methods* A270:578–581.

Gauthier, Ch., Goulon, J., Moguiline, E., Rogalev, A., Lechner, P., Struder, L., Fiorini, C., Longoni, A., Sampietro, M., Besch, H.,

Pfitzner, R., Schenk, H., Tafelmeier, U., Walenta, A., Misiakos, K., Kavadias, S., and Loukas, D. 1996. A high resolution, 6 channels drift detector array with integrated JFET's designed for XAFS spectroscopy: First x-ray fluorescence excitation recorded at the ESRF. *Nucl. Instrum. Methods* A382:524–532.

Heald, S. M. and Stern, E. A. 1977. Anisotropic x-ray absorption in layered compounds. *Phys. Rev. B* 16:5549–5557.

Heald, S. M. and Tranquada, J. M. 1990. X-ray absorption spectroscopy: EXAFS and XANES. *In* Physical Methods of Chemistry, Vol. V (Determination of Structural Features of Crystalline and Amorphous Solids) (B.W. Rossiter and J.F. Hamilton, eds.). pp. 189–272. Wiley-Interscience, New York.

Jacklevic, J., Kirby, J. A., Klein, M. P., Robinson, A. S., Brown, G., and Eisenberger, P. 1977. Fluorescence detection of EXAFS: Sensitivity enhancement for dilute species and thin films. *Sol. St. Commun.* 23:679–682.

Kordesch, M. E. and Hoffman, R. W. 1984. Electron yield extended x-ray absorption fine structure with the use of a gas-flow detector. *Phys. Rev. B* 29:491–492.

Lee, P. A. 1976. Possibility of adsorbate position determination using final-state interference effects. *Phys. Rev. B* 13:5261–5270.

Lee, P. A. and Pendry, J. B. 1975. Theory of the extended x-ray absorption fine structure. *Phys. Rev. B* 11:2795–2811.

Lengeler, B. 1986. Interaction of hydrogen with impurities in dilute palladium alloys. *J. Physique* C8:1015–1018.

Matsushita, T. and Phizackerley, R. P. 1981. A fast x-ray spectrometer for use with synchrotron radiation. *Jpn. J. Appl. Phys.* 20:2223–2228.

McMaster, W. H., Kerr Del Grande, N., Mallett, J. H., and Hubbell, J. H. 1969. Compilation of x-ray cross sections, LLNL report, UCRL-50174 Section II Rev. 1. National Technical Information Services L-3, U.S. Department of Commerce.

Newville, M., Livins, P., Yacoby, Y., Stern, E. A., and Rehr, J. J. 1993. Near-edge x-ray-absorption fine structure of Pb: A comparison of theory and experiment. *Phys. Rev. B* 47:14126–14131.

Pullia, A., Kraner, H. W., and Furenlid, L. 1997. New results with silicon pad detectors and low-noise electronics for absorption spectrometry. *Nucl. Instrum. Methods* 395:452–456.

Rechav, B., Yacoby Y., Stern, E. A., Rehr, J. J., and Newville, M. 1994. Local structural distortions below and above the antiferrodistortive phase transition. *Phys. Rev. Lett.* 69:3397–3400.

Rose, M. E. and Shapiro, M. M. 1948. Statistical error in absorption experiments. *Physiol. Rev.* 74:1853–1864.

Sayers, D. E. and Bunker, B. A. 1988. Data analysis. *In* X-ray Absorption (D. C. Koningsberger and R. Prins, eds.). pp. 211–253. John Wiley & Sons, New York.

Sayers, D. E., Stern, E. A., and Lytle, F. W. 1971. New technique for investigating non-crystalline structures: Fourier analysis of the extended x-ray absorption fine structure. *Phys. Rev. Lett.* 27:1204–1207.

Stern, E. A. 1993. Number of relevant independent points in x-ray-absorption fine-structure spectra. *Phys. Rev. B* 48:9825–9827.

Stern, E. A. and Heald, S. M. 1979. X-ray filter assembly for fluorescence measurements of x-ray absorption fine structure. *Rev. Sci. Instrum.* 50:1579–1582.

Stern, E. A., Ma, Y., Hanske-Pettipierre, O., and Bouldin, C. E. 1992. Radial distribution function in x-ray-absorption fine structure. *Phys. Rev. B* 46:687–694.

Stohr, J. 1988. SEXAFS: Everything you always wanted to know. *In* X-ray Absorption (D. C. Koningsberger and R. Prins, eds.). pp. 443–571. John Wiley & Sons, New York.

Stumm von Bordwehr, R. 1989. A history of x-ray absorption fine structure. *Ann. Phys. Fr.* 14:377–466.

Teo, B. K. 1986. EXAFS: Basic Principles and Data Analysis. Springer-Verlag, Berlin.

Westre, T. E., Dicicco, A., Filipponi, A., Natoli, C. R., Hedman, B., Soloman, E. I., and Hodgson, K. O. 1995. GNXAS, a multiple-scattering approach to EXAFS analysis—methodology and applications to iron complexes. *J. Am. Chem. Soc.* 117:1566–1583.

Zabinsky, S. I., Rehr, J. J., Ankudinov, A., Albers, R. C., and Eller, M. J. 1995. Multiple scattering calculations of x-ray absorption spectra. *Phys. Rev. B* 52:2995–3009.

KEY REFERENCES

Goulon, J., Goulon-Ginet, C., and Brookes, N. B. 1997. Proceedings of the Ninth International Conference on X-ray Absorption Fine Structure, *J. Physique* 7 Colloque 2.

A good snapshot of the current status of the applications and theory of XAFS. See also the earlier proceedings of the same conference.

Koningsberger, D. C. and Prins, P. 1988. X-ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS and XANES. John Wiley & Sons, New York.

A comprehensive survey of all aspects of x-ray absorption spectroscopy. Slightly dated on some aspects of the calculation and analysis of multiple-scattering contributions, but a very useful reference for serious XAFS practitioners.

Stohr, J. 1992. NEXAFS Spectroscopy. Springer-Verlag, New York.

More details on the use and acquisition of near-edge spectra, especially as they apply to surface experiments.

INTERNET RESOURCES

http://ixs.csrri.iit.edu/index.html

International XAFS Society homepage, containing much useful information including upcoming meeting information, an XAFS database, and links to many other XAFS-related resources.

http://www.aps.anl.gov/offsite.html

A list of synchrotron facility homepages maintained by the Advanced Photon Source at Argonne National Laboratory, one of several similar Web sites.

http://www-cxro.lbl.gov/optical_constants/

Tabulation of absorption coefficients and other x-ray optical constants maintained by the Center for X-Ray Optics at the Lawrence Berkeley National Laboratory.

http://www.esrf.fr/computing/expg/subgroups/theory/xafs/xafs_software.html

A compilation of available XAFS analysis software maintained by the European Synchrotron Radiation Facility (ESRF).

STEVE HEALD
Pacific Northwest National Laboratory
Richland, Washington

X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS

INTRODUCTION

Diffuse scattering from crystalline solid solutions is used to measure local compositional order among the atoms, dynamic displacements (phonons), and mean species-dependent static displacements. In locally ordered alloys, fluctuations of composition and interatomic distances break the long-range symmetry of the crystal within local regions and contribute to the total energy of the alloy (Zunger, 1994). Local ordering can be a precursor to a lower temperature equilibrium structure that may be unattainable because of slow atomic diffusion. Discussions of the usefulness of local chemical and displacive correlations within alloy theory are given in Chapter 2 (see PREDICTION OF PHASE DIAGRAMS and COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS). In addition to local atomic correlations, neutron diffuse scattering methods can be used to study the local short-range correlations of the magnetic moments. Interstitial defects, as opposed to the substitutional disorder defects described above, also disrupt the long-range periodicity of a crystalline material and give rise to diffusely scattered x rays, neutrons, and electrons (electron scattering is not covered in this unit; Schweika, 1998).

Use of tunable synchrotron radiation to change the x-ray scattering contrast between elements has greatly improved the measurement of bond distances between the three types of atom pairs found in crystalline binary alloys (Ice et al., 1992). The estimated standard deviation of the first-order (first moment) mean static displacements from this technique approaches ± 0.001 Å (0.0001 nm), which is an order of magnitude more precise than results obtained with extended x-ray absorption fine structure (EXAFS; XAFS SPECTROSCOPY) measurements. In addition, both the radial and tangential displacements can be reliably determined to five or more near-neighbor shells (Jiang et al., 1996). In a binary A-B alloy, the number of A or B near-neighbor atoms to, for example, an A atom can be determined to even less than 1 atom in 100. The second moment of the static displacements, which gives rise to Huang scattering, is also measurable (Schweika, 1998). Measurements of diffuse scattering can also reveal the tensorial displacements associated with substitutional and interstitial defects. This information provides models of the average arrangements of the atoms on a local scale.

An example of chemical local ordering is given in Figure 1, where the probability P^{AB}_{lmn} of finding a B atom out to the sixth lmn shell around an A atom goes from a preference for A atoms (clustering) to a preference for B atoms (short-range order) for a body-centered cubic (bcc) $A_{50}B_{50}$ alloy. The real-space representation of the atom positions is derived from a Monte Carlo simulation of the P^{AB}_{lmn} values (Gehlen and Cohen, 1965) from the measurement of the intensity distribution in reciprocal space (Robertson et al., 1998). In the upper panel, the probability of finding a B atom as the first neighbor to an A atom is 40% (10% clustering; $P^{AB}_{111} = 0.4$). Atoms are located on a (110) plane so that first-neighbor pairs are shown. The middle panel depicts the random alloy where $P^{AB}_{lmn} = -0.5$ for the first six shells (lmn). The lower panel shows the case where $P^{AB}_{lmn} = 0.6$ (a preference for unlike atom pairs). The intensity distribution in the (100) plane of reciprocal space (with the fundamental Bragg maxima removed) is shown in the right column of Figure 1. Note that a preference for like nearest neighbors causes the scattering to be centered near the fundamental Bragg maxima, such as at the origin, 0,0. A preference for unlike first-neighbor pairs causes the diffuse scattering to peak at the superlattice reflections for an ordered structure. Models, such as those shown in Figure 1, are used to understand materials properties and their response to heat treatment, mechanical deformation, and magnetic fields. These local configurations are useful to test advances in theoretical models of crystalline alloys as discussed in COMPUTATION OF DIFFUSE INTENSITIES IN ALLOYS.

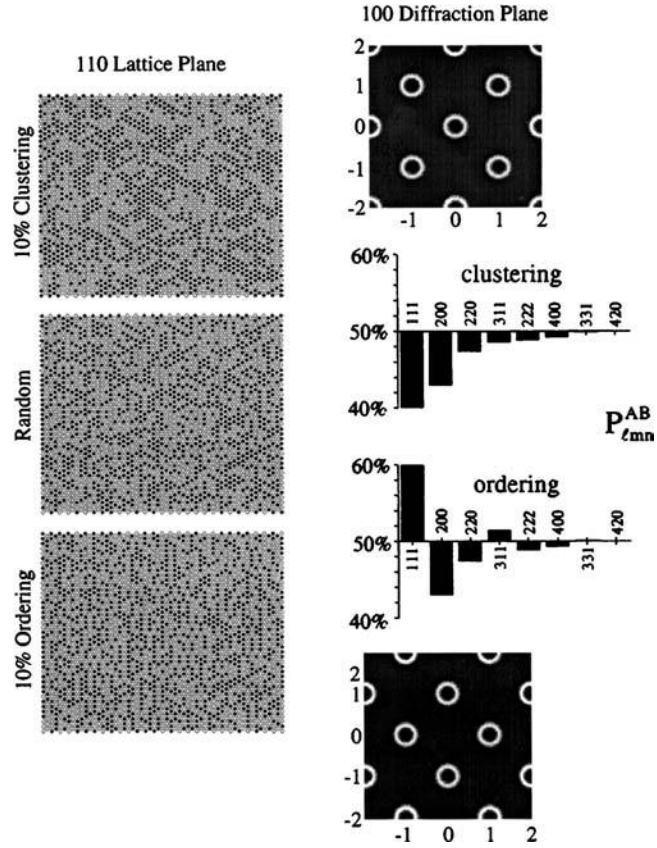


Figure 1. Direct and reciprocal space representations for a clustering, a random, and an ordering $A_{50}B_{50}$ bcc alloy. Courtesy of Robertson et al. (1998).

The diffraction pattern from a crystalline material with perfect periodicity, such as nearly perfect single-crystal Si, consists of sharp Bragg maxima associated with long-range periodicity. With Bragg's law, we can determine the size of the unit cell. Because of thermal motion, atom positions are smeared and Bragg maxima are reduced. In alloys with different atomic sizes, static displacements will also contribute to this reduction. This intensity, which is lost from the Bragg reflections, is diffusely distributed.

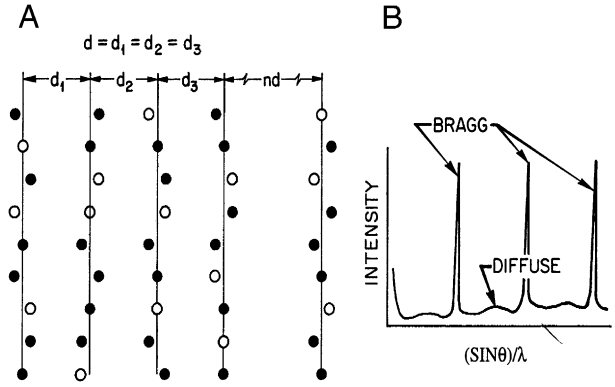


Figure 2. Displacements about the average lattice preserve the regular spacing between atomic planes such that $d = d_1 = d_2 = d_3 = \ldots$. The average lattice is obtained from the positions of the sharp Bragg reflections (**B**). Information about short-range correlations among the atoms is contained in the diffusely distributed intensity between the Bragg peaks. Courtesy of Ice et al. (1998).

Shown schematically in Figure 2A is a solid solution of two kinds of atoms displaced from the sites of the average lattice in such a way that the average plane of atoms is regularly spaced with a constant "*d*" spacing existing over hundreds of planes. As shown schematically in Figure 2B, there is weak diffuse scattering but no broadening of the fundamental Bragg reflections, as would be the case for more extended defects such as stacking faults, high dislocation densities, displacive transformations, and incoherent precipitates, among others (Warren, 1969). In cases where the fundamental Bragg reflections are broadened, our uncertainty in the size of the average lattice increases and the precision of the measured pair separation is reduced.

This unit will concentrate on the use of diffuse x-ray and neutron scattering from single crystals to measure local chemical correlations and chemically specific static displacements. Particular emphasis will be placed on the use of resonant (anomalous) x-ray techniques to extract information on atomic size from binary solid solutions with short-range order. Here the alloys have a well-defined average lattice but have local fluctuations in composition and displacements from the average lattice. In stoichiometric crystals with long-range periodicity, sharp superlattice Bragg reflections appear. If the compositional order is correlated only over short distances, the superlattice reflections are so broadened that measurements throughout a symmetry related volume in reciprocal space are required to determine its distribution. In addition, the displacement of the atom pairs (e.g., the A-A, A-B, and B-B pairs in a binary alloy) from the sites of the average lattice because of different atom sizes also contributes to the distribution of this diffuse scattering. By separating this diffuse intensity into its component parts—that associated with the chemical preference for A-A, A-B, and B-B pairs for the various near-neighbor shells and that associated with the static and dynamic displacements of the atoms from the sites of the average lattice—we are able to recover pair correlation probabilities for the three kinds of pairs in a binary alloy. The interpretation of diffuse scattering associated with dynamic displacements of atoms from their average crystal sites will be discussed only briefly in this unit.

Competitive and Related Techniques

Other techniques that measure local chemical order and bond distances exist. In EXAFS, outgoing photoejected electrons are scattered by the surrounding near neighbors (most from first and to a lesser extent from second nearest neighbors). This creates modulations of the x-ray absorption cross-section, typically extending for ~1000 eV above the edge, and gives information about both local chemical order and bond distances (see XAFS SPECTROSCOPY for details). Usually, the phase and amplitudes for the interference of the photo-ejected electrons must be extracted from model systems, which necessitates measurements on intermetallic (ordered) compounds of known bond distances and neighboring atoms. The choice of an incident x-ray energy specific to an elemental absorption edge makes EXAFS information specific to that elemental constituent. For an alloy of A and B atoms, the EXAFS for an absorption edge of an A atom would be sensitive to the A and B atoms neighboring the A atoms. Separation of the signal into A-A and B-B pairs is typically done by using dilute alloys containing 2 at.% or less of the constituent of interest (e.g., the A atoms). The EXAFS signal is then interpreted as arising from the predominately B neighborhood of an A atom, and analyzed in terms of the number of B first and second neighbors and their bond distances from the A atoms. Claims for the accuracy of the EXAFS method vary between 0.01 and 0.02 Å for bond distance and ~10% for the first shell coordination number (number of atoms in the first shell; Scheuer and Lengeler, 1991). For crystalline solid-solution alloys, the crystal structure precisely determines the number of neighbors in each shell but not the kinds for nondilute alloys. For most alloys, the precision achieved with EXAFS is inadequate to determine the deviations of the interatomic spacings from the average lattice. Whenever EXAFS measurements are applicable and of sufficient precision to determine the information of interest, the ease and simplicity of this experiment compared with three-dimensional diffuse scattering measurements makes it an attractive tool. An EXAFS study of concentrated Au-Ni alloys revealed the kind of information available (Renaud et al., 1988).

Mössbauer spectroscopy is another method for obtaining near-neighbor information (MOSSBAUER SPECTROMETRY). Measurements of hyperfine field splitting caused by changes in the electric field gradient or magnetic field because of the different charge or magnetic states of the nuclear environments give information about the near-neighbor environments. Different chemical and magnetic environments of the nucleus produce different hyperfine structure, which is interpreted as a measure of the different chemical environments typical for first and second neighbors. The quantitative interpretation of Mössbauer spectra in terms of local order and bond distances is often ambiguous. The use of Mössbauer spectroscopy is limited

to those few alloys where at least one of the constituents is a Mössbauer-active isotope. This spectroscopy is complimentary but does not compete with diffuse scattering measurements as a direct method for obtaining detailed information about near-neighbor chemical environments and bond distances (Drijver et al., 1977; Pierron-Bohnes et al., 1983).

Imaging techniques with electrons, such as Z contrast microscopy and other high-resolution electron microscopy techniques (see Chapter 11), do not have the resolution for measuring the small displacements associated with crystalline solid solutions. Imaging for high-resolution microscopy requires a thin sample about a dozen or more unit cells thick with identical atom occupations, and precludes obtaining information about short-range order and bond distances. Electron diffuse scattering measurements are difficult to record in absolute units and to separate from contributions to the diffuse background caused by straggling energy loss processes. Electron techniques provide extremely useful information on more extended defects as discussed in Chapter 11.

Field ion microscopy uses He or Ne gas atoms to image the small radius tip of the sample. Atom probes provide direct imaging of atom positions. Atoms are pulled from a small radius tip of the sample by an applied voltage and mass analyzed through a small opening. The position of the atom can be localized to ~5 Å. Information on the species of an atom and its neighbors can be recovered. Reports of successful analysis of concentration waves and clusters in phase separating alloys have occurred where strongly enriched clusters of like atoms are as small as 5 Å in diameter (Miller et al., 1996). However, information on small displacements cannot be obtained with atom probes. Scanning tunneling (see SCANNING TUNNELING MICROSCOPY) and atomic force microscopy can distinguish between the kinds of atoms on a surface and reveal their relative positions.

PRINCIPLES OF THE METHOD

In this section, we formulate the diffraction theory of diffuse scattering in a way that minimizes assumptions and maximizes the information obtained from a diffraction pattern without recourse to models. This approach can be extended with various theories and models for interpretation of the recovered information. Measurements of diffusely scattered radiation can reveal the kinds and number of defects. Since different defects give different signatures in diffuse scattering, separation of these signatures can simplify recovery of the phenomenological parameters describing the defect.

Availability of intense and tunable synchrotron x-ray sources, which allow the selection of x-ray energies near absorption edges, permits the use of resonant scattering techniques to separate the contribution to the diffuse scattering from different kinds of pairs (e.g., the A-A, A-B, and B-B pairs of a binary alloy). Near an x-ray K absorption edge, the x-ray scattering factor of an atom can change by ~8 electron units (eu) and allows for scattering contrast control between atoms in an alloy. Adjustable contrast, either through resonant (anomalous) x-ray scattering or through isotopic substitution for neutrons, allows for precision measurement of chemically specific local-interatomic distances within the alloy. Figure 3 gives the real-space notation used in describing the atom positions and its reciprocal space notations used in describing the intensity distribution.

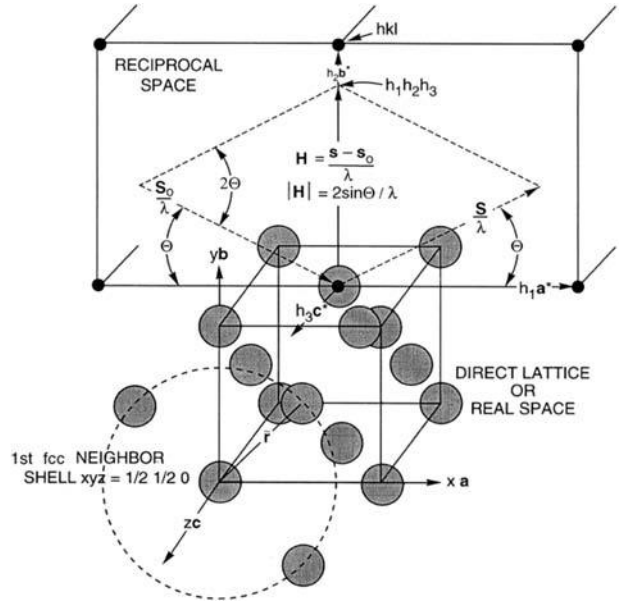


Figure 3. The atom positions for a face-centered cubic (fcc) structure are used to illustrate the notation for the real-space lattice. The unit cell has dimensions $\mathbf{a} = \mathbf{b} = \mathbf{c}$. The corresponding reciprocal space lattice is $\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*$. A position in reciprocal space at which the scattered intensity, $I(\mathbf{H})$, is measured for an incoming x ray in the direction of $\mathbf{S}_0$ of wavelength λ and detected in the outgoing direction of $\mathbf{S}$ would be $\mathbf{H} = (\mathbf{S}-\mathbf{S}_0)/\lambda = h_1\mathbf{a}^* + h_2\mathbf{b}^* + h_3\mathbf{c}^*$. At Bragg reflections, $h_1h_2h_3$ are integers and are usually designated hkl, the Miller indices of the reflection. This notation follows that used in the International Tables for Crystallography. Courtesy of Sparks and Robertson (1994).

X Rays Versus Neutrons

The choice of x rays or neutrons for a given experiment depends on instrumentation and source availability, the constituent elements of the sample, the information sought, the temperature of the experiment, the size of the sample, and isotopic availability and cost, among other considerations. Neutron scattering is particularly useful for measurements of low-Z materials, for high-temperature measurements, and for measurements of magnetic ordering. X-ray scattering is preferred for measurements of small samples, for measurement of static displacements, and for high $\mathbf{H}$ resolution.

Chemical Order

Recovery of the local chemical preference for atom neighbors has been predominately an x-ray diffuse scattering measurement, although x-ray and neutron measurements are complimentary. More than 50 systems have been studied with x rays and around 10 systems with neutrons.

The choice between x-ray and neutron methods often depends upon which one gives the best contrast between the constituent elements and which one allows the greatest control over contrast-isotopic substitution for contrast change with neutrons (Cenedese et al., 1984) or resonant (synonymous with dispersion and anomalous) x-ray techniques with x-ray energies near to absorption edge energies (Ice et al., 1994). In general, x-ray scattering is favored by virtue of its better intensity and collimation, which allow for smaller samples and better momentum-transfer resolution. Neutron diffuse scattering has the advantage of discriminating against thermal diffuse scattering (TDS); there is a significant change in energy (wavelength) when neutrons are scattered by phonons (see PHONON STUDIES). For example, phonons in the few tens of millielectron volt energy range make an insignificant change in the energy of kiloelectron volt x rays but make a significant change in the ~35-meV energy of thermal neutrons, except near Bragg reflections. Thus, TDS of neutrons is easily rejected with crystal diffraction or time-of-flight techniques even at temperatures approaching and exceeding the Debye temperature. With x rays, however, TDS can be a major contribution and can obscure the Laue scattering. Magnetic short-range order can also be studied with neutrons in much the same way as the chemical short-range order. However, when the alloy is magnetic, extra effort is needed to separate the magnetic scattering from the nuclear scattering that gives the information on chemical pair correlations.

Neutrons can be used in combination with x rays to obtain additional scattering contrast. The x-ray scattering factors increase with the atomic number of the elements (since it increases with the number of electrons), but neutron scattering cross-sections are not correlated with the atomic number as they are scattered from the nucleus. When an absorption edge of one of the elements is either too high for available x-ray energies or too low in energy for the reciprocal space of interest, or if enough different isotopes of the atomic species making up the alloy are not available or are too expensive, then a combination of both x-ray and neutron diffuse scattering measurements may be a way to obtain the needed contrast. A more in-depth discussion of the properties of neutrons is given in Chapter 13. Information on chemical short-range order obtained with both x rays and neutrons and a general discussion of their merits is given by Kostorz (1996).

Local chemical order among the atoms including vacancies has been measured for ~70 metallic binary alloys and a few oxides. Only two ternary metallic systems have been measured in which the three independent pair probabilities between the three kinds of atoms have been recovered: an alloy of $Cr_{21}Fe_{56}Ni_{23}$ with three different isotopic contents studied with neutron diffuse scattering measurements (Cenedese et al., 1984) and an alloy $Cu_{47}Ni_{29}Zn_{34}$ studied with three x-ray energies (Hashimoto et al., 1985).

Bond Distances

In binary alloys, recovery of the bond distances between A-A, A-B, and B-B pairs requires measurement of the diffuse intensity at two different scattering contrasts to separate the A-A and B-B bond distances (as shown, later the A-B bond distance can be expressed as a linear combination of the other two distances to conserve volume). Such measurements require two matched samples of differing isotopic content to develop neutron contrast, whereas for x rays, the x-ray energy need only be changed a fraction of a kilovolt near an absorption edge to produce a significant contrast change (Ice et al., 1994). Of course, one of the elemental constituents of the sample needs to have an absorption edge energy above ~5 keV to obtain sufficient momentum transfer for a full separation of the thermal contribution. However, more limited momentum-transfer data can be valuable for certain applications. Thus binary solid solutions where both elements have atomic numbers lower than that of Cr (Z=24) may require both an x-ray and a neutron measurement of the diffuse intensity to obtain data of sufficient contrast and momentum transfer. To date, there have been about 10 publications on the recovery of bond distances. Most have employed diffuse x-ray scattering measurements, but some employed neutrons and isotopically substituted samples (Müller et al., 1989).

Diffuse X-ray (Neutron) Scattering Theory for Crystalline Solid Solutions

In the kinematic approximation, the elastically scattered x-ray (neutron) intensity in electron units per atom from an ensemble of atoms is given by

$$I(\mathbf{H})_{\text{Total}} = \left| \sum_p f_p e^{2\pi i \mathbf{H}\bullet\mathbf{r}_p} \right|^2 = \sum_p \sum_q f_p f_q^* e^{2\pi i \mathbf{H}\bullet(\mathbf{r}_p - \mathbf{r}_q)} \quad (1)$$

Here f_p and f_q^* denote the complex and complex-conjugate x-ray atomic scattering factors (or neutron scattering lengths), p and q designate the lattice sites from 0 to $N-1$, $\mathbf{r}_p$ and $\mathbf{r}_q$ are the atomic position vectors for those sites, and $\mathbf{H}$ is the momentum transfer or reciprocal lattice vector $|\mathbf{H}| = (2 \sin \theta)/\lambda$ (Fig. 3). For crystalline solid solutions with a well-defined average lattice (sharp Bragg reflections), the atom positions can be represented by $\mathbf{r} = \mathbf{R}+\boldsymbol{\delta}$, where $\mathbf{R}$ is determined from the lattice constants and $\boldsymbol{\delta}$ is both the thermal and static displacement of the atom from that average lattice. Equation 1 can be separated into terms of the average lattice $\mathbf{R}$ and the local fluctuations $\boldsymbol{\delta}$,

$$I(\mathbf{H})_{\text{Total}} = \sum_p \sum_q f_p f_q^* e^{2\pi i \mathbf{H}\bullet(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} e^{2\pi i \mathbf{H}\bullet(\mathbf{R}_p - \mathbf{R}_q)} \quad (2)$$

We limit our discussion of the diffraction theory to crystalline binary alloys of A and B atoms with atomic concentration C_A and C_B, respectively, and with complex x-ray atomic scattering factors of f_A and f_B (neutron scattering lengths b_A and b_B). Since an x-ray or neutron beam of even a millimeter diameter intercepts $>10^{20}$ atoms, the double sum in Equation 2 involves $>10^{21}$ first-neighbor atom pairs (one at p, the other at q); the sum over all the atoms is a statistical description, which includes all possible atom pairs that can be formed (i.e., A-A, A-B, B-A, B-B;

Warren, 1969). A preference for like or unlike neighboring pairs is introduced by the conditional probability term P_{pq}^{AB}. This term is defined as the probability for finding a B atom at site q after having found an A atom at site p (Cowley, 1950). The probability for A-B pairs is $C_A\ P_{pq}^{AB}$, which must equal $C_B\ P_{pq}^{BA}$, the number of B-A pairs. Also, $P_{pq}^{BB} = 1 - P_{pq}^{BA}, P_{pq}^{AA} = 1 - P_{pq}^{AB}$; $C_A + C_B = 1$. With the Warren-Cowley definition of the short-range order (SRO) parameter (Cowley, 1950), $\alpha_{pq} \equiv 1 - P_{pq}^{AB}/C_B$. Spatial and time averages taken over the chemically distinct A-A, A-B, or B-B pairs with relative atom positions $p-q$, produce the total elastically and quasielastic (thermal) scattered intensity in electron units for a crystalline solid solution of two components as

$$\begin{aligned} I(\mathbf{H})_{\text{Total}} = \sum_p \sum_q \Big[& (C_A^2 + C_A C_B \alpha_{pq}) |f_A|^2 \langle e^{2\pi i \mathbf{H}\bullet(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} \rangle^{AA} \\ & + C_A C_B (1 - \alpha_{pq}) f_A f_B^* \Big(\langle e^{2\pi i \mathbf{H}\bullet(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} \rangle^{BA} \\ & + \langle e^{2\pi i \mathbf{H}\bullet(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} \rangle^{AB} \Big) + (C_B^2 + C_A C_B \alpha_{pq}) |f_B|^2 \\ & \times \langle e^{2\pi i \mathbf{H}\bullet(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q)} \rangle^{BB} \Big] \times e^{2\pi i \mathbf{H}\bullet(\mathbf{R}_p - \mathbf{R}_q)} \end{aligned} \tag{3}$$

where $|f_A|$ and $|f_B|$ denote the absolute value or moduli of the complex amplitudes.

From the theoretical development given in Appendix A, a complete description of the diffusely distributed intensity through the second moment of the displacements is given as

$$\frac{I(\mathbf{H})_{\text{Diffuse}}}{N} = \frac{I(\mathbf{H})_{\text{SRO}}}{N} + \frac{I(\mathbf{H})_{j=1}}{N} + \frac{I(\mathbf{H})_{j=2}}{N} \tag{4}$$

where

$$\frac{I(\mathbf{H})_{\text{SRO}}}{N} = \sum_{lmn} C_A C_B |f_A - f_B|^2 \alpha_{lmn} \cos \pi (h_1 l + h_2 m + h_3 n) \tag{5}$$

$$\begin{aligned} \frac{I(\mathbf{H})_{j=1}}{N} = & -\text{Re}\big[f_A (f_A^* - f_B^*)\big] \big[h_1 Q_x^{AA} + h_2 Q_y^{AA} + h_3 Q_z^{AA}\big] \\ & + \text{Re}\big[f_B (f_A^* - f_B^*)\big] \big[h_1 Q_x^{BB} + h_2 Q_y^{BB} + h_3 Q_z^{BB}\big] \end{aligned} \tag{6}$$

and

$$\begin{aligned} \frac{I(\mathbf{H})_{j=2}}{N} = & |f_A|^2 \big(h_1^2 R_X^{AA} + h_2^2 R_Y^{AA} + h_3^2 R_Z^{AA}\big) \\ & + f_A f_B^* \big(h_1^2 R_X^{AB} + h_2^2 R_Y^{AB} + h_3^2 R_Z^{AB}\big) \\ & + |f_B|^2 \big(h_1^2 R_X^{BB} + h_2^2 R_Y^{BB} + h_3^2 R_Z^{BB}\big) \\ & + |f_A|^2 \big(h_1 h_2 S_{XY}^{AA} + h_1 h_3 S_{XZ}^{AA} + h_2 h_3 S_{YZ}^{AA}\big) \\ & + f_A f_B^* \big(h_1 h_2 S_{XY}^{AB} + h_1 h_3 S_{XZ}^{AB} + h_2 h_3 S_{YZ}^{AB}\big) \\ & + |f_B|^2 \big(h_1 h_2 S_{XY}^{BB} + h_1 h_3 S_{XZ}^{BB} + h_2 h_3 S_{YZ}^{BB}\big) \end{aligned} \tag{7}$$

Here the individual terms are defined in Appendix A. As illustrated in Equations 4, 5, 6, and 7, local chemical order (Warren-Cowley α's) can be recovered from a crystalline binary alloy with a single contrast measurement of the diffuse scattering distribution, provided the displacement

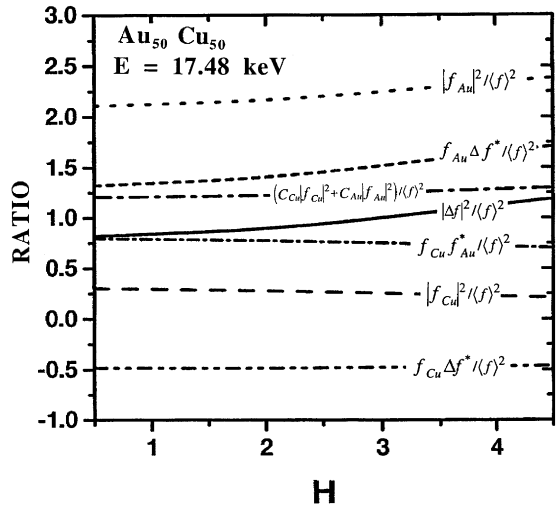


Figure 4. Variation in the ratio of the x-ray atomic scattering factor terms as a function **H**. The divisor $\langle f\rangle^2 = |C_{Cu}f_{Cu} + C_{Au}f_{Au}|^2$ was chosen to reduce the **H** dependence of all the terms for an incident energy of Mo $K_\alpha = 1.748$ keV. The relatively larger x-ray atomic scattering factor of Au, $f_{Au} = 79$ versus Cu, $f_{Cu} = 29$ at $\mathbf{H} = 0$, would require a divisor more heavily weighted with f_{Au}, such as $|f_{Au}|^2$ to reduce the **H** dependence of those terms.

contributions are negligible. This was the early practice until a series of papers used symmetry relationships among the various terms to remove the $I(\mathbf{H})_{j=1}$ term in two dimensions (Borie and Sparks, 1964), in three dimensions (Sparks and Borie, 1965), and to second moment in all three dimensions: $I(\mathbf{H})_{\text{SRO}}$, $I(\mathbf{H})_{j=1}$, $I(\mathbf{H})_{j=2}$ (Borie and Sparks, 1971), henceforth referred to as BS. The major assumption of the BS method is that the x-ray atomic scattering factor terms $|f_A - f_B|^2$, $\text{Re}[f_A(f_A^* - f_B^*)]$, $\text{Re}[f_B(f_A^* - f_B^*)]$, $|f_A|^2$, $|f_B|^2$, and $f_A f_B^*$ of Equations 4, 5, 6, and 7 have a similar **H** dependence so that a single divisor renders them independent of **H**. With this assumption, the diffuse intensity can be written as a sum of periodic functions given by Equation 34. For neutron nuclear scattering, this assumption is excellent; neutron nuclear scattering cross-sections are independent of **H**, and in addition, the TDS terms C and D can be filtered out. Even with x rays, the BS assumption is generally a good approximation. For example, as shown in Figure 4 for Mo K_α x rays and even with widely separated elements such as Au-Cu, a judicious choice of the divisor allows the BS method to be applied as a first approximation over a large range in momentum transfer. In this case, division by $f_{Au}(f_{Au} - f_{Cu}) = f_{Au}\Delta f$ would be a better choice since the Au atom is the major scatterer. Iterative techniques to further extend the BS method have not been fully explored. This variation in the scattering factor terms with **H** has been proposed as a means to recover the individual pair displacements (Georgopoulos and Cohen, 1977).

Equations 5, 6, and 7 are derived from the terms first given by BS, but with notation similar to that used by Georgopoulos and Cohen (1977). There are 25 Fourier series in Equations 5, 6, and 7. For a cubic system with centrosymmetric sites, if we know Q_X^{AA}, then we know Q_Y^{AA} and

Q_Z^{AA}. Similarly, if we know Q_X^{BB}, R_X^{AA}, R_X^{BB}, R_X^{AB}, S_{XY}^{AA}, S_{XY}^{BB}, and S_{XY}^{AB}, then we know all the Q, R, and S parameters. Thus with the addition of the α series, there are nine separate Fourier series for cubic scattering to second order.

As described in Appendix A (Derivation of the Diffuse Intensity), the nine distinct correlation terms from the α, Q, R, and S series can be grouped into four unique **H**-dependent functions, A, B, C, D within the BS approximation. By following the operations given by BS, we are able to recover these unique **H** dependent functions and from these the nine distinct correlation terms. For a binary cubic alloy, one x-ray map is sufficient to recover A, B, C, and D and from A($h_1h_2h_3$), the Warren-Cowley α's. Measurements at two x-ray energies with sufficient contrast are required to separate the A-A and B-B pair contributions to the B($h_1h_2h_3$) terms, and three x-ray energies for the A-A, A-B, and B-B displacements given in Equation 29 and contained in the terms C and D of Equation 34.

In an effort to overcome the assumption of **H** independence for the x-ray atomic scattering factor terms and to use that information to separate the different pair contributions, Georgopoulos and Cohen (1977), henceforth GC, included the variation of the x-ray scattering factors in a large least-squares program. Based on a suggestion by Tibballs (1975), GC used the **H** dependence of the three different x-ray scattering factor terms to separate the first moment of the displacements for the A-A, A-B, and B-B pairs. Results from GC's error analysis (which included statistical, roundoff, x-ray scattering factors, sample roughness, and extraneous backgrounds) showed that errors in the x-ray atomic scattering factors had the largest effect, particularly on the Q terms. They concluded, based on an error analysis of the BS method by Gragg et al. (1973), that the errors in the GC method were no worse than for the BS method and provided correct directions for the first moment displacements.

Improvements in the GC method, with the use of Mo K_α x rays to obtain more data and the use of a Householder transformation to avoid matrix inversion and stabilization with ridge-regression techniques, still resulted in unacceptably large errors on the values of the R and S parameters (Wu et al., 1983). To date, there have been no reported values for the terms R and S that are deemed reliable. However, the Warren-Cowley α's are found to have typical errors of ~10% or less for binary alloys with a preference for unlike first-neighbor pairs with either the BS or GC analysis. For clustering systems, the BS method was reported to give large errors of 20% to 50% of the recovered α's (Gragg et al., 1973). Smaller errors were reported on the α's for clustering systems with the GC method (Wu et al., 1983). With increasing experience and better data from intense synchrotron sources, errors will be reduced for both the BS and GC methods.

Another methodology to recover the pair correlation parameters uses selectable x-ray energies (Ice et al., 1992). Of most practical interest are the α's and the first moment of the static displacements as given in Equations 27 and 28. When alloys contain elements that are near one another in the periodic table, the scattering factor term $f_A - f_B$ can be made to be nearly zero by proper choice of x-ray energy nearby to an x-ray absorption. In this way,

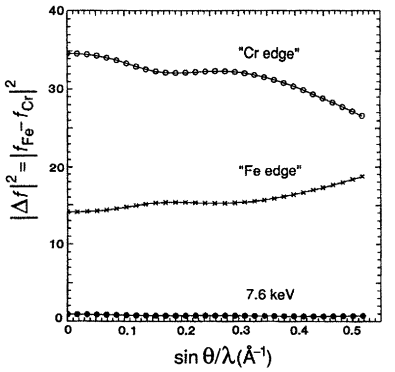


Figure 5. For elements nearby in the periodic table, x-ray energies can be chosen to obtain near null Laue scattering to separate intensity arising from quadratic and higher moments in atomic displacements. Courtesy of Reinhard et al. (1992).

the intensities expressed in Equations 27 and 28 are made nearly zero and only that intensity associated with Equation 29 remains. Then, the term $I(\mathbf{H})_{j=2}$ can be measured and scaled to diffuse scattering measurements made at other x-ray energies (which emphasize the contrast between the A and B atoms) and subtracted off. This leaves only the $I(\mathbf{H})_{SRO}$ term, Equation 5, and the first moment of the static displacements $I(\mathbf{H})_{j=1}$, Equation 6. Shown in Figure 5 are the values of $|f_{Fe} - f_{Cr}|^2$ for x-ray energies selected for maximum contrast at 20 eV below the Fe K and Cr K edges. The near null Laue energy, or energy of minimum contrast, was 7.6 keV.

The major assumption in this null Laue or 3 λ method is that the $I(\mathbf{H})_{j=2}$ and higher moment terms scale with x-ray energy as $|C_A f_A + C_B f_B|^2$, which implies that the A and B atoms have the same second and higher moment displacements or that the different elements have the same x-ray atomic scattering factors. This assumption is most valid for alloys of elements with similar atomic numbers, which have similar masses (similar thermal motion), atom sizes (small static displacement), and numbers of electrons (similar x-ray scattering factors). This 3 λ method has been used to analyze four different alloys, $Fe_{22.5}Ni_{77.5}$ (Ice et al., 1992), $Cr_{47}Fe_{53}$ (Reinhard et al., 1992), $Cr_{20}Ni_{80}$ (Schönfeld et al., 1994), and $Fe_{46.5}Ni_{53.5}$ and recalculated $Fe_{22.5}Ni_{77.5}$ (Jiang et al., 1996). An improvement in the null Laue method by Jiang et al. (1996) removed an iteration procedure to account for the residuals left by the fact that $f_A - f_B$ was not strictly zero over the measured volume.

The same techniques used for x-ray diffuse scattering analysis can also be applied to neutron scattering measurements. Neutrons have the advantage (and complication) of being sensitive to magnetic order as described in Appendix B. This sensitivity to magnetic order allows neutron measurements to detect and quantify local magnetic ordering but complicates analysis of chemical ordering.

Error analysis of the null Laue method has been given by Jiang et al. (1995) and by Ice et al. (1998). The statistical uncertainties of the recovered parameters can be estimated by propagating the standard deviation $\pm\sqrt{n}$ of the total number of counts n for each data point through the

Table 1. Contributions to the Uncertainties in the Short-Range-Order Parameter, of $Fe_{46.5}Ni_{53.5}$[a] ($\pm 1\sigma$)

lmn	α_{lmn} (σ_{Total})	$\sigma(\sqrt{n})$	σ (f') ±0.2 eu	$\sigma(P_0)$ ±1%	σ(RRS) ±1 eu	$\sigma_{Compton}$	$\sigma(C_A)$ ±0.3 at.%
000	1.0000(100)	0.0024	0	0	0	0	0
110	−0.0766 (54)	0.0018	0.0010	0.0048	0	0.0006	0.0011
200	0.0646 (28)	0.0017	0.0003	0.0016	0.0008	0.0013	0.0003
211	−0.0022 (15)	0.0014	0	0.0004	0.0001	0.0002	0.0001
220	0.0037 (14)	0.0013	0.0002	0.0003	0.0003	0.0003	0.0001
310	−0.0100 (11)	0.0011	0.0001	0.0002	0.0001	0.0001	0.0001
222	0.0037 (12)	0.0011	0	0.0002	0.0002	0.0003	0
321	−0.0032 (19)	0.0009	0	0.0001	0.0001	0.0001	0.0001
400	0.0071 (12)	0.0011	0.0002	0.0001	0.0003	0.0004	0
330	−0.0021 (9)	0.0008	0.0001	0	0.0003	0.0001	0
411	0.0007 (7)	0.0007	0	0	0	0.0002	0
420	0.0012 (8)	0.0007	0.0002	0	0.0004	0.0001	0
332	−0.0007 (7)	0.0007	0	0	0	0.0001	0

[a]For statistical and possible systematic errors associated with counting statistics n, the real part of the resonant x-ray scattering factor f' the scaling parameter P_0 to absolute intensities, inelastic resonant Raman scattering (RRS) and Compton contributions, and concentration C_A. Total error is shown in parentheses and 0 indicates uncertainties < 0.00005 Å.

nonlinear least-squares processing of the data. Systematic errors can be determined by changing the values of input variables such as the x-ray atomic scattering factors, backgrounds, and composition; then the data is reprocessed and the recovered parameters are compared.

Because the measured pair correlation coefficients are very sensitive to the relative and to a lesser degree the absolute intensity calibration of data sets collected with varying scattering contrast, the addition of constraints greatly increases reliability and reduces uncertainties. For example, the uncertainty in recovered parameters due to scaling of the measured scattering intensities is determined as input parameters are varied. Each time, the intensities are rescaled so that the I_{SRO} values are everywhere positive and match values at the origin of reciprocal space measured by small-angle scattering. The integrated Laue scattering over a repeat volume in reciprocal space is also constrained to have an average value of $C_A C_B |f_A - f_B|^2$ (i.e., $\alpha_{000} = 1$). These two constraints eliminate most of the systematic errors associated with converting the raw intensities into absolute units (Sparks et al., 1994). The intensities measured at three different energies are adjusted to within ~1% on a relative scale and the intensity at the origin is matched to measured values. For these reasons, the systematic errors for α_{000} are estimated at ~1%.

For the null Laue method, errors on the recovered α's and ΔX's arising from statistical and various possible systematic errors in the measurement and analysis of diffuse scattering data are given in Tables 1 and 2 for the $Fe_{46.5}Ni_{53.5}$ alloy (Jiang et al., 1995; Ice et al., 1998). Details of the conversion to absolute intensity units are given

Table 2. Standard Deviation of $\pm 1\sigma$ of x, y, and z Components of the Pair Fe-Fe Displacements δFe−Fe[a]

lmn	$\Delta X(\sigma_{Total})$ (Å)	$\sigma\sqrt{n}$	$\sigma(f')$ ±0.2 eu	$\sigma(P_0)$ ±1%	σ(RRS) ±1 eu	$\sigma_{Compton}$	$\sigma(C_A)$ ±0.3 at.%
110	0.0211 (25)	0.0002	0.0023	0.0007	0.0002	0.0004	0.0004
200	−0.0228 (14)	0.0004	0.0010	0.0007	0.0002	0.0004	0.0002
211	0.0005 (2)	0.0002	0	0.0001	0.0001	0	0
121	0.0014 (4)	0.0001	0.0003	0.0001	0.0002	0	0
220	0.0030 (7)	0.0002	0.0006	0.0001	0.0003	0.0001	0
310	0.0022 (3)	0.0002	0.0001	0.0001	0.0002	0.0001	0
130	0.0009 (2)	0.0002	0.0001	0	0.0001	0	0
222	0.0003 (3)	0.0002	0.0002	0	0.0001	0	0
321	0.0011 (2)	0.0001	0.0001	0	0.0002	0	0
231	0.0001 (1)	0.0001	0	0	0.0001	0	0
123	0.0008 (4)	0.0001	0.0001	0	0	0	0
400	−0.0019 (6)	0.0004	0.0002	0.0001	0.0003	0.0001	0
330	0.0011 (4)	0.0002	0.0001	0	0.0003	0	0
411	−0.0008 (3)	0.0002	0.0002	0	0.0002	0	0
141	−0.0001 (2)	0.0001	0.0001	0	0.0001	0	0

[a]For the various atom pairs of $Fe_{46.5}Ni_{53.5}$ for statistical and possible systematic errors described in the text. Total error is shown in parentheses and 0 indicates uncertainties <0.00005 Å.

elsewhere (Sparks and Borie, 1965; Ice et al., 1994; Warren, 1969; Reinhard et al., 1992). A previous assessment of the systematic errors, without the constraint of forcing $\alpha_{000}=1$ and keeping the intensity at the origin and fundamentals a positive match to known values, resulted in estimated errors approximately two to five times larger than those reported here (Jiang et al., 1995). Parameters necessary to the analysis of the data (other than well-known physical constants) with our best estimate of their standard deviations and their contributing standard deviations to the α and DX parameters are listed in Tables 1 and 2. From a comparison of theoretical and measured values, we estimate a 0.2-eu error on the real part of the x-ray atomic scattering factors, a 1% error in the P_0 calibration for converting the raw intensities to absolute units (eus), a 1-eu error in separating the inelastic resonant Raman scattering (RRS; Sparks, 1974), a 0- to 1-eu **H** dependent Compton scattering error (Ice et al., 1994), and an error of $\pm$0.3 at.% in composition (Ice et al., 1998). Systematic errors are larger than the statistical errors for the first three shells.

The asymmetric contribution of the first moment of the static displacements, $I_{j=1}$, Equation 13, to the diffuse intensity $I_{\mathrm{SRO}}+I_{j=1}$ for an $Fe_{63.2}Ni_{36.8}$ alloy is displayed in Figure 6. Without static displacements the I_{SRO} maxima would occur at the (100) and (300) superlattice positions. The static atomic displacements for the alloy are similar to those given in Table 2. Such large distortions of the short-range order diffuse scattering caused by displacements of $<$0.02 Å (0.002 nm) emphasizes the sensitivity of this technique. With a change in the x-ray energy from 7.092 to 8.313 keV, f_{Ni} becomes smaller than f_{Fe}. Figure 6 displays a reversal in the shift of the position of the diffuse scattering maxima. Two of these x-ray energies

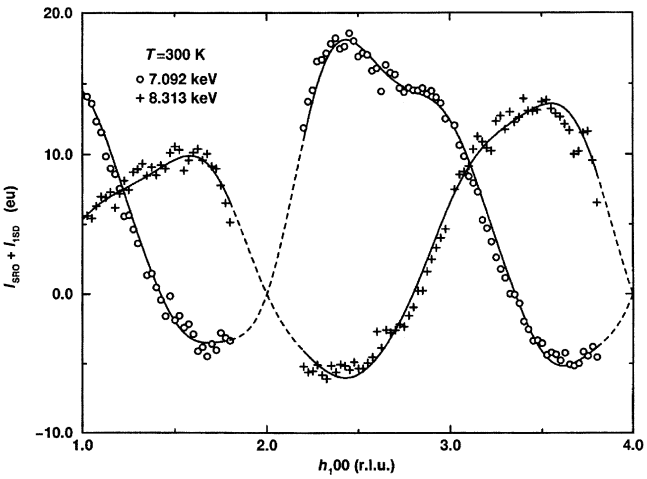


Figure 6. Diffusely scattered x-ray intensity from an $Fe_{63.2}Ni_{36.8}$ Invar alloy associated with the chemical order I_{SRO} and the first moment of the static displacements $I_{J=1}$ versus h in reciprocal lattice units (r.l.u.) along the [h00] direction. A major intensity change is affected by the choice of two different x-ray energies. The solid lines calculated from the α and δ parameters recovered from the 3λ data sets closely fit the observed data given by o and +. The dashed lines are calculated intensity through the fundamental reflections. Courtesy of Ice et al. (1998).

for the 3λ method are chosen to emphasize this contrast and a third nearest the null Laue energy for removal of the TDS. The total estimated standard deviation on the values of the α's and in particular the ΔX's give unprecedented precision for the displacements with errors $\pm$0.003 Å and less.

PRACTICAL ASPECTS OF THE METHOD

Local chemical order (Warren-Cowley α's) from a crystalline binary alloy can be recovered with a single contrast measurement of the diffuse scattering distribution. Recovery of the two terms of the first moment of the static displacements $\langle \Delta X^{\mathrm{AA}}_{lmn} \rangle$ and $\langle \Delta X^{\mathrm{BB}}_{lmn} \rangle$ requires two measurements of sufficiently different contrast in $f_{\mathrm{A}}\Delta f^*$ and $f_{\mathrm{B}}\Delta f^*$ to separate those two contributions (Ice et al., 1992). Scattering contrast can be controlled in at least three ways: (1) by selecting x-ray energies near to and far from an absorption edge energy (resonance or anomalous x-ray scattering, Ice et al., 1994); (2) by measuring diffuse scattering over a wide Q range where there is a significant change in the atomic form factors (Georgopoulos and Cohen, 1977); or (3) with neutrons by isotopic substitution (Cenedese et al., 1984).

The measurement of weak diffuse scattering normally associated with local order and displacements requires careful attention to possible sources of diffusely distributed radiation. Air scatter and other extraneous scattering from sources other than the sample, inelastic contributions such as Compton and resonant Raman, surface roughness attenuation, and geometrical factors associated with sample tilt must be removed. These details are important for placing the measured diffuse scattering in absolute units: a necessary requirement for the recovery of the α's and displacements.

Measurement of Diffuse X-ray Scattering

Methods for collecting diffuse x-ray scattering data from crystalline solid solutions have been discussed by Sparks and Borie (1965), Warren (1969), and Schwartz and Cohen (1987). As demonstrated in Equations 18 and 20, the x-ray scattering intensity from a solid solution alloy contains components arising from long-range periodicity of the crystalline lattice, correlations between the different atom types, and displacements of the atoms off the sites of the average lattice.

Because of the average periodicity of the crystalline solid solution lattice, the x-ray scattering repeats periodically in reciprocal space. The equations for recovering the various components are best conditioned when the data are collected in a volume of reciprocal space that contains at least one repeat volume for the diffuse scattering (Borie and Sparks, 1971). The volume required depends on the significance of the displacement scattering. If static displacements can be ignored and thermal scattering is removed, the smallest repeat volume for SRO is sufficient. This minimum volume increases as higher-order displacement terms become important, but in no case exceeds one-fourth of the volume bounded by $\Delta h_1 = 1$, $\Delta h_2 = 1$, and $\Delta h_3 = 1$.

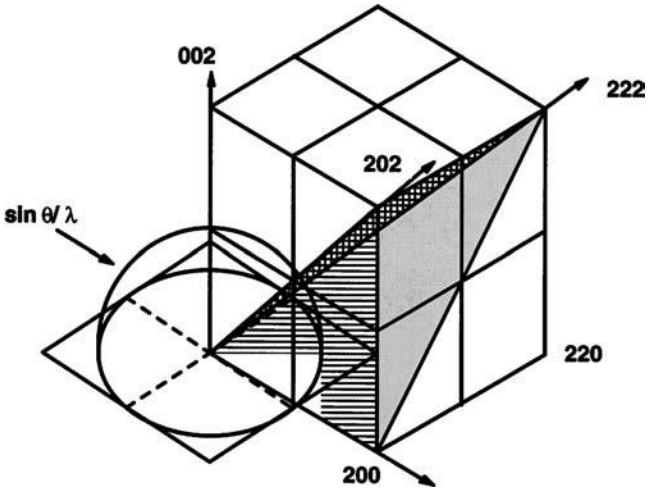


Figure 7. The diffuse intensity is mapped in a volume of reciprocal space bounded by three mirror planes that contain all the information available for a cubic alloy.

A conservative approach is to measure the scattered intensity in a volume of reciprocal space that extends from the origin to as far out in $|\mathbf{H}|$ space as possible, but contains the minimum repeat volume for cubic symmetry. As illustrated in Figure 7, the repeat volume for cubic crystals is 1/48 of the reciprocal space volume limited by the maximum momentum transfer, $\sin\theta/\lambda = 1/\lambda$. This repeat unit contains all the accessible information about the structure of a crystal with average cubic symmetry. As it takes several days to prepare for the experiment and its setup, actual collection of the data is not the time-limiting step with x rays, and as much data as possible should be collected in an effort to accumulate 1000 or more counts at each of several thousand positions.

The various points h_1, h_2, h_3 in reciprocal space are measured by orienting the sample and detector arm with a four-circle diffractometer. Diffuse scattering data are typically collected at intervals of $\Delta h = 0.1$ in a volume of reciprocal space bounded by $4 \geq h_1 \geq h_2 \geq h_3 \geq 1$. Regions of detailed interest are measured at intervals of $\Delta h = 0.05$. There are on the order of 7000 data points collected for a diffuse volume at each x-ray energy. Angular dependence of the absorption corrections is eliminated by measuring diffuse scattering in a bisecting geometry where the redundancy of the three sample-orienting circles is used to maintain the same incident and exit angle for the radiation with respect to the surface.

Shown in Figure 8 is the typical optical train for the measurement of diffuse x-ray scattering. Every effort should be made to ensure that the detector receives the radiation from the sample with the same collection efficiency regardless of the sample orientation. The sample needs to be replaced with a well-calibrated standard to convert the flux incident on the sample to absolute units. As all scattering measurements are made for fixed I_0 monitor counts, any changes to the optical train such as slit positions (excepting scatter slit) and sizes, distances, detectors, and changes in energy require replacing the sample with the scattering standard for recalibration of the incident flux against the I_0 monitor counts.

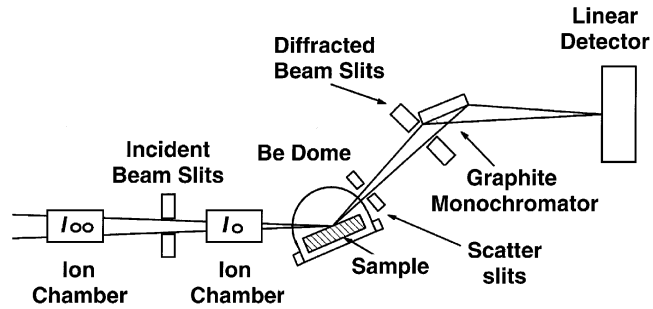


Figure 8. Optical setup for resonant diffuse x-ray scattering measurement.

New challenges arise from the application of resonant (anomalous) x-ray scattering to the study of local order in crystalline solid solutions: (1) the need to work near absorption edges that can create large fluorescent and resonant Raman backgrounds, and (2) the need to know the resonant (anomalous) scattering factors and absorption cross-sections to ~1%, especially at x-ray energies near absorption edges. Background problems due to inelastic scattering are exacerbated. Experimental measurement to recover the elastic scattering from these inelastic contributions (Compton, fluorescence, and resonant Raman) requires a combination of spectroscopy and diffraction.

Removal of Inelastic Scattering Backgrounds

Photoelectric absorption is the dominant x-ray cross-section for elements with $Z > 13$ and x-ray energies $E < 20$ keV. The resultant fluorescence is typically orders of magnitude larger than the diffuse elastic scattering. Fluorescence can be removed by the use of an energy-sensitive detector when the incident x-ray energy is far enough above the photoabsorption threshold that the detector's energy resolution is adequate for separation.

Maximum change in scattering amplitude is obtained when measurements are made with the x-ray energy very near and then far from an absorption edge, as shown in Figure 9. The size of the f' component roughly doubles as the energy gap between the incident x-ray energy E_I and the absorption edge E_K is halved. Near an edge, the size of the inelastic background grows rapidly due to RRS (Sparks, 1974; Eisenberger et al., 1976a,b; Åberg and Crasemann, 1994). RRS is interpreted as the filling of a virtual hole by a bound electron. As with fluorescence, the x-ray energy spectra is distinctive with peaks corresponding to the filling of the virtual hole, say in the K shell, by various higher lying shells: K filled from L, or K filled from M shells, often referred to as K-L and K-M RRS. Unlike fluorescence, the energy of the RRS peaks shift with incident x-ray energy, and the energy of the nearest RRS K-L line is only a few tens of electronvolts from the incident x-ray energy (Åberg and Crasemann, 1994). This large inelastic background must either be removed experimentally or be calculated and subtracted.

The resolution of a solid-state detector at ~150 eV is inadequate to resolve all the RRS K-M component from the elastic peak when excited near the threshold and can

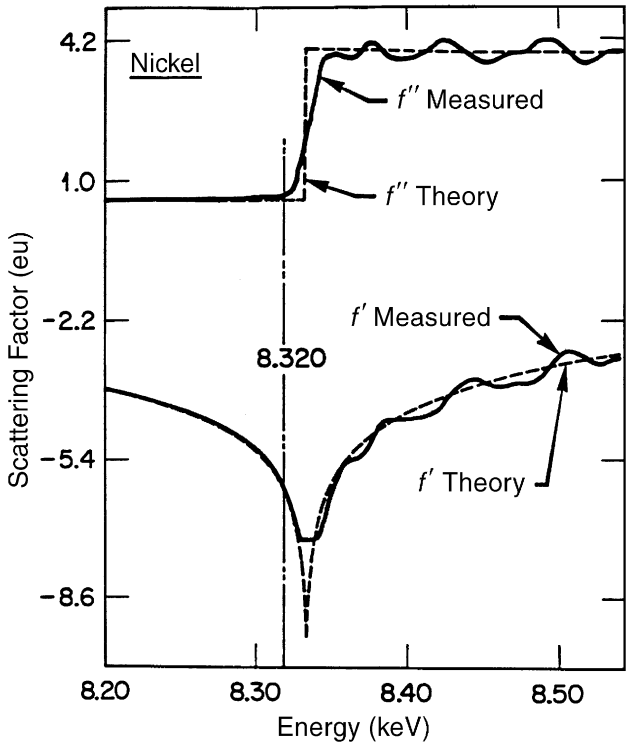


Figure 9. Variation of f' and f'' near the K absorption edge of nickel. Dashed lines are the theoretical calculation from Cromer and Liberman (1970; 1981). Courtesy of Ice et al. (1994).

only resolve the Compton inelastic scattering at high scattering angles. A typical energy spectrum excited by 8.0-keV x rays on a $Ni_{77.5}Fe_{22.5}$ single crystal measured with a Si(Li) detector is shown in Figure 10. Near and below the Fe edge of 7.112 keV, the RRS *K-M* component cannot be resolved from the elastic scattering peak. The RRS *K-M* component can be removed by measuring the RRS *K-L* component and assuming the *K-LK-M* ratio remains constant. The Compton scattering component at low scattering angles is removed by using theoretical tables.

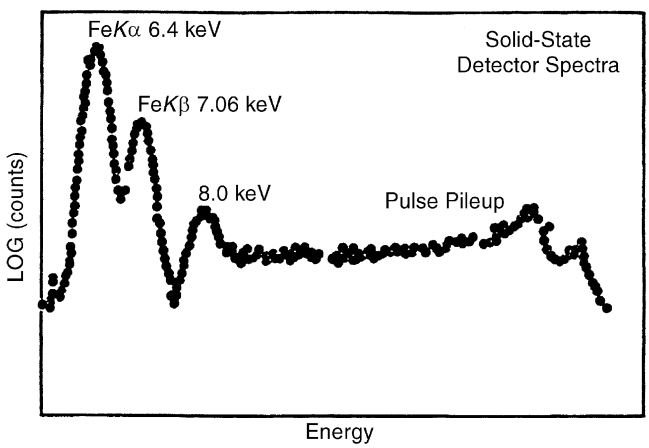


Figure 10. Energy spectrum measured with a solid-state detector from Ni-Fe alloy excited by 8.0-keV x rays. Courtesy of Ice et al. (1994).

Disadvantages to the use of solid-state detectors include the statistical and theoretical uncertainty of the inelastic contributions. Another disadvantage is the large deadtime imposed by the fluorescence K signal in a solid-state detector, which restricts the useful flux on the sample.

Resolution on the order of 10 to 30 eV is necessary to cleanly separate the resonant Raman and Compton scattering from elastic scattering. A crystal spectrometer improves the energy resolution beyond that available with a solid-state detector. Perfect crystal spectrometers are highly inefficient compared to mosaic crystal spectrometers due to their smaller integrated reflectivity. This inefficiency is usually unacceptable for the typical weak diffuse x-ray scattering. A mosaic crystal x-ray spectrometer (Ice and Sparks, 1990) has been found to be a more practical device. The advantages of a mosaic crystal spectrometer is that it is possible to obtain energy resolutions similar to that of a perfect crystal spectrometer, but with an overall angular acceptance and efficiency similar to those of a solid-state detector. A schematic is shown in Figure 11A.

Figure 11B,C illustrates the ability of the graphite spectrometer to resolve Ni K_{α_1} from K_{α_2} and to resolve RRS from elastic scattering near the Ni K edge. Good efficiency is possible if the x-ray energy, E_I, lies within a bandpass, ΔE, set by the crystal Bragg angle θ_B and the rocking curve width $\Delta\omega_R$; $\Delta E = \Delta\omega_R E \cot\theta_B$. The diffracted beam is parafocused onto a linear detector and the beam position is correlated to x-ray energy. The energy resolution and energy scale of the spectrometer are determined by varying the energy of the incident beam and observing the peak position of the elastic peak. At 8 keV, the bandpass, ΔE, of a graphite crystal with a 0.4° full-width at half-maximum (FWHM) mosaic spread is ~250 eV. The energy resolution is limited by the effective source size viewed by the energy analyzer and by imperfections in the crystal. Energy resolutions of 10 to 30 eV are typical with a 0.3- to 1.5-mm high beam in the scattering plane at the sample.

Elastic scattering is resolved from the Compton scattering at higher scattering angles and *K-L* RRS at all angles. When an x-ray energy is selected near but below the absorption edge of the higher Z element, the lower Z element will fluoresce. The mosaic crystal spectrometer can discriminate against this fluorescence and avoid deadtime from this emission. This graphite monochromator gives an overall decrease of 3 to 4 in counting efficiency compared with a solid-state detector, but provides a much cleaner signal with greatly reduced deadtime.

A consideration when using a crystal spectrometer is the sensitivity of the energy resolution to the effective source size. As the source size increases, the energy resolution decreases and increasingly small incident beams are required for good energy resolution (Ice and Sparks, 1990). In addition, the effective source size as viewed by the crystal spectrometer depends on the spread of the incident beam on the sample and the angle of the detector axis to the sample surface. These geometrical factors are governed by the scattering angle, θ, and the chi tilt, χ, as shown in Figure 12.

Diffuse scattering is normally collected in the bisecting mode as intended here. The size and shape of the beam

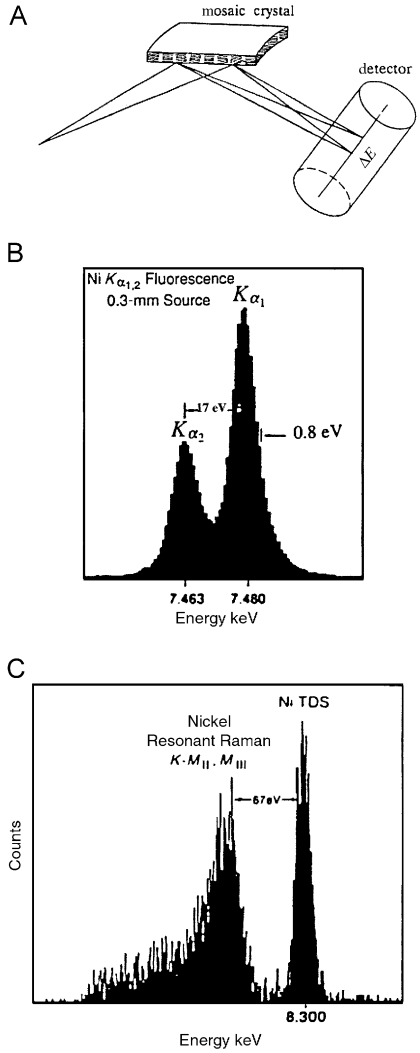


Figure 11. (**A**) Scattered radiation is energy analyzed with a mosaic graphite crystal dispersing radiation along a position sensitive detector to resolve (**B**) fluorescence and (**C**) resonant Raman and Compton scattering. Courtesy of Ice and Sparks (1990).

intercept with the sample surface is determined by the beam height h, width w, and the sample angles θ and ω. As shown in the insert of Figure 12, the intercept is a parallelogram that extends along the beam direction by $h/\sin\theta$ due to the beam height and $w\tan(\omega-\pi/2)/\sin\theta$ due to the beam width. In the reference frame of the detector, the length of the parallelogram is projected into a root-mean-square source height of

$$\sigma_s = \frac{\sqrt{[2w\tan(\chi-\pi/2)\cos\theta]^2 + h^2}}{12} \qquad (8)$$

The measured energy resolution for a 1.5×1.5-mm^2 beam at $\chi = 55^\circ$ is plotted in Figure 12 as a function of scattering angle θ. Deviations from $\chi=90^\circ$, where the sample normal lies in the scattering plane, are held to a minimum by choosing a crystal surface normal centered in the volume of reciprocal space to be measured. This reduces

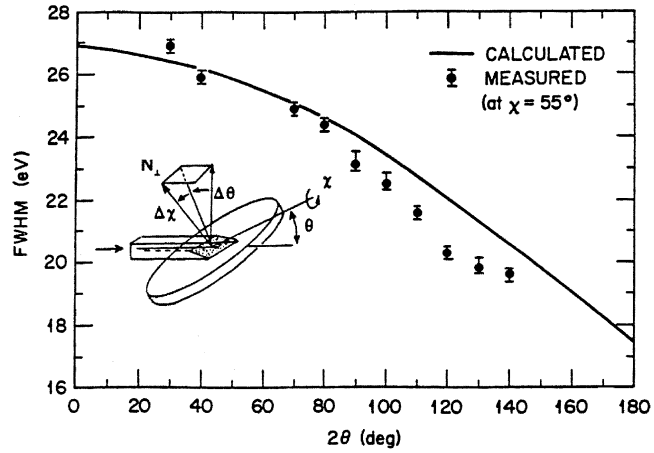


Figure 12. Incident beam spread on the sample depends on χ and θ, which orient the surface normal with respect to the incident beam. As the beam spread on the sample is proportional to the effective source size viewed by the spectrometer, the energy resolution changes as χ and θ change. The measured energy resolution (points) is plotted for $\chi = 55^\circ$ and compared with the theoretical prediction (line).

surface roughness corrections and maintains good energy resolution. Actual spectra collected during a diffuse scattering measurement are shown in Figure 13. Crystal analyzers will perform even better with third-generation synchrotron sources that have smaller source size and higher flux. Measurement of the fluorescent intensity or RRS throughout the volume used for the data with the identical optical train provides a correction for this beam spread, sample roughness, and alignment errors.

A model is used to describe the energy distributions of the RRS shown in Figures 13A,B as observed with the graphite spectrometer. The model contains a truncated Lorentzian centered at the energy for the RRS peak (Sparks, 1974; Ice et al., 1994). The high-energy cutoff is determined from energy conservation. The spectra are corrected for the graphite monochromator efficiency, which has a Gaussian distribution centered on the elastic scattering peak. The spectra are also corrected for the finite spectrometer resolution by convolving the Gaussian shape of the elastically scattered peak. The simple model for the resonant Raman peak shape allows for a good fit to the experimental resonant Raman peak observed with the graphite monochromator, as shown in Figures 13A,B.

Compton scattering can be removed from the elastic scatter by subtracting tabulated theoretical Compton intensities. It is possible to experimentally separate the elastic scattering peak from the Compton peak except at the lowest scattering angles. A correction for overlap of the two peaks at small angles is achieved by modeling the energy dependence of the Compton profile. The doubly differential Compton scattering cross-section is calculated using the impulse approximation (IA; Carlsson et al., 1982; Biggs et al., 1975). The Compton scattering is calculated for each subshell, and energy conservation is used to restrict the high-energy tail from each shell. The total spectrum is determined by adding the contribution from

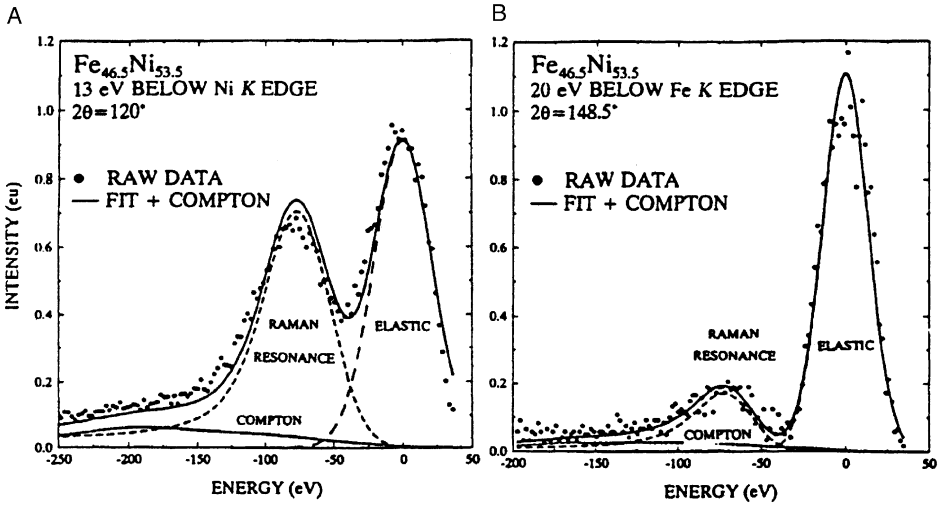


Figure 13. Energy spectrum of scattered radiation when the incident energy is (**A**) 13 eV below the Ni *K* edge and (**B**) 20 eV below the Fe *K* edge of a Fe-Ni crystal. Courtesy of Ice et al. (1994).

each shell and from each atom type. This cross-section is subtracted from the measured data normalized to absolute electron units per atom, which leaves only a resonant Raman peak and an elastic scattering peak. The slight overlap of the Compton peak with the elastic peak is typically small compared with statistical uncertainties.

Comparison of the calculated to the observed Compton spectrum can be achieved with x-ray energies sufficiently removed from an edge that the resonant Raman contribution is negligible. At 8.000 keV, the resonant Raman contribution from an Fe-Ni sample is centered far below the elastic peak and outside the spectrometer window. The IA-calculated Compton profiles are observed to be in qualitative agreement with the observed spectra, but the intensity is overestimated at low angles and underestimated at high angles. Particularly noticeable is a low-energy tail at high scattering angles. A more exact theory without the impulse approximation might improve matters.

Determination of the Resonant Scattering Terms f' and f''

The widely used Cromer-Liberman tabulation (Cromer and Liberman, 1981; Sasaki, 1989) of f' and f'' explicitly ignores the presence of pre-edge unfilled bound states (bound-to-bound transitions), lifetime broadening of the inner-shell hole, and x-ray absorption fine structure (XAFS; XAFS Spectroscopy). These assumptions are justified ~100 eV below and ~1 keV above an edge, but not near an absorption edge (Lengeler, 1994; Chantier, 1994). Of particular concern is the underestimation of the absorption coefficient and f' just below an absorption edge due to the Lorentzian hole width of an inner shell. This is illustrated in Figure 14. An inner-shell hole with a 2-eV broadened lifetime and a *K* edge jump ratio of 8 will increase the absorption cross-section by ~11% at 20 eV below the nominal edge and by 2% at 100 eV below the edge. Theoretical tabulations that ignore the lifetime broadened hole width must be corrected, as they underestimate the absorption coefficient (and f') just below an edge and overestimate f'' above the edge. Theoretical tabulations also do not include unfilled pre-edge states or absorption fine structure that is highly sample dependent. Thus, it is necessary to determine the photoelectric absorption cross-section, μ/ρ, experimentally for each sample, to calculate f'' with the optical theorem (James, 1948), and then to calculate f' from f'' with the Kramers-Kronig relationship.

The practical method of measuring the sample-specific absorption cross-section is to measure the relative absorption cross-section across the edge of each of the elements of the sample over a range of ~100 to 1000 eV. These data are normalized to theoretical data far from the edge (Kawamura and Fukamachi, 1978; Dreier et al., 1983; Hoyt et al., 1984). Measurements are made with a thin foil of the sample in transmission geometry. The measured value of f' is obtained by adding the difference integration to tabulated values of f', as shown in Figure 9.

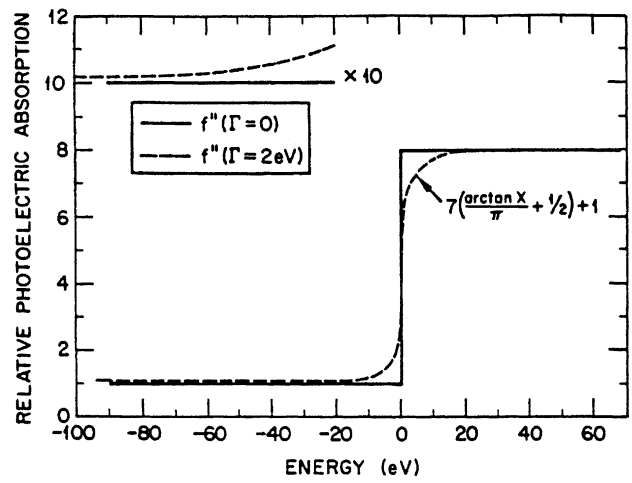


Figure 14. The usual tabulated values of the resonant (anomalous) scattering terms are not corrected for hole width (lifetime), which causes a Lorentzian broadening of the absorption edge and affects the values of f' and f'' near the edge. Courtesy of Ice et al. (1994).

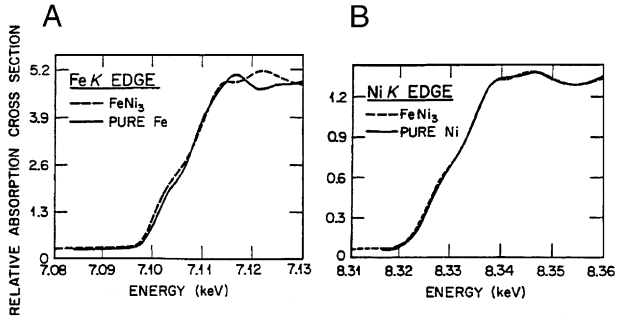


Figure 15. The absorption edge energy shifts are very small for metallic alloys with differing nearest neighbors. Courtesy of Ice et al. (1994).

We find f' is 5 to 10 times less sensitive than f'' to lifetime broadening of the inner-shell hole. For example, the effect of a 2-eV lifetime on f' at 20 eV below the edge is only 2% and at −100 eV the effect is only 0.3%. The value of f' is sensitive to the position of the average inflection point of the absorption edge. A shift of 5 eV results in a 4% to 5% change in f' at 20 eV below the edge. Errors in the absolute energy calibration are removed when the energy of the incident radiation is fixed to the same absorption edge energy as the calculation of f' and f''.

As shown in Figure 15A,B, the absorption edge energies for Fe and Ni in a fully ordered $FeNi_3$ foil do not shift compared to the absorption edge energies of pure Fe or Ni foils. However, the local environment of the Fe atoms in $FeNi_3$ are sufficiently different that calculated or measured values of f' and f'' for pure Fe would be in error close to the Fe absorption edge. For samples where there is a large charge transfer (change in oxidation state), this difference becomes even larger.

Absolute Calibration of the Measured Intensities

Conversions to absolute units depend on previously calibrated standards to place the intensity in electron units (Suortti et al., 1985). Calibrated powder standards account for the monitor efficiency, the beam path transmission, and the efficiency and solid angle of the detector. The largest uncertainty is in the values of the linear x-ray absorption cross-sections, μ, near an absorption edge for both the powder and the sample. This problem is reduced by using a powder similar in elemental composition to the sample or by careful calibration of the sample absorption. Comparison between standardizations with various powder samples are consistent to within 3% to 5% (Suortti et al., 1985).

The relative scaling factors between different energy sets can be refined with great sensitivity by restricting the short-range-order intensity as discussed previously and as described below. For alloys that cluster, the Laue scattering can be obscured by proximity of the fundamental Bragg peaks. Higher-resolution measurements such as small-angle x-ray scattering (SAXS) techniques may then be required to recover the Laue scattering. SAXS can be used to measure the total SRO scattering at the origin, and the relative scaling factors of data sets can be adjusted so that $\alpha_{000}=1$, and the fitted SRO diffuse scattering,

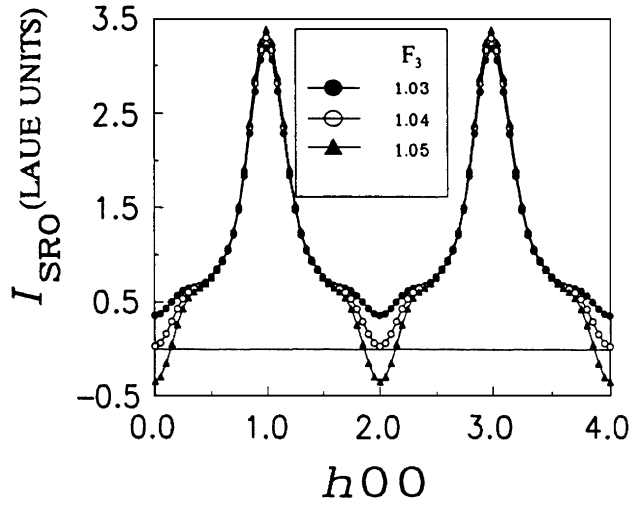


Figure 16. Fitted I_{SRO} along the $h00$ line for three relative scale factors on the near zero contrast data of a $Ni_{77.5}Fe_{22.5}$ sample. With a scale factor of 1.04, I_{SRO} is near zero at the origin and fundamental Bragg peaks as measured by SAXS. Courtesy of Sparks et al. (1994).

approach the SAXS value obtained near the origin. This scaling method is illustrated in Figure 16 for a $Ni_{77.5}Fe_{22.5}$ alloy. As shown in Figure 16, a small change in the relative scaling (here the zero contrast or near null Laue scale factor) makes a big change in the SRO scattering near the origin. Fe-Ni alloys are known to show negligible scattering near the origin (Simon and Lyon, 1991). Scale factors are adjusted to set the SRO scattering to near zero at the Bragg peaks. Nonlinear fitting routines that refine the relative intensities and include these and other restrictions may be possible with more reliable data sets from third-generation synchrotron sources.

DATA ANALYSIS AND INITIAL INTERPRETATION

Most x-ray detectors are count-rate limited, and measured x-ray scattering intensities must be corrected for detector-system deadtime (Ice et al., 1994). With proportional counters and solid-state detectors, the measured deadtime is typically 3 times the amplifier shaping constant. For most measurements, this results in deadtimes of ~1 to 10 μs/count.

Detector survival can also be challenged by intense x-ray beams and requires that the sample orientation and detector position be controlled such that a Bragg reflection from the cyrstal does not enter the detector. Bragg reflections can contain in excess of 10^9 x rays per second at synchrotron sources, which can paralyze or damage detectors. Position-sensitive wire proportional counters are especially vulnerable, and count rates below 10^4 counts/s are generally advisable to prevent damage to the wire or coated filament. Just falshing throug a Bragg reflection when changing orientation can damage the wire anode of a linear-senitive proportional counter, which materially degrades its spatial resolution. Extreme caution is necessary when measurements are taken near Bragg reflections with flux-sensitive detectors.

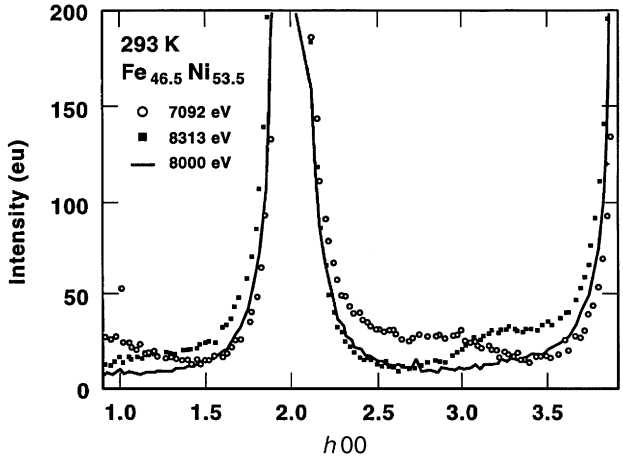


Figure 17. Total elastically scattered x-ray intensity along the $\langle h00 \rangle$ measured at 293 *K* for the three x-ray energies listed. Note the shift in contrast for intensities measured with energies 20 eV below the Fe *K* edge at 7092-eV and the Ni *K* edge at 8313 eV, which changes the sign of $\mathrm{Re}(f_{\mathrm{Ni}}-f_{\mathrm{Fe}})$. The outlying data point at the {100} position is from harmonic energy contamination of the incident radiation and such points are removed before processing. Courtesy of Jiang et al. (1996).

Mirrors and crystal monochromators can pass a significant number of harmonics (multienergy x rays), which are then diffracted by the sample at positions of h/n, k/n, l/n, where n is an integer and hkl are the Miller indices of the Bragg reflections. Any sharp spikes observed in the measurement of diffuse intensities at these positions are suspect and must be removed before processing the data to recover the local correlations. At the position (100) in Figure 17, we note an outlying data point that can be attributed to the Bragg diffraction from the (200) reflection of an x-ray energy twice that of the nominal energy. Such spurious data can also be caused by surface films left from chemical treatment or oxidation.

An example of the raw data measured for three different x-ray energies from an $Fe_{46.5}Ni_{53.5}$ alloy single crystal is shown in Figure 17 (Jiang et al., 1996). The solid line in Figure 17 is the near null Laue measurement, which can be used to remove the quadratic and higher-order displacement terms from the other data sets. The assumption made is that for elements with similar masses and small static displacements the second and higher moment terms are the same for both atoms species so that for $I_{j\geq 2}$, the intensity scales as $\langle f \rangle^2$ This method avoids the need to calculate thermal scattering from a set of force constants with the Born-von Karman central forces model that assumes harmonic vibrations and Hooke's law forces between the atoms independent of their environment (Warren, 1969). Theoretically calculated TDS increasingly deviates from measured TDS on approach to the Brillouin zone boundaries. This is not unexpected as local structure is important near the zone boundaries. A comparison of x-ray null Laue results with neutron diffuse scattering measurements, which are not complicated by TDS, gives very similar α's for Fe_3Ni (Jiang et al., 1996).

An example of the measured diffuse scattering data in the h_1h_20 plane (labeled here as H_1, H_2, 0) for three x-ray energies is shown in Figure 18 for an alloy of $Fe_{27.5}Ni_{77.5}$ (Ice et al., 1992). The x-ray energy at 8.000 keV (Fig. 18B) is the near null Laue energy ($f_{\mathrm{Ni}} - f_{\mathrm{Fe}} \sim 0$), where I_{SRO} is nearly zero compared with the intense SRO maxima such as (100), (110), and (210) at 7.092 keV of Figure 18A. With the data in electron units and with inelastic scattering removed, the data are now processed to recover the α and δ values. As discussed previously (see Principles of the Method), we have the choice of different processing methods.

1. The null Laue method (also referred to as the 3λ method). This method is used when the elements of an alloy are sufficiently near each other in the periodic table that their dynamic displacements are similar and $I_{j\geq 2}$ scales as $(C_{\mathrm{A}}f_{\mathrm{A}}+C_{\mathrm{B}}f_{\mathrm{B}})^2$ for the different energies; $I_{j\geq 2}$ can be experimentally measured at a low x-ray scattering contrast and then substracted from the diffuse scattering with high elemental contrast (Ice et al., 1992; Reinhard et al., 1992; Schoenfield et al., 1994; Jiang et al., 1996). The null Laue method is implemented using a nonlinear least-squares approach.
2. Collection of sufficient data such that the BS separation technique can first be used to recover I_{SRO}, $I_{j=1}$, $I_{j=2}$, and higher terms separately for each of the three x-ray energies. Then a least-squares program is used to recover the α's from I_{SRO}. The α's are used to recover the displacements from $I_{j=1}$ and the second moments from $I_{j=2}$ as given in Equations 4, 5, 6, and 7. As the assumption is made that the x-ray atomic scattering factors are independent of **H**, an interactive technique is required to remove that

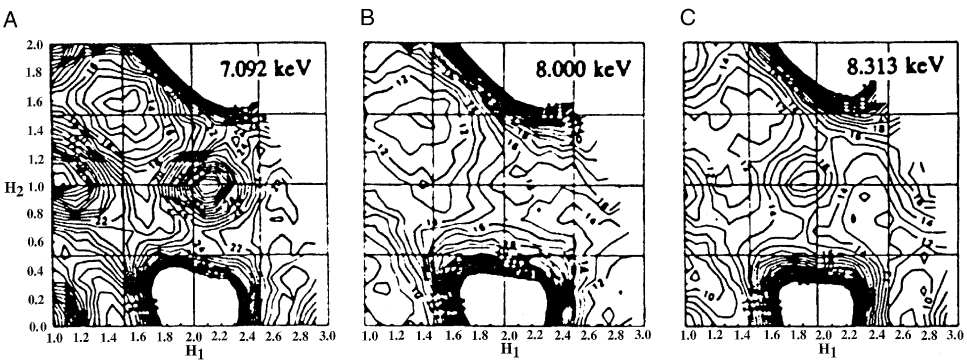


Figure 18. Diffuse x-ray scattering intensities from $Fe_{22.5}Ni_{72.5}$ in the $h_3 = 0$ plane collected with x-ray energies of (**A**) 7.092, (**B**) 8.000, and (**C**) 8.313 keV. Courtesy of Ice et al. (1992).

assumption. This is possibly the most robust of the methods as it does not have the assumption of method 1. Therefore it can be extended to include the $I_{j=3}$ and $I_{j=4}$ terms to account for higher moments of the displacements including second-order TDS, which becomes important for measurements made at temperatures approaching the Debye temperature. Furthermore, data for each energy can be separately into I_{SRO} and displacement scattering and checked for the correct normalization factors before different energy sets are subjected to a simultaneous least-squares program to recover the A-A, A-B, and B-B pair correlations.

3. The GC method of analysis, which uses data measured at only one energy for 25 symmetry-related points for each of the 25 terms expressed in Equations 4, 5, 6, and 7. These symmetry-related points are chosen such that the values of I_{SRO}, and the Q, R, and S terms are of the same or of opposite magnitude. Only their scattering factors and **H** dependence differ. In this way, the terms of Equations 4, 5, 6, and 7 are obtained from a system of linear equations stabilized by a ridge regression technique. These terms are then inverted to recover their Fourier coefficients. Results with this technique have been mixed. An analysis of $AuCu_3$ data (Butler and Cohen, 1989) concluded that the Au-Au bond distance was shorter than that for Cu-Cu. This result is contrary to the experimental findings that ordering of $AuCu_3$ reduces the lattice constant as more first-neighbor Au-Cu pairs are formed and that the addition of the ~14% larger Au atoms to Cu increases the lattice constant because of the larger Au-Au bond distance. Theoretical considerations have also concluded that the Au-Au bond distance is the largest of the three kinds (Chakraborty, 1995; Horiuchi et al., 1995). Apparently, the **H** variation of $f_{Au}\Delta f^*$ and $f_{Cu}\Delta f^*$ shown in Figure 4 is not sufficiently different to provide for a meaningful separation of the Au-Au and Cu-Cu bond distances. In a direct comparison with the 3λ technique on an alloy of $Ni_{80}Cr_{20}$, the GC result gave a Ni-Ni bond distance with a different sign, which was contrary to other information (Schöenfeld et al., 1994). In addition, published GC values for $\langle(\Delta X)^2\rangle$ coefficients are not reliable (Wu et al., 1983). Though first-order TDS is included in the separation, higher-order TDS is calculated from force constants and subtracted.

Interpretation of Recovered Static Displacements

The displacements are defined as deviations from the average lattice and are given by

$$\mathbf{r}_p - \mathbf{r}_q = (\mathbf{R}_p - \mathbf{R}_q) + (\boldsymbol{\delta}_p - \boldsymbol{\delta}_q) \tag{9}$$

As we can move the frame of reference so that its origin always resides on one of the atoms of the pair, such that $\mathbf{r}_p \equiv \mathbf{r}_0$, $\mathbf{R}_p \equiv \mathbf{R}_0$, and $\boldsymbol{\delta}_p \equiv \boldsymbol{\delta}_0$, then

$$\mathbf{r}_p - \mathbf{r}_q = \mathbf{r}_0 - \mathbf{r}_q = \mathbf{r}_0 - \mathbf{r}_{lmn} = -\mathbf{r}_{lmn} \tag{10}$$

and with the atom pair identified by ij

$$\mathbf{r}^{ij}_{lmn} = \mathbf{R}_{lmn} + \boldsymbol{\delta}^{ij}_{lmn} \tag{11}$$

where $\mathbf{R}_{lmn}$ is independent of the kinds of atom pairs since it is defined by the average lattice (i.e., Bragg reflection positions). The average value of the measured $\mathbf{r}_{lmn}$ for all the N pairs contributing to the measured intensity is

$$\langle \mathbf{r}^{ij}_{lmn} \rangle = \frac{1}{N^{ij}} \sum_{ij} \langle \mathbf{R}_{lmn} + \boldsymbol{\delta}^{ij}_{lmn} \rangle = \mathbf{R}_{lmn} + \langle \boldsymbol{\delta}^{ij}_{lmn} \rangle \tag{12}$$

Here $\langle \boldsymbol{\delta}^{ij}_{lmn} \rangle$ is the variable recovered from the diffuse scattering. As shown in Equation 25, we recover the Cartesian coordinates of the average displacement vector,

$$\langle \boldsymbol{\delta}^{ij}_{lmn} \rangle \equiv \langle \Delta X^{ij}_{lmn} \rangle \mathbf{a} + \langle \Delta Y^{ij}_{lmn} \rangle \mathbf{b} + \langle \Delta Z^{ij}_{lmn} \rangle \mathbf{c} \tag{13}$$

For cubic systems, when the atom has fewer than 24 neighboring atoms in a coordination shell (permutations and combinations of $\pm l$, $\pm m$, $\pm n$), $\langle \boldsymbol{\delta}^{ij}_{lmn} \rangle$ must be parallel to the lattice vector $\mathbf{R}_{lmn}$. This maintains the statistically observed long-range cubic symmetry even though on a local scale this symmetry is broken. For lmn multiplicities $\geq$24, the displacements on average need not be parallel to the average interatomic vector $\mathbf{R}_{lmn}$ to preserve cubic symmetry (Sparks and Borie, 1965).

Measurements of diffuse scattering from single crystals provides the components of the atomic displacements $\langle \Delta X \rangle$, $\langle \Delta Y \rangle$, and $\langle \Delta Z \rangle$ whereas the spherical average usually obtained from EXAFS and x-ray measurements on amorphous materials and crystalline powders gives only the magnitude of the radial displacements. Thus, diffuse x-ray scattering from single crystals provides new information about the vector displacements associated with near-neighbor chemistry.

Measured displacements such as those presented in Table 2 provide unique insight into how atoms move off their lattice sites when local symmetry is broken. Local symmetry is broken when a multicomponent crystalline material is above the ordering temperature (with less-than-perfect long-range order) and/or off stoichiometry. With perfect long-range order the atoms are constrained to lie precisely on the sites of the average lattice by balanced forces. In alloys, where the local symmetry is broken, we gain new insights into the chemically distinct bonding, including the interatomic bond distances and whether the displacements have both radial and tangential components. With reference to Figure 23, the displacement for the [110] nearest-neighbor atoms is on average radial with a magnitude given by $|\langle \boldsymbol{\delta}_{110} \rangle| = \sqrt{2}|\Delta X_{110}|$.

We note that the Fe-Fe first-neighbor pair distances given in Table 2 are 0.021(3) Å $\times \sqrt{2} = 0.030(4)$ Å further apart then the average lattice and that second neighbor pairs are closer by (−) 0.023(1) Å. Average bond distances along the interatomic vector between nearest-neighbor pairs for this fcc lattice are obtained by adding the $\sqrt{2}|\Delta X_{110}|$ to the average interatomic vector $\mathbf{R}_{110}$, as defined in Figure 19. The parameter $|\mathbf{R}_{110}|$ is just the

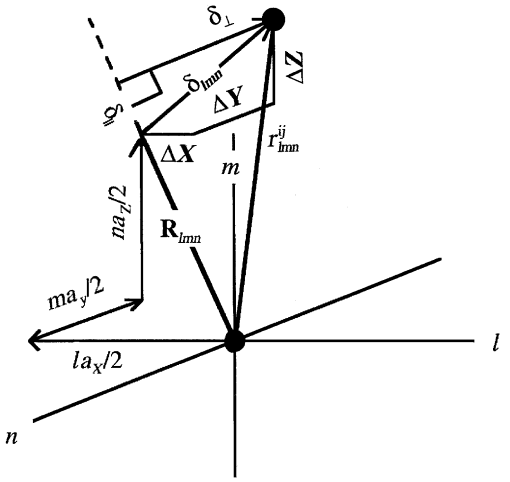


Figure 19. Construction of the vectors recovered from diffuse scattering measurements on single crystals. The parameter $\mathbf{R}_{lmn}$ is obtained from the lattice parameter $|\mathbf{a}|$, and the average components of the displacement δ^{ij}_{lmn} are recovered from measurements of the diffuse scattering. Courtesy of Ice et al. (1998).

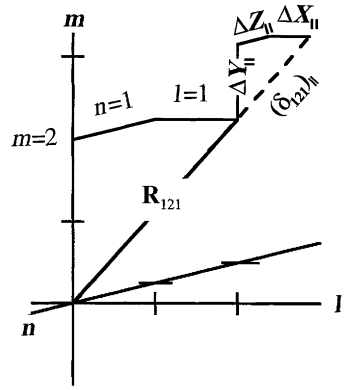


Figure 20. Radial displacements (parallel to the interatomic vector $\mathbf{R}_{lmn}$) between the atom pairs require that the relative magnitudes of the displacement components be in the same proportion as the average lattice vector; $\Delta X:\Delta Y:\Delta Z = l:m:n$. As shown for $lmn = 211$, a radial displacement requires $|\Delta X| = 2|\Delta Y|$ and $|\Delta X| = 2|\Delta Z|$. For $lmn = 121$, $|\Delta X| = |\Delta Y|/2$ and $|\Delta X| = |\Delta Z|$. For a cubic lattice, we can interchange l, m, and n and similarly $\Delta X, \Delta Y$, and ΔZ. Thus there is only one value ΔX for lmn multiplicities <24 (i.e., 110, 200, and 222), two values for ΔX when lmn has multiplicities equal to 24 ($l \neq m$ and $l = m, n$), and three values for ΔX with multiplicities equal to 48. Courtesy of Ice et al. (1998).

cubic lattice constant $|\mathbf{a}|$ times $1/\sqrt{2}$. From the construction shown in Figure 19, it follows that the vector distance between a pair of atoms, $\mathbf{r}^{ij}_{lmn}$, has radial and tangential displacement components with magnitudes given by

$$\left|\boldsymbol{\delta}^{ij}_{lmn}\right|_{||} = \frac{\boldsymbol{\delta}^{ij}_{lmn} \bullet \mathbf{R}_{lmn}}{|\mathbf{R}_{lmn}|} \tag{14}$$

and

$$\left|\boldsymbol{\delta}^{ij}_{lmn}\right|_{\perp} = \sqrt{|\boldsymbol{\delta}^{ij}_{lmn}|^2 - |\boldsymbol{\delta}^{ij}_{lmn}|^2_{||}} \tag{15}$$

The radial (||) and tangential ($\perp$) components of the displacements recovered from diffuse scattering measurements on single crystals are shown in Figure 19.

As the $Fe_{46.5}Ni_{53.5}$ alloy is cubic (face centered), the ΔY and ΔZ displacements are derived from the ΔX's given in Table 2 by permutation of the indices. (Henceforth, we will omit the <> on the displacements for simplicity.) For example, ΔX_{321} has the identical value as ΔY_{231} and ΔZ_{123}, and $\Delta X_{321}=\Delta X_{312}=\Delta Y_{231}=\Delta Y_{132}=\Delta Z_{123}=\Delta Z_{213}$. In addition, $\Delta X_{321}=-\Delta X_{\bar{3}21}$ and similarly for the other combinations as illustrated in Figure 20. The nearest-atom pairs that could have, on average, nonradial components are those in the third neighboring shell, lmn=211 (Sparks and Borie, 1965). If the displacements between atom pairs is on average along their interatomic vector, then $\Delta X_{211} = 2\ \Delta X_{121}$ (Fig. 20). For the Fe-Fe pair displacements given in Table 2, ΔX_{211}= 0.0005(2)Å and $2\,\Delta X_{121}$=0.0028(8)Å; thus the (211) Fe-Fe pair displacements have a significant tangential component. From Equations 14 and 15, the magnitude of the displacement between (211) Fe-Fe pairs along the radial direction $|\boldsymbol{\delta}^{\text{Fe-Fe}}_{211}|_{||}$ is 0.0016 (7)Å and $|\boldsymbol{\delta}^{\text{Fe-Fe}}_{211}|_{\perp}$ tangential is 0.0013(7)Å. Thus the (211) Fe-Fe neighbors have a similar radial and tangential component to their displacements. For the (310) Fe-Fe pair displacements, $\Delta X_{310} \cong 3\ \Delta X_{130}$ within the total estimated error, and on average (310) displacements are predominantly radial. These measured displacements provide new information not obtained in other ways about the local atomic arrangements in crystalline solid solutions.

Results for the few crystalline binary alloys that have had their individual pair displacements measured with this 3λ technique are summarized in Figure 21. Here the ΔX static displacements are plotted as a function of the radial distance $\frac{1}{2}\sqrt{l^2+m^2+n^2}$. When there is more than one value for ΔX, the plots show the various values. Most striking is the observation that for the three ordering alloys the near- neighbor Fe-Ni and Cr-Ni bond distances are the smallest of the three possible pairs (Fig. 21). However, for the clustering $Cr_{47}Fe_{53}$ alloy the Cr-Cr nearest-neighbor bond distances are closest and the Cr-Fe are furthest apart. More details, including the short-range-order parameters α and numerical values of the displacements for each shell, are given in the original papers. These pair displacement observations provide a more rigid test of theoretical predictions than variations of the average lattice parameter with concentrations (Froyen and Herring, 1981; Mousseau and Thorpe, 1992).

METHOD AUTOMATION

Because of the large number of data that is collected (~7000 data points per diffuse scattering volume, each consisting of a multichannel energy spectrum with 200 to 500 channels), data collection is under computer control. The program most widely used for converting the reciprocal space coordinates to Eulerian angles and then stepping the three- (or four-) axis diffractometer to this list of

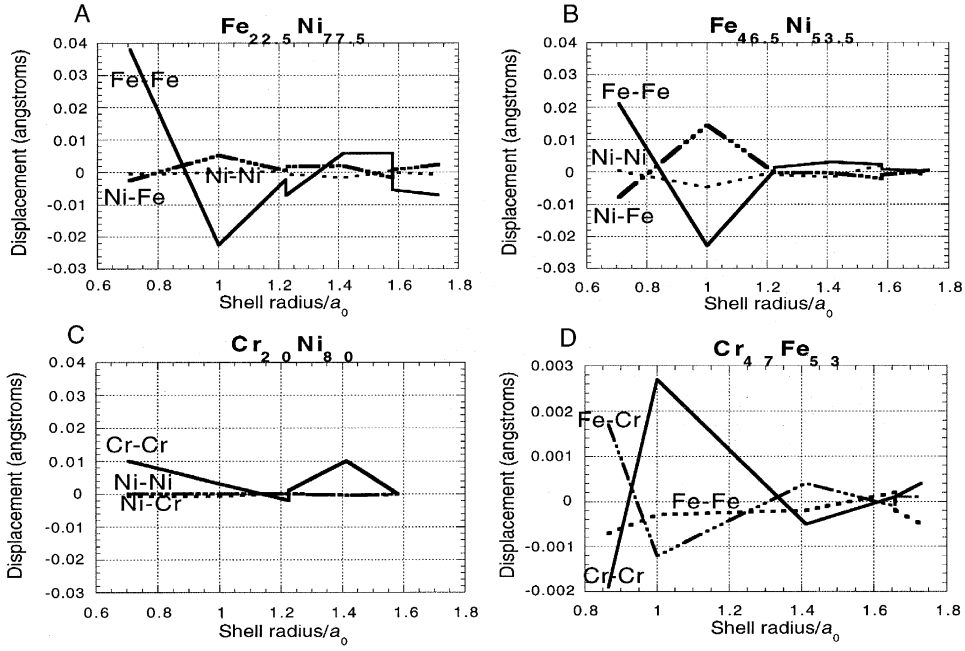


Figure 21. Displacement from the average lattice sites for chemically specific pairs. Shell radius divided by the lattice parameter a_0 becomes 1 for second neighbors (seperated by the cube edge). Courtesy of Ice et al. (1998).

coordinates is SPEC (Certified Scientific Software). Because the diffuse scattering is a slowly varying function without sharp peaks, it is fastest to take the data while the diffractometer is moving and avoid the time-consuming starting and stopping of the diffractometer. The reciprocal space coordinate is estimated from the midrange of the angular scans during data collection. When using the BS separation procedure, it is necessary to have the diffuse scattering data on a uniform cubic grid in reciprocal space. Least-squares procedures do not have this requirement. Since the experiment is enclosed in a radiation hutch, computerized control of all slits, sample positioning, and electronic components in the hutch is desirable.

SAMPLE PREPARATION

The most detailed measurements of diffuse x-ray scattering are made on homogeneous single crystals with polished and etched surfaces. These samples must be carefully prepared and handled to minimize surface roughness effects, surface contamination, and inhomogeneities. To prepare a sample, a single crystal of the intended composition and $\geq$99.9% purity metals should be homogenized for $\geq$3 days at $\sim$50° to 100°C below the crystal melting point in an atmosphere that protects its composition. This step is intended to create a single crystal with uniform chemical composition. For a crystal with cubic symmetry, the crystal should be oriented and then cut with a surface near the <421> normal. This orientation minimizes the goniometer χ range, improves the energy resolution of measurements with a crystal spectrometer, and improves access to a cubic symmetry volume of reciprocal space. The surface of the crystal should be $\sim$20 mm in diameter to ensure that the incident beam is completely intercepted by the sample at all orientations. Electrical discharge machining or chemical sawing minimizes surfaces strains and distortions and is preferred to standard machining. After chemical polishing to remove distorted metal, the crystals should be heat treated to the desired state of thermodynamic equilibrium before quenching. Metallographic polishing is followed by chemical dissolution or electropolishing to produce a mirror smooth surface with nondistorted metal (see SAMPLE PREPARATION FOR METALLOGRAPHY). Verification of negligible distortion is obtained by optical microscopy of subgrain boundary images. Because surface roughness reduces x-ray scattering intensity at low angles, smooth surfaces reduce the corrections to the absolute intensity measurements. The effect of sample roughness on intensity can be checked with fluorescence or RRS measurements to determine if there is any angular variation with respect to **H** in the bisecting geometry (Sparks et al., 1992). The acceptable mosaic spread of the crystals measured from x-ray rocking curves depends on the **H** resolution necessary to resolve the features of interest, but can be as large as $\sim$2 degrees. Verification of the composition to $\pm$ 0.2 at.% or less is needed to reduce the uncertainty in the data analysis. A check of the weight of the raw materials against the weight of the single-crystal ingot will determine if more extensive chemical analysis is necessary.

PROBLEMS

The fundamental problem of diffuse x-ray scattering is that the measurements must be made in absolute units and must be consistent over large ranges in reciprocal space and with different experimental conditions (x ray,

neutron, energy, etc.). Even small systematic uncertainties increase the uncertainties in the recovered local correlations. Special procedures can sometimes be used to calibrate the relative normalization between different experimental conditions, but the reality remains that great care is required to collect meaningful data. For example, near an absorption edge, uncertainties in the scattered intensity can occur because the values of f' and f'' change rapidly with small fluctuations in incident photon energy, as seen in Figure 9. The intensities of RRS also increases as the incident photon energy approaches an absorption edge. As a compromise, the incident energies are usually chosen ~20 eV below an absorption edge. This requires that the x-ray optics, which select the x-ray energy from the white synchrotron spectrum, be very stable so as to control the selected energy fluctuations to less than about ±1 eV.

Similarly, environments must maintain a clean surface free of condensates and oxidation that can add unwanted scattered intensity to measurements. Extremely thin coverages of a few tens of monolayers can contribute noticeable intensity. Equilibrium temperatures can sometimes be difficult to define on quenched samples. For example, quenching to a desired equilibrium local-ordered state may not be possible due to rapid kinetics; quenched-in vacancies enhance diffusion, which can alter chemical order at the actual temperature of interest for local order. Kinetic studies or measurements at the higher temperature are required to ensure that equilibrium order is achieved. Even if diffusion, which changes the α's on cooling, is not a problem, changes in lattice constant on cooling may affect the static displacements. It is better to have the sample in a known state of equilibrium so that the recovered parameters can be directly compared to theory. A discussion of the effect of quenching parameters on the diffuse scattering from an Al-Cu sample has been given by Epperson et al. (1978), and a study of the kinetics of short-range ordering in $Ni_{0.765}Fe_{0.235}$ has been given by Bley et al. (1988).

The complexities of actually performing and interpreting a three-dimensional diffuse scattering experiment require a major commitment of time and resources. To ensure success, we suggest that beginners collaborate with one of the referenced authors who has experience in this science and who has access to the specialized instrumentation and software required for successful diffuse scattering experiments.

ACKNOWLEDGMENTS

We wish to express our appreciation to the early pioneers of diffuse scattering and to our many contemporaries who have contributed to this subject. Research was sponsored by the Division of Materials Sciences, U.S. Department of Energy, under contract DE-AC05-96OR22464 with Lockheed Martin Energy Research. Research was performed at the Oak Ridge National Laboratory Beamline X-14 at the National Synchrotron Light Source, Brookhaven National Laboratory, sponsored by the Division of Materials Sciences and Division of Chemical Sciences, U.S. Department of Energy.

LITERATURE CITED

Åberg, T. and Crasemann, B. 1994. Radiative and radiationless resonant Raman scattering. *In* Resonant Anomalous X-Ray Scattering: Theory and Applications (G. Materlik, C. J. Sparks, and F. Fischer, eds.). pp. 431–448. Elsevier Science Publishing, New York.

Biggs, F., Mendelshohn, L. B., and Mann, J. B. 1975. Hartree-Fock Compton profiles for the element. *Atomic Data Nucl. Tables* 16:201–309.

Bley, F., Amilius, Z., and Lefebvre, S. 1988. Wave vector dependent kinetics of short-range ordering in $^{62}Ni_{0.765}Fe_{0.235}$, studied by neutron diffuse scattering. *Acta Metall. Mater.* 36:1643–1652.

Borie B. and Sparks C. J. 1964. The short-range structure of copper-16 At. % aluminum. *Acta. Crystallogr.* 17: 827–835.

Borie, B. and Sparks, C. J. 1971. The interpretation of intensity distributions from disordered binary alloys. *Acta Crystallogr.* A27:198–201.

Butler, B. D. and Cohen, J. B. 1989. The structure of Cu_3Au above the critical temperature. *J. Appl. Phys.* 65:2214–2219.

Carlsson, G. A., Carlsson, C. A., Berggren, K., and Ribberfors, R. 1982. Calculation of cattering cross sections for increased accuracy in diagnostic radiology. 1. Energy broadening of Compton-scattered photons. *Med. Phys.* 9:868–879.

Cenedese, P., Bley, F., and Lefebvre, S. 1984. Diffuse scattering in disordered ternary alloys: Neutron measurements of local order in stainless steel $Fe_{0.56}Cr_{0.21}Ni_{0.23}$. *Acta Crystallogr.* A40: 228–240.

Chantier, C. T. 1994. Towards improved form factor tables. *In* Resonant Anomalous X-Ray Scattering: Theory and Applications (G. Materlik, C. J. Sparks, and K. Fischer, eds.). pp. 61–78. Elsevier Science Pubishers, New York.

Chakraborty, B. 1995. Static displacements and chemical correlations. *Eur. Phys. Lett.* 30:531–536.

Cowley, J. M. 1950. X-ray measurement of order in single crystals of Cu_3Au. *J. Appl. Phys.* 21:24–30.

Cromer, D. T. and Liberman, D. 1970. Relativistic calculations of anomalous scattering factors for x rays. *J. Chem. Phys.* 53:1891–1898.

Cromer, D. T. and Liberman, D. 1981. Anomalous dispersion calculations near to and on the long-wavelength side of an absorption edge. *Acta. Crystallogr.* A37:267–268.

Dreier, P., Rabe, P., Malzfeldt, W., and Niemann, W. 1983. Anomalous scattering factors from x-ray absorption data by Kramers-Kronig analysis. *In* Proceedings of the International Conference on EXAFS and Near Edge Structure, Frascati, Italy, Sept. 1982 (A. Bianconi, L. Incoccia, and S. Stipcich, eds.). pp. 378–380. Springer-Verlag, New York and Heidelberg.

Drijver, J. W., van der Woude, F., and Radelaar, S. 1977. Mössbauer study of atomic order in Ni_3Fe. I. Determination of the long-range-order parameter. *Phys. Rev. B* 16:985–992.

Eisenberger, P., Platzman, P. M., and Winick, H. 1976a. X-ray resonant Raman scattering: Observation of characteristic radiation narrower than the lifetime width. *Phys. Rev. Lett.* 36:623–625.

Eisenberg, P., Platzman, P. M., and Winick, H. 1976b. Resonant x-ray Raman scattering studies using synchrotron radiation. *Phys. Rev. B.* 13:2377–2380.

Epperson, J. E., Fürnrohr, P., and Ortiz, C. 1978. The short-range-order structure of α-phase Cu-Al alloys. *Acta Crystallogr.* A34:667–681.

Froyen, S. and Herring, C. 1981. Distribution of interatomic spacings in random alloys. *J. Appl. Phys.* 52:7165–7173.

Gehlen, P. C. and Cohen, J. B. 1965. Computer simulation of the structure associated with local order in alloys. *Phys. Rev. A* 139:884–A855.

Georgopoulos, P. and Cohen, J. B. 1977. The determination of short range order and local atomic displacements in disordered binary solid solutions. *J. Phys. (Paris) Colloq.* 38:C7-191–196.

Georgopoulos, P. and Cohen, J. B. 1981. The defect arrangement in (non-stoichiometric) β′-NiAl. *Acta Metall. Mater.* 29:1535–1551.

Gragg. J. E., Hayakawa, M., and Cohen, J. B. 1973. Errors in quantitative analysis of diffuse scattering from alloys. *J. Appl. Crystallogr.* 6:59–66.

Hashimoto, S., Iwasaki, H., Ohshima, K., Harada, J., Sakata, M., and Terauchi, H. 1985. Study of local atomic order in a ternary $Cu_{0.47}N_{0.29}Zn_{0.24}$ alloy using anomalous scattering of synchrotron radiation. *J. Phys. Soc. Jpn.* 54:3796–3807.

Horiuchi, T., Takizawa, S., Tomoo, S., and Mohri, T. 1995. Computer simulation of local lattice distortion in Cu−Au solid solution. *Metall. Mater. Trans. A* 26A:11–19.

Hoyt, J. J., Fontaine, D. D., and Warburton, W. K. 1984. Determination of the anomalous scattering factors for Cu, Ni and Ti using the dispersion relation. *J. Appl. Crystallogr.* 17:344–351.

Ice, G. E. and Sparks, C. J. 1990. Mosaic crystal x-ray spectrometer to resolve inelastic background from anomalous scattering experiments. *Nucl. Instrum. Methods* A291:110–116.

Ice, G. E., Sparks, C. J., Habenschuss, A., and Shaffer, L. B. 1992. Anomalous x-ray scattering measurement of near-neighbor individual pair displacements and chemical order in $Fe_{22.5}Ni_{77.5}$. *Phys. Rev. Lett.* 68:863–866.

Ice, G. E., Sparks, C. J., Jiang, X., and Robertson, L. 1998. Diffuse scattering measurements of static atomic displacements in crystalline binary solid solutions. *J. Phase Equilib.* 19:529–537.

Ice, G. E., Sparks, C. J., and Shaffer, L. B. 1994. Chemical and displacement atomic pair correlations in crystalline solid solutions recovered by resonant (anomalous) x-ray scattering. *In* Resonant Anomalous X-Ray Scattering: Theory and Applications (G. Materlik, C. J. Sparks, and K. Fischer, eds.). pp. 265–294. Elsevier Science Publishing, New York.

James, R.W. 1948. The Optical Principles of the Diffraction of X-Rays. Cornell University Press, Ithaca, N.Y.

Jiang, X., Ice, G. E., Sparks, C. J., Robertson, L., and Zachack, P. 1996. Local atomic order and individual pair displacements of $Fe_{46.5}Ni_{53.5}$ and $Fe_{22.5}Ni_{77.5}$ from diffuse x-ray scattering studies. *Phys. Rev. B* 54:3211–3226.

Jiang, X., Ice, G. E., Sparks, C. J., and Zschack, P. 1995. Recovery of SRO parameters and pairwise atomic displacements in a $Fe_{46.5}Ni_{53.5}$ alloy. *In* Applications of Synchrotron Radiation Techniques to Materials Science. *Mater. Res. Soc. Symp. Proc.* 375:267–273.

Kawamura, T. and Fukamachi, T. 1978. Application of the dispersion relation to determine the anomalous scattering factors. Proceedings of the International Conference on X-ray and VUV Spectroscopies, Sendai, Japan. *J. Appl. Phys.*, Suppl. 17-2:224–226.

Kostorz, G. 1996. X-ray and neutron scattering. *In* Physical Metallurgy, 4th and revised and enhanced edition. (R. W. Cahn and P. Hoosen, eds.). pp. 1115–1199. Elsevier Science Publishing, New York.

Lengeler, B. 1994. Experimental determination of the dispersion correction $f'(E)$ of the atomic scattering factor. *In* Resonant Anomalous X-Ray Scattering: Theory and Applications (G. Materlik, C. J. Sparks, and K. Fischer, eds.). pp. 35–60. Elsevier Science Publishing, New York.

Miller, M. K., Cerezo, A., Hetherington, M. G., and Smith, G. D. W. 1996. Atom Probe Field Ion Microscopy. Oxford University Press, Oxford.

Mousseau, N. and Thorpe, M. F. 1992. Length distributions in metallic alloys. *Phys. Rev. B* 45:2015–2022.

Müller, P. P., Schönfeld, B., Kostorz, G., and Bührer, W. 1989. Guinier-Preston I zones in Al-1.75 at.% Cu single crystals. *Acta Metall. Mater.* 37:2125–2132.

Pierron-Bohnes, V., Cadeville, M. C., and Gautier, F. 1983. Magnetism and local order in dilute Fe-C alloys. *J. Phys. F* 13:1689–1713.

Reinhard, L., Robertson, J. L., Moss, S. C., Ice, G. E., Zschack, P., and Sparks, C. J. 1992. Anomalous-x-ray scattering study of local order in BCC $Fe_{0.53}Cr_{0.47}$. *Phys. Rev B* 45:2662–2676.

Renaud, G., Motta, N., Lançon, F., and Belakhovsky, M. 1988. Topological short-range disorder in $Au_{1-X}Ni_x$ solid solutions: An extended x-ray absorption fine structure spectroscopy and computer-simulation study. *Phys. Rev. B* 38:5944–5964.

Robertson, J. L., Sparks, C. J., Ice, G. E., Jiang, X., Moss, S. C., and Reinhard, L. 1998. Local atomic arrangements in binary solid solutions studied by x-ray and neutron diffuse scattering from single crystals. *In* Local Structure from Diffraction: Fundamental Materials Science Series (M. F. Thorpe and S. Billinge, eds.).. Plenum, New York. In press.

Sasaki, S. 1989. Numerical tables of anomalous scattering factors calculated by the Cromer and Liberman's method. *KEK Report* 88–14.

Scheuer, U. and Lengeler, B. 1991. Lattice distortion of solute atoms in metals studied by x-ray absorption fine structure. *Phys. Rev. B* 44:9883–9894.

Schönfeld, B., Ice, G. E., Sparks, C. J., Haubold, H.-G., Schweika, W., and Shaffer, L.B. 1994. X-ray study of diffuse scattering in Ni-20 at% Cr. *Phys. Status Solidi B* 183:79–95.

Schwartz, L. H. and Cohen, J. B. 1987. Diffraction from Materials. Springer-Verlag, New York and Heidelberg.

Schweika, W. 1998. Disordered Alloys: Diffuse Scattering and Monte Carlo Simulations, Vol. 141. Springer-Verlag, New York and Heidelberg.

Simon, J. P. and Lyon, O. 1991. The nature of the scattering tail in Cu-Ni-Fe and invar alloys investiaged by anomalous small angle x-ray scattering. *J. Appl. Crystallogr.* 24:1027–1034.

Sparks, C. J. 1974. Inelastic resonance emission of x rays: Anomalous scattering associated with anomalous dispersion. *Phys. Rev. Lett.* 33:262–265.

Sparks, C. J. and Borie, B. 1965. Local atomic arrangements studied by x-ray diffraction. *In* AIME Conference Proceedings 36 (J.B. Cohen and J.E. Hilliard, eds.). pp. 5–50. Gordon and Breach, New York.

Sparks, C. J., Ice, G. E., Shaffer, L. B., and Robertson, J. L. 1994. Experimental measurements of local displacement and chemical pair correlations in crystalline solid solutions. *In* Metallic Alloys: Experimental and Theoretical Perspectives (J. S. Faulkner and R. G. Jordan, eds.). pp. 73–82. Kluwer Academic Publishers, Dordrecht, The Netherlands, NATO Vol. 256.

Sparks, C. J. and Robertson, J. L. 1994. Guide to some crystallographic symbols and definitions with discussion of short-range correlations. *In* Resonant Anomalous X-ray Scattering: Theory and Applications (G. Materlik, C. J. Sparks, and K. Fischer,

eds.). pp. 653–664. Elsevier Science North-Holland, Amsterdam, The Netherlands.

Suortti, P., Hastings, J. B., and Cox, D. E. 1985. Powder diffraction with synchrotron radiation. I. Absolute measurements. *Acta Crystallogr. A*41:413–416.

Tibballs, J. E. 1975. The separation of displacement and substitutional disorder scattering: A correction for structure factor ratio variation. *J. Appl. Crystallogr.* 8:111–114.

Warren, B. E. 1969 (reprinted) 1990. *In* X-Ray Diffraction. Dover Publications, New York.

Warren, B. E., Averbach, G. L., and Roberts, B. W. 1951. Atomic size effect in the x-ray scattering by alloys. *J. Appl. Phys.* 22:1493–1496.

Welberry, T. R. and Butler, B. D. 1995. Diffuse x-ray scattering from disordered crystals. *Chem Rev.* 95:2369–2403.

Wu, T. B., Matsubara, E., and Cohen, J. B. 1983. New procedures for quantitative studies of diffuse x-ray scattering. *J. Appl. Crystallogr.* 16:407–414.

Zunger, A. 1994. First-principles statistical mechanics of semiconductor alloys and intermetallic compounds. *In* Statics and Dynamics of Alloy Phase Transitions, Vol. B319 of NATO ASI Series B (P.E.A. Turchi and A. Gonis, eds.). p. 361. Plenum, New York.

KEY REFERENCES

Cowley, J.M. 1975. Diffraction Physics. North-Holland, Amsterdam, The Netherlands.

This reference is a good basic introduction into diffraction physics and x-ray techniques.

Materlik, G., Sparks, C. J., and Fisher, K. (eds.).. 1994. Resonant Anomalous X-ray Scattering. North-Holland, Amsterdam, The Netherlands.

This reference is a collection of recent work on the theory and application of resonant x-ray scattering techniques. It provides the most complete introduction to the application of anomalous (resonant) scattering techniques and how the field has been revolutionized by the availability of intense synchrotron radiation sources.

Schwartz and Cohen, 1987. See above.

This is another introductory text on x-ray diffraction with an emphasis on the application to materials. The section on diffuse x-ray scattering is especially strong and the notation is the same as used in this unit.

Schweika, 1998. See above.

This monograph provides an excellent reference to the interplay between experiment and theory in the field of local atomic order in alloys.

Warren, 1969 (reprinted 1990). See above.

This reference provides an authoritative treatment of all phases of diffuse x-ray scattering, including thermal diffuse scattering, TDS, short-range order, and atomic size displacement scattering. Although it does not include a modern outlook on the importance of resonant scattering, it provides a clear foundation for virtually all modern treatments of diffuse scattering from materials.

APPENDIX A: DERIVATION OF THE DIFFUSE INTENSITY

Our interest is in the diffusely distributed intensity. To separate Equation 3 into an intensity that may be sharply peaked and one that is diffusely distributed, we follow the method of Warren (1969). This method expands $e^{2\pi i\mathbf{H}\bullet(\boldsymbol{\delta}_p-\boldsymbol{\delta}_q)}$ in a Taylor series about $\boldsymbol{\delta}_p - \boldsymbol{\delta}_q$ and examines the displacement terms as the separation of atom pairs becomes large, $p-q\rightarrow\infty$. As the x-ray or neutron beam intercepts many atoms and the atoms undergo many thermal vibrations during the period of the intensity measurement, both a spatial and a time average are taken. These are indicated by $< >$. The spatial and time average of the jth-order Taylor series expansion of the exponential displacement term is

$$\langle e^{e\pi i\mathbf{H}} \bullet (\boldsymbol{\delta}_p - \boldsymbol{\delta}_g)\rangle \equiv \langle e^{iX_{pq}}\rangle = 1 + i\langle X_{pq}\rangle - \frac{\langle X_{pq}^2\rangle}{2} - \frac{i\langle X_{pq}^3\rangle}{3!} + \cdots + \frac{i^j\langle X_{pq}^j\rangle}{j!} \tag{16}$$

The time average for harmonic thermal displacements causes odd-order terms to vanish. With the definition $\langle X_{pq}\rangle = \langle X_p - X_q\rangle$, so that $\langle (X_p - X_q)^2\rangle = \langle X_p^2\rangle + \langle X_q^2\rangle - 2\langle X_pX_q\rangle$, and for sharply peaked Bragg reflections, where $p-q\rightarrow\infty$, the displacements become uncorrelated so that $\langle X_pX_q\rangle = 0$. Therefore, $\langle X_{pq}^2\rangle_{p-q\rightarrow\infty} = \langle X_p^2\rangle + \langle X_q^2\rangle$, and with the harmonic approximation we can estimate the long-range dynamical displacement term by

$$\langle e^{iX_{pq}}\rangle_{p-q\rightarrow\infty} \cong e^{-\frac{1}{2}(\langle x_p^2\rangle + \langle x_q^2\rangle)} = e^{-(M_p+M_q)} \tag{17}$$

Here M_p is the usual designation for the Debye-Waller temperature factor (Warren, 1990, p. 35). The subject is treated by Chen in KINEMATIC DIFFRACTION OF X RAYS. We also include the mean-square static displacements in M. Experience has shown the validity of the harmonic approximation in Equation 17 to account for the reduction in the intensity of the Bragg reflections as a function of $\mathbf{H}$. With the understanding that when there is an A atom at site p or q, M_p or M_q is written as M_A, and similarly M_B when there is a B at p or q, the fundamental Bragg intensity for an alloy is given by the substitution of Equation 17 into Equation 3 as

$$I(\mathbf{H})_{\mathbf{Fund}} = |C_\mathrm{A}f_\mathrm{A}e^{-M_\mathrm{A}} + C_\mathrm{B}f_\mathrm{B}f_\mathrm{B}e^{-M_\mathrm{B}}|^2 \sum_p \sum_q e^{2\pi i\mathbf{H}\bullet(\mathbf{R}_p-\mathbf{R}_q)} \tag{18}$$

This expression accounts for the reduced intensity of the fundamental Bragg reflections due to thermal motion and static displacements of the atoms. Fundamental reflections scale as the average scattering factor and are insensitive to how the chemical composition is distributed on the lattice sites. When the alloy has long-range order among the kinds of atoms, the α_{pq}'s do not converge rapidly with larger p, q and account for the superstructure Bragg reflections that depend on how the atoms are distributed among the sites. We are now concerned with the distribution of this thermal and static scattering. To recover the diffuse intensity, we subtract $I(\mathbf{H})_\mathrm{Fund}$ from $I(\mathbf{H})_\mathrm{Total}$ in Equation 3. To avoid making the harmonic approximation of Equation 17, we subtract $I(\mathbf{H})_\mathrm{Fund}$ term by term and take

the limit as $p-q \rightarrow \infty$. For example, to second order we have

$$\langle e^{iX_{pq}} \rangle_{p-q\rightarrow\infty} = 1 - \frac{1}{2}\left(\langle X_p^2 \rangle + \langle X_2^q \rangle\right) \tag{19}$$

By substitution of Equation 19 for each of the $\langle e^{iX_{pq}} \rangle$ terms in Equation 3 and assignment of the proper atom identity for different pairs, and recalling that as $p-q \rightarrow \infty$, $\rightarrow \infty$, α_{pq}=0, we subtract this expression for $I(\mathbf{H})_{\text{Fund}}$ from $I(\mathbf{H})_{\text{Total}}$ and write the diffuse scattering to second order in the displacements as

$$\begin{aligned} I(\mathbf{H})_{\text{Diffuse}} = \sum_p \sum_q \Big\{ & (C_\text{A}^2 + C_\text{A}C_B\alpha_{pq})|f_\text{A}|^2 \\ & \times \left(1 + i\langle X_p^\text{A} - X_q^\text{A}\rangle - \frac{1}{2}\langle (X_p^\text{A} - X_q^\text{A})^2\rangle\right) \\ & - C_\text{A}^2|f_\text{A}|^2\left(1 - \langle X^2\rangle^\text{A}\right) + C_\text{A}C_\text{B}\,(1-\alpha_{pq})f_\text{A}f_\text{B}^* \\ & \times \left(1 + i\langle X_p^\text{B} - X_p^\text{A}\rangle - \frac{1}{2}\langle (X_p^\text{B} - X_q^\text{A})^2\rangle\right) \\ & - 2C_\text{A}C_\text{B}f_\text{A}f_\text{B}^*\left(1 - \frac{1}{2}\langle X^2\rangle^\text{A} - \frac{1}{2}\langle X^2\rangle^\text{B}\right) \\ & + C_\text{A}C_\text{B}\,(1-\alpha_{pq})f_\text{A}f_\text{B}^* \\ & \times \left(1 + i\langle X_p^\text{A} - X_q^\text{B}\rangle - \frac{1}{2}\langle (X_p^\text{A} - X_q^\text{B})^2\rangle\right) \\ & + (C_\text{B}^2 + C_\text{A}C_\text{B}\alpha_{pq})|f_\text{B}|^2 \\ & \times \left(1 + i\langle X_p^\text{B} - X_q^\text{B}\rangle - \frac{1}{2}\langle (X_p^\text{B} - X_q^\text{B})\rangle\right) \\ & - C_\text{B}^2|f_\text{B}|^2\left(1 - \langle X^2\rangle^\text{B}\right)\Big\} e^{2\pi i\mathbf{H}\bullet(\mathbf{R}_p - \mathbf{R}_q)} \end{aligned} \tag{20}$$

Our use of the double sum requires that the pairs of atoms be counted in both directions, that is, a p, q pair will become a q, p pair such that $i(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q) = -i(\boldsymbol{\delta}_q - \boldsymbol{\delta}_q)$ as shown in Figure 22. As seen in Equation 2, the $I(\mathbf{H})_{\text{Total}}$ double sum is made up of p, q elements that are the product of four complex numbers: the two complex scattering factors and the two complex phase factors. For every p, q pair there is a corresponding q, p pair where the two scattering factors and the two phase factors are the p, q pair complex conjugates; hence the p, q and q, p elements add up to a real number. This means that the terms in the series expansion of the fluctuation displacements must add in

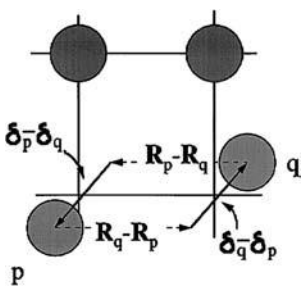


Figure 22. Schematic illustration showing that since all the pairs of atoms are counted as to kind and displacement in both directions, odd-power terms in the displacements are replaced with their negatives $i(\boldsymbol{\delta}_p - \boldsymbol{\delta}_q) = -i(\boldsymbol{\delta}_q - \boldsymbol{\delta}_q)$ and $\mathbf{R}_p - \mathbf{R}_q = -(\mathbf{R}_q - \mathbf{R}_p)$ so that the imaginary terms cancel.

pairs to give real intensity and from Equation 20, the j = odd order displacement terms have a $\sin 2\pi(\mathbf{H}\bullet\mathbf{R})$ $\langle (X_p^\text{A} - X_q^\text{A})^j\rangle$ dependence and the j = even-order terms have a $\cos 2\pi(\mathbf{H}\bullet\mathbf{R})$ $\langle (X_p^\text{A} - X_q^\text{A})^j\rangle$ dependence; the imaginary components cancel.

From the definition of an average lattice, the weighted average of the displacements for all the kinds of pairs formed for any coordination shell is zero (Warren et al., 1951),

$$\begin{aligned} 2(\alpha_{pq} - 1)\langle \boldsymbol{\delta}_p^\text{A} - \boldsymbol{\delta}_q^\text{B}\rangle = & \left(\frac{C_\text{A}}{C_\text{B}} + \alpha_{pq}\right)\langle \boldsymbol{\delta}_p^\text{A} - \boldsymbol{\delta}_q^\text{A}\rangle \\ & + \left(\frac{C_\text{B}}{C_\text{A}} + \alpha_{pq}\right)\langle \boldsymbol{\delta}_p^\text{B} - \boldsymbol{\delta}_q^\text{B}\rangle \end{aligned} \tag{21}$$

If a crystal structure has more than one kind of site symmetry (sublattices with different site symmetries), than Equation 21 may be true for only that sublattice with all the same site symmetries. Different-sized atom species will most likely have a preference for the symmetry of a particular sublattice. This preference could produce long-range correlations and superstructure reflections from which site occupation preferences can be recovered. Disorder among the atoms on any one sublattice and between sublattices will produce scattering from which pair correlations can be recovered. Discussion of this issue is beyond the scope of this unit. A study of a partially ordered non-stoichiometric bcc NiAl crystal, where one of the two sublattices with the same site symmetry is partly occupied by vacancies, has been discussed in Georgopoulos and Cohen (1981). A general review of the application of diffuse scattering measurements to more complicated structures with different site symmetries shows a wider use of models to reduce the number of variables necessary to describe the local pair correlations (Welberry and Butler, 1995).

Equation 20 can be expressed in a more tractable form through three steps: (1) Replace the double sum over atomic sites p and q by N single sums around an origin site where the relative sites are identified by lattice difference lmn such that $\mathbf{R}_p - \mathbf{R}_q = \mathbf{R}_0 - \mathbf{R}_{lmn}$. This approximation neglects surface effects. (2) Use trigonometric functions to simplify the phase factors. (3) Substitute Equation 21 into Equation 20. With these steps, the diffuse intensity can be expressed as,

$$\begin{aligned} \frac{I(\mathbf{H})_{\text{Diffuse}}}{N} = & \sum_{lmn} C_\text{A}C_\text{B}|f_\text{A} - f_\text{B}|^2\alpha_{lmn}\cos 2\pi\mathbf{H}\bullet(\mathbf{R}_0 - \mathbf{R}_{lmn}) \\ & - \sum_{lmn}[(C_\text{A}^2 + C_\text{A}C_\text{B}\alpha_{lmn})\,\text{Re}(f_\text{a}(f_\text{A} - f_\text{B})^*) \\ & \times \langle X_0^\text{A} - X_{lmn}^\text{A}\rangle - (C_\text{B}^2 + C_\text{A}C_\text{B}\alpha_{lmn})\text{Re}(f_\text{B}(f_\text{A} - f_\text{B})^*) \\ & \times \langle X_0^\text{B} - X_{lmn}^\text{B})\rangle]\sin 2\pi\mathbf{H}\bullet(\mathbf{R}_0 - \mathbf{R}_{lmn}) \\ & + \sum_{lmn}\Big[C_\text{A}^2|f_\text{A}|^2\Big(\langle X^2\rangle^\text{A} - \left(1 + \frac{C_\text{B}}{C_\text{A}}\alpha_{lmn}\right) \\ & \times \frac{1}{2}\langle (X_0^\text{A} - X_{lmn}^\text{A})^2\rangle\Big) + C_\text{A}C_\text{B}f_\text{A}f_\text{B}^*(\langle X^2\rangle^\text{A} + \langle X^2\rangle^\text{B} \\ & - (1-\alpha_{lmn})\,\langle (X_0^\text{B} - x_{lmn}^\text{A})^2\rangle) + C_\text{B}^2|f_\text{B}|^2\Big(\langle X^2\rangle^\text{B} \\ & - \left(1 + \frac{C_\text{A}}{C_\text{B}}\alpha_{lmn}\right)\frac{1}{2}\langle (X_0^\text{B} - X_{lmn}^\text{B})^2\rangle\Big)\Big] \\ & \times \cos 2\pi\mathbf{H}\bullet(\mathbf{R}_0 - \mathbf{R}_{lmn}) \end{aligned} \tag{22}$$

Equation 22 is a completely general description of the diffuse scattering from any crystal structure through the second moment of the static and thermal displacements. We now apply this to binary solid solutions, which have received the most attention among crystalline solid solutions.

It is helpful to choose real-space basis vectors **a**, **b**, **c** that reflect the long-range periodicity of the crystal structure. This long-range periodicity in turn is reflected in the intensity distribution in the reciprocal space lattice with basis vectors $\mathbf{a}^* = 1/\mathbf{a}$, $\mathbf{b}^* = 1/\mathbf{b}$, and $\mathbf{c}^* = 1/\mathbf{c}$, as shown in Figure 3. For an alloy that is on average statistically cubic, such as that shown in Figure 3, we define

$$\mathbf{R}_0 - \mathbf{R}_{lmn} \equiv \frac{l}{2}\mathbf{a} + \frac{m}{2}\mathbf{b} + \frac{n}{2}\mathbf{c},\ \mathbf{H} \equiv h_1\mathbf{a}^* + h_2\mathbf{b}^* h_3\mathbf{c}^* \quad (23)$$

so that

$$2\pi\mathbf{H} \bullet (\mathbf{R}_0 - \mathbf{R}_{lmn} = \pi(h_1 l + h_2 m + h_3 n) \quad (24)$$

In addition,

$$\boldsymbol{\delta}_{lmn} \equiv \Delta X_{lmn}\mathbf{a} + \Delta Y_{lmn}\mathbf{b} + \Delta Z_{lmn}\mathbf{c} \quad (25)$$

so that

$$X_0 - X_{lmn} \equiv 2\pi\mathbf{H} \bullet (\boldsymbol{\delta}_0 - \boldsymbol{\delta}_{lmn}) \equiv 2\pi[h_1(\Delta X_0 - \Delta X_{lmn}) + h_2 + (\Delta Y_0 - \Delta Y_{lmn}) + h_3(\Delta Z_0 - \Delta Z_{lmn})] \quad (26)$$

This definition of $\mathbf{R}_0 - \mathbf{R}_{lmn}$ causes the continuous variables $h_1 h_2 h_3$ in reciprocal space to have the integer values of the Miller indices at reciprocal lattice points. We further specify that the site symmetry is cubic such as for the bcc Fe structure and fcc Cu structure. With these definitions, the various diffuse x-ray scattering terms in Equation 22 can be written, starting with the local chemical order term as

$$\frac{I(\mathbf{H})_{\mathrm{SRO}}}{N} = \sum_{lmn} C_\mathrm{A}C_\mathrm{B}|f_\mathrm{A} - f_\mathrm{B}|^2\alpha_{lmn}\cos\pi(h_1 l + h_2 m + h_3 n) \quad (27)$$

which was first given by Cowley (1950). For bcc Fe and fcc Cu structures, $\alpha_{lmn} = \alpha_{\overline{lmn}} = \alpha_{l\bar{m}n} = \alpha_{lm\bar{n}} = \alpha_{l\overline{mn}} = \alpha_{\overline{lmn}} = \alpha_{\overline{l}mn}$ and the cosine term takes the form $\cos\pi h_1 l \cos\pi h_2 m \cos\pi h_3 n$.

For the first-order displacement term,

$$\frac{I(\mathbf{H})_{j=1}}{N} = -\mathrm{Re}\left(f_\mathrm{A}\left(f_\mathrm{A}^* - f_\mathrm{B}^*\right)\right)\left[h_1 Q_x^{\mathrm{AA}} + h_2 Q_Y^{\mathrm{AA}} + h_3 Q_Z^{\mathrm{AA}}\right] + \mathrm{Re}\left(f_\mathrm{B}(f_\mathrm{A}^* - f_\mathrm{B}^*)\right)\left[h_1 Q_X^{\mathrm{BB}} + h_2 Q_Y^{\mathrm{BB}} + h_3^{\mathrm{BB}}\right] \quad (28)$$

where

$$Q_X^{\mathrm{AA}} = 2\pi\sum_{lmn}\left(C_\mathrm{A}^2 + C_\mathrm{A}C_\mathrm{B}\alpha_{lmn}\right)\langle\Delta X_{lmn}^\mathrm{A}\rangle_0^\mathrm{A}\sin\pi h_1 l \times\cos\pi h_2 m\cos\pi h_3 n$$

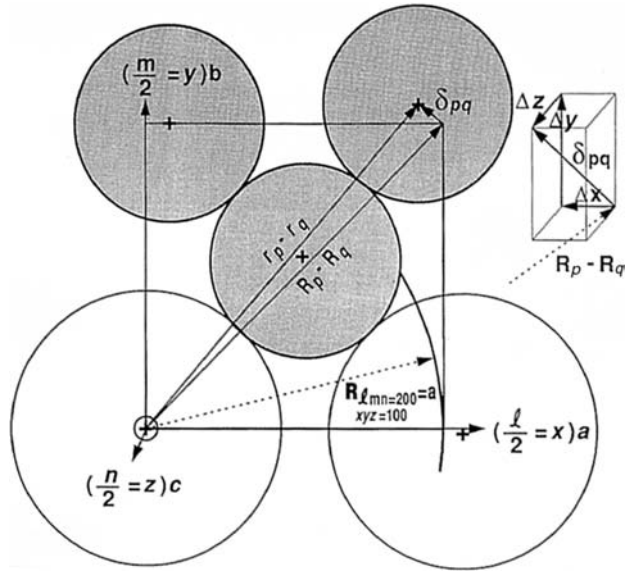


Figure 23. The rectangular square of solid lines is the average lattice about which the atom centers (+) are displaced by the amount $\boldsymbol{\delta}_{pq}$. Shown in the smaller box on the right are the rectangular components of the displacement, $\Delta x, \Delta y$, and Δz. Courtesy of Jiang et al. (1996).

and

$$Q_X^{\mathrm{BB}} = 2\pi\sum_{lmn}\left(C_\mathrm{B}^2 + C_\mathrm{A}C_\mathrm{B}\alpha_{lmn}\right)\langle\Delta X_{lmn}^\mathrm{B}\rangle_0^\mathrm{B}\sin\pi h_1 l \times\cos\pi h_2 m\cos\pi h_3 n$$

and similarly for the other terms as given by Borie and Sparks (1964) and Georgopoulos and Cohen (1977). Where

$$\langle\Delta X_{lmn}^\mathrm{A}\rangle_0^\mathrm{A} = \langle\Delta X_0^\mathrm{A} - \Delta X_{lmn}^\mathrm{A}\rangle$$

Equation 28 is a result given by Borie and Sparks (1964) that avoids an earlier assumption of radial displacements first given by Warren et al. (1951). Shown in Figure 23 is a schematic of the displacements described by Equation 28.

Diffuse scattering from the second-order displacement term can be expanded as,

$$\begin{aligned}\frac{I(\mathbf{H})_{j=2}}{N} = {} & |f_A|^2\left(h_1^2 R_X^{\mathrm{AA}} + h_2^2 R_Y^{\mathrm{AA}} + h_3^2 R_Z^{\mathrm{AA}}\right) \\ & + f_\mathrm{A}f_\mathrm{B}\left(h_1^2 R_X^{\mathrm{AB}} + h_2^2 R_Y^{\mathrm{AB}} + h_3^2 R_Z^{\mathrm{AB}}\right) \\ & + |f_\mathrm{B}|^2\left(h_1^2 R_X^{\mathrm{BB}} + h_2^2 R_Y^{\mathrm{BB}} + h_3^2 R_Z^{\mathrm{BB}}\right) \\ & + |f_\mathrm{A}|^2\left(h_1 h_2 S_{XY}^{\mathrm{AA}} + h_1 h^3 S_{XZ}^{\mathrm{AA}} + h_2 h_3 S_{YZ}^{\mathrm{AA}}\right) \\ & + f_\mathrm{A}f_\mathrm{B}^*\left(h_1 h_2 S_{XY}^{\mathrm{AB}} + h_1 h_3 S_{XZ}^{\mathrm{AB}} + \mathrm{h_2 h_3 S_{YZ}^{AB}}\right) \\ & + |f_\mathrm{B}|^2\left(h_1 h_2 S_{XY}^{\mathrm{BB}} + h_1 h_3 S_{XZ}^{\mathrm{BB}} + h_2 h_3 S_{YZ}^{\mathrm{BB}}\right)\end{aligned} \quad (29)$$

where

$$R_X^{\mathrm{AA}} = 4\pi^2 C_\mathrm{A}^2\sum_{lmn}\left\{\langle\Delta X^2\rangle^\mathrm{A} - \left(1 + \frac{C_\mathrm{B}}{C_\mathrm{A}}\alpha_{lmn}\right) \times\left[\langle(\Delta X^2)_{lmn}^\mathrm{A}\rangle_0^\mathrm{A} - \langle\Delta X_0^\mathrm{A}\Delta X_{lmn}^\mathrm{A}\rangle\right]\right\} \times\cos\pi(h_1 l + h_2 m + h_3 n)$$

and similarly for R_Y and R_Z with Y and Z replacing X, respectively, or $R_X^{\mathrm{AA}}\ (h_1h_2h_3) = R_Y^{\mathrm{AA}}(h_2h_3h_1)$. The R^{BB} terms are given by

$$R_X^{\mathrm{BB}} = 4\pi^2 C_{\mathrm{B}}^2 \sum_{lmn}\left\{\langle\Delta X^2\rangle^{\mathrm{B}} - \left(1+\frac{C_{\mathrm{A}}}{C_{\mathrm{B}}}\alpha_{lmn}\right) \times \left[\langle(\Delta X^2)_{lmn}^{\mathrm{B}}\rangle_0^{\mathrm{B}} - \langle\Delta X_0^{\mathrm{B}}\Delta X_{lmn}^{\mathrm{B}}\rangle\right]\right\} \times \cos\pi(h_1 l + h_2 m + h_3 n) \quad (30)$$

For the R^{AB} terms, we have

$$R_X^{\mathrm{AB}} = 4\pi^2 C_{\mathrm{A}} C_{\mathrm{B}} \sum_{lmn}\Big[\langle\Delta X^2\rangle^{\mathrm{A}} + \langle\Delta X^2\rangle^{\mathrm{B}} - (1-\alpha_{lmn}) \times (\langle(\Delta X^2)_{lmn}^{\mathrm{A}}\rangle_0^{\mathrm{B}} + \langle(\Delta X^2)_0^{\mathrm{B}}\rangle_{lmn}^{\mathrm{A}} - 2\langle\Delta X_0^{\mathrm{B}}\Delta X_{lmn}^{\mathrm{A}}\rangle)\Big] \times \cos\pi(h_1 l + h_2 m + h_3 n) \quad (31)$$

and similar terms for R_Y^{AB} with Y replacing X and for R_Z^{AB} with Z replacing X. For the cross terms

$$S_{XY}^{\mathrm{AA}} = 4\pi^2 C_{\mathrm{A}}^2 \sum_{lmn}\left(1+\frac{C_{\mathrm{B}}}{C_{\mathrm{A}}}\alpha_{lmn}\right)\langle\Delta X_0^{\mathrm{A}} - \Delta X_{lmn}^{\mathrm{A}}\rangle \times \langle\Delta Y_0^{\mathrm{A}} - \Delta Y_{lmn}^{\mathrm{A}}\rangle\cos\pi(h_1 l + h_2 m + h_3 n) \quad (32)$$

and similarly for S_{XZ}^{AA} and S_{YZ}^{AA} with XY replaced with XZ and YZ, respectively. The terms S_{XY}^{BB}, S_{XZ}^{BB}, and S_{YZ}^{BB} are derived from the S^{AA} terms by replacing AA with BB and by replacing C_{A}^2, with C_{B}^2 and $C_{\mathrm{A}}/C_{\mathrm{B}}$ with $C_{\mathrm{B}}/C_{\mathrm{A}}$. For the term S_{XY}^{AB}, we write

$$S_{XY}^{\mathrm{AB}} = 8\pi^2 C_{\mathrm{A}} C_{\mathrm{B}} \sum_{lmn}(1-\alpha_{lmn})\langle(\Delta X_0^{\mathrm{B}} - \Delta X_{lmn}^{\mathrm{A}}) \times (\Delta Y_0^{\mathrm{B}} - \Delta Y_{lmn}^{\mathrm{A}})\rangle\cos\pi(h_1 l + h_2 m + h_3 n) \quad (33)$$

and similarly for S_{XZ}^{AB} and S_{YZ}^{AB}, where XZ and YZ replace XY, respectively. The periodicity of the terms in Equations 27, 28, and 29, and the assumption that the scattering factor terms can be made independent of $\mathbf{H}$, permit us to write their sum as

$$\frac{I(\mathbf{H})_{\mathrm{Diffuse}}}{N|f(\mathbf{H}|^2)} = \mathrm{A}(h_1h_2h_3) + h_1\mathrm{B}(h_1h_2h_3) + h_2\mathrm{B}(h_2h_3h_1) + h_3\mathrm{B}(h_3h_1h_2) + h_1^2\mathrm{C}(h_1h_2h_3) + h_2^2\mathrm{C}(h_2h_3h_1) + h_3^2\mathrm{C}(h_3h_1h_2) + h_1h_2\mathrm{D}(h_1h_2h_3) + h_1h_3\mathrm{D}(h_2h_3h_1) + h_2h_3\mathrm{D}(h_3h_1h_2) \quad (34)$$

where $\mathrm{A}(h_1, h_2, h_3)$ is given by Equation $27 \div |f(\mathbf{H})|^2$, and $\mathrm{B}(h_1h_2h_3)$ contains the two terms $-f_{\mathrm{A}}\Delta f^* Q_X^{\mathrm{AA}} + f_{\mathrm{B}}\Delta f^* Q_X^{\mathrm{BB}}$ given by Equation $28 \div |f(\mathbf{H})|^2$, and likewise for the other terms.

APPENDIX B: NEUTRON MAGNETIC DIFFUSE SCATTERING

The elastic diffuse scattering of neutrons from binary alloys with magnetic short-range order is composed of three parts: the nuclear scattering, the magnetic scattering, and the nuclear magnetic interference scattering. The nuclear scattering length for neutrons is analogous to the x-ray atomic scattering factor for x rays. Thus the information obtained is the same for chemical short-range order and displacements as discussed for x rays. Because neutrons have a magnetic moment, there is also magnetic scattering associated with the unpaired electron spins (see MAGNETIC NEUTRON SCATTERING). If the direction of magnetization is perpendicular to the scattering plane, the magnetic scattering cross-section (in barns) for an A-B alloy of atomic concentration C_{A} of A atoms is given by

$$I_{\mathrm{M}}(\mathbf{H}) = C_{\mathrm{A}} C_{\mathrm{B}} T(\mathbf{H})(0.270)^2 \quad (35)$$

Here, $T(\mathbf{H})$ is the moment-moment correlation function expressed in terms of $\mu_{lmn} f_{lmn}(\mathbf{H})$, where μ_{lmn} is the magnetic moment of the atom on site lmn and $f_{lmn}(\mathbf{H})$ its magnetic form factor (the magnetic scattering comes from the unpaired electrons rather than the nucleus so that it has a scattering angle dependent form factor much like that for x rays):

$$\mu_{lmn} = \mu_{lmn} f_{lmn}(\mathbf{H}) \quad (36)$$

$$C_{\mathrm{A}} C_{\mathrm{B}} T(\mathbf{H}) = \sum_{lmn} e_{lmn}^{2\pi i \mathbf{H}\cdot\mathbf{R}} \langle\mu_{lmn}(\mathbf{H})[\mu_{000}(\mathbf{H}) - \langle\mu(\mathbf{H})\rangle]\rangle \quad (37)$$

The nuclear magnetic interference term $I_{\mathrm{NM}}(\mathbf{H})$ is proportional to a site occupation-magnetic moment correlation $M(\mathbf{H})$. For a magnetization perpendicular to the scattering plane, we have

$$I_{\mathrm{NM}}(\mathbf{H}) = C_{\mathrm{A}} C_{\mathrm{B}} \Delta b(0.540) M(\mathbf{H}) \quad (38)$$

and

$$C_{\mathrm{A}} C_{\mathrm{B}} M(\mathbf{H}) = \sum_{lmn} e_{lmn}^{2\pi i \mathbf{H}\cdot\mathbf{R}} \langle(\alpha_{lmn} - 1)\mu_{000}(\mathbf{H})\rangle \quad (39)$$

Here the quantity in $< >$'s represents the average increase in the moment of the atom at the origin (000) due to the atomic species of the atom located at site lmn.

While the terms $I_{\mathrm{M}}(\mathbf{H})$ and $I_{\mathrm{NM}}(\mathbf{H})$ allow one to study the magnetic short-range order in the alloy, they also complicate the data analysis by making it difficult to separate these two terms from the chemical SRO. One experimental method for resolving the magnetic components is to use polarization analysis where the moment of the incident neutron beam is polarized to be either parallel ($\varepsilon = 1$) or antiparallel ($\varepsilon = -1$) to the magnetization. The total scattering for each case can now be written as

$$I_\varepsilon(\mathbf{H}) = I_{\mathrm{N}}(\mathbf{H}) + \varepsilon I_{\mathrm{NM}}(\mathbf{H}) + I_{\mathrm{M}}(\mathbf{H}) \quad (40)$$

The intensity difference between the two polarization states gives

$$\Delta I_{\mathrm{Total}}(\mathbf{H}) = 2I_{\mathrm{NM}}(\mathbf{H}) \quad (41)$$

and the sum gives

$$\sum_\varepsilon I_\varepsilon(\mathbf{H}) + 2I_{\mathrm{N}}(\mathbf{H}) + 2I_{\mathrm{M}}(\mathbf{H}) \quad (42)$$

If $I_N(\mathbf{H})$ is known from a separate measurement with x rays, all three components of the scattering can be separated from one another.

One of the greatest difficulties in studying magnetic short-range order comes when the moments of the atoms cannot be aligned in the same direction with, for example, an external magnetic field. In the above analysis, it was assumed that the moments are perpendicular to the scattering vector, **H**. The magnetic scattering cross-section is reduced by the sine of the angle between the magnetic moment and the scattering vector. Thus if the magnetization is not perpendicular to the scattering vector, the moments on the atoms must be reduced by the appropriate amount. When the spins are not aligned, the sine of the angle between the moment and the scattering vector for each individual atom must be considered. In this case, it becomes necessary to construct computer models of the spin structure to extract $M(\mathbf{H})$ and $T(\mathbf{H})$. More in-depth discussion is given in MAGNETIC NEUTRON SCATTERING.

GENE E. ICE
JAMES L. ROBERTSON
CULLIE J. SPARKS
Oak Ridge National Laboratory
Oak Ridge, Tennessee

RESONANT SCATTERING TECHNIQUES

INTRODUCTION

This unit will describe the principles and methods of resonant (anomalous) x-ray diffraction as it is used to obtain information about the roles of symmetry and bonding on the electronic structure of selected atoms in a crystal. These effects manifest themselves as crystal orientation-dependent changes in the diffracted signal when the x-ray energy is tuned through the absorption edges for those atoms. Applications of the method have demonstrated it is useful in: (1) providing site-specific electronic structure information in a solid containing the same atom in several different environments; (2) determining the positions of specific atoms in a crystal structure; (3) distinguishing electronic from magnetic contributions to diffraction; (4) isolating and characterizing multipole contributions to the x-ray scattering that are sensitive to the local environment of atoms in the solid; and (5) studying the ordering of $3d$ electron orbitals in the transition metal oxides.

These effects share a common origin with x ray resonant magnetic scattering. Both are examples of anomalous dispersion, both are strongest near the absorption edge of the atoms involved, and both display interesting dependence on x-ray polarization.

Resonant electronic scattering is a microscopic analog of the optical activity familiar at longer wavelengths. Both manifest as a polarization-dependent response of x-ray or visible light propagation through anisotropic media. The former depends on the local environment (molecular point symmetry) of individual atoms in the crystalline unit cell, while the latter is determined by the point symmetry of the crystal as a whole (Belyakov and Dmitrienko, 1989, Section 2). This analogy led to a description of "polarized anomalous scattering" with optical tensors assigned to individual atoms (Templeton and Templeton, 1982). These tensors represent the anisotropic response when scattering involves electronic excitations between atomic core and valence states that are influenced by the point symmetry at the atomic site. In the simplest cases (dipole-dipole), these tensors may be visualized as ellipsoidal distortions of the otherwise isotropic x-ray form factor. Symmetry operations of the crystal space group that affect ellipse orientation can modify the structure factors, leading to new reflections and angle dependencies.

Materials Properties that Can Be Measured

Resonant x-ray diffraction is used to measure the electronic structure of selected atoms in a crystal and for the direct determination of the phases of structure factors. Results to date have been obtained in crystalline materials possessing atoms with x-ray absorption edges in an energy range compatible with Bragg diffraction. As shown by examples below (see Principles of the Method), resonant diffraction is a uniquely selective spectroscopy, combining element specificity (by selecting the absorption edge) with site selectivity (through the choice of reflection). The method is sensitive to the angular distribution of empty states near the Fermi energy in a solid. Beyond distinguishing common species at sites where oxidation and/or coordination may differ, it is sensitive to the orientation of molecules, to deviations from high symmetry, to the influence of bonding, and to the orientation (spatial distribution) of d electron orbital moments when these valence electrons are ordered in the lattice (Elfimov et al., 1999; Zimmermann et al., 1999).

The second category of measurement takes advantage of differences in the scattering from crystallographically equivalent atoms. Because of its sensitivity to molecular orientation, resonant scattering can lead to diffraction at reflections "forbidden" by screw-axis and glide-plane crystal symmetries (SYMMETRY IN CRYSTALLOGRAPHY). Such reflections can display a varying intensity as the crystal is rotated in azimuth about the reflection vector. This results from a phase variation in the scattering amplitude that adds to the position-dependent phase associated with each atom. This effect can lead to the determination of the position-dependent phase in a manner analogous to the multiple-wavelength anomalous diffraction (MAD) method of phase determination.

Secondary applications involve the formulation of corrections used to analyze anomalous scattering data. This includes accounting for orientation dependence in MAD measurements (Fanchon and Hendrickson, 1990) and for the effects of birefringence in measurements of radial distribution functions of amorphous materials (Templeton and Templeton, 1989).

Comparison with Other Methods

The method is complementary to standard x-ray diffraction because it is site selective and can provide direct information on the phase of structure factors (Templeton and

Templeton, 1992). A synchrotron radiation source is required to provide the intense, tunable high-energy x-rays required. Highly polarized beams are useful but not required.

Perhaps the technique most closely related to resonant diffraction is angle-dependent x-ray absorption. However, as pointed out by Brouder (1990), except in special cases, the absorption cross-section reflects the full crystal (point) symmetry instead of the local symmetry at individual sites in the crystal. In diffraction, the position-dependent phase associated with each atom permits a higher level of selectivity. This added sensitivity is dramatically illustrated by (but not limited to) reflections that violate the usual extinction rules.

Resonant diffraction thus offers a distinct advantage over spectroscopy methods that provide an average picture of the material. It offers site specificity, even for atoms of a common species, through selection of the diffraction and/or polarization conditions. When compared to electron probe spectroscopies, the method has sensitivity extending many microns into the material, and does not require special sample preparation or vacuum chambers.

Drawbacks to the method include the fact that signal sizes are small when compared to standard x-ray diffraction, and that the experimental setup requires a good deal of care, including the use of a tunable monochromator and in some cases a polarization analyzer.

General Scope

This unit describes resonant x-ray scattering by presenting its principles, illustrating them with several examples, providing background on experimental methods, and discussing the analysis of data. The Principles of the Method section presents the classical picture of x-ray scattering, including anomalous dispersion, followed by a quantum-mechanical description for the scattering amplitudes, and ends with a discussion on techniques of symmetry analysis useful in designing experiments. Two examples from the literature are used to illustrate the technique. This is followed by a discussion of experimental aspects of the method, including the principles of x-ray polarization analysis. A list of criteria is given for selecting systems amenable to the technique, along with an outline for experimental design. A general background on data collection and analysis includes references to sources of computational tools and expertise in the field.

PRINCIPLES OF THE METHOD

Resonant x-ray diffraction is used to refine information on the location of atoms and the influence of symmetry and bonding for specific atoms in a known crystal structure. Measurements are performed in a narrow range of energy around the absorption edges for the atoms of interest. The energy and crystal structure parameters determine the range of reflections (regions in reciprocal space) accessible to measurement. The point symmetry at the atom location selects the nature (i.e., the multipole character) of the resonance-scattering amplitude for that site. This amplitude is described by a symmetric tensor, and the structure factor is constructed by summing these tensors, along with the usual position-dependent phase factor, for each site in the unit cell. The tensor structure factor is sensitive to the nonisotropic nature of the scattering at each site and this leads to several important phenomena: (1) the usual extinction rules that limit reflections (based on an isotropic response) can be violated; (2) the x-ray polarization may be modified by the scattering process; (3) the intensity of reflections may vary in a characteristic way as the crystal is rotated in azimuth (maintaining the Bragg condition) about the reflection vector; and (4) these behaviors may have a strong dependence on x-ray energy near the absorption edge.

Theory

The classical theory of x-ray scattering, carefully developed in James' book on diffraction (James, 1965), makes clear connections between the classical and quantum mechanical descriptions of anomalous scattering. The full description of resonant (anomalous) scattering requires a quantum-mechanical derivation beginning with the interaction between electrons and the quantized electromagnetic field, and yielding the Thomson and resonant electronic scattering cross-sections. An excellent modern exposition is given by Blume (1985). Several good discussions on the influence of symmetry in anomalous scattering, including an optical model of the anistropy of x-ray susceptibility, are given by Templeton (1991), Belyakov and Dmitrienko (1989), and Kirfel (1994). The quantum mechanical description of symmetry-related effects in anisotropic scattering is included in the work by Blume (1994) on magnetic effects in anomalous dispersion. This has been extended in a sophisticated, general group-theoretical approach by Carra and Thole (1994).

We describe x-ray diffraction by first considering the atomic scattering factor in classical terms and the corrections associated with anomalous scattering. This is followed by the quantum description, and finally the structure factor for resonant diffraction is calculated.

Classical Picture

The scattering of x rays from a single free electron is given by the Thomson scattering cross-section; the geometry is illustrated in Figure 1.

The oscillating electric field (**E**) of the incident x-ray, with wave vector **k** and frequency ω, causes a sinusoidal

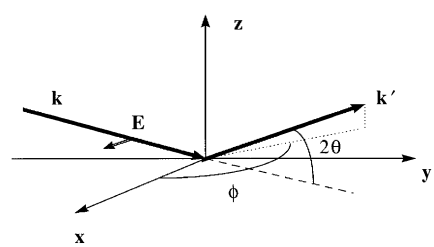


Figure 1. Geometry of x-ray scattering from a single electron. The notation is explained in the text.

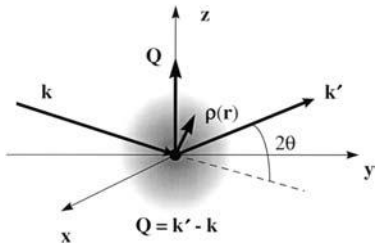


Figure 2. The geometry used in the calculation of the atomic scattering form factor for x rays. The diffraction vector is also defined and illustrated.

acceleration of the electron (with charge $-e$) along the field direction. This produces a time-dependent dipole moment of magnitude $e^2/(m\omega^2)$, which radiates an electromagnetic field polarized so that no radiation is emitted along the direction **E**. This polarization effect (succinctly given in terms if the incident and scattered polarization directions as $\varepsilon \bullet \varepsilon'$) yields a $\sin\phi$ dependence where ϕ is the angle between the incident polarization and the scattered wave vector $\mathbf{k}'$. The Thomson scattered radiation is 180° out of phase with the incident electric field, and has an intensity

$$I_{\text{Thomson}} = I_0[(e^2/mc^2)\sin\phi/R]^2 \tag{1}$$

where the incident intensity $I_0 = |\mathbf{E}|^2$, e, and m are the electron charge and mass, c the speed of light, and R is the distance between the electron and field point where scattered radiation is measured.

When electrons are bound to an atom, the x-ray scattering is changed in several important ways. First, consider an electron cloud around the nucleus as shown in Figure 2. James treats the cloud as a continuous, spherically symmetric distribution with charge density $\rho(\mathbf{r})$ that depends only on distance from the center. The amplitude for x rays scattered from an extended charge distribution into the $\mathbf{k}'$ direction is calculated by integrating over the distribution, weighting each point with a phase factor that accounts for the path length differences between points in the cloud. This is given by

$$f_0 = \int \rho(\mathbf{r}) e^{iQ\cdot\mathbf{r}} dV \tag{2}$$

with **Q** the reflection vector of magnitude $|\mathbf{Q}| = 4\pi\sin\theta/\lambda$, and 2θ and λ the scattering angle and wavelength, respectively.

The atomic scattering form factor, f_0, defined by this integral, is equal to the total charge in the cloud when the phase factor is unity (in the forward scattering direction), and approaches zero for large 2θ. Values of f_0 versus $\sin\theta/\lambda$ for most elements are provided in tables and by algebraic expressions (MacGillavry and Rieck, 1968; Su and Coppens, 1997). The form factor is given in electron units, where one electron is the scattering from a single electron.

An important assumption in this chapter is that the x-ray energy is unchanged by the scattering (i.e., elastic or coherent). Warren (1990) describes the relation between the total intensity scattered from an atom and the coherent scattering. He defines the "modified" (i.e., incoherent or Compton) intensity as the difference between the atomic number and the sum of squares of form factors, one for each electron in the atom.

A second effect of binding to the nucleus must be considered in describing x-ray scattering from atomic electrons. This force influences the response to the incident electric field. James gives the equation of motion for a bound electron as

$$d^2\mathbf{x}/d^2t + \kappa d\,\mathbf{x}/dt + \omega_s^2\mathbf{x} = (e/m)\mathbf{E}e^{i\omega t} \tag{3}$$

where $\mathbf{x}$, κ, ω_s are, respectively, the electron position, a damping factor, and natural oscillation frequency for the bound electron. Using the oscillating dipole model, he gives the scattered electric field at unit distance in the scattering plane (i.e., normal to the oscillator) as

$$A = e^2/(mc^2)[\omega^2 E/(\omega_s^2 - \omega^2 + i\kappa\omega)] \tag{4}$$

The x-ray scattering factor, f, is obtained by dividing A by $-e^2/(mc^2)E$, the scattering from a single electron. This factor is usually expressed in terms of real and imaginary components as

$$f = \omega^2/(\omega^2 - \omega_s^2 - i\kappa\omega) \equiv f' + if'' \tag{5}$$

This expression, and its extension to the many-electron atom discussed by James, are the basis for a classical description of resonant x-ray scattering near absorption edges. The dispersion of x rays is the behavior of f as a function of ω (where the energy is $\hbar\omega$), and "anomalous dispersion" refers to the narrow region near resonance ($\omega = \omega_s$) where the slope $df/d\omega$ is positive. Templeton's review on anomalous scattering describes how f' and f'' are related to refraction and absorption, and gives insight into the analytic connection between them through the Kramers-Kronig relation.

Returning to the description of scattering from all the electrons in an atom, it is important to recognize that the x-ray energies required in most diffraction measurements correspond to $\omega \gg \omega_s$ for all but the inner-shell electrons of most elements. For all but these few electrons, f is unity and their contribution to the scattering is well described by the atomic form factor. The scattering from a single atom is then given by

$$f = f_0 + f' + if'' \tag{6}$$

To this approximation, the dependence of scattering amplitude on scattering angle comes from two factors. The first is the polarization dependence of Thomson scattering, and second is the result of interference because of the finite size of the atomic charge density which effects f_0. This is the classical description assuming the x-ray wavelength is larger than the spatial extent of the resonating electrons. We will find a very different situation with the quantum-mechanical description.

Quantum Picture

The quantum mechanical description of x-ray scattering given by Blume may be supplemented with the discussion by Sakurai (1967). Both authors start with the Hamiltonian describing the interaction between atomic electrons and a transversely polarized radiation field

$$H = \sum_{\mathrm{j}} \left[(\mathbf{p}_{\mathrm{j}} - (e/c)\mathbf{A}(\mathbf{r}_{\mathrm{j}}))^2/(2m) + \sum_{\mathrm{i}} V(\mathbf{r}_{\mathrm{ij}}) \right] \qquad (7)$$

where $\mathbf{p}_{\mathrm{j}}$ is the momentum operator and $\mathbf{r}_{\mathrm{j}}$ is the position of the jth electron, $\mathbf{A}(\mathbf{r}_{\mathrm{j}})$ is the electromagnetic vector potential at $\mathbf{r}_{\mathrm{j}}$, and $V(\mathbf{r}_{\mathrm{ij}})$ is the Coulomb potential for the electron at $\mathbf{r}_{\mathrm{j}}$ due to all other charges at $\mathbf{r}_{\mathrm{i}}$. We have suppressed terms related to magnetic scattering and to the energy of the radiation field. Expanding the first term gives

$$H' = \sum_{\mathrm{j}} [e^2/(2mc^2)\mathbf{A}(\mathbf{r}_{\mathrm{j}})^2 - e/(mc)\mathbf{A}(\mathbf{r}_{\mathrm{j}}) \cdot \mathbf{p}_{\mathrm{j}}] \qquad (8)$$

The interaction responsible for the scattering is illustrated in Figure 3, and described as a transition from an initial state $|\mathbf{I}\rangle = |\ \mathrm{a};\ \mathbf{k},\ \varepsilon\rangle$ with the atom in state a (with energy E_{a}) and a photon of wave vector $\mathbf{k}$, polarization ε, and energy $\hbar\omega_{\mathrm{k}}$ to a final state $|F\rangle$ with the atom in state b (of energy E_{b}) and a photon of wave vector $\mathbf{k}'$, polarization ε', and energy $\hbar\omega_{\mathrm{k}'}$.

The transition probability per unit time is calculated using time-dependent perturbation theory (Fermi's golden rule) to second order in $\mathbf{A}$ (Sakurai, 1967, Sections 2-4,2-5)

$$W = (2\pi/\hbar)\left| e^2/(2mc^2) \sum_{\mathrm{j}} \langle F/\mathbf{A}(\mathbf{r}_{\mathrm{j}}) \cdot \mathbf{A}(\mathbf{r}_{\mathrm{j}})|I\rangle + (e/mc)^2 \right.$$
$$\left. \sum_{\mathrm{n}} \left\{ \frac{\langle F| \sum_{\mathrm{j}} \mathbf{A}(\mathbf{r}_{\mathrm{j}}) \cdot \mathbf{p}_{\mathrm{j}}|n\rangle\langle n| \sum_{\mathrm{l}} \mathbf{A}(\mathbf{r}_{\mathrm{l}}) \cdot \mathbf{p}_{|rml}|I\rangle}{E_{\mathrm{I}} - E_{\mathrm{n}}} \right\} \right|^2 \delta(E_{\mathrm{I}} - E_{\mathrm{F}}) \qquad (9)$$

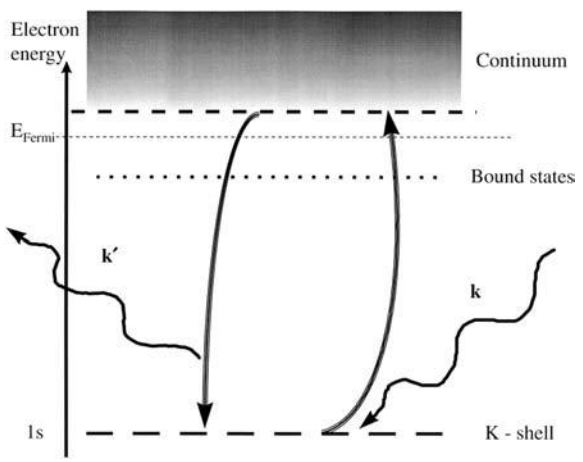


Figure 3. Schematic diagram illustrating the electronic transitions responsible for resonant scattering. An x-ray photon of wave vector $\mathbf{k}$ excites a core electron to an empty state near the Fermi level. The deexcitation releases a photon of wave vector $\mathbf{k}'$.

where the delta function insures that energy is conserved between the initial state of energy $E_{\mathrm{I}} = E_{\mathrm{a}} + \hbar\omega_{\mathrm{k}}$, and the final one with $E_{\mathrm{F}} = E_b + \hbar\omega_{\mathrm{k}'}$, $|n\rangle$ and E_n are the wave functions and energies of intermediate (or virtual) states of the system during the process.

To get the scattered intensity (or cross-section), W is multiplied by the density of allowed final states and divided by the incident intensity. Blume uses a standard expansion for the vector potential, and modifies the energy resonant denominator for processes that limit the scattering amplitude. Sakurai introduces the energy width parameter, Γ, to account for "resonant fluorescence" (Sakurai, 1967). In the x-ray case, the processes that limit scattering are both radiative (such as ordinary fluorescence), and nonradiative (like Auger emission). They are discussed by many authors (Almbladh and Hedin, 1983). When the solid is returned to the initial state ($|a\rangle$), the differential scattering cross-section, near resonance, is

$$d^2\sigma/d\Omega\, dE = e^2/(2mc^2) \left| \langle a| \sum_j e^{i\mathbf{Q}\cdot\mathbf{r}_j}|a\rangle \varepsilon' \cdot \varepsilon - \hbar/m \right.$$
$$\left. \sum_{\mathrm{n}} \left[\frac{\langle a| \sum_{\mathrm{j}} (\varepsilon' \cdot \mathbf{p}_{\mathrm{j}} e^{i\mathbf{k}'\cdot\mathbf{r}} j)|n\rangle\langle n| \sum_{\mathrm{l}} (\varepsilon \cdot \mathbf{p}_{\mathrm{j}} e^{i\mathbf{k}\cdot\mathbf{r}_{\mathrm{l}}})|a\rangle}{E_{\mathrm{a}} - E\mathrm{n} + \hbar\omega_{\mathrm{k}} - i(\Gamma/2)} \right] \right|^2 \qquad (10)$$

where $\mathbf{Q} = \mathbf{k} - \mathbf{k}'$. It is useful to express the momentum operator in terms of a position operator using a commutation relation (Baym, 1973, Equations 13-96). The final expression, including an expansion of the complex exponential is given by Blume (1994) as

$$d^2\sigma/(d\Omega\, dE)$$
$$= e^2/(2mc^2) \left| \langle a| \sum_{\mathrm{j}} e^{i\mathbf{Q}\cdot\mathbf{r}_{\mathrm{j}}}|a\rangle \varepsilon' \cdot \varepsilon \frac{m}{\hbar^2} \frac{\sum_{\mathrm{n}} (E_{\mathrm{a}} - E_{\mathrm{n}})^3}{\hbar\omega} \right.$$
$$\left. \times \frac{\langle a| \sum_{\mathrm{j}} (\varepsilon' \cdot \mathbf{r}_{\mathrm{j}})(1 - i\mathbf{k}' \cdot \mathbf{r}_{\mathrm{j}}/2)|n\rangle\langle n| \sum_{\mathrm{l}} (\varepsilon \cdot r_{\mathrm{l}})(1 + i\mathbf{k} \cdot \mathbf{r}_{\mathrm{l}}/2)|a\rangle}{E_{\mathrm{a}} - E_{\mathrm{n}} + \hbar\omega_{\mathrm{k}} - i(\Gamma/2)} \right|^2 \qquad (11)$$

The amplitude, inside the squared brackets, should be compared to the classical scattering amplitude given in Equation 6. The first term is equivalent to the integral which gave the atomic form factor; the polarization dependence is that of Thomson scattering. The second term, the resonant amplitude, is the same form as Equation 5

$$\omega^2/\omega_{\mathrm{k}}^2 - \omega_{\mathrm{ca}}^2 - i(\Gamma/\hbar\omega)_{\mathrm{k}} \qquad (12)$$

Now, however, the ground-state expectation value,

$\langle \mathrm{a}|$ transition operator $|\mathrm{a}\rangle$, can have a complicated polarization and angular dependence.

The transition operator is a tensor that increases in complexity as more terms in the expansion of the exponentials (Equation 10) are kept. Blume (1994) and Templeton (1998) identify the lowest-order term (keeping only the 1 yields a second-rank tensor) as a dipole excitation and dipole deexcitation. Next in complexity is the dipole-quadrupole terms (third-rank tensors); this effect has

been observed by Templeton and Templeton (1994). The next order (fourth-rank tensor) in scattering can be dipole-octapole or quadrupole-quadrupole; the last term has been measured by Finkelstein (Finkelstein et al., 1992). An important point is that different multipole contributions make the scattering experiment sensitive to different features of the resonating atoms involved. The atom site symmetry, the electronic structure of the outer orbitals, and the orientational relationship between symmetry equivalent atoms combine to influence the scattering.

A METHOD FOR CALCULATION

Two central facts pointed out by Carra and Thole (1994) guide our evaluation of the resonant scattering. The first is a selection rule; the scattering may exist only if the transition operator is totally symmetric (i.e., if it is invariant under all operations which define the symmetry of the atomic site). This point is explained in Landau and Lifshitz (1977) and many books on group theory (Cotton, 1990). The second point is that the Wigner-Eckart Theorem permits separation of the scattering amplitude into products of angle- and energy-dependent terms. The method outlined below involves an evaluation of this angular dependence. We cannot predict the strength of the scattering (the energy-dependent coefficients); this requires detailed knowledge of electronic states that contribute to the scattering operator.

The first task is to determine the multipole contributions to resonant scattering that may be observable at sites of interest in the crystal; the selection rule provides this information. The rule may be cast as an orthogonality relation between two representations of the point symmetry. Formula 4.3–11 in Cotton (1990) gives the number of times the totally symmetric representation (A_1) exists in a representation of this point symmetry. It is convenient to choose spherical harmonics as basis functions for the representation because harmonics with total angular momentum L represent tensors of rank L.

First calculate the trace $\chi(\mathbf{R}i)$ of each symmetry operation, $\mathbf{R}i$, of the point group. Because the trace of any symmetry operation of the totally symmetric representation equals 1, Cotton's formula reduces to

$$N(A_1) = (1/h)\sum_{\mathrm{i}}[\chi(\mathbf{R}i)] \tag{13}$$

where $N(A_1)$ is the number of times the totally symmetric representation is found in the candidate representation, h is the number of symmetry operations in the point group, and the sum is over all group operations. If the result is nonzero, this number equals the number of independent tensors (of rank L) needed to describe the scattering.

The second task is to construct the angle-dependent tensors that describe the resonant scattering. There are a number of approaches (Belyakov and Dmitrienko, 1989; Kirfel, 1994; Blume, 1994; Carra and Thole, 1994), but in essence this problem is familiar to physicists, being analogous to finding the normal modes of a mechanical system, and to chemists as finding symmetry-adapted linear combinations of atomic orbitals that describe states of a molecule. To calculate where, in reciprocal space, this scattering may appear, and to predict the polarization and angular dependence of each reflection, we use spherical harmonics to construct the angle-dependent tensors. This choice is essential for the separation discussed above, and simplifies the method when higher-order (than dipole-dipole) scattering is considered.

Our method uses the projection operator familiar in group theory (see Chapter 5 in Cotton, 1990; also see Section 94 in Landau and Lifshitz, 1977). We seek a tensor of rank L, which is totally symmetric for the point group of interest. The projection operator applied to find a totally symmetric tensor is straightforward; we average the result of applying each symmetry operation of the group to a linear combination of the $2L + 1$ spherical harmonics with quantum number L, i.e.,

$$f(L) = (1/h)\sum_{\mathrm{n}}\left[\mathbf{R}(n)\left\{\sum_{\mathrm{m}} a_{\mathrm{m}} Y(L,m)\right\}\right] \tag{14}$$

where a_{m} is a constant which may be set equal to 1 for all m, and $Y(L,m)$ is the spherical harmonic with quantum numbers L and m. If our "first task" indicated more than one totally symmetric contribution to the scattering amplitude, the other tensors must be orthogonal to $f(L)$ and again totally symmetric.

The third task is to calculate the contribution this atom makes to the structure factor. In calculating $f(L)$, we assumed an oriented coordinate system to evaluate the symmetry operations. If this orientation is consistent with the oriented site symmetry for this atom and space group in the International Tables (Hahn, 1989), then the listed symmetry operations can be used to transform the angle-dependent tensor from site to site in the crystal. The structure factor of the atoms of site i, for reflection $\mathbf{Q}$ is written

$$F(i;\mathbf{Q}) = \sum_{\mathrm{j}}\left[\left\{f_{0\mathrm{j}} + \sum_{\mathrm{p}} f_{\mathrm{pj}}(L)\right\} e^{i\mathbf{Q}\bullet\mathbf{r}_{\mathrm{j}}}\right] \tag{15}$$

where $f_{0\mathrm{j}}$ is the nonresonant form factor for this atom, and $f_{\mathrm{pj}}(L)$ is the pth resonant scattering tensor for the atom at $\mathbf{r}_{\mathrm{j}}$; both terms are multiplied by the usual position-dependent phase factor. The tensor transformations are familiar in Cartesian coordinates as a multiplication of matrices (Kaspar and Lonsdale, 1985, Section 2.4). Here, having used spherical harmonics, it is easiest to use standard formulas (Sakurai, 1985) for rotations and simple rules for reflection and inversion given in APPENDIX A of this unit.

When more then one atom contributes to the scattering, these structure factors are added together to give a total $F(\mathbf{Q})$.

The final task is to calculate the scattering as a function of polarization and crystal orientation. For a given L, our resonant $F_{\mathrm{L}}(\mathbf{Q})$ is a compact expression

$$F_{\mathrm{L}}(\mathbf{Q}) = \sum_{\mathrm{m}}[b_{\mathrm{m}} Y(L,m)] \tag{16}$$

with b_m the sum of contributions to the scattering with angle dependence $Y(L,m)$. To evaluate the scattering, we express each $Y(L,m)$ as a sum of products of the vectors in the problem. The products are obtained by standard methods using Clebsch-Gordan coefficients, as described in Sakurai (1985, Section 3.10). For $L = 2$ (dipole-dipole) scattering, the two vectors are the incident and scattered polarization, for $L = 3$ a third vector (the incident or outgoing wave vector) is included, and for $L = 4$ both wave vectors are used to characterize the scattering. The crystal orientation (relative to these defining vectors) is included by applying the rotation formula (see APPENDIX C in this article) to $F_L(\mathbf{Q})$.

Illustrations of Resonant Scattering

The following two examples illustrate the type of information available from these measurements; the first describes an experiment sensitive to $L = 2$ scattering, while the second presents results from $L = 4$ work.

***L* = 2 measurements**. Wilkinson et al. (1995) illustrate how resonant scattering can be used to separately measure parameters connected to the electronic state of gold at distinct sites. The diffraction measurements were performed at the gold L_{III} edge in Wells' salt ($Cs_2[AuCl_2][AuCl_4]$). Both sites, I and III in Figure 4, have D_{4h} symmetry, but the first has a linear coordination of neighbors (along the tetragonal c axis), while the coordination is square planar (parallel to the crystalline a-b plane) at the second site.

Application of Equation 13 shows that one $L = 2$ tensor is required (for each of the two sites) to describe the anisotropic scattering. The construction formula, Equation 14, gives the angular dependence, which is proportional to $Y(2,0)$. This agrees with the anisotropic (the traceless symmetric) part of the matrix given (in Cartesian form) by Wilkinson et al. (1995) in their Equation 1. Subtracting one-third of the trace of this matrix from each component gives $f_{xx} = f_{yy} = -2 \cdot f_{zz}$, which is equivalent to $Y(2,0) \propto 2z^2 - (x^2 + y^2)$. Further discussion on the angular dependence for dipole scattering is found in APPENDIX B of this article. The angular function at each gold site is represented by the lobe-shaped symbols in Figure 4. This angular dependence is then used to calculate the structure factor, which is a sum of three terms: one for each of the two gold atomic sites and a third for the remaining atoms in the unit cell. The 16 symmetry operations of this space group (#139 in Hahn, 1989) leave the angular function unchanged, so the structure factor for the two gold sites may be written

$$\mathbf{F_{gold}} = [\mathbf{Q}(h,k,l)] \propto \cos[(\pi/2)(h+k+l)] \times \{2f_0\mathbf{I}\cos[\pi l/2] + \mathbf{F}_{\mathrm{I}}e^{-i(\pi l/2)} + \mathbf{F}_{\mathrm{III}}e^{+i(\pi l/2)}\} \quad (17)$$

where $\mathbf{F}_{gold}$, $\mathbf{F}_{\mathrm{I}}$, $\mathbf{F}_{\mathrm{III}}$, and $\mathbf{I}$ are second-rank tensors and the reciprocal lattice vector $\mathbf{Q}(h,k,l) = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$. $\mathbf{I}$, the diagonal unit tensor, represents the isotropic nature of the gold atom form factor. Wilkinson et al. (1995) point out that the resonant scattering is proportional to the sum (difference) of $\mathbf{F}_{\mathrm{I}}$, $\mathbf{F}_{\mathrm{III}}$ for reflections with l = even (odd), so that separation of the two quantities requires data for both types of reflections. Figure 4 shows that $\mathbf{F}_{\mathrm{I}}$ and $\mathbf{F}_{\mathrm{III}}$ share a common orientation in the unit cell. The tensor structure factor is thus easily described in a coordinate system that coincides with the reciprocal lattice. Wilkinson et al. (1995) use σ for the direction along **c*** and π for directions in the basal plane. To avoid confusion with notation for polarization, we will use s and p instead of σ and π.

With the crystal oriented for diffraction, the resonant scattering amplitude depends on the orientation of the tensors with respect to the polarization vectors. $\mathbf{F_{gold}}(\mathbf{Q})$ should be re-expressed in the coordinate system where these vectors are defined (see APPENDIX C in this article). The amplitude is then written

$$A = \boldsymbol{\varepsilon}' \cdot \mathbf{F}_{\mathrm{lab}}(\mathbf{Q}) \cdot \boldsymbol{\varepsilon} \quad (18)$$

where $\boldsymbol{\varepsilon}'$ and ε are the scattered and incident polarization vectors and matrix multiplication is indicated. The amplitude expressed in terms of the four (2-incoming, 2-outgoing) polarization states is shown in Table 1, where F_{ij} are cartesian tensor components in the new (laboratory) frame, and 2θ is the scattering angle.

This scheme for representing the scattering amplitude makes it easy to predict basic behaviors. Consider a vertical scattering plane with **Q** pointing up, and the incident

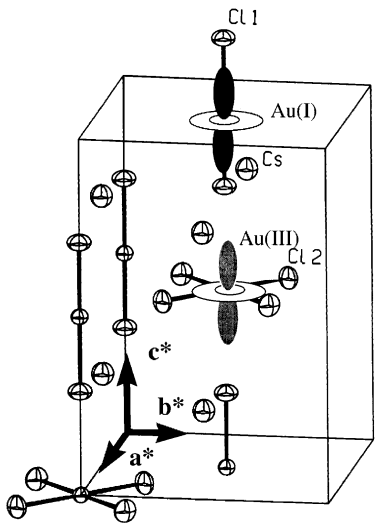


Figure 4. The crystal structure of Wells' salt contains two sites for gold; the labeling indicates the formal valence of the two atoms. At each of these two sites, the lobe-shaped symbol indicates the angular dependence for resonant scattering of rank $L = 2$.

Table 1. Amplitude Expressed in Terms of the Four (2-Incoming, 2-Outgoing) Polarization States[a]

	$\boldsymbol{\varepsilon} = \boldsymbol{\sigma}$	$\boldsymbol{\varepsilon} = \boldsymbol{\pi}$
$\boldsymbol{\varepsilon}' = \boldsymbol{\sigma}'$	F_{11}	F_{13}
$\boldsymbol{\varepsilon}' = \boldsymbol{\pi}''$	$F_{31}\cos(2\theta)$-$F_{33}\cos(2\theta)$	$F_{21}\sin(2\theta)$-$F_{23}\sin(2\theta)$

[a] $\boldsymbol{\sigma}(\boldsymbol{\pi})$ points normal (parallel) to the scattering plane.

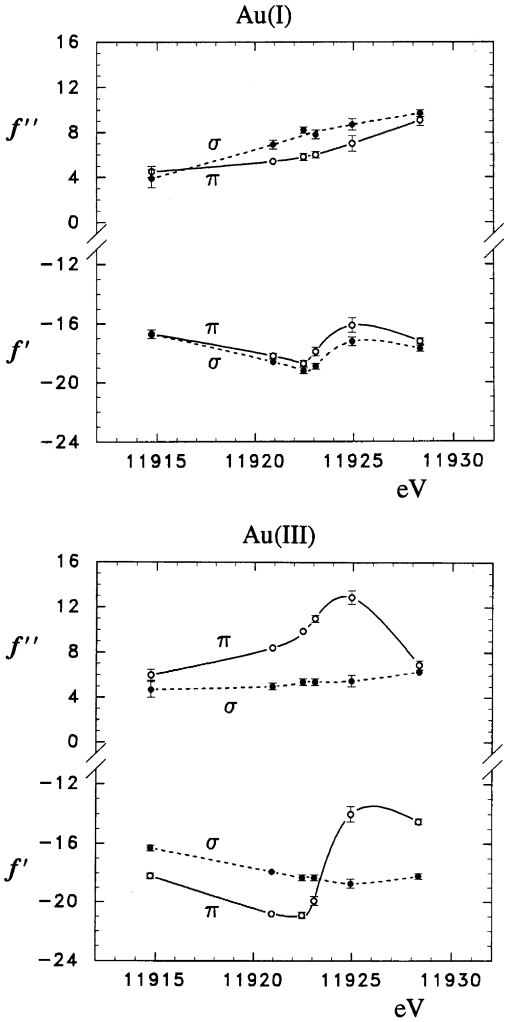


Figure 5. The energy dependence of the real and imaginary parts of the anomalous scattering from the two gold sites in Wells' salt is given in electron units. The symbols σ and π (referred to in the text as *s* and *p*) correspond to the tensor components along the $\mathbf{c}^*$ axis and in the basal plane in Figure 4.

polarization (**σ**) normal to the plane. For any $\mathbf{Q}(h,k,l)$, one can rotate the crystal around that reflection vector until the donut shaped *p* lobe of the tensor is along the incident polarization; the measurement is then most sensitive to this feature. Only for $\mathbf{Q}(h,k,0)$ reflections can scattering from the *s* lobe be isolated. This analysis also suggests the benefits of selectively measuring the polarization of the scattered beam. By looking only at the **σ′**-scattered polarization, one can, in the first case, completely separate out *p*-lobe scattering; in the second case both contributions can be determined (by the azimuth scan; see Practical Aspects of the Method) at a single reflection. By selecting the π′-scattered polarization, one eliminates the Thomson (f_0) contribution at allowed reflections.

Wilkinson et al. (1995) do not use polarization analysis, but instead measure a series of reflections, at angular orientations that favor either tensor component, and extract four values for the anomalous scattering at each gold site (Templeton and Templeton, 1982, 1985, 1992). Their results, plotted versus energy in Figure 5, show a clear distinction between the two sites. The gold (III) site exhibits a strongly anisotropic near-edge behavior (the difference $|f_p - f_s|$ is >15% of the total resonant scattering) which the authors suggest may result from an electronic transition between the $2p_{3/2}$ and empty $3d_{x^2-y^2}$ orbitals. The gold(I) site on the other hand would (formally) have a filled *d* shell, and in fact shows a much smaller anisotropic difference (<4%).

***L* = 4 measurements**. The second example illustrates $L = 4$ scattering where polarization analysis is critical to measurement success.

The measurements were performed on a single crystal of hematite (Fe_2O_3), at x-ray energies corresponding to the iron K-edge (Finkelstein et al., 1992). The results show: (1) resonant scattering is a sensitive measure of crystal field splitting of the iron 3*d* level; (2) an intensity variation with azimuth identifies the symmetry at the iron site and separates chemical from magnetic contributions to the scattering; and (3) that polarization analysis is a powerful tool for selection and quantitative evaluation of these signals.

The structure of hematite is trigonal (space group #167; Hahn, 1989), with a single iron site of symmetry C_3, positioned at *c*/6 separation along the crystalline *c* axis. The structure, illustrated in Figure 6, indicates that the oxygen atoms of each molecule are mirror reflected from site to site by 3-fold vertical glide planes.

A single reflection, the $(00.3)_{\text{Hexagonal}}$ (probing lattice planes of *c*/3 spacing normal to the *c* axis), was measured

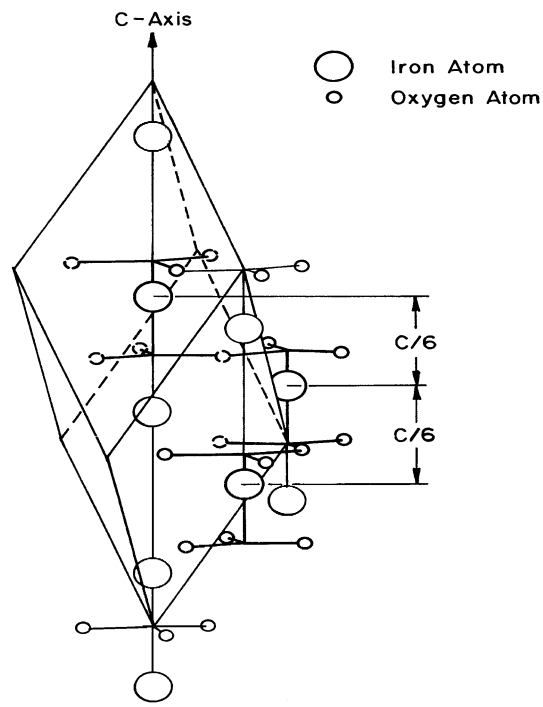


Figure 6. The crystal structure of hematite is displayed with a rhombohedral unit cell. Each iron atom is surrounded by an approximate octahedral arrangement of oxygen atoms.

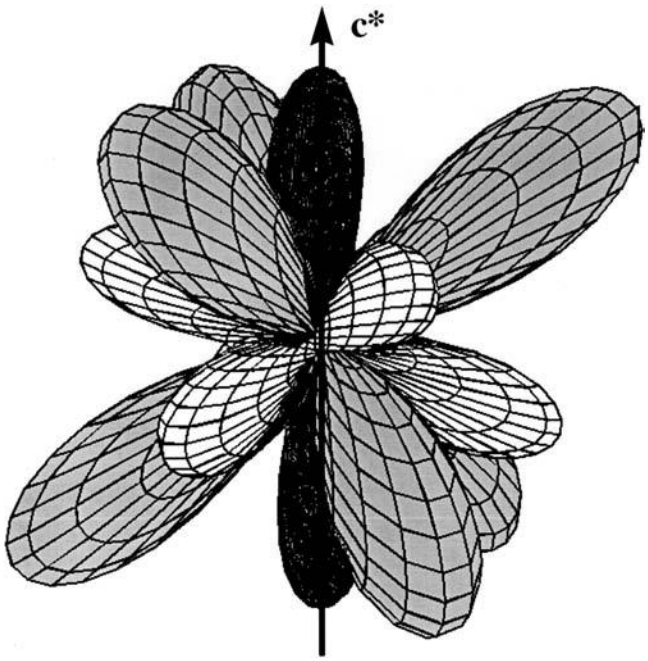


Figure 7. A geometrical representation of the angular dependence for $L = 4$ resonant scattering in O_h symmetry. The object has 3-fold rotational symmetry about the $\mathbf{c}^*$ axis. The black lobes are rotationally symmetric; the scattering amplitude associated with gray and white lobes is opposite in sign.

at a series of energies and azimuth angles around this axis. This reflection is "forbidden" by general conditions of the space group, but magnetically "allowed" because the iron atoms are antiferromagnetically ordered.

The iron has a formal valence of +3, implying a half-filled $3d$ shell. The site symmetry is very close to cubic (O_h), and to this approximation the $3d$ levels are crystal field split into e_g bonding states (with electron density directed toward oxygen neighbors), and t_{2g} antibonding states (with the density directed between these neighbors). $1s \rightarrow 3d$ resonant scattering is sensitive to the filling and directionality of these states.

In O_h symmetry Equation 10b.3.13, excludes $L = 2$ and 3, but permits $L = 4$ scattering with one tensor contributing. Equation 14 gives an angle-dependent tensor proportional to

$$\sqrt{(7/10)}Y(4,0) + Y(4,3) - Y(4,-3) \qquad (19)$$

The angular properties of this tensor are illustrated in Figure 7.

The shape of this object is a geometrical measure of the sensitivity which this type of scattering has to electronic transitions between the $1s$ and $3d$ states. By considering this object positioned at each iron site, the explanation for 3 out of 4 experimental observations is immediately apparent: (1) the "forbidden" reflection is allowed because the orientation of oxygen atoms changes the sign of the scattering amplitude from site to site; (2) this amplitude changes sign every 60° as the crystal is rotated about the c axis; (3) this behavior depends on electronic transitions (with $\Delta L = 2$; quadrupole excitation and deexcitation) and it therefore occurs only at a well-defined resonance energy.

A fourth experimental fact, that the resonant behavior was visible as a rotation of the x-ray polarization (from primarily σ to π'), can be explained when the tensor is expanded using the methods in Sakurai (1985). The quantitative details, worked out by Hamrick (1994) and by Carra and Thole (1994), show that the energy dependences of this resonance are exceedingly sensitive to the magnitude of the crystal field splitting. The resonant scattering amplitude was equivalent to less than about 0.08 electrons! However, by analyzing the scattered beam for the rotated polarization, the signal intensity was found to be 45 times that of the background.

PRACTICAL ASPECTS OF THE METHOD

Instrumentation

These measurements are made with synchrotron radiation because an intense, energy-tunable incident x-ray beam is required. It is useful to work in a vertical scattering geometry to take advantage of the small divergence and high degree of horizontal, linear polarization available from synchrotron sources. A standard silicon (111) double crystal monochromator is commonly used for energy tuning. Focusing optics, when used, can deliver substantially more flux to small sample crystals, but can also lead to several problems. In particular, strong horizontal focusing (i.e., sagittal focusing) increases the beam divergence normal to the scattering plane and therefore increases the likelihood of multiple-beam (Renninger) reflection contamination in the scattered signal. In addition, when polarization analysis is employed to improve sensitivity, the analyzer may not fully accept the increased divergence. The use of a vertical reflection mirror for harmonic rejection and focusing (if available) is generally an advantage. Many experiments have been performed with bend magnet radiation, and most of these with an incident energy resolution not better than $1 - 5 \times 10^{-4}$. The ideal monochromator would be tunable, and provide a beam with an energy width and stability matched to the core-hole energy associated with the absorption edge under study.

A four-circle diffractometer is essential for access to a wide range of reflections from one sample. With appropriate software, it permits flexible scanning of reciprocal space, including the important azimuthal scan discussed below. For more challenging experiments, the diffractometer may need to accommodate a polarization analyzer. The technique of polarization analysis is also discussed in this section. It is often useful to simultaneously monitor fluorescence from the sample as diffraction data are collected. In so doing, one may: (1) reduce errors caused by energy drifts (through periodic energy scans that monitor the position of the absorption edge) and (2) conveniently obtain a measurement of f'' useful for signal normalization; also (3) as discussed earlier, measurement of the angular behavior of f'' gives an essentially independent measure of the average site anisotropy for the absorbing species (Brouder, 1990).

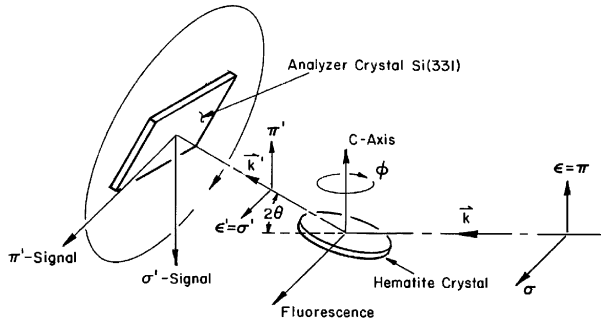


Figure 8. The experimental arrangement for resonant scattering measurements may include an analyzer for selecting polarization in the signal beam.

Figure 8 illustrates the geometry of the scattering experiment, including the placement of polarization analyzer and fluorescence detector. The samples are usually single crystals mounted on a eucentric goniometer for convenient orientation of crystal faces. The azimuth scan is performed to measure the anisotropic nature of the atomic environment. The sample is oriented so that the Bragg angle is maintained as it is rotated about the axis of the reflection vector. Because tensor scattering effects are small, one often chooses weak or space-group forbidden reflections for these studies. In this case an appropriate choice of azimuth is important for eliminating the influence of Renninger reflections, which are localized and predictable (Cole et al., 1962) in both energy and angle.

Polarization Analysis

The principles and methods of x-ray polarization analysis are discussed by several authors (Blume and Gibbs, 1988; Belyakov and Dmitrienko, 1990; Shen and Finkelstein, 1993; Finkelstein et al., 1994). Although tensor scattering effects can appear even with an unpolarized incident beam, it is important to know the beam polarization for accurate analysis of the data. Further, it has been shown (Finkelstein et al., 1992) that polarization analysis of the diffraction signal can significantly improve sensitivity and provide a detailed characterization of the scattering effects. In particular, resonant scattering is one of only a few effects that rotate the incident polarization; thus a search for resonant scattering is facilitated by a polarized incident beam and an analyzer set to accept the orthogonal polarization state.

The linear polarization analyzer, illustrated in Figure 9, is basically a 3-circle diffractometer mounted on the 2-theta (detector) arm of the 4-circle instrument. The heart of the analyzer is a crystal that intercepts signal from the sample and diffracts it into a detector set near 90° to the signal beam. The analyzer can be rotated about the beam by changing chi. When the analyzer and sample diffraction planes are parallel, the measurement is sensitive to σ polarization. When rotated 90° about chi, the analyzer passes the rotated or π polarized radiation. The linear polarization state (relative to these planes) is characterized by the Stokes-Poincaré parameter

$$P_1 = (I_\sigma - I_\pi)/(I_\sigma + I_\pi) \qquad (20)$$

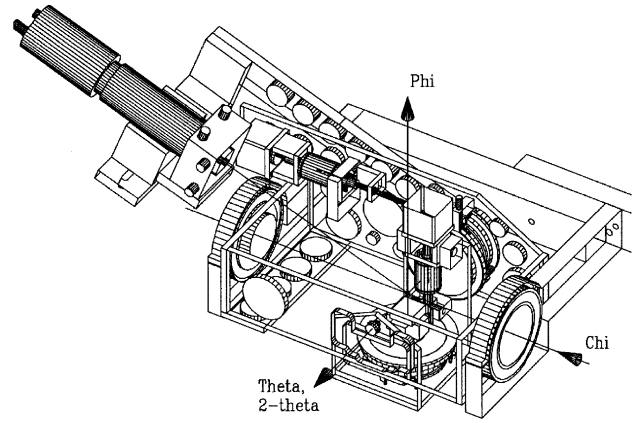


Figure 9. An x-ray polarimeter uses diffraction, with a Bragg angle (theta) near 45° to measure the direction of linear polarization in a beam. The device is rotated (in Chi) about the beam for this purpose.

where P_1 is +1 for a beam fully σ-polarized, and P_1 is equal to −1 if the beam is fully π polarized. Three parameters are required to completely characterize the beam polarization, but for the purposes of these experiments, knowledge of the incident and scattered P_1 is usually sufficient.

Besides possessing high reflectivity, the polarization-analyzing crystal should have an angular acceptance matched to the signal beam and a Bragg angle close to 45° for the energy being measured. If the first condition is not met, the analyzer must be scanned through its reflection (rocking) curve to ensure accurate measurement. When set to pass π-polarized radiation, the analyzer is considerably less sensitive to noise sources. This may help in the search for resonant signals, but is of only limited value if the incident beam horizontal divergence is large. In the case where an energy scan, at constant sample **Q** (also known as a DAFS scan; Stragier, et al., 1992) is required, the monochromator, sample and detector arm (2-theta) angles, as well as the analyzer crystal angle must move in synchronization.

Experiment Preparation

Resonant scattering experiments are generally nontrivial to perform and require significant advanced planning. No well established techniques exist for estimating signal size, but the method for calculation outlined earlier can provide a map for candidate reflections. Several recent attempts have demonstrated how difficult it is to explore higher multipole scattering processes without employing the methods of polarization analysis. Some general rules, gleaned from the literature, may help define appropriate characteristics for systems under consideration for study.

Sample size and crystal perfection have not generally limited these measurements. Crystal structures with compound symmetry elements (glide planes, and screw axes) are good candidates for study. The resonant scattering from atoms related by these elements may differ because of differences in orientation from atom to atom.

Structure-factor relationships are an important experimental consideration. Forbidden reflections, which are excited at resonance (by the previous mechanism), are a sensitive measure of the local environment. Weak allowed reflections reveal tensor scattering effects through azimuthal and polarization anisotropy. $L = 3$ scattering may occur only from sites with no center of symmetry. Templeton (1994) has shown that this leads to an interesting selectivity of the scattering process.

The width, strength, and orientational dependence of absorption edge features influence the resonant scattering amplitude. Because absorption (f'') and refraction (f') are related through the Kramers-Kronig dispersion relation, x-ray dichroism (polarization-dependent absorption) is a good predictor of birefringence (polarization-dependent scattering). In general, the higher the edge energy, the broader the intrinsic (core-hole) width and the weaker the resonant response. M and L edges have yielded the largest signals, followed by first row transition metal K edges. In this regard, another predictor of resonant behavior in diffraction comes from examining the optical activity of the molecules that compose the crystal. The highest-order multipole scattering observed corresponds to $L = 4$. This effect was measured to be less than 0.08 electrons in amplitude, but was visible through polarization analysis. $L = 3$ scattering has ranged in size from 0.5 to 3 electrons, while $L = 2$ scattering amplitudes (including the isotropic component) as large as 30 electrons have been observed.

Multiple beam reflections must be accounted for in experiment design and analysis. It is important to note that this effect can also be responsible for a rotation of the scattered beam polarization.

Experiment Design

A reasonable approach to the design of these experiments is to begin by constructing maps of reciprocal space. Strong, weak, and forbidden nonresonant reflections should be noted. This is followed by adding the location of resonant reflections and their associated multipole order. The map is then sectioned into regions consistent with the resonant energy and range of scattering angles available. Principle directions (faces of the crystal) should be indicated to help define orientations for azimuthal scanning.

For polarization analysis, the appropriate analyzer crystal will have a strong reflection with approximately a 45° Bragg angle at the resonant energy. The angular acceptance of this crystal should be consistent with beam divergence and it should not fluoresce at the energies of interest. The analyzer should be used with a high count rate detector (ion chamber) to measure the incident beam polarization and can also determine energy resolution and divergence, as well as the absolute energy (*http://www.chess.cornell.edu/technical/hardware/special_setups/energy_analyzer.html*). The incident beam is also used to align the detector for accurate measurement of both polarization states. A low-count-rate detector (scintillation counter) is then installed for the diffraction measurements.

METHOD AUTOMATION

Single-crystal data are collected with a 4-circle diffractometer and an energy scanning monochrometer. Automated procedures for controlling the diffractometer allow the experimenter to: (1) collect a large numbers of reflections, one energy at a time; (2) perform a scan in azimuth around a reciprocal lattice vector set to maintain the Bragg diffraction condition; and (3) coordinate the diffractometer and monochromator in such a way that an energy scan is performed while one reflection is excited (i.e., at constant **Q**). These capabilities are available through software available at most synchrotron radiation beamlines set up for 4-circle diffraction, with an energy-tunable monochromator.

The first data collection scheme, applied to resonant scattering in work at a synchrotron, was by Phillips (Phillips, et al., 1978) using a CAD-4 diffractometer. Methods of azimuth scanning were developed for surface x-ray diffraction studies (Mochrie, 1988) and are built into 4-circle control programs such as FOURC (Certified Scientific Software). An implementation of this method, used with the FOURC program, is documented on the CHESS World Wide Web site available at *http://www.chess. cornell.edu/CHESS_spec/azimesh.html*. The third data collection mode (sometimes referred to as a DAFS scan) is again built into some commercial control programs. In FOURC the monochromator energy scan ("Escan") permits the optional argument "fixQ" which moves the 4-circle to maintain a constant reflection condition. For information on this scan see *http://www.chess.cornell.edu/Certif/spec/help/mono.html*.

DATA ANALYSIS AND INITIAL INTERPRETATION

The tools required for a preliminary analysis of resonant scattering measurements have generally been developed by individual investigators. A good, general-purpose scientific graphics and analysis package, which includes a capability for nonlinear least-squares fitting, will suffice for all but the most specialized data manipulations.

Basic corrections to the data include the scaling of signals to account for: (1) changes of incident beam intensity with time and energy; (2) sharp variations in sample absorption because measurements are always taken near absorption edges; (3) standard geometrical corrections to the scattering from different reflections; and (4) efficiency (throughput) of the polarization analyzer. It is also crucial to set up criteria for eliminating data that are too weak (often due to misalignments in monochromator, diffractometer, or analyzer) or too strong (a typical effect when working close to multiple-beam reflections) to be considered reliable.

Accurate measurements of absorption, obtained by monitoring fluorescence from the sample, are extremely useful for several purposes. First, because they may be combined with numerical values for absorption coefficients far from these edges (calculated in programs like those by Cromer and Liberman (Cromer, 1983) available on the WWW at *http://www.cindy.lbl.gov/optical_constants/*)

to calculate the energy dependence of f'. These methods are discussed in the Templeton (1991) review of anomalous scattering. Second, the angular dependence of f'' (Brouder, 1990) is an important source of supplementary information.

Problems with multiple beam reflections (when measuring weak or "forbidden" reflections) can be reduced by mapping their locations before data are collected.

The work on hematite, discussed earlier, shows how important it is to know the polarization content and energy resolution of the incident beam. The former is required when an analyzer is used to determine polarization-dependent cross-sections. The latter is needed for determining information from resonance widths.

Finally, the angular dependence of azimuthal scans at resonant reflections may be compared to models derived using the calculational method (discussed earlier) by nonlinear least-squares analysis.

ACKNOWLEDGMENTS

The author is indebted to B.T. Thole, R. Lifshitz, M. Hamrick, M. Blume, and S. Shastri for helping him appreciate the theory of resonant scattering. He also thanks A.P. Wilkinson, L.K. Templeton, D.H. Templeton, A. Kirfel, Q. Shen, and E. Fontes, for contributions to the manuscript, and is grateful to E. Pollack, L. Pollack, and A.I. Goldman for their guidance in improving the manuscript. This work is supported by the National Science Foundation through CHESS, under Grant No. DMR 93-11772.

LITERATURE CITED

Almbladh, C. and Hedin, L. 1983. Beyond the one-electron model: Many-body effects in atoms, molecules, and solids. *In* Handbook on Synchrotron Radiation, Vol. 1B, (E. E. Koch, ed.) North-Holland Publishing, New York.

Baym, G. 1973. Lectures on Quantum Mechanics. Benjamin/Cummings, Reading, Mass.

Belyakov, V. A. and Dmitrienko, V. E. 1989. Polarization phenomena in x-ray optics. *Sov. Phys. Usp*. 32:697–718.

Blume, M. 1985. Magnetic scattering of x-rays. *J. Appl. Phys*. 57:3615–3618.

Blume, M. 1994. Magnetic effects in anomalous scattering. *In* Resonant Anomalous X-Ray Scattering (G. Materlik, C. J. Sparks, and K. Fischer, eds.). pp. 495–512. North-Holland, Amsterdam, The Netherlands.

Blume, M. and Gibbs, D. 1988. Polarization dependence of magnetic x-ray scattering. *Phys. Rev. B* 37:1779–1789.

Brouder, C. 1990. Angular dependence of x-ray absorption spectra. *J. Phys. Condens. Matter* 2:701–738.

Busing, W. R. and Levy, H. A. 1967. Angle calculations for 3- and 4-circle x-ray and neutron diffractometers. *Acta Crystallogr.* 22:457–464.

Carra, P. and Thole, B. T. 1994. Anisotropic x-ray anomalous diffraction and forbidden reflections. *Rev. Mod. Phys*. 66:1509–1515.

Cole, H., Chambers, F. W., and Dunn, H. M. 1962. Simultaneous diffraction: Indexing Umweganregung peaks in simple cases. *Acta Crystallogr.* 15:138–144.

Cotton, F. A. 1990. Chemical Applications of Group Theory 3rd Edition. John Wiley & Sons, New York.

Cromer, D. T. 1983. Calculation of anomalous scattering factors at arbitrary wavelengths, *J. Appl. Crystallogr*. 16:437.

Elfimov, I. S. Anisimov, V. I., and Sawatzky, G. A. 1999. Orbital ordering, Jahn-Teller distortion, and anomalous x-ray scattering in manganates. *Phys. Rev. Lett*. 82:4264–4267.

Fanchon, E. and Hendrickson, W. A. 1990. Effects of the anisotropy of anomalous scattering on the MAD phasing method. *Acta Crystallogr.* A46:809–820.

Finkelstein, K. D., Shen, Q., and Shastri, S. 1992. Resonant x-ray diffraction near the iron k-edge in hematite (Fe_2O_3). *Phys. Rev. Lett*. 69:1612–1615.

Finkelstein, K., Staffa, C., and Shen, Q. 1994. A multi-purpose polarimeter for x-ray studies. *Nucl. Instr. Meth*. A 347:124–127.

Hahn, T. (ed.) 1989. International Tables for Crystallography, Vol. A: Space-Group Symmetry, 2nd ed. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Hamrick, M. D. 1994. Magnetic and Chemical Effects in X-ray Resonant Exchange Scattering in Rare Earths and Transition Metal Compounds. Ph.D. thesis, Rice University, Houston, Tex.

James, R. W. 1965. The Optical Principles of the Diffraction of X-rays. Cornell University Press, Ithaca, N.Y.

Kaspar, J. S. and Lonsdale, K. (eds.). 1985. International Tables for X-Ray Crystallography, Vol. II. Reidel Publishing Company, Dordrecht, The Netherlands.

Kirfel, A. 1994. Anisotropy of anomalous scattering in single crystals in Resonant Anomalous X-ray Scattering: Theory and Applications, (G. Materlik, C. J. Sparks, and K. Fischer, eds.).. Elsevier Science Publishing, New York.

Landau, L. D. and Lifshitz, E. M. 1977. Quantum Mechanics (Non-relativistic Theory). Pergamon Press Ltd., Elmsford, New York.

MacGillavry, C. H. and Rieck, G. D. (eds.). 1968. International Tables for X-ray Crystallography, Vol. III: Physical and Chemical Tables. Kynoch Press, Birmingham, U.K.

Mochrie, S. G. J. 1988. Four-circle angle calculations for surface diffraction. *J. Appl. Crystallogr.* 21:1–3.

Nowick, A. S. 1995. Crystal Properties Via Group Theory. Cambridge University Press, Cambridge.

Phillips, J. C., Templeton, D. H., Templeton, L. K., and Hodgson, K. O. 1978. LIII-edge anomalous x-ray scattering by cesium measured with synchrotron radiation. *Science* 201:257–259.

Sakurai, J. J. 1967. Advanced Quantum Mechanics, Chapter 2. Addison-Wesley, Reading, Mass.

Sakurai, J. J. 1985. Modern Quantum Mechanics. Benjamin/Cummings, Menlo Park, Calif.

Shen, Q. and Finkelstein, K. 1993. A complete characterization of x-ray polarization state by combination of single and multiple bragg reflections. *Rev. Sci. Instrum*. 64:3451–3455.

Stragier, H., Cross, J. O., Rehr, J. J., and Sorensen, L. B. 1992. Diffraction anomalous fine structure: A new x-ray structural technique. *Phys. Rev. Lett*. 69:3064–3067.

Su, Z. and Coppens, P. 1997. Relativistic x-ray elastic scattering factors. *Acta. Crystallogr.* A53:749–762.

Templeton, D. H. 1991. Anomalous Scattering. *In* Handbook on Synchrotron Radiation, Vol. 3, (G. Brown and D. E. Moncton, eds.).. Elsevier Science Publishing, New York.

Templeton, D. H. 1998. Resonant scattering tensors in spherical and cubic symmetries. *Acta Crystallogr.* A54:158–162.

Templeton, D. H. and Templeton, L. K. 1982. X-ray dichroism and polarized anomalous scattering of the uranyl ion. *Acta Crystallogr.* A38:62–67.

Templeton, D. H. and Templeton, L. K. 1985. Tensor x-ray optical properties of the bromate ion. *Acta Crystallogr.* A41:133–142.

Templeton, D. H. and Templeton, L. K. 1989. Effects of x-ray birefringence on radial distribution functions for amorphous materials. *Phys.Rev. B* 40:6506–6508.

Templeton, D. H. and Templeton, L. K. 1992. Polarized dispersion, glide-rule forbidden reflections and phase determination in barium bromate monohydrate. *Acta Crystallogr.* A48:746–751.

Templeton, D. H. and Templeton, L. K. 1994. Tetrahedral anisotropy of x-ray anomalous scattering. *Phys. Rev.* B49:14850–14853.

Warren, B. E. 1990. X-ray Diffraction, Dover Publications, New York.

Wilkinson, A. P., Templeton, L. K., and Templeton, D. H. 1995. Probing local electronic anisotropy using anomalous scattering diffraction: $Cs_2[AuCl_2][AuCl_4]$. *J. Solid State Chem.* 118:383–388.

Zimmermann, M. v., Hall, J. P., Gibbs, Doon, Blume, M., Casa, D., Keimer, B., Murakami, Y., Tomioka, Y., and Tokura, Y. 1999. Interplay between charge, orbital, and magnetic order in $Pr_{1-x}Ca_xMnO_3$. *Phys. Rev. Lett.* 83:4872–4875.

KEY REFERENCES

Belyakov and Dmitrienko, 1989. See above.

A good early review of resonant scattering and more general polarization-related x-ray phenomena. Section 2 provides physical insight and a general discussion on anisotropy of x-ray susceptibility for atoms in diffracting crystals.

Blume, 1994. See above.

This presentation is comprehensive enough to include all the basic concepts and references. All tensor components up to fourth rank are given in a Cartesian (as opposed to spherical) representation.

Sakurai, 1967. See above.

An excellent, primarily nonrelativistic description of the quantum theory used to describe scattering of the type discussed here. Of particular relevence are the discussions on the Kramers-Heisenberg formula and radiation damping.

INTERNET RESOURCES

CHESS World Wide Web site

http://www.chess.cornell.edu/technical/hardware/special_setups/ energy_analyzer.html

LBNL's X-Ray Interactions with Matter

http://cindy.lbl.gov/optical_constants

APPENDIX A: TRANSFORMING THE TENSOR STRUCTURE FACTORS

The tensor structure factor is constructed by adding the partial structure factors for each unique site in the unit cell. Each of these contributions has a form given by Equation 15, above. To calculate each contribution, the tensor derived using Equation 14 is transformed from site to site according to the space group symmetry operations.

Assume a tensor is written as a sum of spherical harmonics ($|L,m\rangle$) and defined in a Cartesian coordinate system (x,y,z). A general rotation is performed using the rotation operator $\mathbf{D}(\alpha,\beta,\gamma)$ and Euler angles (α,β,γ). See Sakurai (1985), Sections 3.3, 5, and 9. This is represented as

$$\mathbf{D}(\alpha,\beta,\gamma)|L,m\rangle = \sum_{\mathrm{m}'} \mathrm{D}^{(\mathrm{L})}_{\mathrm{m}',\mathrm{m}}(\alpha,\beta,\gamma)|L,m'\rangle \qquad (21)$$

where

$$\begin{aligned}\mathrm{D}^{(\mathrm{L})}_{\mathrm{m}',\mathrm{m}}(\alpha,\beta,\gamma) &= e^{-i(m'\alpha+m\gamma)} \sum_k (-1)^{k-m-m'} \\ &\times \frac{\sqrt{[(L+m)!(L-m)!(L+m')!(L-m')!]}}{(L+m-k)!k!(L-k-m')!(k-m+m')!} \\ &\times (\cos\beta/2)^{2\mathrm{L}-2\mathrm{k}+\mathrm{m}-\mathrm{m}'}(\sin\beta)^{2\mathrm{k}-\mathrm{m}-\mathrm{m}'} \qquad (22)\end{aligned}$$

For inversions the rule is

$$|L,m\rangle \rightarrow (-1)^{\mathrm{L}}|L,m\rangle \qquad (23)$$

For reflections in a plane normal to the z axis: $z \rightarrow -z$, so

$$|L,m\rangle \rightarrow (-1)^{(L+M)}|L,m\rangle \qquad (24)$$

For reflections in a plane parallel to the z axis and at angle to the x axis, the azimuthal angle of the harmonic $\phi \rightarrow 2\delta - \phi$, so

$$|L,m\rangle \rightarrow (-1)^{\mathrm{m}} e^{\mathrm{i}2\mathrm{m}\delta}|L,-m\rangle \qquad (25)$$

APPENDIX B: ANGULAR DEPENDENT TENSORS FOR L=2

It should be pointed out that the angular function $Y(2,0)$ describes $L = 2$ anisotropic scattering for all seven tetragonal and all twelve trigonal/hexagonal point groups. In fact, one can find the form of $L = 2$ angular dependence for all 32 point groups in character tables that list the "binary products of coordinates" for the totally symmetric representation. Such tables indicate there are four other possible $L = 2$ angular functions, one for each of the remaining families (triclinic, monoclinic, orthorhombic, and cubic). One may also see that the $L = 2$ angular dependence for all cubic groups is isotropic (i.e., $\propto x^2 + y^2 + z^2$), which means that resonant scattering from these sites cannot display anisotropy. More information on these functions can be found in books that discuss the properties of crystals via group theory (Nowick, 1995).

APPENDIX C: TRANSFORMATION OF COORDINATES

Describing the scattering experiment involves transforming the resonant scattering structure factor, usually

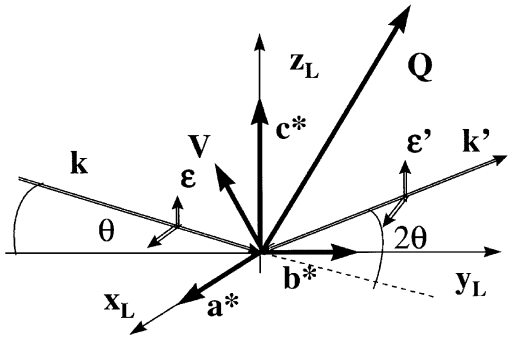


Figure 10. Two vectors describe the orientation of a crystal. If the reciprocal lattice of the crystal parallels the "laboratory" coordinate system, then the transformation formula (below) will reorient $\mathbf{Q}$ to be along $\mathbf{z}_L$ for diffraction and set $\mathbf{V}$ at azimuth ψ.

defined in a coordinate system fixed to the crystal, into a system fixed with respect to the scattering geometry (i.e., the "laboratory frame"). The methods involved are found in references used to calculate diffractometer angle settings in 4-circle crystallography (Busing and Levy, 1967). Here, these ideas merit special consideration because the tensor nature of the structure factors lead to an important additional dependence of the scattering on crystal orientation.

Figure 10 shows a rectilinear reciprocal lattice (a crystal coordinate system) coincident with a laboratory coordinate system. The orientation of vectors representing the scattering geometry are tied to the laboratory system.

Two vectors in the crystal system are needed to describe its orientation; the reflection vector $\mathbf{Q}(h,k,l)$, and a noncolinear reference vector $\mathbf{V}(h_V,k_V,l_V)$. The condition for Bragg diffraction ($\sin\theta_B = \lambda Q/2$) and the constraint that $\mathbf{Q}$ points along $\mathbf{z}_L$, allows any orientation of $\mathbf{V}$ on a cone about $\mathbf{Q}$. The azimuth angle (ψ) fixes the crystal orientation by specifying the projection of $\mathbf{V}$ normal to $\mathbf{z}_L$; $\psi = 0$ when this projection is along $-\mathbf{y}_L$ and increases with clockwise rotation about $\mathbf{Q}$.

The tensor structure factor (usually derived in crystal coordinates) is expressed in the laboratory frame through a coordinate transformation based on Euler angles (Sakari 1985, Section 3.3). In matrix form the transformation is written

$$\begin{bmatrix} \cos\alpha\cos\beta\cos\gamma & \sin\alpha\cos\beta\cos\gamma & -\sin\beta\cos\gamma \\ -\sin\alpha\sin\gamma & +\cos\alpha\sin\gamma & \\ -\cos\alpha\cos\beta\sin\gamma & -\sin\alpha\beta\sin\gamma & \sin\gamma\sin\beta \\ -\sin\alpha\cos\gamma & +\cos\alpha\cos\gamma & \\ \sin\beta\cos\alpha & \sin\beta\sin\alpha & \cos\beta \end{bmatrix} \quad (26)$$

where α and β (the usual azimuthal and polar angles of spherical coordinates) describe the orientation of $\mathbf{Q}$ relative to crystal axes. These angles may be expressed in components of the reflection vectors as $\cos\beta = 1c^*/Q$, and $\tan\alpha = (k\,b^*)/(h\,a^*)$. The angle γ is related to the experimental azimuth, ψ, through the relation

$$\sin(\gamma - \psi) = (l_v c^* - (\mathbf{V}\cdot\mathbf{Q}/Q)\cos\beta)/[(\mathbf{V}\cdot\mathbf{Q}/Q - V)\sin\beta] \quad (27)$$

KENNETH D. FINKELSTEIN
Cornell University
Ithaca, New York

MAGNETIC X-RAY SCATTERING

INTRODUCTION

For an x ray incident on a solid, the dominant scattering process is Thomson scattering, which probes the charge density of that solid. It is this interaction that provides the basis for the vast majority of x-ray scattering experiments, supplying a wealth of information on crystal structures and the nature of order in solid-state systems (see, e.g., XAFS SPECTROSCOPY, X-RAY AND NEUTRON DIFFUSE SCATTERING and RESONANT SCATTERING TECHNIQUES). However, in addition to such scattering, there are other terms in the cross-section that are sensitive to the magnetization density. Scattering associated with these terms gives rise to new features in the diffraction pattern from magnetic systems. This is magnetic x-ray scattering. For antiferromagnets, for which the magnetic periodicity is different from that of the crystalline lattice, such scattering gives rise to magnetic superlattice peaks, well separated from the charge Bragg peaks. For ferromagnets, the magnetic scattering is coincident with the charge scattering from the underlying lattice, and difference experiments must be performed to resolve the magnetic contribution.

The magnetic terms in the cross-section are much smaller than the Thomson scattering term, and pure magnetic scattering is typically 10^4 to 10^6 times weaker than the charge Bragg peaks. Nevertheless, this scattering is readily observable with synchrotron x-ray sources, and successful experiments have been performed on both antiferromagnets and ferromagnets. The technique has undergone a rapid growth in the last decade as a result of the availability of synchrotrons and the consequent access to intense, tunable x-ray beams of well-defined energy and polarization.

The basic quantity measured in a magnetic x-ray scattering experiment is the instantaneous magnetic moment-moment correlation function. In essence, the technique takes a snapshot of the Fourier transform of the magnetic order in a system. This means that the principal properties measured by magnetic scattering are quantities connected with the magnetic structure—including the ordering wave-vector, the magnetic moment direction, the correlation length, and the order parameter characterizing the phase (for example, this latter might be the staggered magnetization in the case of an antiferromagnet or the uniform magnetization for a ferromagnet). These are obtained from the position, intensity, and lineshape of

the magnetic peaks. The scattering is very weak and therefore not affected by extinction effects, so that the temperature dependence of the order parameter may be measured very precisely, leading, for example, to an accurate determination of the order parameter critical exponent, β. However, used alone, x-ray magnetic scattering can typically determine the order parameter only to within an arbitrary scale factor, because of the difficulty in obtaining an absolute scale for the scattered intensities.

There are two regimes of importance in magnetic x-ray scattering, characterized by the incident photon energy relative to the positions of the absorption edges in the system under study. These are the nonresonant and resonant limits to the cross-section.

Nonresonant magnetic scattering is observed when the incident photon energy is far from any absorption edges. It is energy independent, and has the advantages of being (in principle) present for all magnetic ions and of having a simple cross-section. The non-resonant cross-section contains terms due to both the orbital, $\vec{Z}$, and spin, $\vec{S}$, components of the magnetic moment. These each contribute, with different polarization dependences, to the scattering, and it is therefore possible to separate the two contributions empirically. This ability is one of the technique's unique strengths. However, nonresonant magnetic scattering is extremely weak, and for the signal to be observable samples of high crystallographic quality are required in order to match the high collimation of the synchrotron source. The size of the signal varies as the square of the moment; therefore large-moment systems are good candidates for study.

Conversely, resonant magnetic scattering is observed when the incident photon energy is tuned to an absorption edge of an ion with which there is an associated magnetic polarization. It arises from second-order scattering processes. Under certain circumstances, enhancements in the magnetic scattering may be observed at such edges. These enhancements can be very large; increases of a factor of 50 have been observed at the L edges of the rare earths, and of several orders of magnitude at the M edges of the actinides. Such enhancements make the technique particularly useful for the study of weakly scattering systems—e.g., thin films, very small samples, critical fluctuations, and magnetic scattering from surfaces. Resonant scattering is sensitive to the total moment at a given site and it is not possible to extract $\vec{L}$ and $\vec{S}$ in the same way as it is for non-resonant scattering.

The resonant energy depends on the atomic species, and therefore resonant magnetic scattering is element specific—by tuning the incident photon energy to the absorption edge of a particular element, one can study the magnetic order of that sublattice with very little contribution from other magnetic sublattices. This is a particular strength of resonant scattering, in contrast to magnetic neutron scattering (MAGNETIC NEUTRON SCATTERING), which is sensitive to the total moment of ++ all sublattices contributing at a particular wave-vector. In addition, resonant scattering is also orbital specific—i.e., the magnetic polarization of different electronic orbitals may be probed independently. For example, in the elemental rare earths, both the (itinerant) $5d$ orbitals and the (localized) $4f$ orbitals carry an ordered moment. At an L_2 or L_3 edge, these states are reached through dipole or quadrupole transitions, respectively. Typically the dipole and quadrupole resonances differ in energy by 5 to 10 eV (Bartolomé et al., 1997), and have different polarization dependencies. They may therefore be separated empirically in a relatively straightforward manner. Recent experiments, for example, have exploited these features to elucidate the RKKY interaction in multicomponent systems (Pengra et al., 1994; Everitt et al., 1995; Langridge et al., 1999; Vigliante et al., 1998).

Both nonresonant and resonant magnetic scattering techniques offer extremely high reciprocal space resolution as a result of the intrinsic collimation of the synchrotron radiation. Typically, the resolution widths are $\sim 10^{-4}$ Å^{-1}. Thus, magnetic x-ray scattering is useful for detailed quantitative measurements of such properties as incommensurabilities, correlation lengths, and scattering lineshapes. In addition, the poor energy resolution of x-ray scattering experiments, relative to the relevant magnetic excitations, ensures that the quasi-elastic approximation is rigorously fulfilled and that the instantaneous correlation function is measured.

In many ways, x-ray magnetic scattering is complementary to the more established technique of neutron magnetic scattering (MAGNETIC NEUTRON SCATTERING). This latter probe measures many of the same properties of a system, and each technique has its own strengths and weaknesses. It is frequently of value to perform experiments using both techniques on the same sample.

Neutron scattering offers bulk penetration and a large magnetic cross-section, making it more generally applicable than x-ray magnetic scattering. In addition, inelastic magnetic neutron scattering is readily observable (in samples of sufficient volume) allowing the study of spin dynamics. Absolute intensities may be obtained, so it is also possible to measure ordered moments. As a result, unless there are special circumstances, it is generally preferable to solve the magnetic structure using neutron scattering. Here, powder neutron diffraction (NEUTRON POWDER DIFFRACTION) is particularly powerful. In contrast, powder magnetic x-ray diffraction is, typically, prohibitively small. However, neutron scattering is not inherently element specific, nor can it resolve $\vec{L}$ and $\vec{S}$ contributions simply, or study small-volume samples. Further, it requires a reactor or spallation source to perform the experiments. The reciprocal space resolution is typically lower by a factor of 10 than that achievable with x-rays, and there are some elements that have high neutron absorption cross-sections (e.g., Gd and Sm), which are difficult to study with neutrons but that have relatively large x-ray magnetic cross-sections. In addition, in some cases, the incident neutron flux will activate the sample, which can cause problems for subsequent experiments on the same sample. Finally, for high-quality samples, the magnetic elastic intensity can be affected by extinction, making quantitative determinations of the order parameter and lineshape difficult.

In summary, magnetic x-ray scattering is particularly suitable when high-resolution, or elemental or orbital-specific information is required, or when small samples

are involved, or if one needs to separate the $\vec{L}$ and $\vec{S}$ contributions to the ordered moment.

There are other techniques that can measure quantities related to those obtained by x-ray magnetic scattering. These include x-ray magnetic circular dichroism (XMCD), discussed in X-RAY MAGNETIC CIRCULAR DICHROISM, which is closely related to resonant magnetic x-ray scattering. XMCD is particularly useful for obtaining moment directions and $\vec{L}$ and $\vec{S}$ contributions in ferromagnets. Like resonant magnetic scattering it is both elemental and orbital specific. In addition, bulk techniques (Chapter 6) such as SQUID (superconducting quantum interference device; THERMOMAGNETIC ANALYSIS) magnetometry, and heat capacity measurements can provide insight into the magnetic behavior of a particular system, including the total (uniform) magnetization and the existence of phase transitions. These techniques also have the advantage of being available in a home laboratory, and are less demanding in their crystallographic quality requirements. However, such techniques cannot provide the microscopic information on the magnetic order that is available from scattering (x-ray and neutron) techniques.

In any discussion of magnetic x-ray scattering, one is inevitably led to considering the polarization states of the photon. The two important cases are linear and circular polarization. The degree of linear polarization is characterized by

$$P_\zeta = \frac{I_{\text{hor}} - I_{\text{vert}}}{I_{\text{hor}} + I_{\text{vert}}} \quad (1)$$

where I_{hor} is the horizontally polarized and I_{vert} the vertically polarized intensity. Similarly, the degree of circular polarization is defined by

$$P_\eta = \frac{I_{\text{left}} - I_{\text{right}}}{I_{\text{left}} + I_{\text{right}}} \quad (2)$$

where I_{left} is the intensity of the left circularly polarized light and I_{right} the intensity of the right circularly polarized light. In the case of linear polarization, the polarization state is often described relative to the scattering plane. Polarization perpendicular to the scattering plane is referred to as σ polarization; polarization parallel to the scattering plane is called π polarization.

It is also useful to review the nomenclature for absorption edges, which are labeled by the core hole created in the absorption process. The principal quantum number ($n = 1,2,3,4\ldots$) of the core-hole is labeled by K,L,M,N,.... and the angular momentum s, $p_{1/2}$, $p_{3/2}$, $d_{3/2}$, $d_{5/2}$,... by 1,2,3,4,5... Thus the L_3 edge corresponds to the excitation of the $2p_{3/2}$ electron.

This unit provides an overall background to the theoretical and experimental principles of nonresonant and resonant magnetic x-ray scattering. The more general applications of the resonant terms in the cross-section will not be discussed here; some of them are covered in RESONANT SCATTERING TECHNIQUES and X-RAY MAGNETIC CIRCULAR DICHROISM. The experimental techniques are very similar to standard synchrotron x-ray diffraction experiments; therefore discussion of the beamlines and spectrometers will be limited, with only those elements particular to magnetic scattering, such as polarization analyzers, discussed in any detail. At high incident photon energy, it is possible to perform inelastic experiments utilizing the nonresonant magnetic scattering terms in the cross-section. This allows for the measurement of magnetic Compton profiles (projections of the momentum density of the magnetic electrons, along a given direction). Magnetic Compton scattering may only be applied to ferromagnets, because of the incoherent nature of the inelastic scattering, and the technique differs in many aspects of its practical implementation from the diffraction experiments discussed here. It will therefore not be covered in this unit. Reviews may be found in Schülke (1991) and Lovesey and Collins (1996).

The history of the field will not be covered in this unit. Brief reviews have been given in Brunel and de Bergevin (1991), Gibbs (1992), and McMorrow et al. (1999).

PRINCIPLES OF THE METHOD

Theoretical Concepts

As mentioned in the introduction, magnetic x-ray scattering arises from terms in the scattering cross-section that are dependent on the magnetic moment. For nonresonant scattering, these terms can be most simply be understood using a classical picture of scattering from free electrons (de Bergevin and Brunel, 1981a,b). The different classical scattering mechanisms are shown in Figure 1. Charge (Thomson) scattering is illustrated in the uppermost (a) diagram of Figure 1, in which the electric field of the

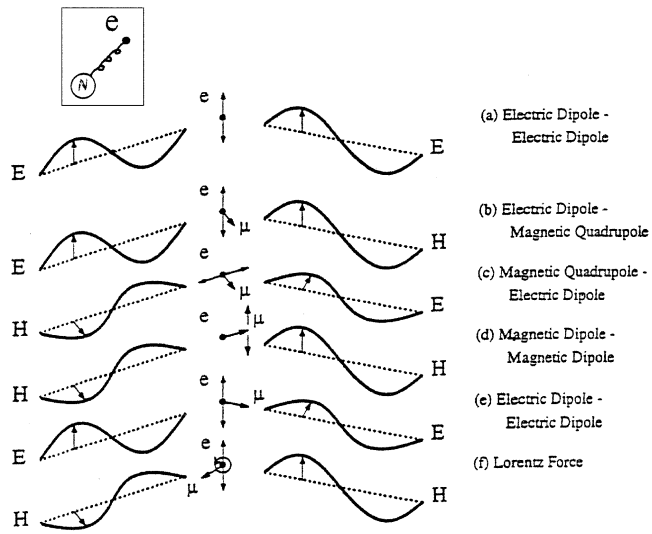


Figure 1. Classical picture of the dominant mechanisms for an x-ray scattering from an electron (*e*). In (**a**) the x-ray is scattered from the electronic charge, while in (**b-f**) the magnetic moment of the electron is involved. Mechanisms (**b-d**) result in spin-dependent scattering, while (**e-f**) are for a moving magnetic dipole with finite orbital angular momentum. Mechanism (**a**) represents resonant scattering from an electron bound to an atomic nucleus (*N*) as shown in the inset. Figure adapted from McMorrow et al. (1999).

incident photon accelerates the electronic charge, leading to electric dipole radiation in the form of the scattered photon. The terms that give rise to nonresonant magnetic scattering are illustrated in Figure 1, diagrams b to f. These arise from the (weaker) interaction between the electric and magnetic fields of the photon and the magnetic moment of the electron. The scattering due to nonresonant interactions is reduced relative to charge scattering, by a factor proportional to $(\hbar\omega/mc^2)^2$, which is $\approx 10^{-4}$ for an incident photon energy of $\hbar\omega = 10$ keV (here h is Planck's constant, $\hbar = h/2\pi$, ω is the frequency, m is the electron mass, and c is the speed of light).

Resonant scattering arises from the fact the electrons are not free, but bound to the nucleus. The resulting quantized bound states give rise to well defined transition energies, and resonant scattering results when these transitions are selectively excited. Maintaining the classical picture, this may be visualized with a ball and spring model and its associated resonant frequencies (see insert over diagram "a" of Fig. 1).

The scattering mechanisms of Figure 1 may be calculated quantum mechanically within second-order perturbation theory with the appropriate choice of interaction Hamiltonian (Blume, 1985, 1994).

The cross-section for elastic scattering of x-rays is

$$\left(\frac{d\sigma}{d\Omega}\right) = r_0^2 \left| \sum_n e^{i\vec{Q}\cdot\vec{r}_n} f_n(\vec{k},\vec{k}',\hbar\omega) \right|^2 \tag{3}$$

where $\vec{Q}$ is the momentum transfer in the scattering process, $r_0 = 2.8 \times 10^{-15}$ m is the classical electron radius, $\vec{k}$ and $\vec{k}'$ are the incident and scattered wave-vectors of the photon, $d\sigma$ is the differential cross-section of the scattering center for scattering into the solid angle element $d\Omega$. The scattered amplitude from site $\vec{r}_n$ is f_n. It may be expressed as the sum of several contributions

$$f_n(\vec{k},\vec{k}',\hbar\omega) = f_n^{\text{charge}}(\vec{Q}) + f_n^{\text{non-res}}(\vec{Q},\vec{k},\vec{k}') + f_n^{\text{res}}(\vec{k},\vec{k}',\hbar\omega) \tag{4}$$

where, for incident and scattered photon polarization states described by the unit vectors $\hat{\varepsilon}$ and $\hat{\varepsilon}'$, respectively

$$f_n^{\text{charge}}(\vec{Q}) = -\rho_n(\vec{Q})\hat{\varepsilon}\cdot\hat{\varepsilon}' \tag{5}$$

is the Thomson scattering, which is proportional to the Fourier transform of the charge density, $\rho(\vec{Q})$. The dot product of the polarization vectors requires that the polarization state not be rotated by the scattering. The nonresonant magnetic scattering amplitude, $f_n^{\text{non-res}}$ and the resonant magnetic scattering amplitude, which is contained in the anomalous (energy dependent) terms, f_n^{res}, are considered in the following section.

Nonresonant Scattering. The nonresonant scattering amplitude may be written (Blume, 1985)

$$f_n^{\text{non-res}}(\vec{Q}) = -i\left(\frac{\hbar\omega}{mc^2}\right)\left[\frac{1}{2}\vec{L}_n(\vec{Q})\cdot\vec{A} + \vec{S}_n(\vec{Q})\cdot\vec{B}\right] \tag{6}$$

where $\frac{1}{2}\vec{L}_n(\vec{Q})$ and $\vec{S}_n(\vec{Q})$ are proportional to the Fourier transforms of the orbital and spin magnetization densities, respectively. The vectors $\vec{A}$ and $\vec{B}$ contain the polarization dependences of the two contributions to the scattering. Importantly, these two vectors are not equal and have distinct $\vec{Q}$ dependences. Hence, by measuring the polarization of a scattered beam for a number of magnetic reflections, one may obtain the ordered $\vec{L}$ and $\vec{S}$ moments for the system. For synchrotron diffraction experiments with linear incident polarization, $\vec{A}$ and $\vec{B}$ are most usefully expressed in terms of linear polarization basis states (Blume and Gibbs, 1988). In this basis, the charge scattering may be written as

$$\langle M_{\text{charge}}\rangle = \rho(\vec{Q})\begin{pmatrix} 1 & 0 \\ 0 & \cos 2\theta \end{pmatrix} \tag{7}$$

where 2θ is the angle between $\vec{k}$ and $\vec{k}'$. Note, this matrix is diagonal, reflecting the fact that the polarization state is not rotated by the charge scattering process. In this same basis the non-resonant magnetic scattering operator is

$$\langle M_{\text{non-res}}\rangle = \begin{pmatrix} \langle M_{\sigma\sigma}\rangle\langle M_{\pi\sigma}\rangle \\ \langle M_{\sigma\pi}\rangle\langle M_{\pi\pi}\rangle \end{pmatrix} = -i\left(\frac{\hbar\omega}{mc^2}\right)\times \begin{pmatrix} \sin 2\theta S_2 & -2\sin^2\theta[\cos\theta(L_1+S_1) - \sin\theta S_3] \\ 2\sin^2\theta[\cos\theta(L_1+S_1) + \sin\theta S_3] & \sin 2\theta[2\sin^2\theta L_2 + S_2] \end{pmatrix} \tag{8}$$

in the coordinate system of Figure 2. Here and throughout this unit, the $\langle\ \rangle$ symbol represents the ground state expectation value of an operator. Note that non-resonant magnetic scattering is only sensitive to the components of the orbital angular momentum perpendicular to the

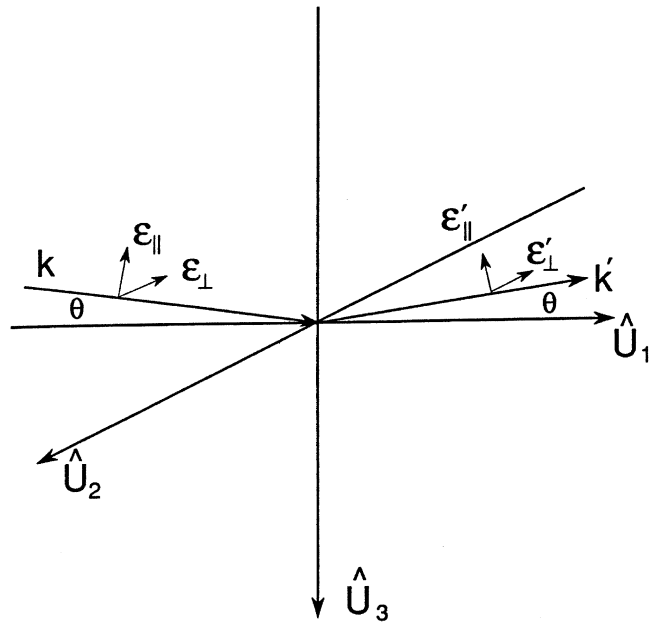


Figure 2. The coordinate system used in this unit, after Blume and Gibbs (1988). Note that the momentum transfer $\vec{Q}$ is along $-\hat{U}_3$.

momentum transfer (i.e., the components L_1 and L_3). For pure magnetic scattering, then

$$\frac{d\sigma}{d\Omega} = r_o^2 \left(\frac{\hbar\omega}{mc^2}\right)^2 \mathrm{tr}\langle M_{\text{non-res}}\rangle \rho \langle M^{\dagger}_{\text{non-res}}\rangle \tag{9}$$

where tr is the trace operator, $\langle M^{\dagger}_{\text{non-res}}\rangle$ is the Hermitian adjoint of $\langle M_{\text{non-res}}\rangle$ and ρ is the density matrix for the incident polarization. For example, for linear incident polarization

$$\rho = \frac{1}{2}\begin{pmatrix} 1+P_\zeta & 0 \\ 0 & 1-P_\zeta \end{pmatrix} \tag{10}$$

The general case is given in Blume and Gibbs (1988). With this form, the $\sigma \rightarrow \sigma$ scattering is given by $|\langle M_{\sigma\sigma}\rangle|^2$, the $\sigma \rightarrow \pi$ by $|\langle M_{\sigma\pi}\rangle|^2$ and so on.

Equation 9 allows for the straightforward determination of the polarization dependence of the scattering for a given scattering geometry and magnetic structure. As an illustrative example, consider the case of MnF_2, an Ising antiferromagnet for which the spins are oriented along the c axis, and the orbital moment is quenched, $L_1 = L_2 = L_3 = 0$. The ordering wave vector is perpendicular to the c axis. If the scattering geometry is arranged such that the c axis is perpendicular to the scattering plane, then $S_2 = S$ and $S_1 = S_3 = 0$. The scattering matrix, Equation 8, is then diagonal, and for a σ incident beam, all the scattering is into the σ channel. If the sample is then rotated about the scattering vector such that the spins are then in the scattering plane, i.e., $S_1 = S$ and $S_2 = S_3 = 0$, then the diagonal elements are zero and all the scattering rotates the photon polarization state by 90°, i.e., the scattering is entirely $\sigma \rightarrow \pi$ and $\pi \rightarrow \sigma$. Further, the scattered intensity is reduced by a factor of $\sin^2\theta$, relative to the previous geometry. In typical experiments, this is a small number, and, as a result, non-resonant magnetic scattering is largely sensitive to the component of the spin perpendicular to the scattering plane.

These predictions have been largely experimentally verified in MnF_2 (Goldman et al., 1987; Hill et al., 1993; Brückel et al., 1996).

An estimate of the size of pure non-resonant magnetic scattering, relative to that of charge scattering (ignoring polarization factors) may be obtained from Blume (1985)

$$\frac{\sigma_{\text{mag}}}{\sigma_{\text{charge}}} = \left(\frac{\hbar\omega}{mc^2}\right)^2 \left(\frac{N_{\text{m}}}{N}\right)^2 \left(\frac{f_{\text{m}}}{f}\right)^2 \langle \mu^2 \rangle \tag{11}$$

where N_{m}/N is the ratio of the number of magnetic electrons to the total number of electrons, f_{m}/f is the ratio of the magnetic to charge form factors, and μ^2 is the average magnetic moment. For Fe and 10 keV photons

$$\frac{\sigma_{\text{mag}}}{\sigma_{\text{charge}}} \sim 4 \times 10^{-6} \langle \mu^2 \rangle \tag{12}$$

In fact, this ratio may be further reduced for a particular Bragg magnetic reflection because the magnetic form factor falls off faster than the charge form factor. While, to a large extent, the small amplitude of the non-resonant scattering may be countered by the use of the intense incident x-ray flux available at synchrotrons, it is generally true that samples of relatively small mosaic spread (e.g., $<0.1°$) are required in order to make full use of this incident intensity. Nevertheless, even quite small moments have been observed. For example, non-resonant magnetic scattering from the spin density wave in chromium, which has an ordered moment of 0.4 μ_B/atom, gave rise to count rates on the order of 10 counts per sec at a bending magnet beamline at the National Synchrotron Light Source (Hill et al., 1995a) and >1000 counts per second at an undulator beamline at the European Synchrotron Radiation facility (Mannix, 2001).

Resonant Scattering. The resonant terms arise from second-order scattering processes which may, in a loose sense, be thought of as the absorption of an incident photon, the creation of a short lived intermediate (excited) state, and the decay of that state, via the emission of an elastically scattered photon, back into the ground state. For Bragg scattering, this process is coherent and the phase relationship between the incident and scattered photons is preserved. A schematic energy level diagram illustrating this process for a rare-earth ion is shown in Figure 3.

The resonant scattering amplitude is (Blume, 1994)

$$f_n^{\text{res}}(\vec{k},\vec{k}',\hbar\omega) = -\frac{1}{m}\sum_c \left(\frac{E_a - E_c}{\hbar\omega}\right) \frac{\langle a|O_\lambda(\vec{k})|c\rangle\langle c|O^{\dagger}_{\lambda'}(\vec{k}')|a\rangle}{E_a - E_c - \hbar\omega} + \frac{1}{m}\sum_c \left(\frac{E_a - E_c}{\hbar\omega}\right) \frac{\langle a|O^{\dagger}_{\lambda'}(\vec{k}')|c\rangle\langle c|O_\lambda(\vec{k})|a\rangle}{E_a - E_c + \hbar\omega - i\frac{\Gamma}{2}} \tag{13}$$

where $E_a, |a\rangle$, and $E_c, |c\rangle$, are the energies and wavefunctions of the initial and intermediate states, respectively, Γ is the inverse lifetime of the intermediate state, and the operator $O_\lambda(\vec{k})$ is given by

$$O_\lambda(\vec{k}) = \sum_n e^{i\vec{k}\cdot\vec{r}_n}(\vec{P}_n \cdot \hat{\varepsilon}_\lambda - i\hbar\hat{\varepsilon}_\lambda \cdot (\vec{k} \times \vec{S}_n)) \tag{14}$$

The two terms of Equation 13 give rise to anomalous dispersion, i.e., they contain all the energy dependence of the scattering. They have come to play an increasingly important role in x-ray scattering experiments with the availability of energy tunable x-ray sources at synchrotron beamlines. Some of these applications are discussed in the units on resonant scattering techniques (RESONANT SCATTERING TECHNIQUES). In the limit of zero momentum transfer, these terms give rise to XMCD effects (X-RAY MAGNETIC CIRCULAR DICHROISM).

In a resonant experiment, the incident photon energy, $\hbar\omega$ is tuned to the absorption edge, such that $\hbar\omega = E_c - E_a$, and the second of the two terms in Equation 13 is enhanced. In this limit, it is common to ignore the first, non-resonant term of Equation 13 (Blume, 1994). That resonant effects may contain some sensitivity to the magnetization was first suggested by Blume (1985) and observed by Namikawa et al. (1985). It was explored in

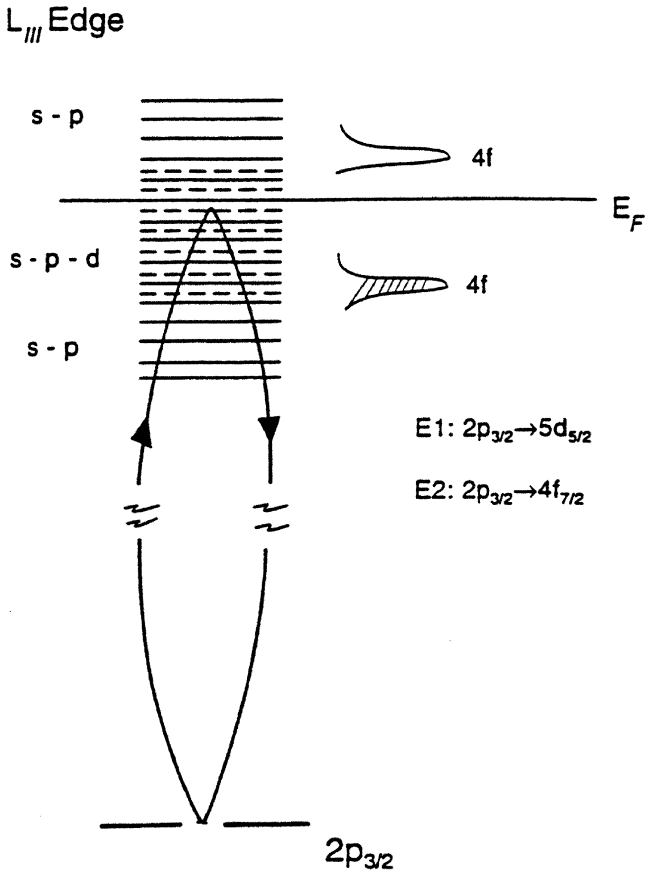


Figure 3. A schematic single-electron picture of the resonant scattering process, illustrated for a rare earth, at the L_3 edge. A $2p_{3/2}$ electron is excited into an unoccupied state above the Fermi level by the incident photon. The $5d$ states are reached through electric dipole transitions and the $4f$ through electric quadrupole transitions. In the elemental rare earths, the $5d$ states form delocalized bands, polarized through an exchange interaction with the localized, magnetic $4f$ electrons. Resonant elastic scattering results when the virtually excited electron decays by filling the core hole and emitting a photon. Figure taken from Gibbs (1992).

detail by Hannon et al. (1988) following the experimental work of Gibbs et al. (1988).

The magnetic sensitivity arises from spin-orbit correlations, which must be present in at least one of the two levels involved in the resonant process, and from exchange effects. "Exchange effects" refers to any phenomena originating from the anti-symmetry of the wave function under the exchange of electrons. This includes the exclusion principle, which only allows transitions to unoccupied spin orbitals in partially filled spin polarized valence states, as well as the exchange interaction between electrons in different orbitals. Resonant magnetic scattering is often referred to as x-ray resonant exchange scattering (XRES), as a result of the requirement that exchange effects be present.

Almost all of the experimental phenomena observed in resonant magnetic scattering can be explained in terms of a one-electron picture of electric multipole transitions (Hannon et al., 1988; Luo et al., 1993). Magnetic multipole transitions are smaller by a factor of $\hbar\omega/mc^2$ and have not been considered in the literature to date. In this picture, the size of the resonant enhancement of the magnetic scattering depends on the radial matrix elements ($\langle c|r^{\mathrm{L}}|a\rangle$, where $L = 1, 2..$ for dipole, quadrupole... transitions), the magnetic polarization of the intermediate state, the value of the intermediate-state lifetime, and the energy width of the incident photon. Thus the largest enhancements are expected when large matrix elements couple states with strong polarizations.

Such a situation occurs at the M_4 and M_5 edges in the actinides, for which dipole allowed transitions couple to the magnetic $5f$ electrons. Enhancements of several orders of magnitude are typical at such edges, allowing extremely small moments to be studied. For example, in URu_2Si_2, scattering from an ordered moment of $0.02\mu_B$ was observed (Isaacs et al., 1990).

Conversely, at the L_2 and L_3 edges of the rare-earths, the dipole transitions are $2p \leftrightarrow 5d$. The $5d$ electrons are polarized through an exchange interaction with the localized, magnetic $4f$ electrons, and the induced moment is relatively small, $\approx 0.2\mu_B$. In addition, the radial matrix elements are 5 to 10 times smaller than at the $M_{4,5}$ edges (Hamrick, 1994) and smaller enhancements are therefore expected. Enhancement factors of $\sim$50 were observed at the Ho L_3 edge (Gibbs et al., 1988, 1991). Quadrupolar transitions at the rare-earth $L_{2,3}$ edges, however, couple to the magnetic $4f$ levels, $2p \leftrightarrow 4f$. To some extent the greater polarization of the $4f$ orbitals compensates for the much smaller quadrupole matrix elements, and this scattering is also observable. In Ho, it is $\approx$5 times weaker than the dipole scattering at the first harmonic, after correction for absorption (Gibbs et al., 1988). Table 1 summarizes the absorption edges for various elements and approximate enhancement factors, together with a typical reference.

One property of resonant scattering is that it does not contain a magnetic form factor, in contrast with non-resonant scattering and magnetic neutron diffraction. In essence, this is because of the two-step nature of the scattering event. The lack of a magnetic form factor is in fact beneficial when observing magnetic reflections at large $\vec{Q}$, something that is often necessary when studying off-specular peaks in a reflection geometry.

In order to extract quantities such as ordered moments and moment directions, the cross-section needs to be considered in more detail (Hannon et al., 1988; Carra et al., 1993; Luo et al., 1993). Absolute moments of one species have been estimated in multicomponent systems by comparison with signals from a known moment in the other species, measured in the same experiment, e.g., in alloy systems such as Dy/Lu (Everitt et al., 1995). However, the large number of estimates and approximations that are required, especially when the intermediate states are delocalized and band structure calculations must be invoked, mean that this method is at best semiquantitative. Work along these lines in Ho/Pr alloys (Vigliante et al., 1998) highlights another problem with extracting such quantitative information, i.e., a breakdown of the one-electron, atomic picture. This is manifest most

Table 1. Resonant Energies and Approximate Enhancements for Various Atomic Absorption Edges

Series	Edge	Energy Range	Transition	Enhancement	Reference
3*d* transition metals	$L_{2,3}$	400–1000 eV	E1 $2p \rightarrow 3d$	$\geq 10^2$	Tonnerre et al., 1995
	K	4.5–9.5 keV	E1 $1s \rightarrow 4p$	~ 2	Namikawa et al., 1985
			E2 $1s \rightarrow 3d$	~ 2	Hill et al., 1996; Neubeck et al., 2001
5*d* transition metals	$L_{2,3}$	5.8–14 keV	E1 $2p \rightarrow 5d$	$\sim 10^2$	de Bergevin et al., 1992
Rare earths	$L_{2,3}$	5.8–10.3 keV	E1 $2p \rightarrow 5d$	50	Gibbs et al., 1988
			E2 $2p \rightarrow 4f$		
Actinides	$M_{4,5}$	3.5–4.7 keV	E1 $3d \rightarrow 5f$	10^7	Isaacs et al., 1989
	M_3	4.3 keV	E1 $3p \rightarrow 6d$	< 2	McWhan et al., 1990
			E2 $3p \rightarrow 5f$		
	$L_{2,3}$	7–20 keV	E1 $2p \rightarrow 6d$	$\sim 10^2$	—[a]
		7–20 keV	E2 $2p \rightarrow 5f$		

[a] D. Wermeille, C. Vettier, N. Bernhoeft, N. Stunault, P. Lejay, and J. Flouquet, unpub. observ.

obviously in the branching ratio (the ratio of scattered intensities obtained at a pair of resonances resulting from excitations of two spin-orbit split core-holes, e.g., the L_2 and L_3 or the M_4 and M_5 edges). In the simplest theories (Hannon et al., 1988), these are predicted to be near unity. In fact, empirically, they are known to vary across the rare-earth series, with the heavy rare earths supporting $L_2/L_3 < 1$ and the light rare earths having $L_2/L_3 > 1$ (Gibbs et al., 1991; Hill et al., 1995b; D. Watson et al., 1996; Vigliante et al., 1998). Further, the ratio has even been observed to be temperature dependent. These results point to a need for more detailed theoretical work (e.g., van Veenendaal et al., 1997). However, until a full understanding of the scattered amplitude (and electronic structure used in the calculations) is obtained, the resonant scattered intensity must be interpreted with some caution. A very recent example of an attempt along these lines is the work of Neubeck et al. (2001).

The polarization dependence of the resonant scattering is a function of moment direction and scattering geometry. For dipole (E1) scattering, it is

$$f_{n\mathrm{E1}}^{\mathrm{X\,RES}} = F^{(0)}\begin{pmatrix} 1 & 0 \\ 0 & \cos 2\theta \end{pmatrix} - iF^{(1)}\begin{pmatrix} 0 & z_1\cos\theta + z_3\sin\theta \\ z_3\sin\theta - z_1\cos\theta & -z_2\sin 2\theta \end{pmatrix} + F^{(2)}\begin{pmatrix} z_2^2 & -z_2(z_1\sin\theta - z_3\cos\theta) \\ +z_2(z_1\sin\theta + z_3\cos\theta) & -\cos^2\theta(z_1^2\tan^2\theta + z_3^2) \end{pmatrix} \tag{15}$$

in the same coordinate system and basis as previously (Hannon et al., 1988; Pengra et al., 1994; Hill and McMorrow, 1996). The z_i are the components of a unit vector aligned along the quantization axis at site n and the $F^{(L)}$ are the resonant matrix elements. The first term contains no dependence on the magnetic moment and therefore only contributes to the charge Bragg peaks. The second term is linear in z and for an antiferromagnet will produce new scattering at the magnetic wave-vector. The final term is quadratic in z. This will produce scattering at twice the ordering wave vector. For an incommensurate antiferromagnet, this will give rise to a second-order resonant harmonic.

By way of an example, consider the case of a basal plane spiral, such as occurs in Ho, Tb, and Dy. In this structure the moments are confined to the *a*-*b* plane of the hcp crystal structure, in ferromagnetic sheets. The direction of the magnetization rotates from basal plane to basal plane, creating a spiral structure propagating along the *c* axis. The modulation vector then lies along (00L). The magnetic moment takes the form

$$\hat{z}_n = \cos(\tau \cdot \mathbf{r}_n)\hat{U}_1 + \sin(\tau \cdot \mathbf{r}_n)\hat{U}_2 \tag{16}$$

where τ is the period of the spiral, $\mathbf{r}_n$ the coordinate of the nth atom, and $\hat{U}_1$ and $\hat{U}_2$ are unit vectors defined in Figure 2.

Now, the full x-ray scattering cross-section may be written as

$$\frac{d\sigma}{d'} = \sum_{\lambda\lambda'} P_\lambda |\langle\lambda'|\langle M_{\mathrm{charge}}\rangle|\lambda\rangle - i\frac{\hbar\omega}{mc^2}\langle\lambda'|\langle M_{\mathrm{non-res}}\rangle|\lambda\rangle + \langle\lambda'|\langle M_{E1}^{\mathrm{X\,RES}}\rangle|\lambda\rangle + \langle\lambda'|\langle M_{\mathrm{E2}}^{\mathrm{X\,RES}}\rangle|\lambda\rangle + \cdots I^2 \tag{17}$$

where λ, λ' are incident and scattered polarization states and P_λ the probability for incident polarization λ. As before, $\langle\ \rangle$ represents the ground-state expectation value of the respective operators i.e.,

$$\langle M_{\mathrm{E1}}^{\mathrm{X\,RES}}\rangle = \langle a|\sum_n e^{i\mathbf{Q}\cdot\mathbf{r}_n} f_{n\mathrm{E1}}^{\mathrm{X\,RES}}|a\rangle \tag{18}$$

and E1 and E2 are the dipole and quadrupole terms, respectively. For the purposes of this example, we assume that the dipole contribution is dominant and ignore any interference effects. This will be valid at resonance and away from charge Bragg peaks. Then from Equation 15

$$M_{\mathrm{E1}}^{\mathrm{X\ RES}} = \sum_n e^{i\mathbf{Q}\cdot\mathbf{r}_n} \begin{pmatrix} z_2^2 F^{(2)} & -iz_1\cos\theta F^{(1)} - z_2 z_1 \sin\theta F^{(2)} \\ iz_1\cos\theta F^{(1)} + z_2 z_1 \sin\theta F^{(2)} & iz_2 \sin 2\theta F^{(1)} - z_1^2 \sin^2\theta F^{(2)} \end{pmatrix} \tag{19}$$

For simplicity, we take $P_\lambda = \delta_{\lambda\sigma}$, i.e., an incident beam of linear polarization perpendicular to the scattering plane. Then, only $\sigma \rightarrow \sigma$ and $\sigma \rightarrow \pi$ terms contribute and

$$\frac{d\sigma}{d\Omega} = |\langle a| \sum_n e^{i\mathbf{Q}\cdot\mathbf{r}_n} z_2^2 F^{(2)} |a\rangle|^2 + |\langle a| \sum_n e^{i\mathbf{Q}\cdot\mathbf{r}_n} (iz_1\cos\theta F^{(1)} + z_2 z_1 \sin\theta F^{(2)} |a\rangle|^2 \tag{20}$$

Substituting $z_1 = \cos(\tau \cdot \mathbf{r}_n), z_2 = \sin(\tau \cdot \mathbf{r}_n)$ and writing the sine and cosine terms as the sums of complex exponentials, we obtain the cross-section for a flat spiral with incident radiation perfectly polarized perpendicular to the scattering plane

$$\left.\frac{d\sigma}{d\Omega}\right)_{\mathrm{E1}}^{\mathrm{X\ RES}} = \frac{1}{4}F^{2(2)}\delta(\mathbf{Q}-\mathbf{G}) + \frac{1}{4}\cos^2\theta F^{2(1)}\delta(\mathbf{Q}-\mathbf{G}\pm\tau) + \frac{1}{16}(1+\sin^2\theta)F^{2(2)}\delta(\mathbf{Q}-\mathbf{G}\pm 2\tau) \tag{21}$$

where $\mathbf{G}$ is a reciprocal lattice vector. The intensity of the observed scattering will only be proportional to the above expression if both polarization components of the scattered beam are collected with equal weight. This is the case if no analyzer crystal is employed. If one is used, then the polarization component that is in the scattering plane of the analyzer must be weighted by a factor $\cos^2 2\theta_{\mathrm{A}}$, where θ_{A} is the Bragg angle of the analyzer crystal. From Equation 21, we see that in addition to producing scattering at the Bragg peak, the resonant dipole contribution produces two magnetic satellites at $\pm\tau$ and $\pm 2\tau$ on either side of the Bragg peak. The scattering at $\pm 2\tau$ are the resonant harmonics. Note that these are distinct from any diffraction harmonics that may arise from a nonsinusoidal modulation of the magnetization density.

For quadrupole transitions, the situation is still more complex (Hannon et al., 1988; Hill and McMorrow, 1996) and terms up to $O(z^4)$ are obtained, giving rise to up to four orders of resonant harmonics. The presence of third and fourth resonant harmonics is a clear indication of quadrupole transitions.

Finally, the original formulation of the resonant cross-section (Hannon et al., 1988) has been re-expressed in terms of effective scattering operators following theoretical work performed for the XMCD cross-section (Thole et al., 1992; Carra et al., 1993; Luo et al., 1993; Carra and Thole, 1994). These scattering operators are constituted from spin-orbital moment operators of the valence shell involved in the scattering process and are valid in the short collision time approximation. They offer the possibility of relating resonant elastic and inelastic scattering measurements directly to physically interesting properties of the system.

Observation of Ferromagnetic Scattering. Given the presence of terms in the cross-section sensitive to the magnetization density, the observation of pure magnetic scattering is, in principle, straightforward for antiferromagnets because the superlattice peaks are well separated from the much stronger charge Bragg peaks of the underlying chemical lattice. For ferromagnets, however, the situation is complicated by the fact that the magnetic and charge scattering are superposed.

At first glance, the situation might seem hopeless, since one is apparently required to resolve a signal of one part in 10^6. However, the presence of the charge scattering at the same position does allow for the possibility of observing the interference between the charge and magnetic scattering amplitudes. For example, if the charge and magnetic scattering were in phase, then

$$\frac{d\sigma}{d\Omega} = |f_c + f_m|^2 = f_c^2 + f_m^2 + 2f_c f_{\mathrm{m}} \tag{22}$$

where $f_{\mathrm{m}}(f_{\mathrm{c}})$ is the magnetic (charge) scattering amplitude. The interference term is linear in f_{m} and not quadratic as for pure magnetic scattering, and this scattering is therefore much stronger, on the order of 10^{-3} of the charge scattering. In addition, it changes sign when the sign of the magnetization is changed, allowing the magnetic scattering to be distinguished from the charge scattering in a difference experiment. Unfortunately, the charge and magnetic scattering are in fact precisely 90° out of phase for plane-polarized incident light (in centrosymmetric structures), and thus there is no interference between them.

The interference can be recovered, however, if there is a phase shift between the magnetic and charge amplitudes—e.g., by utilizing resonant effects, or, in non-centrosymmetric structures, by use of circularly polarized light. For non-resonant magnetic scattering, the last option is the only generally applicable alternative. Circularly polarized light also allows for the possibility of reversing the sign of the magnetic scattering by "flipping" the helicity of the incident beam, as well as by reversing the magnetization.

In practice, difference experiments are performed in which the scattered intensity is measured in each of the two configurations and the asymmetry ratio

$$R = \frac{I_+ - I_-}{I_+ + I_-} \tag{23}$$

is calculated, where I_+ and I_- are the intensities measured in the two configurations. This approach eliminates many sources of systematic errors.

For the observation of non-resonant ferromagnetic scattering, the preferred incident beam has some degree of linear polarization as well as a finite circular polarization. The asymmetry ratio is then maximized by scattering in

the plane of the linear polarization of the beam at a scattering angle of $2\theta = 90°$, thus minimizing the charge scattering related to the linear polarization of the incident beam. In this limit, for non-resonant magnetic scattering (Lovesey and Collins, 1996)

$$R = -2\left(\frac{\hbar\omega}{mc^2}\right)\frac{P_\eta}{1+P_\eta}\left(\frac{F_S}{F_c}\sin\alpha + \frac{F_L}{F_c}(\sin\alpha + \cos\alpha)\right) \quad (24)$$

where F_S, F_L, and F_c are the structure factors for the spin, orbital and charge densities respectively, α is the angle between the spin and orbital moments (assumed collinear) and the incident beam, P_ζ is the degree of linear polarization (also see Equation 1), and P_η is the degree of circular polarization (also see Equation 2). Since the scattering is within the plane of linear polarization, $P_\zeta < 0$, and therefore R is maximized by maximizing P_ζ while maintaining a non-zero circular polarization. By measuring the asymmetry ratio for two different angles, α, and taking the ratio of these quantities, then the only unknown is the ratio of the spin to orbital structure factors. This data may then be used together with polarized beam neutron experiments, which determine $(F_S + F_L)$, to obtain F_S and F_L separately, without the need to characterize the polarization state.

Observation of Ferromagnetic Scattering: Resonance Effects. The anomalous dispersion effects associated with absorption edges open up new possibilities for inducing interference between charge and magnetic scattering amplitudes, and thus for studying ferrimagnetism and ferromagnetism. Such techniques can be placed into two classes: the interference of non-resonant magnetic scattering with resonant charge scattering and the interference of resonant magnetic scattering with non-resonant charge scattering. Lovesey and Collins (1996) refer to these as magnetic-resonant charge interference scattering and resonant magnetic-charge interference scattering, respectively. In each case, the phase shift is induced through the anomalous terms, and thus circularly polarized light is not required. In the former case, for which the non-resonant magnetic scattering terms are utilized, it is still possible, in principle, to separate the $\vec{L}$ and $\vec{S}$ contributions to the scattering.

In order for the results of such experiments to be cleanly interpreted, it is important that the experiment be performed in the appropriate limit of the magnetic scattering terms (i.e., resonant or non-resonant). For example, in order to ensure that non-resonant magnetic scattering is dominant, but still be able to take advantage of the resonant phase shift in the charge scattering amplitude, one could use an absorption edge of a nonmagnetic ion, or move far enough above an edge associated with a weak enhancement (such as a K edge in a transition metal), so that the resonant magnetic scattering is negligible. This latter approach was utilized in the first magnetic scattering experiments on ferromagnets (de Bergevin and Brunel, 1981a,b; Brunel, 1983).

Alternatively, the different polarization dependence of resonant and non-resonant magnetic scattering may be used to select a particular channel, utilizing the fact that for interference to occur it must be allowed by both magnetic and charge scattering. Recall that charge scattering (Equation 7) permits only $\sigma \rightarrow \sigma$ and $\pi \rightarrow \pi$ processes and that in a difference experiment, the only terms which contribute to the observed signal are odd in z_n. Now, for dipole resonant scattering, $\sigma \rightarrow \sigma$ processes are forbidden in the first-order term (Equation 15) and therefore resonant magnetic-charge interference is only observed in the $\pi \rightarrow \pi$ channel (ignoring second-order terms). Conversely, non-resonant magnetic scattering may contribute to both $\sigma \rightarrow \sigma$ and $\pi \rightarrow \pi$ channels (Equation 15) and so, for example, the $\sigma \rightarrow \sigma$ channel may be used to select the non-resonant magnetic-charge interference.

PRACTICAL ASPECTS OF THE METHOD

Experimental Hardware

Although the first x-ray magnetic scattering experiments were performed with laboratory sources, the inherent advantages of synchrotron sources—the high flux and collimation, energy tunability, and the possibility of manipulating the incident polarization state—make such facilities nearly ideal for the study of magnetic phenomena with x rays. For all practical purposes, therefore, a synchrotron source is a requirement for a successful experiment. There are three sources of synchrotron radiation in current storage rings. These are the dipole bending magnet, which applies a circular acceleration to the electron beam, and the so-called insertion devices: wigglers and undulators. These latter are linear periodic arrays of alternating bending magnets. A wiggler of N poles will provide N times the flux of a single bending magnet of the same field strength, with the same spectral distribution. An undulator is similar to a wiggler except that the amplitude of the oscillations is much smaller. The amplitude is constrained such that, when viewed far from the source, the radiation from each oscillation is coherent with the others. In such a limit, the flux scales as N^2, and emission is peaked around harmonics of a fundamental (Brown and Lavender, 1991).

The experimental hardware may be broken down into three broad classes of components:

1. *The beamline.* This delivers the x-rays to the sample and is comprised of various optical elements, including monochromator and (typically) focusing mirror, together with evacuated beam transport tubes and possibly a further optical element to alter the polarization state, such as a phase plate.
2. *The spectrometer.* This instrument manipulates the sample and the detector, providing the ability to determine the angle of incidence and exit of the x rays, to a precision of $\sim 0.001°$, allowing the experimenter to explore reciprocal space, often as a function of incident energy. The spectrometer may also include a polarization analyzer for detecting the polarization state of the scattered beam.
3. *The sample environment.* Frequently in magnetic experiments this is a cryostat of some kind.

In the following, the practical aspects of the beamline and the spectrometer are discussed.

The Beamline. Synchrotron beamlines are large, expensive pieces of equipment. They must accept the incident white light spectrum and its associated heat load from the x-ray source and deliver a monochromatic, stable beam to the sample. Optical elements upstream of and including the monochromator must be cooled to prevent problems associated with heating, such as degradation of the optical elements and the loss of flux delivered to the sample due to mismatches of Bragg reflecting crystals that are experiencing different heat loads. At second-generation sources this is typically accomplished with water cooling. The third-generation sources such as the European Sychrotron Radiation Facility (ESRF) in France, the Advanced Photon Source (APS) in the U.S.A., and the Spring-8 synchrotron in Japan) experience greater heat loads and power densities, and cooling with liquid nitrogen is a common solution.

Typical beamlines suitable for magnetic x-ray scattering are composed of a number of components. The mirror (or pair of mirrors) acts as a low-pass filter, eliminating the high-energy component of the white beam; the mirrors are frequently used as focusing elements. They are operated in a grazing incidence geometry such that the angle of incidence is below the critical angle of reflection for the x-ray energies of interest, and these energies are passed with a reflectivity approaching unity. However, for the higher energies, which have a lower critical angle, the angle of incidence will exceed the corresponding critical angle and photons of these wavelengths will be increasingly attenuated.

The high-energy part of the spectrum is particularly troublesome for a magnetic scattering experiment because of the property whereby the monochromator passes higher-order components of the beam. A typical monochromator, e.g., Si(111) or Ge(111), set to pass λ will also pass $\lambda/2$, $\lambda/3\ldots$ (in varying amounts), since the Bragg conditions for these energies will be met simultaneously by the (222), (333)... reflections. These components of the incident beam will produce charge scattering at apparent fractional Bragg positions, when indexed based on a wavelength of λ. Such harmonics may superpose magnetic scattering for commensurate magnetic systems. Typically, this problem is overcome through a combination of one or more of the following techniques. Firstly, the mirror and inherent energy dependence of the incident spectrum reduce the incident intensity at the higher energies. Secondly, the (222) reflection is all but forbidden in the diamond structure, so that after two reflections from a Si(111) or Ge(111) double-bounce monochromator, and (if used) a (111) reflection in an analyzer crystal, it is greatly reduced. Thirdly, the energy resolution of the detector can further discriminate between the λ and λ/n components (though if the harmonic content of the signal is too great, the detector will saturate, and this approach will not work). Fourth, the monochromator may be detuned. This final method is based on the fact that the rocking widths (Darwin curves) of the higher-order reflections are sharper than that of the fundamental. Thus, if one operates the monochromator with the second crystal slightly rotated away from the θ_B, the reflectivity for the higher-order components may be greatly reduced with only a slight loss of intensity of the fundamental.

For resonant scattering experiments, the monochromator needs to be a scanning, fixed exit monochromator. The energy resolution of the monochromator is crucial in determining the size of any resonant enhancement, and, to maximize the enhancement, it should be approximately matched with the width of the resonance, i.e., to the core-hole lifetime of the associated transition. For $3d$ transition metal K edges, this is 1 to 2 eV, and for the rare earth L edges and the actinide M edges it is ≈ 5 eV.

Other elements in the beamline include various defining slits. Those upstream of the monochromator and mirrors determine the fraction of the full beam accepted by the optical elements. This is useful in quantitative polarization analysis experiments, where the vertical and horizontal divergences must be closely controlled. A final set of slits before the sample are used to trim the beam of diffuse scattering arising from imperfections in the optical elements and to define the footprint of the beam on the sample. This is of importance when measuring integrated intensities of a number of Bragg peaks. The last component of the beamline, immediately preceding the sample, is an incident beam monitor that is used to normalize scattered count rates.

Bending magnets, wigglers, and undulators are all planar devices—i.e., the electron orbit is only perturbed in the horizontal plane, and the radiation produced is therefore horizontally polarized in the same plane. The degree of polarization represents a figure of merit for performing polarization analysis experiments. It is largely dependent on the type and stability of the source, although subsequent optics can modify it. For example, for a bending magnet source at the National Synchrotron Light Source (NSLS; Brookhaven National Laboratory, U.S.A.), $P_\zeta = 0.9$. For a wiggler $P_\zeta = 0.99$, and for an undulator source at the ESRF, $P_\zeta = 0.998$. This greater degree of polarization, together with smaller horizontal and vertical divergences at the third-generation sources, make undulator beamlines significantly better for quantitative polarization analysis. The degree of linear polarization can be further improved by choosing a monochromator Bragg reflection close to 45°—or by inserting a second (polarizing) monochromator just upstream of the sample, also with a $\theta_B \approx 45°$. This will further reduce the π component of the incident polarization. This approach has been employed successfully in the study of ferromagnets.

While a linearly polarized beam is often sufficient for magnetic scattering, as discussed above (see Principles of the Method, Theoretical Concepts) the observation of ferromagnetic order via interference between non-resonant magnetic and charge scattering requires circular polarization. There are three main methods for obtaining circularly polarized light in the hard x-ray regime at a synchrotron (see also X-RAY MAGNETIC CIRCULAR DICHROISM).

1. Firstly, although bending magnet radiation is horizontally polarized on-axis, it is circularly polarized out of the plane of the synchrotron. This results from the fact that the electron motion appears elliptical when viewed from above or below the plane, and thus the off-axis radiation is elliptically polarized, with opposite polarization states being

obtained above and below the plane. Techniques based on this fact have been relatively popular, because of the ready access to bending magnet sources. The degree of circular polarization increases rapidly from zero to near unity as the linear, horizontal polarization drops to zero. The disadvantage of this method is that the flux also drops very rapidly, and even at the small viewing angles involved (a few tenths of a milliradian) intensity losses at factors of 5 to 10 must be suffered to gain appreciable circular polarization. In addition, the technique is more sensitive to beam motion, since such motions cause a change in both intensity and polarization. Finally, in order to switch the handedness of the polarization, the whole experiment must be realigned from above the plane to below the plane. This is a cumbersome procedure at best, and one liable to incur systematic errors.

2. A second source of circularly polarized light in the hard x-ray regime is to use purpose-built insertion devices. Planar wigglers and undulators do not produce circularly polarized light when viewed from out of the plane of the orbit, because they are comprised of equal numbers of left and right oscillations and the net helicity is zero. However, it is possible to design periodic magnet structures such that circularly polarized light is produced on-axis. These include crossed undulator pairs (i.e., a vertical undulator followed by a horizontal undulator close enough together for the radiation to remain coherent) and elliptical wigglers for which the electron motion is helical. These devices produce intense beams with high degrees of polarization, combined with the ability, in some cases, to rapidly switch the polarization state. However, they are expensive, and it can be difficult to preserve the circular polarization through the optical elements. To date there have been only a few magnetic scattering beamlines built around such devices in the hard x-ray regime.

3. A third device converts linear polarization into circular polarization, with some cost in intensity. This is a quarter wave plate. In the optical regime, such devices are constituted from birefringent materials—i.e., optically anisotropic materials whose thickness is chosen such that a linearly polarized wave sent in at 45° to the optical axes has its two components phase-shifted by $\pi/2$ relative to one another, thus producing circularly polarized radiation. In the hard x-ray regime, a similar result may be achieved by utilizing the fact that dynamical diffraction is birefringent. In this case, birefringent refers to the fact that the refractive index is different for polarizations parallel or perpendicular to the diffracting planes. Thus, by arranging the Bragg planes to be at 457° with respect to the plane of linear polarization, an x-ray phase plate may be constructed. The most effective devices employ a transmission Bragg geometry, for which the phase shift between the π and σ components is inversely proportional to the angle of offset from the Bragg peak. The resulting transmitted beam is circularly polarized, and the circularly polarized flux is maximized by minimizing absorption losses. X-ray phase plates have been constructed from silicon, germanium, and diamond (Hirano et al., 1993, 1995). For sufficiently thin crystals, transmission factors of 10% to 20% are achievable. Since these devices may be employed on-axis of insertion device beamlines, a relatively high circularly polarized flux is obtained. These devices are particularly well suited for use with undulator radiation, where the collimation of the beam does not compromise the polarization by providing a spread in angles incident on the phase plate. For example, a diamond (111) phase plate 0.77 mm thick, installed at the undulator beamline, ID10, at the ESRF, achieved a degree of circular polarization $P_\eta = 0.96$, with a transmission of 17% (Sutter et al., 1997). The crystal was offset 0.016° from the Bragg peak. Typically, such phase plates are the last optical component before the sample.

Finally, we note that for experiments performed at the M edges of the actinides, the problem of absorption becomes acute, and this must be reflected in the beamline design. The number and total thickness of all x-ray windows should be minimized. At 3.7 keV, the absorption length of Be is $\Lambda = 0.5$ mm, and for Kapton $\Lambda = 0.05$ mm. For comparison, the same numbers at 8 keV are $\Lambda(\text{Be}) = 5$ mm and $\Lambda(\text{Kapton}) = 0.7$ mm. This puts constraints on the design of the experiment. For example, in their work on uranium arsenides, McWhan et al. (1990) utilized a beamline configuration with only one Be window and a further 500 μm of Be and 39 μm of Kapton between the source and detector. In addition, a helium-filled bag was placed over the spectrometer, so that the beam path was either through vacuum or helium.

The Spectrometer. Often, no special equipment is required for the magnetic scattering spectrometer. For the study of antiferromagnets, in particular, a standard x-ray diffraction spectrometer, such as those manufactured by Huber GmbH (*http://www.xhuber.com*) is sufficient. Larger, more robust spectrometers have also been made for magnetic scattering beamlines by, for example, Franke GmbH (X22C, NSLS; *http://www.franke-gmbh.de*) and Micro-Control (ID20, ESRF). These have the advantage that they can support large cryostats, for example. The latter design also allows for diffraction in both σ-incident and π-incident geometries with the same spectrometer.

Two types of detector are commonly used. The NaI photomultiplier tube is frequently the workhorse of such experiments, with a relatively low dark current (<0.1 counts per sec) and a high efficiency in the regimes of interest. However, it has poor energy resolution ($\Delta E/E \sim 0.3$). This is a particular problem for resonant experiments if an analyzer is not used, because of the large background arising from the fluorescent decays associated with the creation of the core hole. In addition, harmonic rejection of such detectors is not always sufficient. In such circumstances, a solid-state detector is used. This has an energy

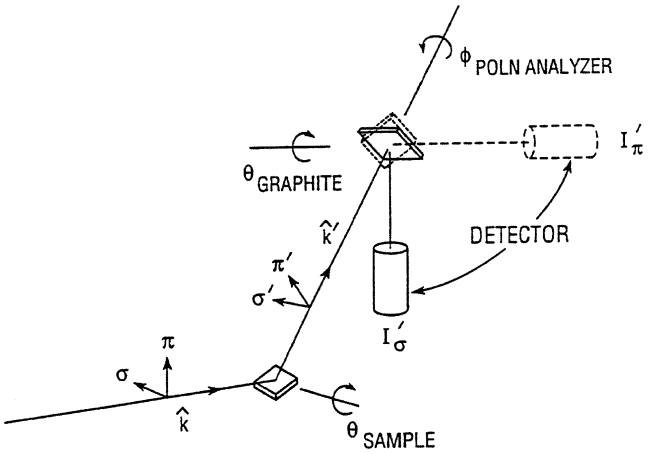


Figure 4. Linear polarization analyzer. When the analyzer, with a Bragg angle of 45°, is set to diffract in the scattering plane (solid lines), the σ' component of the scattered beam is detected and the π' component suppressed. Rotating the analyzer crystal by 90° about the scattered photon wave-vector, $\vec{k}'$, ($\phi_{poln\ analyzer} = 90°$) selects the π' component of the scattered beam (dotted lines). Figure taken from Gibbs et al. (1988).

resolution, $\Delta E/E$, of approximately 0.02, which allows for the removal of the fluorescent background with electronic discrimination. Often the fluorescence is in fact recorded simultaneously with the elastic signal by means of a multichannel analyzer. This then provides an on-line calibration of the energy. In addition, this signal is directly proportional to the intensity incident on the sample and may be used to normalize the scattered count rate. This method is of particular value when measuring the integrated intensity for a number of Bragg peaks with different beam footprints on the sample.

For polarization analysis experiments, a polarimeter must be mounted on the 2θ arm of the spectrometer. A number of ways of measuring the polarization have been developed, including methods to determine the complete polarization state with multiple beam diffraction methods (Shen et al., 1995). However, for analyzing linear polarization, the simplest method is to choose an analyzer crystal with a Bragg angle of $\theta_A = 45°$. This geometry is illustrated in Figure 4, taken from Gibbs et al. (1988). With the analyzer crystal set to diffract in the scattering plane, the linear component perpendicular to the diffraction plane (σ component) is reflected, while the parallel (π) component is suppressed. If the analyzer crystal is rotated by 90° about the scattered beam, the π component is reflected and the σ component is suppressed. Several devices have been built based on such an approach (see, e.g., Gibbs et al., 1989, and Brückel et al., 1996).

In order to perform quantitative polarization analysis, one must measure the integrated intensities of both the σ and π scattered components. How this is achieved depends on the relative divergences in the vertical and horizontal directions (which depend on the sample and beamline optics) and the mosaic of the analyzer crystal. For example, in the early work on holmium (Gibbs et al., 1991), the perpendicular and parallel divergences of the scattered beam were 0.05° and 0.03°, respectively,

Table 2. Selected Crystals and the Energy for Which $\theta_{anal} = 45°$

Crystal	Energy (keV)	Crystal	Energy (keV)
Al(111)	3.75	PG(006)	7.85
LiF(200)	4.35	Ge(333)	8.05
Cu(200)	4.85	Si(333)	8.39
LiF(220)	6.16	Cu(222)	8.41
Cu(220)	6.86	LiF(400)	8.71
Be(11.0)	7.64	Cu(400)	9.70

and a pyrolytic graphite crystal was used as an analyzer ($\theta_A = 45°$ at $E = 7.85$ keV). This has a mosaic of 0.3° and therefore effectively integrates over the scattered beam. In this limit, it is sufficient to rock the sample in both configurations to obtain integrated intensities.

The choice of analyzer crystal is dictated by the need to obtain a reflection close to 45° at the energy of interest, and to match the experimental conditions. A number of crystals are listed in Table 2.

Inevitably, a compromise must be made, typically in the Bragg angle of the crystal. The leakage rate, i.e., the fraction of intensity from the other polarization channel that is reflected, is equal to $\cos^2(2\theta_A)$. For a Bragg angle of $\theta_A = 43°$, the leakage is $\approx 0.5\%$, so such a crystal is relatively tolerant of errors (the actual leakage also depends on the analyzer mosaic and the beam divergence).

Specific Experimental Examples

X-ray magnetic scattering offers a wealth of information on the magnetic order of a system, and the combination of resonant and non-resonant techniques make it widely applicable. It is therefore not possible to adequately represent all the practical aspects of the various implementations of the principles discussed above (see Principles of the Method). In the following, four broad classes of experiments are identified. In each case, a few examples are given to illustrate some of the strengths of the particular technique, together with some of the different experimental realizations.

Non-resonant Scattering: Antiferromagnets. The implementation for non-resonant magnetic scattering from antiferromagnets is relatively straightforward. The experiments are done at fixed incident energy, well away from any absorption edges; the magnetic signal is well separated from charge scattering (even for long period incommensurate structures) and the scattering may be interpreted directly in terms of magnetic densities. The principal difficulty lies in the weakness of the signal and the need to separate it from the background (diffuse) charge scattering. This is simplest to achieve in samples of high crystallographic quality, well matched to the incident beam divergence. For systems with long-range magnetic order, this will maximize the peak count rate. The situation may be further improved by employing an analyzer crystal on the 2θ arm, in front of the detector. This sharpens the longitudinal resolution and discriminates against any broad diffuse background. In addition, the

effective energy resolution is then ~10 eV, and this will eliminate any fluorescence background. For samples of reasonable quality (mosaic <0.1°), the use of a Ge(111) analyzer, for example, will dramatically improve the signal-to-noise ratio of a magnetic peak with long-range order, at a cost of a factor of ~5 in the counting rate. If there is a significant amount of charge scattering present (for example, when studying thin films, reflectivity from the various interfaces may be the dominant contribution to the background), then it may be preferable to employ a polarization analyzer and record only the photons of rotated polarization. Depending on the degree of incident polarization, as well as the magnetic structure, this can provide gains in data quality. For non-resonant scattering, one is free to pick the energy at which the experiment is performed, such that an appropriate analyzer crystal has a Bragg angle of exactly 45°.

As mentioned in the introduction, the weakness of the scattering can also be an advantage, because it ensures that extinction effects, which can prevent reliable determination of intensities, are negligible. This has been exploited in a number of experiments in which precise measurements of the evolution of the order parameter are required.

One such example is the series of experiments performed on dilute antiferromagnets in an applied field. These are believed to be realizations of the random field Ising model and have attracted a great deal of interest, including numerous neutron scattering investigations. These latter, however, are hindered by extinction effects and by an inability to study the long length scale behavior. In addition, the samples must be doped with non-magnetic ions in order to induce the required phenomena and concentration gradients in the samples can then be a serious problem, masking the intrinsic behavior close to the phase transition. This illustrates another strength of x-ray magnetic scattering for these experiments, that is the small penetration depths (~5 μm in this case), which significantly reduce the effects of any concentration gradients. The experiments were performed on MnF_2 and FeF_2, doped with Zn, utilizing a superconducting magnet to apply fields up to 10 T (Hill et al., 1993, 1997). The cryostat constrained the scattering plane to be horizontal, that is a π incident geometry was utilized.

One problem in this work that is worth mentioning (Hill et al., 1993) arose because the crystals order as commensurate antiferromagnets, but have two magnetic ions per unit cell, so that the periodicity is not doubled by the magnetic order. Magnetic scattering can then only be observed at reflections for which the charge structure factor is identically zero. Under these conditions, multiple charge scattering can be a problem (see Problems), and efforts must be taken to reduce such scattering.

Non-resonant experiments to separate $\vec{L}$ and $\vec{S}$ contributions to the ordered moment in an antiferromagnet have also been performed. These require measuring the integrated magnetic intensities, in both scattered polarization channels, for a number of Bragg peaks. Integrated intensities are obtained by rocking the sample and/or the polarization analyzer in an appropriate manner (see, e.g., Langridge et al., 1997). These must be corrected for any differences in the rocking widths of the analyzer in the two configurations. Any charge scattering background may be subtracted off by performing the same measurements above the Nel temperature. The integrated intensity must also be corrected for absorption. The magnetic structure must already be known and Equation 6 is then used to relate the intensities to the quantities $\vec{L}$ and $\vec{S}$. Typically, intensities are not put on an absolute scale because of the difficulty in making extinction corrections for the strong charge Bragg peaks. Thus, often it is the ratio $\vec{L}/\vec{S}$ which is determined, or more precisely, $(\mu_L f_L(\vec{Q}))/(\mu_S f_S(\vec{Q}))$, where $f_{L,S}(\vec{Q})$ are the orbital and spin form factors, and $\mu_{L,S}$ the atomic orbital and spin moments. Such experiments require a high degree of incident polarization in order to simplify the analysis and thus benefit greatly from the third-generation sources. Examples of measurements of this type that have been carried out on antiferromagnets include work on holmium (Gibbs et al., 1988, 1991; C. Sutter, unpublished observation), uranium arsenide (McWhan et al., 1990; Langridge et al., 1997), and NiO (Fernandez et al., 1998).

The non-resonant cross-section, (Equation 6) is energy independent, and this may be exploited to work at very high incident photon energies ($\hbar\omega \geq 80$ keV). While experiments at these energies bring new difficulties, there are obvious advantages, primarily the greater penetration depth (~mm rather than ~μm), which ensures that bulk-like properties are measured. In addition, the larger penetration depth can provide enhancements in the scattered count rate, relative to that at medium x-ray energies. This is because for weak reflections that are limited by absorption, such as magnetic peaks, the integrated scattered intensity is proportional to the scattering volume. Thus, quite large increases can be realized in going from medium to high energies due to this volume enhancement (in contrast to the case for strong charge reflections, for which the intensity is limited by extinction effects). However, care must be taken to match the instrumental resolution (which is typically extremely good at high x-ray energies) with the quality of the sample to take advantage of any increases.

The principles of high-energy x-ray magnetic scattering experiments are identical to those described previously; however, there are some differences in implementation. Firstly, the need for specialized x-ray windows is eliminated and simple Al windows on cryostats may be used (e.g., at 80 keV, 10 mm of Al results in a 40% loss of intensity). This can bring improvements in parameters such as temperature stability. The experiments are typically performed in transmission, and the need to worry about sample surface preparation is eliminated. At high energies, the Bragg angles are correspondingly reduced; for low index reflections, the non-resonant cross-section greatly simplifies. Neglecting terms containing $\sin^2\theta$, then

$$\langle M_{\mathrm{m}} \rangle = i\left(\frac{\hbar\omega}{mc^2}\right)\begin{pmatrix} \sin 2\theta S_2 & 0 \\ 0 & \sin 2\theta S_2 \end{pmatrix} \qquad (25)$$

and the experiments are only sensitive to the spin component perpendicular to the scattering plane. For linear incident polarization, the scattering does not alter the

polarization state. In this limit, one cannot distinguish the magnetic scattering from charge scattering through the use of polarization analysis. On the other hand, one obtains the spin density, without the need for polarization analysis to separate the orbital contribution. High-energy experiments can then be combined, in principle, with neutron diffraction experiments to derive $\vec{L}$ and $\vec{S}$.

A disadvantage of working at such high energies is the limited number of beamlines capable of providing a high flux at the relevant wavelengths. In addition, higher-order contamination can be a significant problem because of the lack of x-ray mirrors, and multiple scattering, if present, is significantly worse at high energies. Nevertheless, successful experiments have been performed (Lippert et al., 1994; Strempfer et al., 1996; 1997; Vigliante et al., 1997; Strempfer et al., 2001).

Non-resonant Scattering: Ferromagnets. Non-resonant magnetic scattering can be used to determine $\vec{L}$ and $\vec{S}$ in ferromagnets. One particular implementation for such an experiment, first proposed by Laundry et al. (1991), utilizes a white, elliptical polarized beam incident on a single crystal and an energy dispersive detector positioned at a scattering angle of 90°. The advantages of this method are that it is relatively insensitive to sample motion upon reversal of the magnetization, that the scattering is performed at the optimum angle of 90°, and that several reflections may be collected at once. The disadvantages are that strong fluorescence lines may superpose some Bragg peaks and that high count rates may saturate the detector, giving rise to dead-time errors. If measurements are made for two different magnetization angles, α (Equation 24), then the $\vec{L}$ and $\vec{S}$ contributions may be separated. In an experiment on $HoFe_2$, Collins et al. (1993) chose $\alpha = 0$ and 90°, as shown in Figure 5. The ratio of the asymmetry ratios then determines the spin-orbital ratio

$$\frac{F_S}{F_L} = \frac{R(\alpha = 90^\circ)}{R(\alpha = 0)} - 1 \tag{26}$$

The measurements of Collins et al. (1993) were in approximate agreement with band-structure calculations. The disadvantage of this method is the need for two easy magnetization axes 90° apart and the requirement that the material be magnetically soft. It should also be mentioned that in cases for which absorption edges are accessible, the technique of XMCD (X-RAY MAGNETIC CIRCULAR DICHROISM) can provide an alternative to such approaches and can determine $\langle L_z \rangle$ and $\langle S_z \rangle$ independently, in an element-specific manner, in ferromagnets.

Resonant Scattering: Antiferromagnets. The principal difference between a resonant and a non-resonant magnetic scattering experiment is the need for a scanning monochromator. The incident energy resolution should match the width of the resonance to attain the greatest enhancement, or exceed it, if information on lifetimes is required. In order to compare the experimentally measured resonances with the theoretical predictions of Equation 15, integrated magnetic intensities must be recorded at each energy and corrected for absorption. While calculations for the energy dependence of the absorption exist (Cromer and Lieberman, 1981), these are not reliable in the vicinity of the edge, where they are most important. Ideally, to obtain the appropriate corrections, the absorption would be measured directly through a thin film of the sample material. By reference to calculations made far from the edge, such data may be brought to an absolute scale and $\mu(E)$ obtained. However, this is not always possible, and various schemes have been attempted to circumvent this problem. One is to obtain μ from measurements of the fluorescence intensity, I_f using the relation

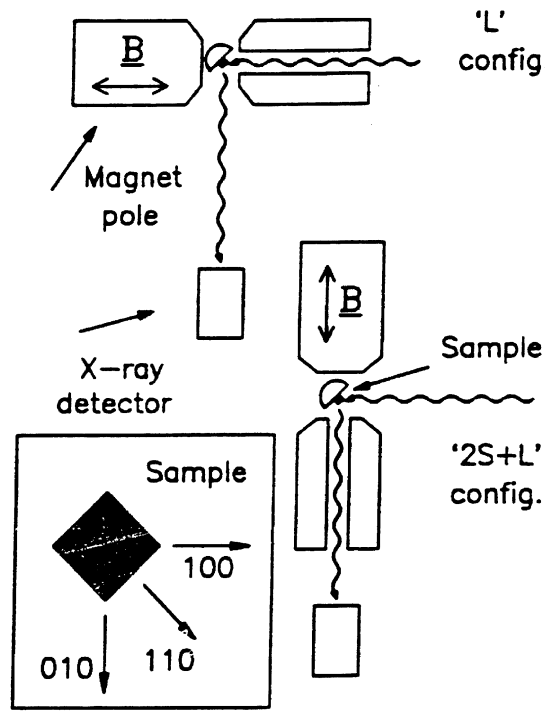


Figure 5. Schematic of the diffraction geometry used in Collins et al. (1993), for scattering from ferrimagnetic $HoFe_2$. In the top configuration, with the sample magnetized along the incident beam direction, the scattering is only sensitive to the orbital moment density. In the second configuration (bottom), the sample is magnetized perpendicular to the incident beam and contributions from both $\vec{L}$ and $\vec{S}$ are observed. The sample is a single crystal and the incident beam is "white" (unmonochromatized).

$$I_f(E) = C \frac{\mu_R(E)}{\mu_S(E_f) + \mu_S(E) \dfrac{\sin\psi}{\sin\phi}} \tag{27}$$

where $\mu_S(E)$ is the absorption of the sample at photon energy E, E_f is the fluorescence energy, μ_R is the absorption of the resonance ion and ψ and ϕ are the angles of incidence and fluorescent beams, respectively. The fluorescence is measured as a function of incident energy by placing an open detector at 90° to the scattering plane (to minimize charge scattering) and with the discriminator set to select the fluorescent energy. The constant, C, may

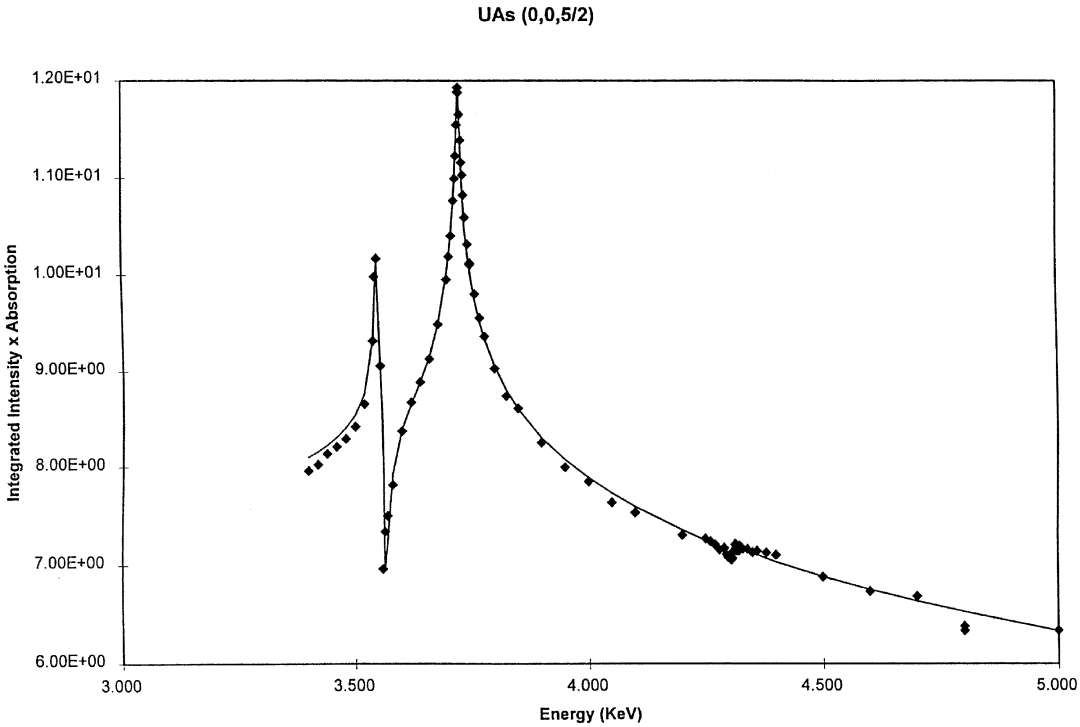


Figure 6. The energy dependence of the magnetic scattering at the (0,0,2.5) antiferromagnetic Bragg peak in uranium arsenide as the incident energy is scanned through the U M_5, M_4, and M_3 edges. Each data point represents an integrated intensity and is corrected for absorption. The solid line is a fit to three dipole oscillators (McWhan, 1998).

be obtained by normalizing the data to the calculated values far from the edges (see, e.g., McWhan et al., 1990; McWhan, 1998). This method has its limitations, and other approaches have also been tried (e.g., Tang et al., 1992; Sanyal et al., 1994). For a Bragg peak along the surface normal, and a flat sample, the absorption correction corresponds to simply multiplying the integrated intensity by μ. The angular factors of the cross-section, Equation 15, will also contribute some energy dependence, together with the Lorentz factor $1/\sin 2\theta$.

Figure 6 shows the absorption-corrected energy dependence of the (0,0,2.5) magnetic Bragg peak in uranium arsenide on scanning the incident energy through the U M_5, M_4, and M_3 edges (McWhan et al., 1990). The solid line is a fit to three coherent dipole resonances, together with the appropriate angular factors. These data were taken with an incident energy resolution of 1 eV, and the authors were thus able to extract branching ratios and lifetimes for the various resonances, though these are sensitive to the absorption correction and are therefore only semiquantitative. The overall energy dependence illustrated in Figure 6 is also sensitive to the exchange splitting present in the excited state, and these data led McWhan and co-workers to conclude that it was smaller than 0.1 eV.

The position of the resonance is a function of atomic species and valence state, and thus this technique also provides a means for investigating magnetic order in mixed valent compounds. For example, McWhan et al. (1993) studied TmSe. In this material both Tm^{2+} and Tm^{3+} configurations exist, each of which may carry a magnetic moment. The resonance energies of the two ions differ by 7 eV, as determined by x-ray absorption spectroscopy (XAS) which is comparable to the energy width of the resonances. Thus, by tuning the incident photon energy to one or other of the dipole transitions of the two ionic configurations, the dominant contribution to the cross-section may be varied from one configuration to the other, and the magnetic order of the two valences may be studied separately. The (003) magnetic reflection showed two peaks as a function of energy, demonstrating that both valence states exhibit long-range magnetic order.

Both non-resonant and resonant magnetic x-ray scattering may be used to determine moment directions in ordered structures, and this has been demonstrated for a number of antiferromagnets. To do this, integrated intensities of a number of magnetic Bragg peaks are measured, and then the Q dependence of the intensities is fit to expressions such as Equation 9 and Equation 15. Detlefs et al. (1996, 1997) demonstrated this approach in studies of RNi_2B_2C, with R = Gd, Nd, and Sm. For R=Gd and Sm, the structure could not be solved using neutron scattering techniques because of the high neutron absorption of the Gd and Sm, together with the B. The different polarization dependences of the non-resonant, resonant dipole, and resonant quadrupole scattering were exploited to solve the structures from a relatively small data set. Before a direct comparison with the cross-section can be made, a variety of other angular factors must be included. For a Bragg peak at an angle α away from the surface

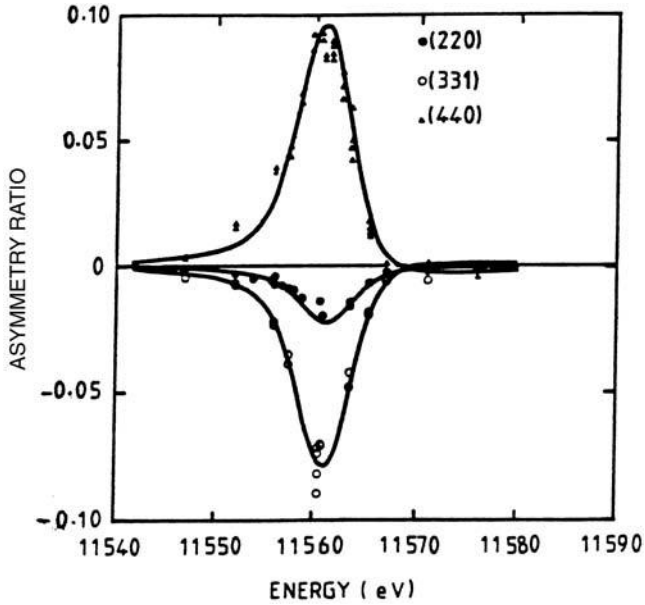


Figure 7. The energy dependence of the asymmetry ratios for the (220), (331) and (440) Bragg peaks of the ferromagnet, $CoPt_3$. Data show a large resonant enhancement at the Pt L_3 edge and are well fit by a single $2p \rightarrow 5d$ dipole transition. Taken from de Bergevin et al. (1992).

normal and scattering from a flat fully illuminated sample, then

$$I \propto \left| \sum_n e^{i\vec{Q}\cdot\vec{r}_n} f_n^{res} \right|^2 \frac{\sin(\theta+\alpha)\sin(\theta-\alpha)}{2\mu\sin\theta\,\cos\alpha\,\sin 2\theta} \qquad (28)$$

where the Lorentz factor has also been included. A more detailed study has been carried out on pure Nd (D. Watson et al., 1998).

These measurements are made simpler by the lack of a magnetic form factor for resonant scattering, and the fact that the $F^{(L)}$ of Equation 15 are independent of the scattering angle. Apart from an overall scale factor, then, the only unknowns are the spin components, z_1, z_2, and z_3. A coordinate transformation, which depends on θ, is then required to rotate into the coordinate system of the sample, obtaining the moment directions relative to the crystallographic axes.

Other experiments on antiferromagnets have studied thin films, ultra-small samples, multilayers, and compounds with more than one magnetic species. These take advantage of the small penetration depth, high-resolution, and element specificity of the technique. However, in implementation, they do not involve concepts different from those already discussed.

Resonant Scattering: Ferromagnets. The first experimental observation of magnetic scattering with x-rays from ferro- and ferrimagnets is credited to de Bergevin and Brunel (1981a,b), who utilized Cu K_α radiation to study powdered samples of zinc-substituted magnetite, $Fe_{2-x}\ Zn_xO_4$, and Fe. Interference was induced through the proximity to the iron K edge. Studies of interfacial magnetism were first performed in ferromagnetic multilayers, namely Gd-Y (Vettier et al., 1986), utilizing non-resonant magnetic-charge interference scattering. The interference was induced by tuning through the Gd L_2 and L_3 edges.

More recently, de Bergevin et al. (1992) carried out a study of $CoPt_3$ utilizing resonant magnetic scattering at the Pt L_3 edge. The asymmetry ratio was determined for three Bragg peaks, and as a function of energy (Fig. 7). The data are well fit by a single dipole ($2p \leftrightarrow 5d$) resonance, with a peak scattering amplitude of $0.8r_0$.

Surface Magnetic Scattering. The large resonant enhancements in the cross-section allow the measurement of the very weak signals associated with the surface magnetization density profile. To a first approximation, the experimental techniques for such work are identical to those developed for surface x-ray scattering (SURFACE X-RAY DIFFRACTION), with the additional complication of a strong energy dependence to the scattering.

Briefly, for a semi-infinite crystal, the diffraction pattern consists of rods of scattering, known as truncation rods, which pass through the Bragg points and which are parallel to the surface normal (Fig. 8A). Characterizing the intensity along such rods provides information on the decay of the charge density near the surface. Similarly, magnetic truncation rods are present when a magnetic lattice is terminated, and the decay of the magnetic intensity is related to the profile of that magnetic interface. For a step function interface, the diffraction pattern is the convolution of a reciprocal lattice of delta function Bragg peaks with the Fourier transform of the step function, $1/q$. This gives a $1/(\Delta q_z)^3$ broadening of each Bragg point, where Δq_z is the deviation from the Bragg peak, along the surface normal. If a broader interface is present, the Fourier transform will be sharper than $1/q$ and the intensity fall-off around each Bragg point will be faster (Robinson, 1991). Thus, measuring the intensity fall-off measures the Fourier transform of the magnetic interface profile, which may in principle be different from the chemical interface.

For an antiferromagnet, there is both a magnetic contribution to the charge rods, as well as pure magnetic rods, as shown in Figure 8A, taken from G. Watson et al. (1996b; 2000). In addition to scans around Bragg points, the surface magnetic behavior may also be probed explicitly by employing the grazing incidence scattering geometry shown in Figure 8B. α_j and α_s are the angles of the incident and scattered beams with respect to the surface, respectively. In the figure, most of the momentum transfer is in the plane of the crystal $\vec{q}_\parallel$, with a small component along the surface normal, $\vec{q}_z$. For example, such a situation occurs for scans along the (0,1,L) magnetic rod, with L small. In such a geometry, it is possible to take advantage of the well known enhancement in the scattering, when $\alpha_i = \alpha_s = \alpha_c$, the critical angle. Such techniques have been demonstrated in the antiferromagnet UO_2, working at the U M_4 edge (G. Watson et al., 1996b; 2000). The energy dependence of scans taken along a magnetic rod (0,1,L) are shown in Figure 9 (G. Watson et al., 1996b; 2000). These data illustrate that the critical angle, for which the intensity is a maximum, is also a function of the incident energy, since $\alpha_c^2 \sim (Z + f'')\rho\lambda$, where Z is the total charge, ρ is the average atomic density and, f'' the

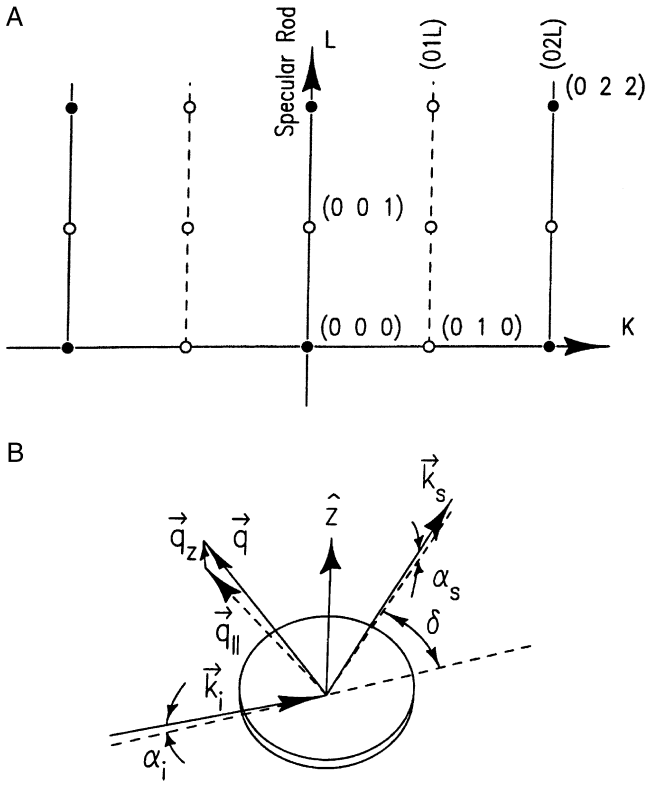


Figure 8. (**A**) Schematic of reciprocal space for a semiinfinite antiferromagnet, illustrated for UO_2. In addition to the charge (closed circles) and antiferromagnetic (open circles) bulk Bragg peaks, there are rods of scattering connecting the Bragg points and running parallel to the surface (vertical lines). The dotted lines are pure magnetic rods of scattering. The specular rod (00L) corresponds to scattering in which there is no in-plane momentum transfer, and the angle of incidence equals the angle of reflection. (**B**) Scattering geometry for surface magnetic scattering. α_i and α_S are the angles of incidence and exit, normal to the surface, and δ is the in-plane scattering angle. In this diagram, the momentum transfer $\vec{q}$ is largely in-plane, $\vec{q}_{\|}$, with a small component normal to the surface, $\vec{q}_z$. Figure taken from G. Watson et al. (1996b).

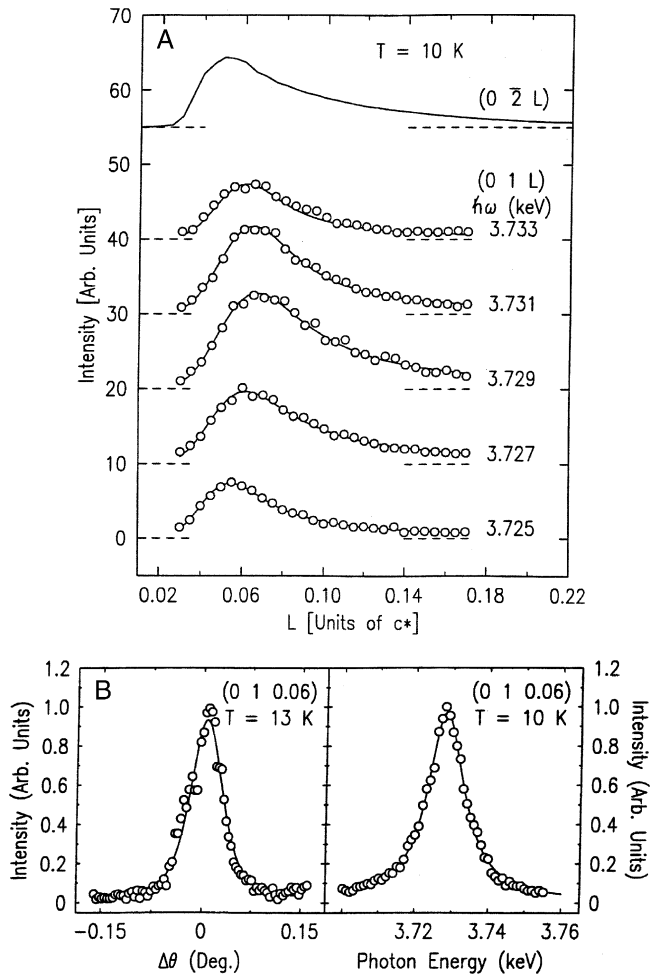


Figure 9. (**A**) Energy dependence of scans along the (0,1,L) magnetic rod in UO_2. The data exhibit an enhancement at an energy dependent critical angle, α_c. For comparison, the same scan is shown for along the (0,2,L) charge rod, solid line. (**B**) Transverse scan across the (0,1,L) magnetic rod. (**C**) Intensity of the magnetic rod as a function of incident photon energy. Figure taken from G. Watson et al. (1996b).

real part of $f_n^{\rm res}$ of Equation 4. Measurements of α_c thus provide an independent measure of f'. Scans across the magnetic rod provide information on the disorder of the magnetic layer normals. In UO_2, such scans were found to be identical to those of the charge rods (Fig. 9). The scattered intensity may be described by

$$\frac{I(k_i,k_s)}{I_0} = \frac{A}{A_0}\sin\alpha_i\,\sin\alpha_s|T(\alpha_i)|^2|T(\alpha_s)|^2 \times \left|\sum_{n=0}^{\infty} r_0 f_n^{\rm res}\rho_n^{\rm M}(q_{\|})e^{iq_z z_n}\right|^2 \quad (29)$$

where $T(\alpha)$ is the Fresnel transmission coefficient, A/A_0 the ratio of the illuminated area to the cross-sectional area of the beam, I_0 the incident flux, $r_0 f_n^{\rm res}$ the magnetic scattering amplitude, and $\rho_n^{\rm M}$ the magnetic density in the nth layer with wave-vector $q_{\|}$. The fits to this form in Figure 9A assume a perfectly sharp magnetic interface, and describe the data well. This technique can be used to determine the temperature and depth dependence of the magnetic order parameter profile. By varying L, the effective penetration depth varies. For example, in UO_2, for $L = 0.075, 0.15$, and 1, the corresponding penetration depth was 50, 120, and 850Å, respectively.

Similar studies of surface magnetism have been carried out in ferromagnets, for which the asymmetry ratio is measured along the truncation rod. This was first done in the soft x-ray regime at the Fe $L_{2,3}$ edges (Kao et al., 1990) and Co $L_{2,3}$ edges (Kao et al., 1994) in which the specular magnetic reflectivity was measured (i.e., (0,0,L) scans). This describes the magnetic order normal to the surface. Recent studies of Co multilayers have focused on the off-specular diffuse scattering, which describes the roughness of the interface (Mackay et al., 1996). In this work, the magnetic interfaces were found to be less rough than the structural interfaces. Finally, working at the Pt L_3 edge, Ferrer et al. (1996) have measured complete magnetic truncation rods in ferromagnetic $Co_3Pt(111)$ and Co/Pt(111) ultra-thin films (Ferrer et al., 1997). The advantage of resonant mag-

netic scattering in such studies is the technique's element specificity, allowing multicomponent systems to be resolved and buried interfaces to be studied, in contrast to scanning probe techniques. In addition, it is essential to characterize the structural interface(s) in order to fully understand the magnetic interface. X-ray scattering techniques are able to characterize both the structural and magnetic interfaces with the same depth sensitivity and resolution, in the same experiment.

DATA ANALYSIS AND INITIAL INTERPRETATION

Quantities such as ordering wave vector, ordered moment, and moment direction may be obtained from the position and integrated intensity of magnetic reflections. For more detailed information on the magnetic correlations, e.g., to extract forms for the moment-moment correlation function, as well as parameters such as correlation lengths and critical exponents, then it is necessary to fit the lineshapes of the magnetic peaks. The observed intensity, $I\vec{Q}$, is given by the convolution of the cross-section, $d\sigma(\vec{Q})/d$, and the instrumental resolution $R(\vec{q})$

$$I(\vec{Q}) = \int R(\vec{q}) \frac{d\sigma(\vec{Q}-\vec{q})}{d\Omega} d\vec{q} \qquad (30)$$

A common problem in the analysis of x-ray data is the accurate determination of $R(\vec{q})$. Ideally, a perfect δ-function scatterer (e.g., a silicon single crystal) is placed in the sample position and the resolution is measured at the relevant momentum transfers. The problem is that it is, in general, not possible to find a suitable test sample. A number of approaches are therefore adopted. For studies of critical fluctuations in the vicinity of a phase transition, it is common to take the scattering below T_N to be long-range ordered, i.e., resolution limited. Fits to scans through this Bragg peak can then be used to parameterize the resolution function and in fitting the experimental data, following Equation 30. In other cases, nearby Bragg peaks can be used, again as a lower bound on the resolution function (the intrinsic resolution is equal to or better than the apparent widths of the Bragg peaks). The longitudinal resolution width (i.e., along the momentum transfer) is controlled by the monochromator and analyzer, if present, or the angular acceptance of the detector, if not. For a given monochromator-analyzer combination, the longitudinal resolution is a minimum at the nondispersive condition, i.e., when the scattering vectors of the monochromator, sample, and analyzer are equal, and in alternating directions. For a beamline with a double crystal monochromator and analyzer, the in-plane resolution functions are often Lorentzian squared functions. The out-of-plane resolution is determined by collimating slits and can be approximated by a triangular function. In such a case, with the assumption that the resolution function is separable

$$R(\vec{q}) = \left(\frac{1}{w_x^2 + q_x^2}\right)^2 \left(\frac{1}{w_y^2 + q_y^2}\right)^2 \left(1 - \left|\frac{q_z}{w_z}\right|\right) \qquad (31)$$

where the x and y directions are transverse and longitudinal to the momentum transfer, and q_z is perpendicular to the scattering plane. For example, in a study of the critical fluctuations on holmium, Thurston et al. (1994) employed a Ge(111) analyzer and double bounce Ge(111) monochromator, and obtained resolution half-widths of 2.9×10^{-4}, 4.5×10^{-4}, and 4.3×10^{-3}Å^{-} in the transverse, longitudinal, and out-of-plane directions, respectively, at $Q = 1.92$Å^{-1}.

Approximating the out-of-plane resolution with a triangular function, as in Equation 31, has the advantage that the integration in the q_z direction may be performed analytically, for certain cross-sections, reducing the fitting of the data to a 2-D numerical integration and lessening the fitting time. Note that it is crucial that a full 3-D convolution procedure be carried out in fitting the data in order to obtain meaningful correlation lengths, even if the data themselves are only obtained in one direction.

Strictly, the cross-section is proportional to the dynamic structure factor, $S(\vec{Q}, \omega)$, where ω is the energy transfer of the scattering process. However, typically the energy resolution of an x-ray experiment is relatively broad (e.g., 1 to 10 eV) compared to the relevant magnetic energy scales, and all the magnetic fluctuations are integrated over, reducing $S(\vec{Q}, \omega)$ to the static structure factor, $S(\vec{Q})$. This is the quasi-elastic approximation. In turn, this structure factor may be expressed as

$$\frac{d\sigma}{d\Omega} = S(\vec{Q}) \propto \langle f(\vec{Q}) f(-\vec{Q}, t=0) \rangle = \langle |f(\vec{Q})|^2 \rangle \qquad (32)$$

i.e., the instantaneous correlation function. For non-resonant magnetic scattering, this may be directly related to a particular moment-moment correlation function for a given scattering geometry. In the MnF_2 example (see Principles of the Method), for which $\vec{L} = 0$, then for σ incident polarization, $S(\vec{Q}) = \langle \vec{S}_\perp(\vec{Q}) \vec{S}_\perp(-\vec{Q}) \rangle$, where $\vec{S}_\perp$ is the component of the spin perpendicular to the scattering plane, and we have ignored terms of order $\sin^2\theta$ in the scattering amplitude. For resonant scattering, the situation is more complex, (see e.g., Luo et al., 1993; Hill and McMorrow, 1996), though for first-order dipole scattering, and σ incident polarization, $f(\vec{Q}) \propto \hat{z}(\vec{Q}) \cdot \hat{k}_f$, the projection of the moment direction along the scattered photon direction.

The form of $S(\vec{Q})$ depends on the system under study. Thermal critical fluctuations, for example, are expected to give a Lorentzian-like behavior

$$S(\vec{Q}) \sim \frac{k_B T \chi(T)}{1 + (Q - Q_0)^2 / \kappa^2} \qquad (33)$$

where Q_0 is the Bragg position, and $\kappa = 1/\xi$ is the inverse correlation length. Fits to the temperature dependence of the parameters, $\chi = \chi_0((T - T_N)/T_N)^{-\gamma}$ and $\xi = \xi_0((T - T_N)/T_N)^{-\nu}$ allow for the determination of the critical exponents γ and ν. The order parameter exponent, β, is determined from fits to the temperature dependence of the integrated intensity below T_N, i.e., $I = I_0((T_N - T)/T_N)^{2\beta}$. Conversely, static random disorder may give rise to Lorentzian-squared lineshapes. One of the strengths of high-resolution x-ray scattering is the ability to distinguish between such behavior, even at relatively long length scales, e.g., close to a phase transition. This ability

has revealed new phenomena; for example the critical fluctuations immediately above a magnetic phase transition have been shown to exhibit two length scales (Thurston et al., 1993, 1994).

SAMPLE PREPARATION

The range of experimental conditions and techniques described in this unit make it very difficult to make any useful general statements about what is required by way of sample preparation. However, perhaps the most important point to emphasize is that the better the sample is characterized before any synchrotron beamtime, the higher the chances of success. Particularly important properties to be characterized include the sample orientation and crystallographic quality, and the magnetic properties, through, for example bulk magnetization measurements to obtain transition temperatures, magnetic anisotropies etc. (see MAGNETROMETRY, THERMOMAGNETIC ANALYSIS and TECHNIQUES TO MEASURE MAGNETIC DOMAIN STRUCTURES). For reflection experiments, surface (and near-surface) quality can be all-important. Careful polishing of the appropriate surface (ideally one with a surface normal parallel to the direction of interest in reciprocal space) can provide large rewards in terms of signal to noise. However, the appropriate polishing method is extremely sample dependent, with electro-polishing or the use of diamond paste (with perhaps a 1 mm grit size) among the most common techniques used (see SAMPLE PREPARATION FOR METALLOGRAPHY).

PROBLEMS

Other Sources of Weak Scattering

One of the biggest problems in x-ray magnetic scattering experiments is distinguishing the weak magnetic scattering from other possible sources of scattering that may produce peaks coincident with it. In general, there is no single test that is always definitive, and it is necessary to use a combination of tests, together with prior knowledge of the system. Possible sources of scattering are discussed under the following subheadings.

Multiple Scattering. This is the scattering that results from two (or more) successive Thomson scattering events. It occurs when a second reciprocal lattice point, $\vec{q}_2$, intercepts the Ewald sphere, in addition to the point of interest, $\vec{q}_1$. Then double scattering of the form $\vec{q}_2 + \vec{q}_3$ occurs, where $\vec{q}_3 = \vec{q}_1 - \vec{q}_2$ is another allowed reciprocal lattice vector. It is a particular problem in systems for which the magnetic unit cell is the same size as the chemical unit cell. At synchrotrons, multiple scattering count rates can exceed 10^4 photons per sec. The problem can be worse in crystals of lower crystallographic quality, or in lower-resolution configurations for which the requirements for making two successive Bragg reflections are less stringent. It is also aggravated at higher incident x-ray energies, because the radius of the Ewald sphere is larger and the probability of more than one reciprocal lattice point intercepting it becomes correspondingly larger. There are two possible means of eliminating multiple scattering. First, by rotating the Ewald sphere around the momentum transfer, $\vec{q}_1$, thus moving the $\vec{q}_2$ lattice point off the Bragg condition, while leaving the scattering vector unchanged. Typically, only small rotations are required and the scattering geometry relative to the moment directions need not be changed significantly. The second alternative is to vary the incident photon energy. This alters the radius of the Ewald sphere, sweeping it through reciprocal space, thus moving off the $\vec{q}_2$ lattice point. Such a tactic is not an option for resonant experiments.

Unless there is a structural phase transition associated with the magnetic ordering, the multiple scattering is likely to be temperature independent.

Higher Harmonic Contamination of the Beam. This was already discussed in the section on beamlines. It can produce scattering at non-integer, but commensurate reciprocal, lattice points, but may be readily identified by the energy of the scattered photons, even with a NaI detector. It is typically temperature independent.

Magnetoelastic Scattering. This is charge scattering that arises from a lattice modulation induced by the magnetic order. A magnetoelastic coupling can occur when, for example, there is a distance dependence to the exchange interaction, though other mechanisms are possible (see e.g., McMorrow et al., 1999). A magnetization density wave can then induce a displacement wave in the lattice with an amplitude which can depend only on the relative change in magnitude of the magnetization wave. In such a situation, it is straightforward to show (e.g., Lovesey and Collins, 1996) that for a transverse magnetization density wave, a distortion is induced at twice the wave-vector (i.e., half the wavelength) with an amplitude that varies as the magnetization squared. The diffracted intensity then varies as the moment to the fourth power. Thus, although the scatterings from both modulations will go to zero at the same temperature, they will have very different temperature dependences, and may be distinguished in this manner. In addition, of course, the charge scattering does not rotate the polarization state of the photon.

In general, more complicated modulations are possible, and the magnetoelastic scattering need not occur at twice the wave-vector of the magnetic scattering. Such is the case for the spin structure in holmium, for example (Gibbs et al., 1988, 1991). Below $T_N = 132$ K, holmium forms a basal plane spiral. At low temperatures, the magnetic wave-vector, τ_m, exhibits a tendency to lock into rational fractions of the c axis reciprocal lattice vector, $c* = 2\pi/c$. This is a result of a bunching (in pairs) of the moments in each successive ab plane about the six easy axes of magnetization, 60° apart. These pairs are known as doublets. If only a single spin (layer) is associated with a given easy axis, then this is known as a singlet, or spin slip. Ordered arrays of such spin slips give rise to the magnetic wave-vectors observed. In addition, at each spin slip, the phase of the magnetic modulation takes a jump in going from layer to layer and thus the magneto-elastic coupling is changed. This can give rise to a lattice distortion, with a new periodicity, τ_l, where $\tau_l = 12\tau_m - 2$ (Bohr et al., 1986). In general then, the magnetoelastic scattering can appear close to the magnetic scattering, and can even be

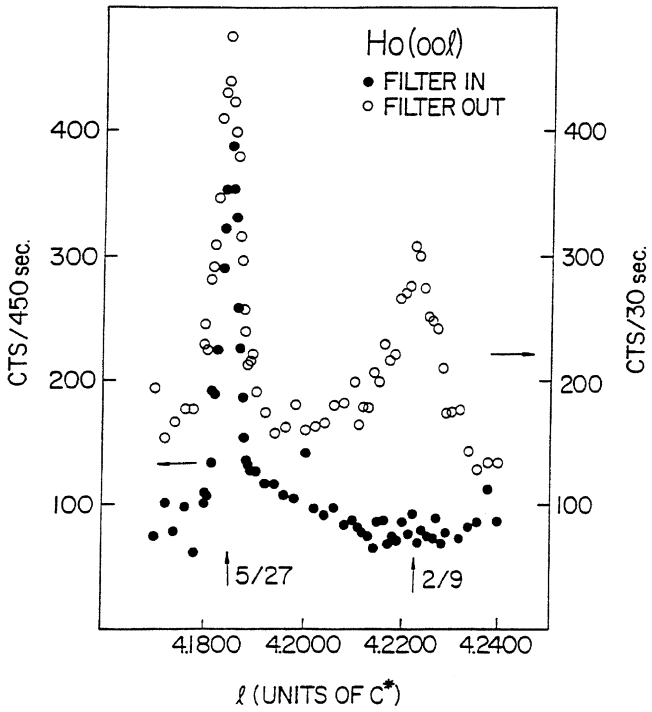


Figure 10. Magnetic and charge scattering in holmium. Data were taken at 7847 eV (off resonance). The open circles are recorded with no polarization analyzer and exhibit two peaks, at $\tau_m = 5/27$ and $\tau_I = 2/9$. With the polarization analyzer in, and set to record the $\sigma \rightarrow \pi$ scattering processes, only the τ_m peak is observed (closed circles). This is consistent with τ_m being magnetic scattering, and τ_I arising from charge scattering due to a lattice distortion induced by the magnetic order (Gibbs et al., 1985).

coincident with it (e.g., if $\tau_m = 2/11$). Polarization analysis is therefore required to distinguish the scattering arising from these two modulations, as illustrated in Figure 10, from Gibbs et al. (1985). The magnetic scattering observed at 5/27 is observed both with and without the analyzer. Conversely, the 2/9 peak is not seen in the $\sigma \rightarrow \pi$ channel and therefore corresponds to charge scattering. This peak is broader than the magnetic peak because of disorder in the spin slip structure, which is not reflected in the magnetic spiral.

Finally, there may be other couplings induced by the magnetic order that give rise to charge scattering. For example, in non-resonant magnetic scattering studies of $Fe_{0.75}Co_{0.25}TiO_3$, Harris et al. (1997) observed a quadratic increase in the scattering at the antiferromagnetic Bragg position (1,1,2.5), as a function of applied field, H. This effect was not seen in neutron data, and was attributed to a magneto-elastic distortion of the same wave-vector as the antiferromagnetic order. This can arise from terms in the free energy of the form $\rho_s M_s M$, where ρ_s is the staggered charge density, M_s the staggered magnetization, and M the uniform magnetization. If the cost of the distortion due to the elastic energy, $1/2\rho_s^2$, is included and the total minimized with respect to ρ_s, then it follows that $\rho_s \sim M_s M$. The intensity of the scattering from the charge density is then $I_{charge} \sim (M_s M)^2$ and that from the magnetic density, $I_{mag} \sim M_s^2$. Therefore, in this case, the two intensities have the same temperature dependence. However, in an applied field, the uniform magnetization will grow linearly (for small fields) and so the charge scattering will increase as H^2 (note that for $H = 0$, the charge scattering should go to zero). In addition, polarization analysis should be able to distinguish between scattering mechanisms in this case. The magnetoelastic scattering was not seen in the neutron diffraction experiments because of the weakness of the coupling between the neutrons and the charge density waves, relative to the strength of the magnetic interaction.

Near-Surface Effects

With the exception of high energy non-resonant scattering, all x-ray magnetic scattering experiments are performed in a reflection geometry, and penetration depths vary from perhaps ~10 μm in some non-resonant experiments to ~1000 Å, or less, in some resonant cases. While this surface sensitivity can be a boon, it can also cause varying degrees of complications in studying "bulk-like" behavior. The near-surface region can differ from the bulk in strain field densities, stochiometry, and different lattice connectivities (due to surface roughness), in addition to the intrinsic surface effects related to the termination of the lattice. Such effects can in turn alter the magnetic behavior of the near-surface region. This can be relatively benign, e.g., producing a single-domain state in cases where there would otherwise be a random distribution amongst energetically equivalent wave-vectors (e.g., Hill et al., 1995a; D. Watson et al., 1996), or more severe, introducing entirely new phenomena, such as the second length scale in magnetic critical fluctuations (e.g., Thurston et al., 1993, 1994; G. Watson et al., 1996a) or altering the observed magnetic state entirely (e.g., Hill et al., 1993). To date, little systematic work has been performed to assess the effect of surface preparation on "bulk" magnetic behavior and there is no general prescription for the correct surface preparation for a given x-ray experiment: It is something that must be approached on a case-by-case basis, and is beyond the scope of this unit. The purpose of this section is simply to emphasize the problem and to encourage careful characterization of the near-surface region for each experiment.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Division of Materials Science under contract no. DE-AC02-98CH10886.

LITERATURE CITED

Bartolomé, F., Tonnerre, J. M., Sève, L., Raoux, D., Chaboy, J., García, L. M., Krisch, M., and Kao, C.-C. 1997. Identification of quadrupolar excitation channels at the L_3 edge of rare-earth compounds. *Phys. Rev. Lett.* 79:3775–3778.

Blume, M. 1985. Magnetic scattering of x-rays. *J. Appl. Phys.* 57:3615–3618.

Blume, M. 1994. Magnetic effects in anomalous dispersion. *In* Resonant Anomalous X-ray Scattering. (G. Materlik, C.

Sparks, and K. Fischer, eds.). pp. 513–528. Elsevier Science Publishing, New York.

Blume, M. and Gibbs, D. 1988. Polarization dependence of magnetic x-ray scattering. *Phys. Rev. B* 37:1779–1789.

Bohr, J., Gibbs, D., Moncton, D. E., and D'Amico, K. L. 1986. Spin slips and lattice modulations in holmium: A magnetic x-ray scattering study. *Physica A* 140:349–358.

Brown, G. and Lavendar, W. 1991. Synchrotron radiation spectra. *In* Handbook of Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.). pp. 37–61. North Holland Publishing, New York.

Brückel, Th., Lippert, M., Köhler, Th., Schneider, J. R., Prandl, W., Rilling, V., and Schilling, M. 1996. The non-resonant magnetic x-ray scattering cross-section of MnF_2. Part I. Medium x-ray energies from 5 to 12 keV. *Acta Crystallogr.* A52:427–437.

Brunel, M. and de Bergevin, F. 1991. Magnetic scattering. *In* Handbook of Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.). pp. 535–564. Elsevier Science Publishing, New York.

Carra, P. and Thole, B. T. 1994. Anisotropic x-ray anomalous diffraction and forbidden reflections. *Rev. Mod. Phys.* 66:1509–1515.

Carra, P., Thole, B. T., Altarelli, M., and Wang, X. 1993. X-ray circular dichroism and local magnetic fields. *Phys. Rev. Lett.* 70:694–697.

Collins, S. P., Laundy, D., and Guo, G. H. 1993. Spin and orbital magnetic x-ray diffraction in $HoFe_2$. *J. Phys. Condens. Matter* 5:L637–642.

Cromer, D. T. and Liberman, D. A. 1981. Anomalous dispersion calculations near to and on the long-wavelength side of an absorption edge. *Acta Cryst.* A37:267–68.

de Bergevin, F. and Brunel, M. 1981a. Diffraction of x-rays by magnetic materials. I. General formulae and measurements on ferro- and ferrimagnetic compounds. *Acta Crystallogr.* A37:314–324.

de Bergevin, F. and Brunel, M. 1981b. Diffraction by magnetic materials. II. Measurements on antiferromagnetic Fe_2O_3. *Acta Crystallogr.* A37:324–331.

de Bergevin, F., Brunel, M., Galéra, R. M., Vettier, C., Elkïm, E., Bessière, M., and Lefèbvre, S. 1992. X-ray resonant scattering in the ferromagnet, CoPt. *Phys. Rev. B* 46:10772–10776.

Detlefs, C., Goldman, A. I., Stassis, C., Canfield, P. C., Cho, B. K., Hill, J. P., and Gibbs, D. 1996. Magnetic structure of $GdNi_2B_2C$ by resonant and non-resonant x-ray scattering. *Phys. Rev. B* 53:6355–6361.

Detlefs, C., Islam, A. H. M. Z., Goldman, A. I., Stassis, C., Canfield, P. C., Cho, B. K., Hill, J. P., and Gibbs, D. 1997. Determination of magnetic moment directions using x-ray resonant exchange scattering. *Phys. Rev. B* 55:R680–683.

Everitt, B. A., Salamon, M. B., Park, B. J., Flynn, C. P., Thurston, T. R., and Gibbs, D. 1995. X-ray magnetic scattering from nonmagnetic Lu in a Dy/Lu alloy. *Phys. Rev. Lett.* 75:3182–3185.

Fernandez, V., Vettier, C., de Bergevin, F., Giles, C., and Neubeck, W. 1998. Observation of orbital moment in NiO. *Phys. Rev.* B57:7870–7876.

Ferrer, S., Fajardo, P., de Bergevin, F., Alvarez, J., Torrelles, X., van der Vegt, H. A., and Elgens, V. H., 1996. Resonant surface magnetic x-ray diffraction from $Co_3Pt(111)$. *Phys. Rev. Lett.* 77:747–750.

Ferrer, S., Alvarez, J., Lundgren, E., Torrelles, X., Fajardo, P., and Borscherini, F. 1997. Surface x-ray diffraction from Co/Pt(111) ultra thin films and alloys: Structure and magnetism. *Phys. Rev. B* 56:9848–9857.

Gibbs, D. 1992. X-ray magnetic scattering. *Synchr. Radiat. News* 5:18–23.

Gibbs, D., Moncton, D. E., D'Amico, K. L., Bohr, J., and Grier, B. H. 1985. Magnetic x-ray scattering studies of holmium using synchrotron radiation. *Phys. Rev. Lett.* 55:234–237.

Gibbs, D., Harshmann, D. R., Isaacs, E. D., McWhan, D. B., Mills, D., and Vettier, C. 1988. Polarization and resonant properties of magnetic x-ray scattering on holmium. *Phys. Rev. Lett.* 61:1241–1244.

Gibbs, D., Blume, M., Harshmann, D. R., and McWhan, D. B. 1989. Polarization analysis of magnetic x-ray scattering. *Rev. Sci. Instrum.* 60:1655–1660.

Gibbs, D., Grübel, G., Harshmann, D. R., Isaacs, E. D., McWhan, D. B., Mills, D., and Vettier, C. 1991. Polarization and resonant studies of x-ray magnetic scattering in holmium. *Phys. Rev. B* 61:1241–1244.

Goldman, A. I., Mohanty, K., Shirane, G., Horn, P. M., Greene, R. L., Peters, C. J., Thurston, T. R., and Birgeneau, R. J. 1987. Magnetic x-ray scattering measurements on MnF_2. *Phys. Rev. B* 36:5609–5612.

Hamrick, M. D. 1994. Magnetic and chemical effects in x-ray resonant exchange scattering in rare earths and transition metal compounds. Ph.D. thesis, Rice University, Houston, Tex.

Hannon, J. P., Trammell, G. T., Blume, M., and Gibbs, D. 1988. X-ray resonant exchange scattering. *Phys. Rev. Lett.* 61:1245–1249, *Phys. Rev. Lett.* 62:2644–2647.

Harris, Q. J., Feng, Q., Lee, Y. S., Birgeneau, R. J., and Ito, A. 1997. Random fields and random anisotropies in the mixed Ising-XY magnet $Fe_xCo_{1-x}TiO_3$. *Phys. Rev. Lett.* 78:346–349.

Hill, J. P. and McMorrow, D. F. 1996. X-ray resonant exchange scattering: Polarization dependence and correlation functions. *Acta Crystallogr.* A52:236–244.

Hill, J. P., Feng, Q., Birgeneau, R. J., and Thurston, T. R. 1993. Magnetic x-ray scattering study of random field effects in $Mn_{0.75}Zn_{0.25}F_2$. *Z. Phys. B* 92:285–305.

Hill, J. P., Helgesen, G. H., and Gibbs, D. 1995a. X-ray scattering study of charge and spin density waves in chromium. *Phys. Rev. B* 51:10336–10344.

Hill, J. P., Vigliante, A., Gibbs, D., Peng, J. L., and Greene, R. L. 1995b. Observation of x-ray magnetic scattering in Nd_2CuO_4. *Phys. Rev. B* 52:6575–6580.

Hill, J. P., Kao, C.-C., and McMorrow, D. F. 1996. K-edge resonant x-ray magnetic scattering from a transition metal oxide: NiO. *Phys. Rev. B* 55:R8662–8665.

Hill, J. P., Feng, Q., Harris, Q., Birgeneau, R. J., Ramirez, A. P., and Cassanho, A. 1997. Phase transition behavior in the random field antiferromagnet $Fe_{0.5}Zn_{0.5}F_2$. *Phys. Rev. B* 55:356–369.

Hirano, K., Ishikawa, T., and Kikuta, S. 1993. Perfect crystal phase plate retarders. *Nucl. Instrum. Methods A* 336:343–353.

Hirano, K., Ishikawa, T., and Kikuta, S. 1995. Development and application of phase plate retarders. *Rev. Sci. Instrum.* 66:1604–1609.

Isaacs, E. D., McWhan, D. B., Peters, C., Ice, G. E., Siddons, D. P., Hastings, J. B., Vettier, C., and Vogt, O. 1989. X-ray resonant exchange scattering in UAs. *Phys. Rev. Lett.* 63:1671–1674.

Isaacs, E. D., McWhan, D. B., Keliman, R. N., Bishop, D. J., Ice, G. E., Zschack, P., Gaulin, B. D., Mason, T. E., Garret, J. D., and Buyers, W. J. L. 1990. X-ray magnetic scattering in antiferromagnetic URu_2Si_2. *Phys. Rev. Lett.* 65:3185–3188.

Kao, C.-C., Hastings, J. B., Johnson, E. D., Siddons, D. P., Smith, G. C., and Prinz, G. A. 1990. Magnetic resonance exchange

scattering at the iron L_2 and L_3 edges. *Phys. Rev. Lett.* 65:373–376.

Kao, C.-C., Chen, C. T., Johnson, E. D., Hastings, J. B., Lin, H. J., Ho, G. H., Meigs, G., Brot, J.-M., Hulbert, S. L., Idzerda, Y. U., and Vettier, C. 1994. Dichroic interference effects in circularly polarized soft x-ray resonant magnetic scattering. *Phys. Rev. B* 50:9599–9602.

Langridge, S., Lander, G. H., Bernhoeft, N., Stunault, A., Vettier, C., Grübel, G., Søutter, C., de Bergevin, F., Nuttall, W. J., Stirling, W. G., Mattenberger, K., and Vogt, O. 1997. Separation of the spin and orbital moments in antiferromagnetic UAs. *Phys. Rev. B* 55:6392–6398.

Langridge, S. Paixão, J. A., Bernhoeft. N., Vettier, C., Lander, G. H., Gibbs, D., Sørensen, S. Aa, Stunault, A., Wermeille, D, and Talik, E. 1999. Changes in 5*d* band polarization in rare-earth coumpounds. *Phys. Rev. Lett.* 82:2187–2190.

Laundry, D., Collins, S., and Rollason, A. J. 1991. Magnetic x-ray diffraction from ferromagnetic iron. *J. Phys. Condens. Matter* 3:369–372.

Lippert, M., Brückel, T., Kohler, T., and Schneider, J. R. 1994. High resolution bulk magnetic scattering of high energy synchrotron radiation. *Europhys. Lett.* 27:537–541.

Lovesey, S. W. and Collins, S. P. 1996. X-ray Scattering and Absorption by Magnetic Materials. Oxford University Press, New York.

Luo, J., Trammell, G. T., and Hannon, J. P. 1993. Scattering operator for elastic and inelastic resonant x-ray scattering. *Phys. Rev. Lett.* 71:287–290.

Mackay, J. F., Teichert, C., Savage, D. E., and Lagally, M. G. 1996. Element specific magnetization of buried interfaces probed by diffuse x-ray resonant magnetic scattering. *Phys. Rev. Lett.* 77:3925–3928.

Mannix, D., de Camargo, P. C., Giles, C., de Oliveira, A. J. A., Yokaichiya, F., and Vettier, C. 2001. The chromium spin density wave: Magnetic x-ray scattering studies with polarization analysis. *Eur. Phys. J. B* 20(1):19–25.

McMorrow, D. F., Gibbs, D., and Bohr, J. 1999. X-ray scattering studies of the rare earths. *In* Handbook on the Physics and Chemistry of the Rare Earths, vol. 26 (K. A. Gschneidner Jr. and L. Eyring, eds.).. Elsevier, North-Holland, Amsterdam.

McWhan, D. B. 1998. Synchrotron radiation as a probe of actinide magnetism. *J. Alloys Compounds* 271–273:408:413.

McWhan, D. B., Vettier, C., Isaacs, E. D., Ice, G. E., Siddons, P., Hastings, J. B., Peters, C., and Vogt, O. 1990. Magnetic x-ray scattering study of uranium arsenide. *Phys. Rev. B* 42:6007–6017.

McWhan, D. B., Isaacs, E. D., Carra, P., Shapiro, S. M., Thole, B. T., and Hoshino, S. 1993. Resonant magnetic x-ray scattering study of mixed valence TmSe. *Phys. Rev. B* 47:8630–8633.

Namikawa, K., Ando, M., Nakajima, T., and Kawata, H. 1985. X-ray resonance magnetic scattering. *J. Phys. Soc. Jpn.* 54:4099–4102.

Neubeck, W., Vettier, C., de Bergevin, F., Yakhov, F., Mannix, D., Bengone, O., Alouani, M., and Barbier, A. 2001. Probing the 4*p* electron-spin polarization in NiO. *Phys. Rev. B* 63:134430–134431.

Pengra, D. B., Thoft, N. B., Wulff, M., Feidenhans'l, R., and Bohr, J. 1994. Resonance-enhanced magnetic x-ray diffraction from a rare-earth alloy. *J. Phys. Condens. Matter* 6:2409–2422.

Robinson, I. K. 1991. Surface crystallography. *In* Handbook of Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.). pp. 221–266. North Holland Publishing, New York.

Sanyal, M. K., Gibbs, D., Bohr, J., and Wulff, M. 1994. Resonance magnetic x-ray scattering study of erbium. *Phys. Rev. B* 49:1079–1085.

Schülke, W. 1991. Inelastic scattering by electronic excitations. *In* Handbook of Synchrotron Radiation, Vol. 3 (G. S. Brown and D. E. Moncton, eds.). pp. 565–637. North Holland Publishing, New York.

Shen, Q., Shastri, S., and Finkelstein, K. D. 1995. Stokes polarimetry for x-rays using multiple beam diffraction. *Rev. Sci. Instrum.* 66:1610–1613.

Strempfer, J., Brückel, Th., Rütt, U., Schneider, J. R., Liss, K.-D., and Tschentscher, Th. 1996. The non-resonant magnetic x-ray cross-section of MnF_2. Part 2. High energy x-ray diffraction at 80 keV. *Acta Crystallogr.* A52:438–449.

Strempfer, J., Rütt, U., and Jauch, W. 2001. Absolute spin magnetic moment of FeF_2 from high-energy photon diffraction. *Phys. Rev. Lett.* 86:3152–3155.

Sutter, C., Grübel, G., Vettier, C., de Bergevin, F., Stunault, A., Gibbs, D., and Giles, C. 1997. Helicity of magnetic domains in holmium studied with circularly polarized x rays. *Phys. Rev. B* 55:954–959.

Tang, C. C., Stirling, W. G., Lander, G. H., Gibbs, D., Herzog, W., Carra, P., Thole, B. T., Mattenberger, K., and Vogt, O. 1992. Resonant magnetic scattering in a series of uranium compounds. *Phys. Rev. B* 46:5287–5297.

Thole, B. T., Carra, P., Sette, F., and van der Laan, G. 1992. X-ray circular dichroism as a probe of orbital magnetization. *Phys. Rev. Lett.* 68:1943–1947.

Thurston, T. R., Helgesen, G., Gibbs, D., Hill, J. P., Gaulin, B. D., and Shirane, G. 1993. Observation of two length scales in the magnetic critical fluctuations of holmium. *Phys. Rev. Lett.* 70:3151–3154.

Thurston, T. R., Helgesen, G., Hill, J. P., Gibbs, D., Gaulin, B. D., and Simpson, P. 1994. X-ray and neutron scattering measurements of two length scales in the magnetic critical fluctuations of holmium. *Phys. Rev. B* 49:15730–15744.

Tonnerre, J. M., Séve, L., Raoux, D., Soullié, G., Rodmacq, B., and Wolfers, P. 1995. Soft x-ray resonant magnetic scattering from a magnetically coupled Ag/Ni multilayer. *Phys. Rev. Lett.* 75:740–743.

van Veenendaal, M., Goedkoop, J. B., and Thole, B. T. 1997. Branching ratios of the circular dichroism at rare earth $L_{2,3}$ edges. *Phys. Rev. Lett.* 78:1162–1165.

Vettier, C., McWhan, D. B., Gyorgy, E. M., Kwo, J., Buntschuh, B. M., and Batterman, B. W. 1986. Magnetic x-ray scattering study of interfacial magnetism in a Gd-Y superlattice. *Phys. Rev. Lett.* 56:757–760.

Vigliante, A., von Zimmerman, M., Schneider, J. R., Frello, T., Andersen, N. H., Madsen, J., Buttrey, D. J., Gibbs, D., and Tranquada, J. M. 1997. Detection of charge scattering associated with stripe order in $La_{1.775}Sr_{0.225}NiO_4$ by hard x-ray diffraction. *Phys. Rev. B* 56:8248–8251.

Vigliante, A., Christensen, M. J., Hill, J. P., Helgesen, G., Sorensen, A. Aa., McMorrow, D. F., Gibbs, D., Ward, R. C. C., and Wells, M. R. 1998. Interplay between structure and magnetism in Ho_xPr_{1-x} alloys II: Resonant magnetic scattering. *Phys. Rev. B* 57:5941.

Watson, D., Forgan, E. M., Nuttall, W. J., Stirling, W. G., and Fort, D. 1996. High resolution magnetic x-ray diffraction from neodymium. *Phys. Rev. B* 53:726–730.

Watson, D., Nuttall, W. J., Forgan, E. M., Perry, S., and Fort, D. 1998. Refinement of magnetic structures with x rays: Nd as a test case. *Phys. Rev. B Condensed Matter* 57(14):R8095–R8098.

Watson, G., Gaulin, B. D., Gibbs, D., Lander, G. H., Thurston, T. R., Simpson, P. J., Matzke, H. J., Wong, S., Dudley, M., and Shapiro, S. M. 1996a. Origin of the second length scale found above T_N in UO_2. *Phys. Rev. B* 53:686–698.

Watson, G., Gibbs, D., Lander, G. H., Gaulin, B. D., Berman, L. E., Matzke, H. J., and Ellis, W. 1996b. X-ray scattering study of the magnetic structure near the (001) surface of UO_2. *Phys. Rev. Lett.* 77:751–754.

Watson, G., Gibbs, D., Lander, G. H., Gaulin, B. D., Berman, L. E., Matzke, H. J., and Ellis, W. 2000. Resonant x-ray scattering studies of the magnetic structure near the surface of an antiferromagnet. *Phys. Rev. B* 61:8966–8975.

KEY REFERENCES

Als-Nielsen, J. and McMorrow, D. 2001. Elements of Modern X-Ray Physics. John Wiley & Sons, New York.

Excellent introductory text for modern x-ray diffraction techniques.

de Bergevin and Brunel, 1991. See above.

Review of early experiments and non-resonant formalism.

Lovesey and Collins, 1996. See above.

Provides a general overview of theoretical background to magnetic x-ray scattering, and reviews some important experiments.

McMorrow et al., 1999. See above.

Excellent review article of the x-ray magnetic scattering work to date on the rare-earths, and rare-earth compounds. Includes review of the theoretical ideas and experiments.

J. P. HILL
Brookhaven National Laboratory
Upton, New York

X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS

INTRODUCTION

X-ray diffraction (see SYMMETRY IN CRYSTALLOGRAPHY, X-RAY POWDER DIFFRACTION and SURFACE X-RAY DIFFRACTION) and x-ray-excited fluorescence analysis are powerful techniques for the nondestructive measurement of crystal structure and chemical composition. X-ray fluorescence analysis is inherently nondestructive, with orders-of-magnitude lower power deposited for the same detectable limit as with fluorescence excited by charged particle probes (Sparks, 1980). X-ray diffraction analysis is sensitive to crystal structure with orders-of-magnitude greater sensitivity to crystallographic strain than electron probes (Rebonato et al., 1989; Chung and Ice, 1999). When a small-area x-ray microbeam is used as the probe, chemical composition ($Z > 14$), crystal structure, crystalline texture, and crystalline strain distributions can be determined. These distributions can be studied both at the surface of the sample and deep within the sample (Fig. 1). Current state-of-the-art can achieve an ~1-μm diameter x-ray microprobe and an ~0.1 μm diameter x-ray microprobe has been demonstrated (Bilderback et al., 1994; Yun et al., 1999).

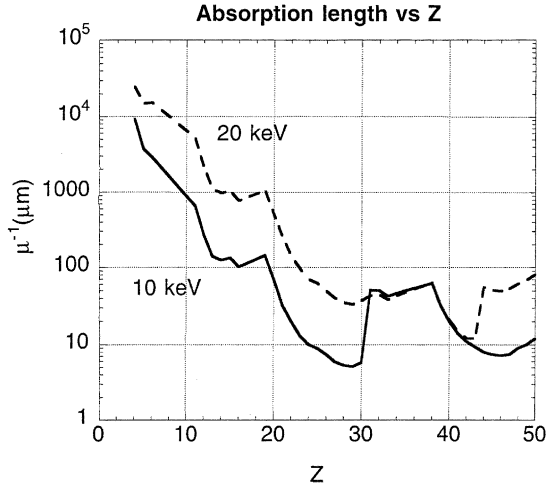


Figure 1. Absorption depth for 10- and 20-keV x rays as a function of elemental composition, *Z*.

Despite their great chemical and crystallographic sensitivities, x-ray microprobe techniques have until recently been restricted by inefficient x-ray focusing optics and weak x-ray sources; x-ray microbeam analysis was largely superseded by electron techniques in the 1950s. However, interest in x-ray microprobe techniques has now been revived (Howells and Hastings, 1983; Ice and Sparks, 1984; Riekel, 1992; Thompson et al., 1992; Chevallier and Dhez, 1997; Hayakawa et al., 1990; APS Users Meeting Workshop, 1997) by the development of efficient x-ray focusing optics and ultra-high-intensity synchrotron x-ray sources (Buras and Tazzari, 1984; Shenoy et al., 1988). These advances have increased the achievable microbeam flux by more than 12 orders of magnitude (Fig. 2; also see Ice, 1997); the flux in a tunable 1 μm-diameter beam on a so-called third-generation synchrotron source such as the Advanced Photon Source (APS) can exceed the flux in a fixed-energy mm^2 beam on a conventional source. These advances place x-ray microfluorescence and x-ray microdiffraction analysis techniques among the most powerful techniques available for the nondestructive measurement of chemical and crystallographic distributions in materials.

This unit reviews the physics, advantages, and scientific applications of hard x-ray ($E > 3$ keV) microfluorescence and x-ray microdiffraction analysis. Because practical x-ray microbeam instruments are extremely rare, a special emphasis will be placed on instrumentation, accessibility, and experimental needs which justify the use of x-ray microbeam analysis.

Competitive and Related Techniques

Despite their unique properties, x-ray microprobes are rare and the process of gaining access to an x-ray

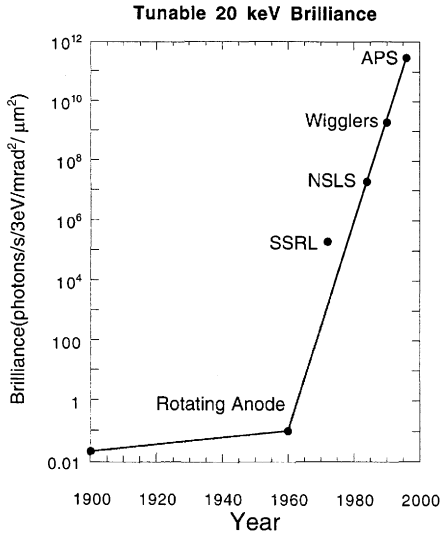


Figure 2. X-ray brilliance over the last 100 years shows an increase of more than 12 orders of magnitude since the use of hard x-ray synchrotron radiation sources began in the late 1960s. Both x-ray microdiffraction and x-ray microfluorescence have brilliance (photons/s/eV/mm^2/mrad^2) as the figure of merit (Ice, 1997).

microprobe can be difficult. For many samples, alternative techniques exist with far greater availability. Destructive methods such as laser ionization with mass spectrometry, as well as atom probe methods (Miller et al., 1996) can yield information on composition distributions. Atom probe measurements in particular can measure the atom-by-atom distribution in a small volume but require extensive sample preparation.

Auger spectroscopy (AUGER ELECTRON SPECTROSCOPY) and Rutherford backscattering techniques (HEAVY-ION BACKSCATTERING SPECTROMETRY) are other possible methods for determining elemental distributions. Auger analysis is inherently surface-sensitive, whereas Rutherford backscattering measurements can probe below the sample surface. Although these techniques are generally considered very sensitive, x-ray analysis can be even more sensitive.

The most directly comparable analysis techniques are charged particle microprobes such as electron or proton microprobes (Sparks, 1980; LOW-ENERGY ELECTRON DIFFRACTION and NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION). Whereas proton microprobes are almost as rare as x-ray microprobes, electron microprobes are widely used to excite fluorescence for chemical analysis. Electron microbeams are also used to measure crystallographic phase and texture, and strain resolution to 2×10^{-4} has been demonstrated (Michael and Goehner, 1993). Advanced electron microbeams can deliver 10^{12} to 10^{15} electrons/μm^2 and can be focused to nanometer dimensions. Electron microbeams are available at many sites within the United States and around the world.

Sparks (1980) has compared the relative performance of x ray, proton, and electron microprobes for chemical analysis. He specifically compared the intrinsic ability of x ray and charged particle microprobes to detect trace elements in a matrix of other elements. He also compared their relative abilities to resolve small-dimensioned features. In summary he finds that x rays have major advantages: (1) x rays are very efficient at exciting electrons in atoms, thereby creating inner-shell holes; (2) x-ray excitation produces very low backgrounds; (3) beam spreading with x rays is low; and (4) x-ray microprobe analysis requires minimum sample preparation.

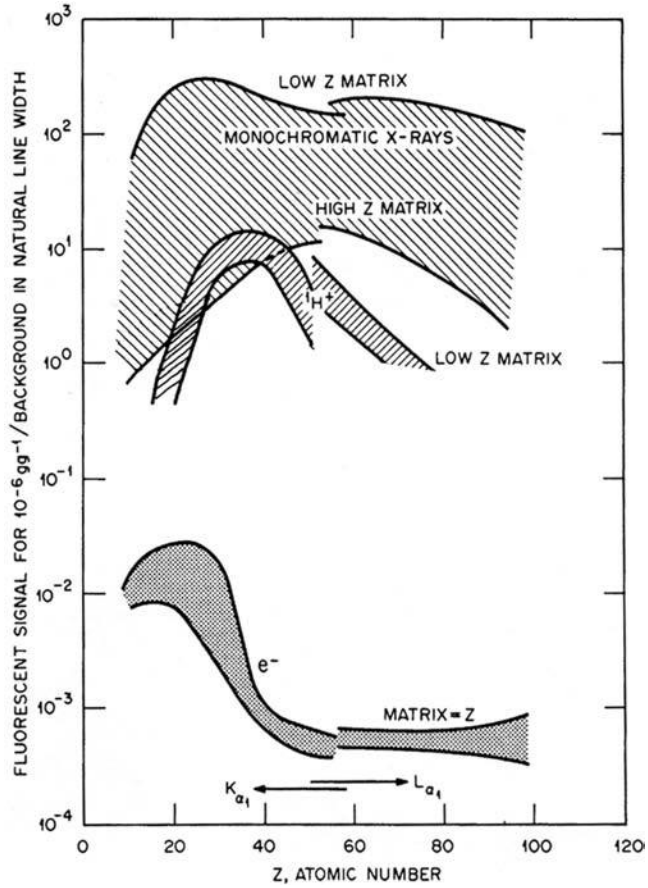


Figure 3. Comparison of the fluorescence signal-to-background ratio for various excitation probes at a concentration of 10^{-6} gg^{-1} for an x-ray detection system with an energy resolution of the natural line width (Sparks, 1980).

A comparison of the signal-to-background ratio for various probes is shown in Figure 3. Monochromatic x-ray excitation produces the best fluorescence signal-to-background. Proton excitation produces signal-to-background ratios between that of x-ray- and electron-induced fluorescence. For low-Z elements, proton microbeams can sometimes approach the signal-to-background of x-ray microbeams.

A rather direct comparison of the elemental sensitivity of various microbeam probes can be made by comparing their minimum-detectable limits (MDL). We adopt the MDL definition of Sparks (1980) for fluorescence analysis

$$C_{\mathrm{MDL}} = 3.29\, C_{\mathrm{Z}} \frac{\sqrt{N_{\mathrm{b}}}}{N_{\mathrm{s}}} \qquad (1)$$

Here C_{MDL} is the minimum detectable limit, C_{Z} is the mass fraction in a calibrated standard, N_{b} is the

Table 1. Estimated MDL/s from (Sparks, 1980), Scaled to 10^{12} Monochromatic Photons/s/μm^2

Probe	MDL (ppm/s)[a]
Proton	~10 to 100
Electron	~5 to 30
Filtered x ray	~1 to 8
Monochromatic x ray	~0.005 to 0.08

[a]This comparison assumes that the matrix and trace element have similar Z and that an advanced multielement solid-state detector is used where deadtime does not limit performance. With a low-Z matrix or with an advanced crystal spectrometer, the MDL can be lower.

background counts beneath the fluorescence signal, and N_s is the net counts at the fluorescence energy. Lowest MDL results when the ratio $N_s/\sqrt{N_b}$ is large (good signal-to-background and high flux), and when N_s/C_Z is large (efficient inner-shell hole production and high flux).

Electron and proton microprobes cannot match the achievable MDL of an advanced x-ray microprobe; compared to x rays, the inner-shell hole production cross-section and signal-to-background of an electron probe are too low and the proton probe flux density on the sample is too low. An advanced x-ray microprobe with 10^{12} photons/s/μm^2 has about a 10^3 lower MDL for most elements than a charged particle probe and can achieve the same MDL as electron probes with 10^4 less energy deposited in the sample (Table 1).

For thin samples, the spatial resolution of electron probes is far better than that of either x-ray or proton microprobes. In terms of their practical spatial resolution for thick samples, however, x-ray microprobes are competitive or superior to charged particle probes. Although electron probes can be focused to nanometer dimensions, beam spreading in thick samples degrades their effective resolution. For example, in a thick Al sample, a nanometer electron probe spreads to an effective size of ~2-μm diameter. In Cu and Au samples the same beam spreads to 1- and 0.4-μm diameter, respectively (Goldstein, 1979; Ren et al., 1998). Proton microprobes can be made very small, but the fluxes are so low that few instruments exist with probe dimensions approaching 1-μm diameter (Lindh, 1990; Doyle et al., 1991).

X-ray beams are now so intense that their flux density is approaching the maximum that can usefully be applied to most samples. For example, the estimated thermal rise of a thin target under an advanced x-ray microbeam is shown in Figure 4. Existing x-ray microbeams have highly monochromatic flux densities approaching 10^{12} photons/s/μm^2 and can go to 10^{13} to 10^{14} photons/s/μm^2 with larger band-pass optics. With 10^{14} photons/s/μm^2, the flux must be attenuated for most samples. Alternatively, the probe area can be decreased or the dwell time on the sample can be reduced to prevent sample melting.

Despite its many advantages, however, x-ray microbeam analysis remains an emerging field with very little available instrumentation. The effort required to gain access to x-ray microbeam facilities must therefore be weighed against the benefits. X-ray microfluorescence analysis becomes justified when the MDL from other techniques is inadequate. It may also be justified when the probe must penetrate the sample surface, when the analysis must be highly nondestructive, when the analysis must be quantitative, or when the measurement must be done in the presence of an air, water, or other low-Z overlayer.

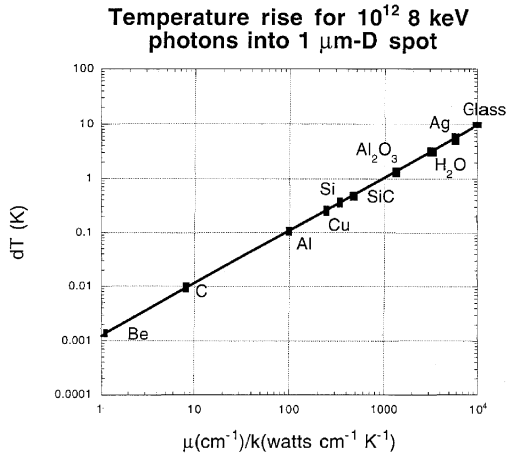


Figure 4. Thermal rise as a function of thermal conductivity K and absorption coefficient, μ (Ice and Sparks, 1991). Note that existing x-ray microbeams have achieved 10^{12} photons/s/μm^2 and are anticipated to reach 10^{13}–10^{14} photons/s/μm^2 for some applications.

Average crystallographic grain orientations can be studied with large probes such as standard x-ray beams or neutron beams. Individual grain orientations can also be determined to ~1° with electron back-reflection measurements and strain resolution to 2×10^{-4} has been reported (Michael and Goehner, 1993). However, large beam probes cannot study individual grain-grain correlations, and electron probes are predominantly surface sensitive, and in most cases have both strain and orientation precision two orders of magnitude worse than for x-ray microdiffraction. Microdiffraction analysis therefore becomes justified when the strain resolution $\Delta d/d$ must be better than $\sim 5 \times 10^{-4}$. Microdiffraction analysis is also justified for the study of crystallographic properties beneath the surface of a sample, for the measurement of texture in three dimensions or for nondestructive analysis of insulating samples where charge buildup can occur.

PRINCIPLES OF THE METHOD

A typical x-ray microbeam experiment involves three critical elements: (1) x-ray condensing or aperturing optics on a high-brilliance x-ray source, (2) a high-resolution sample stage for positioning the sample, and (3) a detector system with one or more detectors (Fig. 5). The x-ray beam axis and focal position is determined by the optics of the particular arrangement. Different locations on the specimen are characterized by moving the specimen under the fixed x-ray microprobe beam. Details of the detector arrangement and the principles involved depend strongly on the particular x-ray microprobe and on whether elemental distributions or crystallographic information is to be collected.

Both x-ray microdiffraction and x-ray microfluorescence are brilliance (photons/s/eV/mm^2/mrad^2) limited (Ice, 1997). For x-ray microdiffraction, momentum

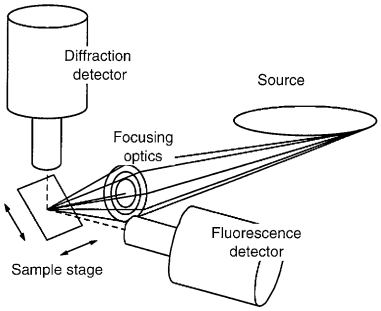


Figure 5. Key elements of an x-ray microbeam experiment.

transfer resolution is limited by spread in wavelength and angular divergence on the sample. Count rate is limited by flux per unit area. For x-ray microfluorescence, best signal-to-background occurs when the x-ray bandwidth $\Delta E/E$ is $\leq$3% (Sparks, 1980). Spatial resolution for thick samples degrades when the divergence of the beam exceeds $\sim$10 mrad. The principles of x-ray microfluorescence and x-ray microdiffraction analysis are briefly outlined below.

X-Ray Microfluorescence Analysis

The unique advantages of x-ray probes arise from the fundamental interaction of x rays with matter. Below the electron-positron pair production threshold, the interaction of x rays with matter is dominated by three processes: photoabsorption (photoelectric effect), elastic scattering, and inelastic (Compton) scattering (Veigele et al., 1969). Of these three processes, photoabsorption has by far the largest cross-section in the 3- to 100-keV range.

Photoabsorption and Compton scattering are best understood in terms of a particle-like interaction between x rays and matter (Fig. 6). The quantized energy of an x-ray photon excites an electron from a bound state to an unbound (continuum) state, while the photon momentum is transferred either to the atomic nucleus (photoeffect) or

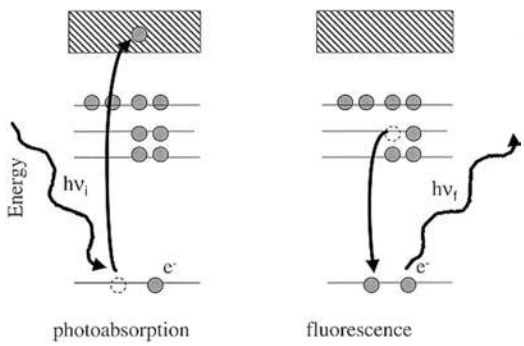


Figure 6. Schematic of x-ray photoabsorption followed by fluorescence. An x-ray photon of energy $h\nu_i$ is absorbed by the atom, which ejects an electron from an inner shell of the atom. The atom fills the electron hole by emitting an x ray with an energy $h\nu_f$; $h\nu_f$ is characteristic to the atom and has an energy equal to the energy difference between the initial and final hole-state energies. Alternatively the atom fills the inner-shell hole by emitting an energetic electron (Auger effect; Bambynek et al., 1972 also see AUGER ELECTRON SPECTROSCOPY).

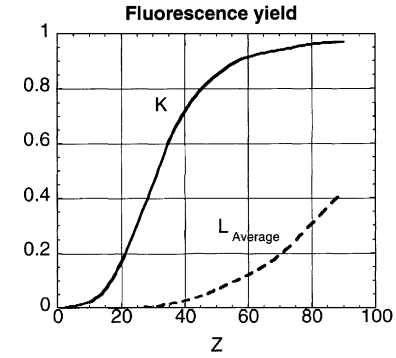


Figure 7. Fit to experimental fluorescence yields for K and L holes. (Bambynek et al., 1972.)

to the electron (Compton scattering). X-ray fluorescence is the name given to the elementally distinct or "characteristic" x-ray spectrum that is emitted from an atom as an inner-shell hole is filled.

Fluorescence Yields. An inner-shell hole can also be filled by nonradiative mechanisms (Auger and Coster-Kronig effects; Bambynek et al., 1972). Here the singly ionized atom fills the inner shell hole with a higher-energy electron and emits an energetic (Auger/Coster-Kronig) electron with a characteristic energy that is determined by the initial and final energy of the atom. The fraction of holes that are filled by x-ray fluorescence decay is referred to as the fluorescence yield. In general, nonradiative processes become increasingly likely as the initial-hole binding energy decreases (Fig. 7). We note that for K holes with binding energies above 5 keV, the x-ray fluorescence yields are more than 20%. For L holes with binding energies greater than 5 keV, all fluorescence yields exceed 10%. The fluorescence yields of deep inner-shell holes can exceed 90%.

Characteristic Radiation. The characteristic x-ray energies emitted when the initial hole decays by fluorescence serve as a "fingerprint" for the element and are quite distinct. Fluorescence spectra are labeled according to the electron hole being filled and the strength of the decay channel. For example, as shown in Figure 8, $K\alpha_1$ fluorescence arises when an L_{III} ($2P_{3/2}$) electron fills a K hole.

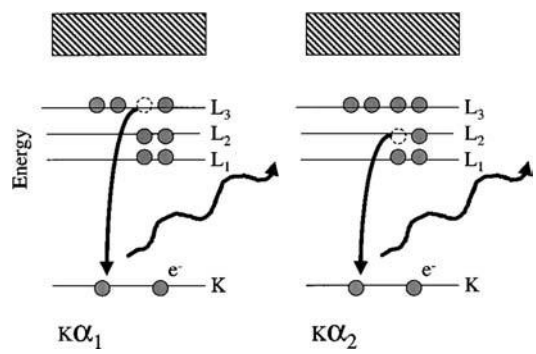


Figure 8. Fluorescence decay channels for K holes.

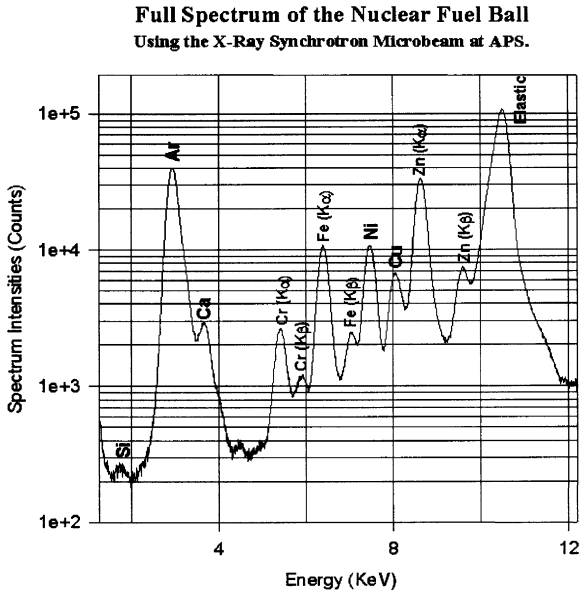


Figure 9. Fluorescence spectra from the SiC shell of an advanced nuclear fuel particle. The trace elements in the sample emit characteristic x-ray lines when excited with x rays.

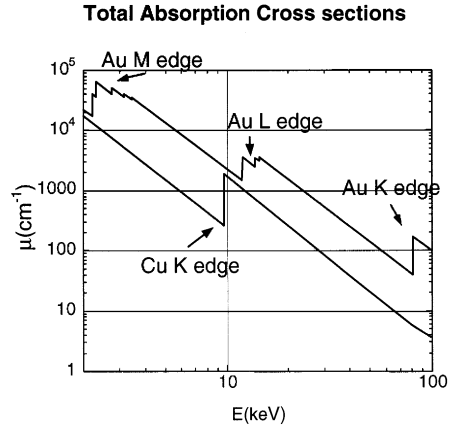


Figure 10. Photoabsorption cross-section for Cu and Au.

This transition is the strongest fluorescence decay channel for K holes. A similar nomenclature is used for L holes, etc.

Because chemical effects on inner-shell wavefunctions are small, x-ray absorption cross-sections, fluorescence yields, and characteristic x-ray spectra are virtually unchanged by the sample environment except very near threshold. However, the measured-fluorescence signal can strongly depend on absorption and secondary excitation due to the sample matrix.

A typical spectrum from a low-Z matrix with trace elements is shown in Figure 9. The characteristic lines have a natural bandwidth of a few eV that is smeared by the energy resolution of the detector. Even in this case where the trace elements are nearby in the periodic table, the fluorescence signature of each element is distinct. Crystal spectrometers with higher energy resolution can be used in more complicated cases with overlapping L lines to resolve nearby fluorescence lines.

Photoabsorption Cross-Sections. To a first approximation, only x-rays with sufficient energy to excite an electron above the Fermi level (above the occupied electron states) can create an inner-shell hole. As a consequence, the photoabsorption cross-section for x-rays has thresholds that correspond to the energy needed to excite an inner-shell electron into unoccupied states. As shown in Figure 10, the photoabsorption cross-section exhibits a characteristic saw-tooth pattern as a function of x-ray energy. Highest x-ray efficiency for the creation of a given hole is just above its absorption edge energy. Maximum elemental sensitivity with minimum background results when the x-ray microprobe energy is tuned just above the absorption edge of the element of interest (Sparks, 1980). Because x rays are highly efficient at creating inner-shell holes, x-ray excited fluorescence has a very high signal-to-noise ratio (Fig. 3) and the energy deposited in the sample is low for a given signal.

Micro-XAFS. The saw-tooth pattern of Figure 10 shows the typical energy dependence of x-ray absorption cross-sections over a wide energy range, but does not include lifetime broadening of the inner shell hole, the density of unfilled electron states near the Fermi energy, or the influence of photoelectron backscattering. These various processes lead to fine structure in the photoabsorption cross-section which can be used to determine the valence state of an element, its local neighbor coordination, and bond distances. Near-edge absorption fine structure (NEXAFS) is particularly sensitive to the valence of the atom. Extended x-ray absorption fine structure (EXAFS) is sensitive to the near-neighbor coordination and bond distance. Fluorescence measurements have the best signal-to-background ratio for XAFS of trace elements. It is therefore possible to use XAFS techniques with an x-ray microprobe to determine additional information about the local environment of elements within the probe region. More detail about XAFS techniques is given in XAFS SPECTROSCOPY.

Penetration Depth. The effective penetration depth of a fluorescence microprobe depends on the energy of the incident and fluorescent x rays, the composition of the sample, and the geometry of the measurement. As shown in Figure 10, x-ray absorption decreases with an $\sim E^{-3}$ power dependence between absorption edges. For a low-Z matrix, the penetration depth of an x-ray microfluorescence beam can be tens of millimeters into the sample, while for a high-Z matrix the penetration can be only a few microns (Fig. 1).

With a thick sample, the fluorescence signal and effective depth probed depends on the total scattering angle and on the asymmetry of the incident-to-exit beam angles (Sparks, 1980). To a first approximation, the fluorescence signal is independent of total scattering angle but depends on the asymmetry between the incident angle ψ and the exit angle ϕ (Fig. 11).

$$C_Z \propto \left(\mu_i + \mu_f \frac{\sin\psi}{\sin\phi}\right) \qquad (2)$$

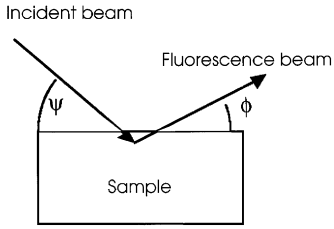


Figure 11. Depth penetrated by a fluorescence microprobe depends on the total absorption cross-section of the incident and fluorescence radiation and on the incident and exit angles with respect to the sample surface.

Here μ_i is the linear absorption coefficient for the incident beam and μ_f is the total absorption coefficient for the fluorescence beam. The approximation of Equation 2 is only valid for uniformly smooth sample surfaces. Where sample granularity and surface roughness are large, the fluorescence signal decreases as the glancing angle decreases (Campbell et al., 1985; Sparks et al., 1992). The effective depth probed depends both on the total scattering angle $\psi + \phi$ and on the asymmetry between ψ and ϕ.

The characteristic depth of an x-ray microprobe is given by

$$x = \left(\frac{\mu_i}{\sin\psi} + \frac{\mu_f}{\sin\phi}\right) \quad (3)$$

Shallow depth penetration can be achieved with glancing angle and asymmetric geometries. With smooth surfaces, even greater surface sensitivity can be achieved by approaching or achieving total external reflection from the surface (Brennan et al., 1994). With total external reflection, the surface sensitivity can approach 1 nm (10 Å) or better (Fig. 12).

Backgrounds. Background signal is generated under fluorescence lines by various scattering and bremstrahlung processes. With a white or broad bandpass incident beam, elastically scattered x-rays including those with the same energy as the fluorescent line can be directly scattered into the detector. This scattering can be greatly reduced by operating at a 2θ scattering angle of 90° in the plane of the synchrotron storage ring where linear beam polarization inhibits elastic and Compton scattering. A much better way to reduce background is through the use of a monochromatic x-ray beam. As shown in Figure 2, monochromatic x-ray beams produce the best signal-to-background because the background under a fluorescence peak must arise from multiple scattering events, bremstrahlung from photoelectrons, or other cascade processes.

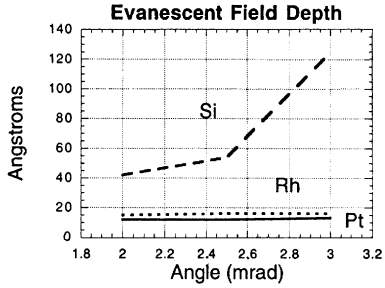


Figure 12. Evanescent wave depth of 10 keV x rays into Si, Rh, or Pt as a function of glancing angle.

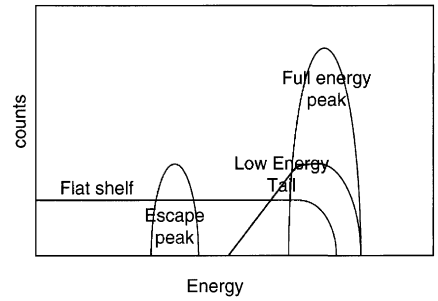


Figure 13. Schematic of low-energy tails in a solid state detector.

Detectors. Two kinds of detectors are used for observing fluorescence. Solid-state detectors are most widely used because they are efficient and can simultaneously detect many elements in the sample. A state-of-the-art solid-state detector with a 1-cm^2 active area has ~130-eV resolution at 5.9 keV with a 5000 counts-per-second (cps) counting rate. Much higher counting rates are possible by compromising the energy resolution and by using multiple detector arrays. For example, 30-element arrays with 500,000 cps counting rates/element can achieve 15,000, 000 cps. One drawback with solid-state detectors is additional background for trace elements, which is introduced by high intensity peaks in the spectra. As shown in Figure 13, the response of a solid-state detector to an x ray of a single energy typically has both short- and long-range low-energy tails due to insufficient charge collection (Cambell, 1990). These can be the dominant contribution to background under a trace element.

Wavelength dispersive spectrometers can also be used to measure fluorescence (Ice and Sparks, 1990; Koumelis et al., 1982; Yaakobi and Turner, 1979). These detectors have much poorer collection efficiency but much better energy resolution (lower background) and are not paralyzed by scattering or by fluorescence from the major elements in the sample. Because wavelength dispersive detectors only count one fluorescent energy at a time, they are not count-rate limited by fluorescence or scattering from the major elements in the sample matrix.

X-Ray Microdiffraction Analysis

X-ray microdiffraction is sensitive to phase (crystal structure), texture (crystal orientation), and strain (unit cell distortion) (Ice, 1987; Rebonato et al., 1989). Diffraction (Fig. 14) is best understood in terms of the wave-like nature of x rays. Constructive and destructive interference from x-ray scattering off the charge-density distribution varies the x-ray scattering efficiency as a function of angle and wavelength. This so-called diffraction pattern can be Fourier transformed to recover charge-density

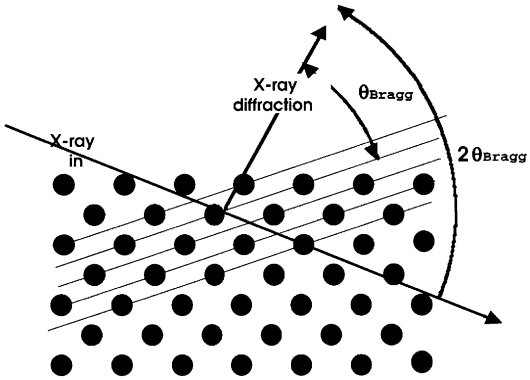


Figure 14. X-ray diffraction from a crystal becomes large at Bragg angles where the crystal lattice spacing, d, the wavelength, λ, and the angle of the incident beam with respect to the crystal lattice planes, θ_{Bragg}, satisfy Bragg's law (Equation 4).

information. Although the basic diffraction process is identical for all diffraction probes (e.g., x rays, electrons, or neutrons), x rays have three very favorable attributes for the characterization of crystal structure: (1) the wavelength is similar to the atomic spacing of matter; (2) the cross-section is sufficiently low that multiple scattering effects are often small; and (3) the cross-section of elastic scattering is a large fraction of the total interaction cross-section of x rays, which contributes to low noise. In addition, x-ray scattering contrast can be adjusted by tuning near to x-ray absorption edges (Materlik et al., 1994).

Strong diffraction occurs when the incident beam satisfies Bragg's law

$$2d \sin\theta = n\lambda \tag{4}$$

Here d is a crystal lattice spacing, λ is the x-ray wavelength, and θ is the so-called "Bragg angle" between the incident beam and the crystal plane. X-ray microdiffraction can yield detailed information about the sample unit cell and its orientation. If the x-ray wavelength is known, the angle between the incident and an intense exit beam (2θ) determines the spacing, d. The relative orientation of crystals (mosaic spread or texture) can be determined by observing the rotation angles of a sample needed to maximize scattered intensity. These two methods are useful for studies of polycrystalline materials that have a strong preferred orientation.

For unknown crystal orientation, the unit-cell parameters and crystallographic orientation of a single crystal can be determined from the x-ray energy and the angles of three noncolinear reflections (Busing and Levy, 1967). In standard crystallography, reflections are found by rotation of the sample under the beam. With microdiffraction x-ray measurements on polycrystalline samples, however, the imprecision of mechanical rotations will cause the sample to translate relative to the beam on a micron scale. In addition, for complex polycrystalline samples with many grains, the penetration of the x-ray beam into the specimen ensures that the sample volume (and therefore microstructure) will vary as the sample rotates. The changing grain illumination makes a standard crystallographic solution impossible. These two problems can be overcome by the use of Laue diffraction with white or broad-bandpass x-ray beams. With Laue diffraction, no sample rotations are required. Development is in progress to automatically index the overlapping reflections of up to 10 crystals (Marcus et al., 1996; Chung, 1997; Chung and Ice, 1998; Wenk et al., 1997).

X-ray microdiffraction also allows for three-dimensional imaging of crystal structure as demonstrated in some first experiments (Stock et al., 1995; Piotrowski et al., 1996). For these measurements, the sample-to-detector distance is changed and the origin of the reflecting crystal along the microprobe beam is determined by ray tracing. Spatial resolution of less than 2 μm along the microbeam direction has recently been demonstrated (Larson et al., 1999).

PRACTICAL ASPECTS OF THE METHOD

Sources

Only second- and third-generation synchrotron sources have sufficient x-ray brilliance for practical x-ray microprobe instrumentation. Worldwide, there are only three third-generation sources suitable for hard x-ray microprobes: the 6-GeV European Synchrotron Radiation Facility in Grenoble France (Buras and Tazzari, 1984), the 7-GeV Advanced Photon Source at Argonne, Illinois (Shenoy et al., 1998), USA, and the 8-GeV Spring-8 under construction in Japan. Although third-generation sources are preferred, x-ray microprobe work can also be done on second-generation sources like the National Synchrotron Light Source (NSLS), Brookhaven, New York. A list of contacts at synchrotron radiation facilities with x-ray microbeam instrumentation is given in Table 2.

Optics

Tapered Capillaries. The development of intense synchrotron x-ray sources, with at least 12 orders of magnitude greater brilliance than conventional x-ray sources (Fig. 2), has revived interest in x-ray optics. At least three microbeam-forming options have emerged with various strengths and weaknesses for experiments (Ice, 1997). Tapered capillary optics have produced the smallest x-ray beams (Stern et al., 1988; Larsson and Engström, 1992; Hoffman et al., 1994). Bilderbach et al. (1994) have reported beams as small as 50 nm full width at half-maximum (FWHM). This option appears to be the best for condensing beams below 0.1 μm. One concern with capillary optics, however, is their effect on beam brilliance. Ray tracing and experimental measurements have found that the angular divergence following a capillary has a complex annular distribution. This distribution arises from roughness inside the capillary and from the nonequal number of reflections of different rays as they are propagated along the capillary.

Zone Plates. Hard x-ray zone plates are a rapidly emerging option for focusing synchrotron radiation to μm dimensions (Bionta et al., 1990; Yun et al., 1992). This

Table 2. Worldwide Dedicated X-ray Microbeam Facilities[a]

Facility	Beamline	hν (keV)	Spot Size (μm^2)	$\Delta E/E$	Total Flux (photons/s)	Local Contact
APS	7-ID	5–20	0.5	2×10^{-4} to 1×10^{-1}	10^8–10^{11}	Walter Lowe *Lowe@mhattcat.howard.edu*
APS	2-ID-CD	5–20	~0.03	2×0^{-4}	5×10^{10}	Barry Lai *Lai@aps.anl.gov*
APS	13-ID	5–25	~0.7	2×10^{-4} to 1×10^{-2}	1×10^{13}	Steve Sutton *Sutton@cars.uchicago.edu*
ESRF	ID-13	6–16	~40	2×10^{-4}	2×10^{11}	Christian Riekel *riekel@esrf.fr*
ESRF	ID-30	6–20	~4	—	—	—
ESRF	ID-22	4–35	~2	2×10^{-4}	10^9–10^{12}	Anatoly Snigirev *snigirev@esrf.fr*
ALS	10.3.2	5–12	~0.6	2×10^{-4}	—	Howard Padmore *Padmore@lbl.gov*
ALS	10.3.1	6–12	~2	5×10^{-2}	2×10^{10}	Scott McHugo *samchugo@lbl.gov*
LNLS	—	2–14	100	2×10^{-4} white	3×10^7 3×10^9	Helio C.N. Tolentino *helio@lnls.br*
CHESS	B2	5–25	0.001	2×10^{-2}	10^6	Don Bilderbach *dhb2@cornell.edu*
Photon Factory	BL4-A	5–15	25	2×10^{-4} 5×10^{-2}	10^8 10^{10}	A. Iida
Hasylab	L	4–80	9	—	—	Thomas Wroblewski *wroblewt@mail.desy.de*
Hasylab	BW-1	10	1	1×10^{-2}	6×10^7	Thomas Wroblewski *wroblewt@mail.desy.de*
DCI LURE	D15	6,10,14	~1–100	2×10^{-2}	10^5–10^7	P. Chevallier *chevallier@lure.u-psud.fr*
SSRL	—	—	—	—	—	J. Patel *Patel@ssrl01.Slac.stanford.edu*
NSLS	X16C	5–20	4 μm^2	White	10^{10}	M. Marcus
NSLS	X26A	5–20	~10–200 μm^2	10^{-4} white	7×10^7–7×10^{10}	Steve Sutton *Sutton@cars.uchicago.edu*

[a]The top four facilities are on hard x-ray third-generation sources. The ALS x-ray microprobe is on a third-generation VUV ring with a small emittance beam which acts like a high-performance second generation ring for hard x-rays. The remaining beamlines are on second-generation rings.

option appears especially promising for focusing monochromatic radiation ($\Delta E/E = \sim 10^{-4}$). These devices are simple to align, allow good working distance between the optics and the sample, and have already achieved submicron spots. Although zone plates are inherently chromatic, they can in principle be used with tunable radiation by a careful translation along the beam direction. State-of-the-art zone plates provide the most convenient optics for monochromatic experiments even though their focusing efficiency has not reached the 40% to 60% efficiency promised by more advanced designs.

Kirkpatrick-Baez Mirrors. Kirkpatrick-Baez (KB) mirrors provide a third highly promising option for focusing synchrotron radiation (Yang et al., 1995; Underwood et al., 1996). A Kirkpatrick-Baez mirror pair consists of mirrors that condense the beam in orthogonal directions. Both multilayer and total-external-reflection mirrors have been used for focusing synchrotron radiation. Multilayer mirrors appear most suitable for fluorescence measurements with a fixed wide-band-pass beam and where large divergences can be accepted. Total-external-reflection mirrors appear to offer the best option for focusing white beams to μm dimensions. The key challenge with KB mirrors is achieving low figure and surface roughness with elliptical surfaces. There are numerous parallel efforts currently underway to advance mirror figuring for advanced KB focusing schemes.

Refractive Lenses. In addition to the focusing schemes mentioned above, there are two new options that have recently emerged from experiments at the European Synchrotron Radiation Facility (ESRF). These are compound refractive lenses and Bragg-Fresnel optics. Compound refractive lenses are very interesting because they are relatively easy to manufacture (Snigirev et al., 1996). Estimates of their theoretical efficiency, however, indicate that they cannot compete with the theoretical efficiency of KB mirrors or zone plates.

Bragg-Fresnel Optics. Bragg-Fresnel optics are also an interesting option used both at LURE (Chavallier et al., 1995, 1996) and the ESRF (Aristov, 1988). With Bragg-Fresnel optics, the phase-contrast steps are

lithographically etched into a multilayer or monolithic Si substrate. This offers a rugged coolable substrate and serves to simultaneously focus and monochromatize the beam. Some of the alignment simplicity of zone plates is lost with these devices because the beam is deflected, but they have the advantage that the 0th order (direct beam) is spatially removed from the focus and can therefore be easily stopped.

MICROBEAM APPLICATIONS

Microprobe analysis has already been applied to many problems with second- and third-generation x-ray beams (Jones and Gordon, 1989; Rebonato et al., 1989; Langevelde et al., 1990; Thompson et al., 1992; Wang et al., 1997). Studies include measurement of strain and texture in integrated circuit conduction paths (Wang et al., 1996, 1997), the measurement of buried trace elements in dissolution reactions (Perry and Thompson, 1994), and determination of chemistry in small regions. Three simple examples are given to illustrate possible applications.

Example 1: Trace Element Distribution in a SiC Nuclear Fuel Barrier

TRISO-coated fuel particles contain a small kernel of nuclear fuel encapsulated by alternating layers of C and SiC as shown in Figure 15A. The TRISO-coated fuel particle is used in an advanced nuclear fuel designed for passive containment of the radioactive isotopes. The SiC layer provides the primary barrier for radioactive elements in the kernel. The effectiveness of this barrier layer under adverse conditions is critical to containment.

Shell coatings were evaluated to study the distribution and transport of trace elements in the SiC barrier after being subjected to various neutron fluences (Naghedolfeizi et al., 1998). The C buffer layers and nuclear kernels of the coated fuel were removed by laser drilling through the SiC and then leaching the particle in acid (Myers et al., 1986). Simple x-ray fluorescence analysis can detect the presence of trace elements, but does not indicate their distribution. Trace elements in the SiC are believed to arise at least in part as daughter products from the fission process; lower trace-element concentrations are found in unirradiated samples. The radial distribution of these elements in the SiC shells can be attributed to diffusion of elements in the kernel due to thermal and radiation-enhanced diffusion. Other elements in the shells may originate in the fabrication of the TRISO (three-layer coatings of pyrolytic carbon and SiC) particles.

Linear x-ray microprobe scans were made through the SiC shell. X-ray fluorescence is an ideal tool for this work because it is nondestructive (no spread of contamination), it is sensitive to heavy elements in a low-Z matrix, and it provides a picture of the elemental distribution. Results of a simple line scan through a leached shell are shown in Figure 15B. This scan was made with 2-μm steps and an ~1-μm probe. As seen in Figure 15B, very localized Fe-rich regions <1-μm broad are observed throughout the shell. This behavior is typical of other trace elements observed in the shell.

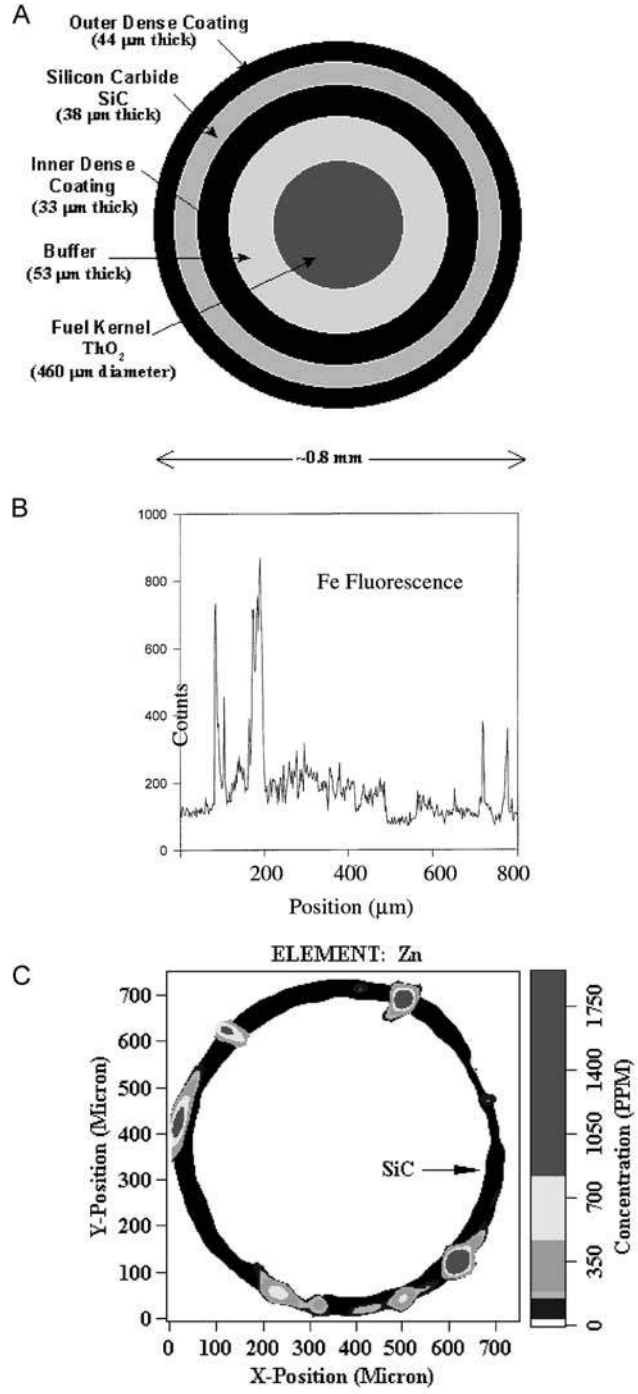


Figure 15. (**A**) Schematic of TRISO fuel element. (**B**) The Fe fluorescence during a line scan through a SiC shell shows a complicated spatial distribution with sharp features less than 1 μm wide. (**C**) Zn distribution in a SiC shell after reconstruction from x-ray microfluorescence data (Naghedolfeizi et al., 1998). The spatial resolution was limited because time restricted the number of steps. **See color figure.**

The spatial distribution can be further investigated by extending the x-ray microprobe analysis to x-ray microfluorescence tomography (Biosseau, 1986; Naghedolfeizi et al., 1998). Although x-ray fluorescence tomography is

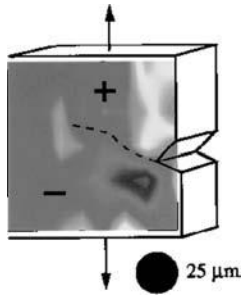


Figure 16. Strain map in a Mo single crystal (Rebonato et al., 1989) showing $\Delta d/d$ for the (211) planes when the crystal is pulled. A region of contraction (below the dashed line) and a region of elongation (above the dashed line) are observed. The reduction in d beneath the dashed line arises from the orientation of the observed Bragg plane with respect to the free surface. The probe size of ~25 μm was barely small enough to detect the strain features. **See color figure.**

inherently time-consuming, the method yields three-dimensional distributions of trace elements. For example, the Zn distribution in a plane of the SiC shell is illustrated in Figure 15C. Note that the measurements are quantitative.

Example 2: Change in the Strain Distribution Near a Notch During Tensile Loading

Early experiments with second-generation sources demonstrated some of the features of x-ray microprobe–based diffraction: good strain resolution, ability to study strain in thick samples, ability to study dynamics of highly strained samples, and the ability to distinguish lattice dilation from lattice rotations. For example, a measurement of differential strain induced by pulling on a notched Mo single crystal (Rebonato et al., 1989) shows lattice dilation above the notch and a lattice contraction below the notch (Fig. 16). This behavior would be totally masked with a topographic measurement because of the high density of initial dislocations. In addition, the measurement allowed the lattice rotations to be separated from the lattice dilations. Finally, the measurements are quantitative, and even in this early experiment yielded strain sensitivity $\Delta d/d$ of $\sim 5 \times 10^{-4}$.

Example 3: Strain Distribution in an Advanced Ferroelectric Sample

$BaTiO_3$ has a tetragonal structure in the ferroelectric state. For a single-crystal thin film deposited on Si, the tetragonal axis can lie either in the plane of the thin film or normal to the surface. The direction of the tetragonal axis is referred to as the "poled" direction. In general, the poling is in the plane of the surface, but electrical measurements indicate that the poled state of the $BaTiO_3$ can be switched if the size of the $BaTiO_3$ film is small enough. X-ray microdiffraction provides a means to study the distribution of poled $BaTiO_3$ in small circular pads. The measurement sampled regions with $\sim 1 \times 10$ μm^2 spatial resolution. In the experiment, an x-ray microbeam was scanned across monolithic films of the $BaTiO_3$ film (reference film) and across small ~60-μm diameter mesasurements (Fig. 17A). The diffraction pattern from the film was recorded on a charge-coupled device (CCD) x-ray detector.

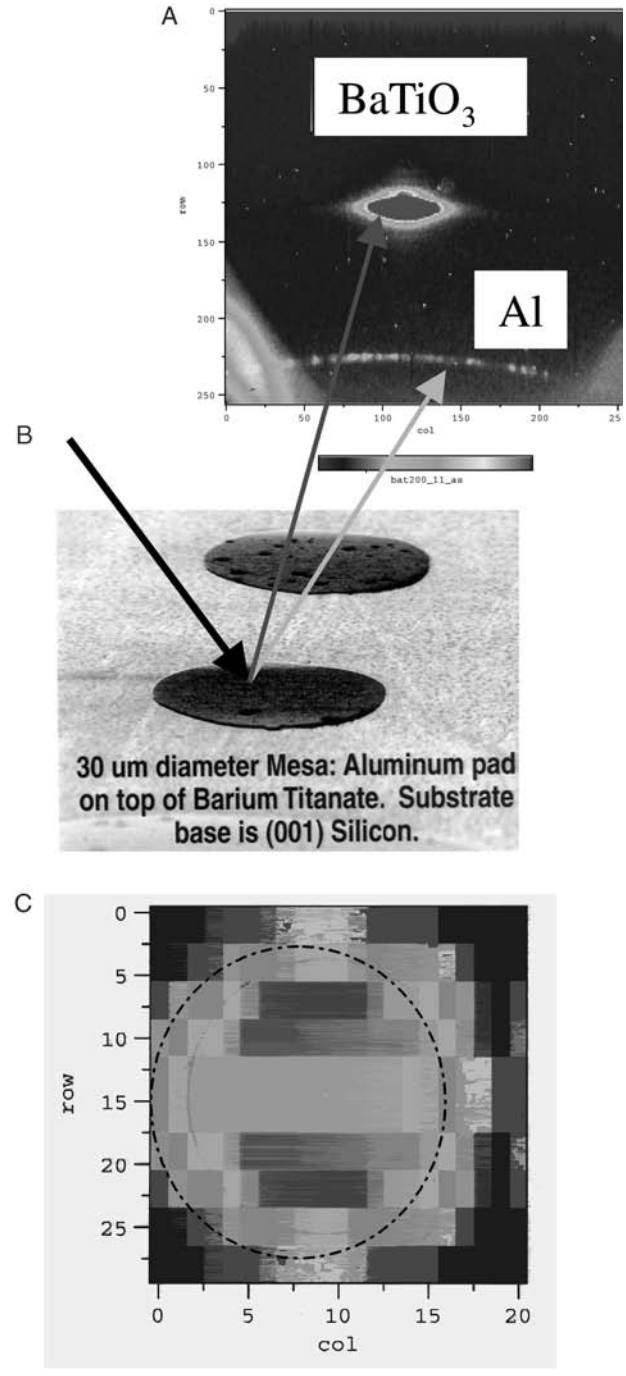


Figure 17. (**A**) Mesas of $BaTiO_3$ show interesting poling behavior. (**B**) Schematic of microdiffraction process with x-ray microdiffraction image from an advanced thin film. The CCD image shows the single-crystal reflection and the powder-like image from the Al overlayer. (**C**) Strain map across the mesa. The strain map has a range of only 5×10^{-4} and an uncertainty of $\sim 1 \times 10^{-5}$. **See color figure.**

A typical diffraction pattern, shown in Figure 17B, illustrates many of the powerful attributes of microdiffraction. The $BaTiO_3$ 002/200 Bragg peak appears as a lenticular intense pattern in the CCD image even though it is covered with a 200-nm cap of aluminum. The strain in the thin film can be measured to $\Delta d/d = \sim 1 \times 10^{-5}$. This is sufficient to easily resolve the ~1% difference in the lattice parameter between 002 or 200 poled $BaTiO_3$. Texture, particle size, and strain of the powder-like aluminum overlayer can be inferred directly from the 111 and 200 Debye rings as seen in the lower part of Figure 17B. Even better strain resolution can be obtained by measurement of higher order reflections.

This example clearly indicates the need for an x-ray diffraction microprobe. Small spot size is required to study the distribution of poled material near small dimensioned features of the film. Good strain resolution is required to differentiate between the various poling options. A penetrating probe is required to measure the film in an unaltered state below a 200-nm cap of aluminum.

METHOD AUTOMATION

X-ray microfluorescence and x-ray microdiffraction data collection at synchrotron sources occurs remotely in a hostile, high-radiation environment. Experiments are located inside shielded hutches that protect the researcher from the x rays near the experiment, but which restrict "hands-on" manipulation of the experiment. Fluorescence and simple diffraction data are typically collected by step-scanning the sample under the beam. A record is made of the fluorescence spectrum or the diffraction pattern from the sample at hundreds to thousands of positions. Fluorescence data collection and analysis is typically performed using specially designed one-of-a-kind programs. Some recent efforts have been directed to incorporate commercial fluorescence data analysis and collection software, but no standard has yet evolved. The general-purpose program SPECT has however been adopted on several beamlines worldwide for general purpose data collection and is used on a number of beamlines worldwide for collection of both microdiffraction and microfluorescence data.

Because the field has been revolutionized by new sources and instrumentation, there is increased urgency to further automate and standardize the data collection of both microfluorescence and microdiffraction analysis. The actual data collection program can be quite simple for both microfluorescence and microdiffraction. However, data analysis can be complicated, especially for microdiffraction. Improved computer processing and storage power is certain to make a major impact on the way data is collected, analyzed and stored. For example, to avoid the storage of huge data files, fluorescence spectra are often stored as "regions of interest." This procedure works well in many cases, but can lose important information for overlapping characteristic lines. A better solution is to record complete spectra or to fit the spectra "on the fly" to separate overlapping peaks. The availability of large storage devices and faster computers make these intrinsically superior options increasingly attractive.

Until recently, there has been little or no specialized software for x-ray microdiffraction. With monochromatic microdiffraction experiments, a key problem is the sphere of confusion of the diffractometer. Recent efforts have resulted in the development of automated methods of mapping the sphere of confusion of a diffractometer, to greatly reduce the experimental uncertainty of the relative beam/sample position (Noyan et al., 1999).

With microdiffraction experiments, a key challenge is automated indexing of Laue reflections. This quite general problem appears feasible for wide-band-pass x-ray beams and for samples where the unit cell is known (but not the grain orientation or strain) (Marcus et al., 1996; Chung, 1997; Wenk et al., 1997; Chung and Ice, 1999). Recent advances in specialized computer software for pattern recognition and microdiffraction data analysis has now demonstrated the ability to automatically separate the overlapping Laue patterns from 5 to 20 grains and to automatically determine strain to 2×10^{-5} to 2×10^{-4} and to localize the diffracting grain position along the incident beam to ~2 μm. The key elements of the so called ORDEX (Oak Ridge inDEXing) program used for automated indexing of overlapping Laue patterns on beamline 7 of the Advanced Photon Source have been described by Chung and Ice (1999). Parallel software development efforts are also underway at the European Synchrotron Radiation Facility and at the Advanced Light Source.

SAMPLE PREPARATION

One major advantage of x-ray microprobe analysis is the minimal sample preparation required. In general, samples need to be small due to the short working distances between the focusing optics and the focal plane. Samples should also be mounted on low background-producing materials to reduce x-ray scatter and fluorescence (e.g., Mylar or Kapton for fluorescence and single-crystal Si for diffraction). In most cases, no additional preparation is required, and samples can be run in air, under water, or coated with low-Z barriers for radioactive containment or to prevent oxidation. Of course, what the x-ray microbeam measures is the elemental and crystallographic distributions in the sample as it is mounted. If, for example, a sample is machined, its surface will include a heavily damaged and chemically contaminated layer that may or may not be of interest to the experiment. In addition, because of the sensitivity of x-ray microfluorescence to trace elements, particular care must be taken during sample mounting to prevent contamination of the sample. For example, samples should be handled with low-Z plastic tweezers to avoid metal contamination from metallic tweezers.

Surface roughness and variations in sample thickness can also complicate the data analysis. In general, the analysis of fluorescence data is easier if the sample is homogeneous, smooth, and of uniform thickness. If the sample is very thick (semi-infinite) or very thin (negligible absorption), the data analysis is particularly easy. For most microdiffraction experiments to date, relative reflectivities are sufficient, so absorption corrections are not critical to the success of the experiment.

To aid in sample throughput, some microbeam instruments have mounting systems that allow the sample to be mounted on an optical microscope remote from the x-ray microprobe. The position of the sample with respect to a fiducial in the remote microscope holder can be accurately reproduced on the x-ray microbeam positioning stage. Features of interest can be identified off line and their coordinates noted for later x-ray microcharacterization.

DATA ANALYSIS AND INITIAL INTERPRETATION

Microfluorescence

For many samples, x-ray microflurorescence can yield rather direct information about trace element distributions. X-ray fluorescence data can be placed on an absolute scale if the approximate composition of the sample is understood and the incident flux known/measured. With a solid-state detector, the major components of the sample are readily identified for $Z > 12$. The composition can be used to estimate the linear absorption coefficient and therefore the sample volume probed by the x-ray beam. If the matrix is known, the absorption can be calculated from standard tables or can be found on various web sites. Because of uncertainties in the absorption coefficients of x-rays, first-principles methods are only good to ~5% (Sparks, 1980; Lachance and Claisse, 1995). More precise absolute measurements can be obtained with prepared standards. Standard samples can be obtained from the National Institutes of Standards and Technology (NIST) or by careful dilution of trace elements in a matrix of known composition.

Initial fluorescence data analysis can be made by simply selecting a region of interest that incorporates the fluorescence lines of interest and nearby regions for background subtraction. The elemental concentration in the probe volume can be approximated from the estimated linear absorption coefficients of the incident and fluorescence beams and the net counts in the line. Sparks (1980) gives a general formula for the mass concentration of element z, C_z, as a function of the experimental geometry

$$C_z = \frac{I_Z 4\pi R^2(\mu_{S,0} + \mu_{S,i}\sin\psi/\sin\phi)\rho_S}{P_0 D_Z \sigma_{zij}\langle 1 - \exp\{1[(\mu_{S,0} + \mu_{S,i}\sin\psi/\sin\phi)/\rho_S]\rho_S T\sin\psi\}\rangle} \quad (5)$$

Here R is the specimen-to-detector distance, I_Z is the fluorescence intensity per area at the detector, $\mu_s,_0/\rho_s$ and $\mu_{s,i}/\rho_s$ are the sample mass absorption coefficients for the incident and fluorescence radiation, P_0 is the incident power, D_Z is an absorption correction for attenuation between the sample and the detector, σ_{zij} is the fluorescence cross-section for the exciting energy, and $\rho_s T$ is the mass per unit area of the sample.

The precision of the measurement can be improved by fitting the fluorescence lines with linear (assumed shape) or nonlinear least-squares techniques. All fluorescence lines have a natural Lorenztian shape which is a consequence of the lifetime of the inner-shell holes that give rise to the fluorescence. For most measurements, however, the Lorenztian shape is dominated by the Gaussian resolution of the wavelength- or energy-dispersive spectrometer. Fluorescence lines are therefore typically fit to Gaussian or Voight functions. The Lorentzian tail of a fluorescence line is mainly of importance for estimating the background it contributes under a weak line.

Microdiffraction

With a monochromatic x-ray beam, relative texture and strain information can be obtained rather directly from measured microdiffraction scattering angle and the x-ray wavelength (see X-RAY POWDER DIFFRACTION, XAFS SPECTROSCOPY and RESONANT SCATTERING TECHNIQUES). X-ray microdiffraction can measure absolute strain to better than 1 part in 10^5, but is often used for high-precision measurements of strain or rotation differences, rather than absolute measurements. If the sample grain size is sufficiently small compared to the probe volume, standard powder diffraction methods can be used to determine the lattice constants of the sample (X-RAY POWDER DIFFRACTION). However, it is useful to utilize a position-sensitive area detector to collect a large fraction of the Debye ring for data analysis. Each powder ring can be fit to optimize the statistical determination of the d spacing. The shape and intensity distribution of each ring also can be used to help determine the local texture and strain tensor of the sample. Standard samples are particularly useful for putting strain measurements on an absolute scale.

With white-beam microdiffraction, each grain intercepted by the beam produces a Laue pattern. This ensures that the deviatoric strain tensor information can be recovered for each grain, but complicates the analysis, since many grains simultaneously contribute to the Laue spectra. Often grains from various layers can be identified according to characteristics of the layer (i.e., intense sharp peaks from large grains, fuzzy weak peaks from small grains). Whatever the process, the key step is to identify which reflections are associated with a single grain and to index the reflections for each grain. This problem has now been solved for many systems with an automated indexing program that can simultaneously index ~5 to 20 grains or subgrains (Chung and Ice, 1999). With this program, the detector sample alignment is first calibrated using a perfect crystal of Ge, then the program automatically determines the grain orientation and strain. Typically, the orientation can be determined to better than 0.01° and the strain can be determined to less than 1 part in 10^4. This precision already is better than for virtually all other probes, and can be further refined in some directions by moving the detector back both to localize the origin of the diffracted beam and also to gain angular precision in the measurement.

PROBLEMS

As described previously, the two key problems of x-ray microbeam analysis are the limited number of useful sources and the difficulty of fabricating efficient x-ray

optics. In addition to these two problems, there are several annoying problems that complicate the use of x-ray microbeam techniques. One difficulty with x-ray microbeam experiments is identification of the x-ray beam position on the sample. Because of the small beam size, even intense x-ray microbeams are difficult to view on fluorescent screens with visible light optics, and even when they are visualized, the transfer from beam position on a fluorescence screen to beam position on a sample can be difficult. For some experiments, it is necessary to place markers on the sample to help in locating the x-ray beam. For example, a cross-hair of a fluorescing heavy metal can be used to locate the beam position with respect to an interesting region.

Another annoying problem with x-ray microbeam analysis is the difficulty of monitoring the absolute beam intensity on the sample. Because the distance from the focusing optics to the focus is typically short, there is little room to install a transmission beam monitor. Even when such a monitor is installed, great care must be taken to avoid contamination of the monitor signal due to backscatter from the sample.

The problem of backscatter contamination into a transmission monitor is but one example of a general class of shielding problems that arise due to the proximity of sample, detector, and optics. In general, great care is required to reduce parasitic backgrounds associated with the beam path through x-ray optics and any air path to the sample. Typically, scatter from beam-defining slits and upstream condensing optics can swamp a CCD detector unless care is taken to shield against such direct scattering. The short working distance between the optics and the sample also require care to avoid collisions.

ACKNOWLEDGMENTS

Research sponsored in part by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory and the Division of Material Sciences, U.S. Department of Energy under contract DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation. Research carried out in part on beamline 10.3.1 at the ALS and on beamline 2ID at the APS. Both facilities are supported by the U.S. Department of Energy.

LITERATURE CITED

APS Users Meeting Workshop. 1997. Technical report: Making and using very small X-ray beams. APS, Argonne, Ill.

Aristov, V. V. 1988. X-ray Microscopy II. pp. 108–117. Bogorodski Pechatnik Publishing Company, Chernogolovka, Moscow Region.

Bambynek, W., Crasemann, B., Fink, R. W., Freund, H. U., Mark, H., Swift, C. D., Price, R. E., and Venugopala Rao, P. 1972. X-ray fluorescence yields, Auger and Coster-Kronig transition probabilties. *Rev. Mod. Phys.* 44:716–813.

Bilderback, D. H., Hoffman, S. A., and Thiel, D. J. 1994. Nanometer spatial resolution achieved in hard X-ray imaging and Laue diffraction experiments. *Science* 263:201–202.

Bionta, R. M., Ables, E., Clamp, O., Edwards, O. D., Gabriele, P. C., Miller, K., Ott, L. L., Skulina, K. M., Tilley, R., and Viada, T. 1990. Tabletop X-ray microscope using 8 keV zone plates. *Opt. Eng.* 29:576–580.

Boisseau, P. 1986. Determination of three dimensional trace element distributions by the use of monochromatic X-ray microbeams. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Mass.

Brennan, S., Tompkins, W., Takaura, N., Pianetta, P., Laderman, S. S., Fischer-Colbrie, A., Kortright, J. B., Madden, M. C., and Wherry, D. C. 1994. Wide band-pass approaches to total-reflection X-ray fluorescence using synchrotron radiation. *Nucl. Instrum. Methods A* 347:417–421.

Buras, B. and Tazzari, S. 1984. European Synchrotron Radiation Facility Report of European Synchrotron Radiation Project. Cern LEP Division, Geneva, Switzerland.

Busing, W. R. and Levy, H. A. 1967. Angle calculations for 3- and 4-circle X-ray and neutron diffractometers. *Acta Crystallogr.* 22:457–464.

Cambell, J. L. 1990. X-ray spectrometers for PIXE. *Nucl. Instrum. Methods* B 49:115–125.

Campbell, J. R., Lamb, R. D., Leigh, R. G., Nickel, B. G., and Cookson, J. A. 1985. Effects of random surface roughness in PIXE analysis of thick targets. *Nucl. Instrum. Methods B* 12:402–412

Chevallier, P. and Dhez, P. 1997. Hard X-ray microbeam: Production and application. *In* Accelerator Based Atomic Physics Techniques and Applications (S. M. Shafroth and J. Austin, eds.). pp. 309–348. AIP Press, New York.

Chevallier, P., Dhez, P., Legrand, F., Idir, M., Soullie, G., Mirone, A., Erko, A., Snigirev, A., Snigireva, I., Suvorov, A., Freund, A., Engstrom, P., Als. Nielsen, J., and Grubel, A. 1995. Microprobe with Bragg-Fresnel multilayer lens at ESRF beam line. *Nucl. Instum. Methods A* 354:584–587

Chevallier, P., Dhez, P., Legrand, F., Irko, A., Agafonov, Y., Panchenko, L. A., and Yakshin, A. 1996. The LURE-IMT photon microprobe. *J. Trace Microprobe Techniques* 14:517–540.

Chung, J.-S. 1997. Automated indexing of wide-band-pass Laue images. 1997. APS Users Meeting Workshop on Making and Using Very Small X-ray Beams. APS, Argonne, Ill.

Chung, J.-S. and Ice, G. E. 1998. Automated indexing of Laue images from polycrystalline materials. MRS Spring Symposium, San Francisco, Calif.

Chung, J. S. and Ice, G. E. 1999. Automated indexing for texture and strain measurement with broad-bandpass X-ray microbeams. *J. Appl. Phys.* 86:5249–5256.

Doyle, B. L., Walsh, D. S., and Lee, S. R. 1991. External micro-ion-beam analysis (X-MIBA). *Nucl. Instrum. Methods B* 54:244–257.

Goldstein. J. I. 1979. Principles of thin film x-ray microanalysis. *In* Introduction to Analytical Electron Microscopy (J. J. Hern, J. I. Goldstein, and D. C. Joy, eds.). pp. 83–120. Plenum, New York.

Hayakawa, S., Gohshi, Y., Iida, A., Aoki, S., and Ishikawa, M. 1990. X-ray microanalysis with energy tunable synchrotron x-rays. *Nucl. Instrum. Methods B* 49:555.

Hoffman, S. A., Thiel, D. J., and Bilderback, D. H. 1994. Applications of single tapered glass capillaries-submicron X-ray imaging and Laue diffraction. *Opt. Eng.* 33:303–306.

Howells, M. R. and Hastings, J. B. 1983. Design considerations for an X-ray microprobe. *Nucl. Instrum. Methods* 208:379–386.

Ice, G. E. 1987. Microdiffraction with synchrotron radiation. *Nucl. Instrum. Methods B* 24/25:397–399.

Ice, G. E. 1997. Microbeam-forming methods for synchrotron radiation. *X-ray Spectrom.* 26:315–326.

Ice, G. E. and Sparks, C. J., Jr. 1984. Focusing optics for a synchrotron X-radiation microprobe. *Nucl. Instrum. Methods* 222:121–127.

Ice, G. E. and Sparks, C. J. 1990. Mosaic crystal X-ray spectrometer to resolve inelastic background from anomalous scattering experiments. *Nucl. Inst. Methods* A291:110–116.

Ice, G. E. and Sparks, C. J. 1991. MICROCAT proposal to APS. APS, Argonne, Ill.

Jones, K. W. and Gordon, B. M. 1989. Trace element determinations with synchrotron-induced X-ray emission. *Anal. Chem.* 61:341A–356A.

Koumelis, C. N., Londos, C. A., Kavogli, Z. I., Leventouri, D. K., Vassilikou, A. B., and Zardas, G. E. 1982. *Can. J. Phys.* 60:1241–1246.

Lachance, G. R. and Claisse, F. 1995. Quantitative X-ray Fluorescence Analysis: Theory and Application. John Wiley & Sons, New York.

Langevelde, F. V., Tros, G. H. J., Bowen, D. K., and Vis, R. D. 1990. The synchrotron radiation microprobe at the SRS, Daresbury (UK) and its applications. *Nucl. Instrum. Methods B* 49:544–550.

Larson, B. C., Tamura, N., Chung, J.-S., Ice, G. E., Budai, J. D., Tischler, J. Z., Yang, W., Weiland, H., and Lowe, W. P. 1999. 3-D measurement of deformation microstructure in Al(0.2%) Mg using submicron resolution white X-ray microbeams. MRS Fall Symposium Proceedings. Materials Research Society, Warrendale, Pa.

Larsson, S. and Engström, P. 1992. X-ray microbeam spectroscopy with the use of capillary optics. *Adv. X-ray Anal.* 35:1019–1025.

Lindh, U. 1990. Micron and submicron nuclear probes in biomedicine. *Nucl. Instrum. Methods B* 49:451–464.

Marcus, M. A., MacDowell, A. A., Isaacs, E. D., Evans-Lutterodt, K., and Ice, G. E. 1996. Submicron resolution X-ray strain measurements on patterned films: Some hows and whys. *Mater. Res. Soc. Symp.* 428:545–556.

Materlik, G., Sparks, C. J., and Fischer, K. 1994. Resonant Anomalous X-ray Scattering. North Holland Publishers, Amsterdam, The Netherlands.

Michael, J. R. and Goehner, R. P. 1993. Crystallographic phase identification in scanning electron microscope: Backscattered electron Kikuchi patterns imaged with a CCD-based detector. *MSA Bull.* 23:168–175

Miller, M. K., Cerezo, A., Hetherington, M. G., and Smith, G. D. W. 1996. Atom Probe Field Ion Microscopy. Oxford Science Publications, Oxford.

Myers, B. F., Montogmery, F. C., and Partain, K. E. 1986. The transport of fission products in SiC. Doc. No. 909055, GA Technologies. General Atomics, San Diego.

Naghedolfeizi, M., Chung, J.-S., and Ice, G. E. 1998. X-ray fluorescence microtomography on a SiC nuclear fuel ball. *Mater. Res. Soc. Symp.* 524:233–240.

Noyan, I. C., Kaldor, S. K., Wang, P.-C., Jordan-Sweet, J. 1999. A cost-effective method for minimizing the sphere-of-confusion error of x-ray microdiffractometers. *Rev. Sci. Inst.* 70:1300–1304.

Perry, D. L. and Thompson, A. C. 1994. Synchrotron induced X-ray fluorescence microprobe studies of the copper-lead sulfide solution interface. *Am. Chem. Soc. Abstr.* 208:429.

Piotrowski, D. P., Stock, S. R., Guvenilir, A., Haase, J. D., and Rek, Z. U. 1996. High resolution synchrotron X-ray diffraction tomography of large-grained samples. *Mater. Res. Soc. Symp. Proc.* 437:125–128.

Rebonato, R., Ice, G. E., Habenschuss, A., and Bilello, J. C. 1989. High-resolution microdiffraction study of notch-tip deformation in Mo single crystals using X-ray synchrotron radiation. *Philos. Mag. A* 60:571–583.

Ren, S. X, Kenik, E. A., Alexander. K. B., and Goyal, A. 1998. Exploring spatial resolution in electron back-scattered diffraction (EBSD) experiments via Monte Carlo simulation. *Microsc. Microanal.* 4:15–22.

Riekel, C. 1992. Beamline for microdiffraction and micro small-angle scattering. *SPIE* 1740:181–190.

Riekel, C., Bosecke. P., and Sanchez del Rio, M. 1992. 2 high brilliance beamlines at the ESRF dedicated to microdiffraction, biological crystallography and small angle scattering. *Rev. Sci. Instrum.* 63:974–981.

Shenoy, G. K., Viccaro, P. J., and Mills, D. M. 1988. Characteristics of the 7 GeV advanced photon source: A guide for users. ANL-88-9 Argonne National Laboratory, Argonne, Ill.

Snigirev, A., Kohn, V., Snigireva, I., and Legeler, B. 1996. A compound refractive lens for focusing high-energy X-rays. *Nature (London)* 384:49–51.

Snigirev, A., Snigireva, I., Bosecke, P., Lequien, S., and Schelokov, I. 1997. High energy X-ray phase contrast microscopy using a circular Bragg-Fresnel lens. *Opt. Commun.* 135:378–384.

Sparks, C. J., Jr. 1980. X-ray fluorescence microprobe for chemical analysis. *In* Synchrotron Radiation Research (H. Winick and S. Doniach, eds.). pp. 459–512. Plenum, New York.

Sparks, C. J., Kumar, R., Specht, E. D., Zschack, P., Ice, G. E., Shiraishi, T., and Hisatsune, K. 1992. Effect of powder sample granularlity on fluorescent intensity and on thermal parameters in X-ray diffraction rietveld analysis. *Adv. X-ray Anal.* 35:57–62.

Stern, E. A., Kalman, Z., Lewis, A., and Lieberman, K. 1988. Simple method for focusing X-rays using tapered capillaries. *Appl. Opt.* 27:5135–5139.

Stock, S. R., Guvenilir, A., Piotrowski, D. P., and Rek, Z. U. 1995. High resolution synchrotron X-ray diffraction tomography of polycrstalline samples. *Mater. Res. Soc. Symp.* 375:275–280.

Thompson, A. C., Chapman, K. L., Ice, G. E., Sparks, C. J., Yun, W., Lai, B., Legnini, D., Vicarro, P. J., Rivers, M. L., Bilderback, D. H., and Thiel, D. J. 1992. Focusing optics for a synchrotron-based X-ray microprobe. *Nucl. Instrum. Methods A* 319:320–325.

Underwood, J. H., Thompson, A. C., Kortright, J. B., Chapman, K. C., and Lunt, D. 1996. Focusing x-rays to a 1 μm spot using elastically bent, graded multilayler coated mirrors. *Rev. Sci. Instrum. Abstr.* 67:3359.

Veigele, W. J., Briggs, E., Bates, L., Henry. E. M., and Bracewell, B. 1969. X-ray Cross Section Compilation from 0.1 keV to 1 MeV. Kaman Sciences Report No. DNA 2433F.

Wang, P. C., Cargill, G. S., Noyan, I. C., Liniger, E. G., Hu, C. K., and Lee, K. Y. 1996. X-ray microdiffraction for VLSI. *Mater. Res. Soc. Symp. Proc.* 427:35.

Wang, P. C., Cargill G. S., III, Noyan, I. C., Liniger, E. G., Hu, C. K., and Lee, K. Y. 1997. Thermal and electromigration strain distributions in 10 μm-wide aluminum conductor lines measured by x-ray microdiffraction. *Mater. Res. Soc. Symp. Proc.* 473:273.

Wenk, H. R., Heidelbach, F., Chadeigner, D., and Zontone, F. 1997. Laue orientation imaging. *J. Synchrotron Rad.* 4:95–101.

Wilkinson, A. J. 1997. Methods for determining elastic strains from electron backscattering diffraction and electron channeling patterns. *Mater. Sci. Tech.* 13:79.

Yaakobi, B. and Turner, R. E. 1979. Focusing X-ray spectrograph for laser fusion experiments. *Rev. Sci. Inst.* 50:1609–1611.

Yang, B. X., Rivers, M., Schildkamp, W., and Eng, P. J. 1995. GeoCARS microfocusing Kirkpatric-Baez bender development. *Rev. Sci. Instrum.* 66:2278–2280.

Yun, W. B., Lai, B., Legnini, D., Xiao, Y. H., Chrzas, J., Skulina, K. M., Bionta, R. M., White, V., and Cerrina, F. 1992. Performance comparison of hard X-ray zone plates fabricated by two different methods. *SPIE* 1740:117–129.

KEY REFERENCES

Boisseau, 1986. See above.

Most complete discussion of x-ray microfluorescence tomography with some early examples.

Chevallier and Dhez, 1997. See above.

Recent overview of x-ray microbeam science and hardware.

Chung and Ice, 1999. See above.

Derives the mathematical basis for white-beam microdiffraction measurements of the deviatoric and absolute strain tensor and includes a description of the ORDEX program, which automatically indexes multiple overlapping Laue patterns.

Ice, 1997. See above.

Recent overview of x-ray focusing optics for x-ray microprobes with methods for comparing the efficiency of various x-ray focusing optics.

Sparks, 1980. See above.

Quantitative comparison of x-ray microfluorescence analysis to electron and proton microprobes.

GENE E. ICE
Oak Ridge National Laboratory
Oak Ridge, Tennessee

X-RAY MAGNETIC CIRCULAR DICHROISM

INTRODUCTION

X-ray magnetic circular dichroism (XMCD) is a measure of the difference in the absorption coefficient ($\mu_c = \mu^+ - \mu^-$), with the helicity of incident circularly polarized light parallel (μ^+) and antiparallel (μ^-) to the local magnetization direction of the absorbing material. These measurements are typically made at x-ray energies spanning an absorption edge, as illustrated in Figure 1, where a deep atomic core electron is photoexited into the valence band. Because the binding energy of these core electrons is associated with specific elements, XMCD can be used to separately measure the magnetic contributions from individual constituents in a complex material, by tuning the incident beam energy to a specific edge. Furthermore, different absorption edges of the same element provide information on the magnetic contributions of different kinds of valence electrons within a single element.

XMCD measurements yield two basic types of magnetic information, both of which are specific to the element and orbital determined by the choice of measured absorption edge. First, the strength of the dichroic signal is proportional to the projection of the magnetic moment along the incident x-ray beam direction. Thus, the signal can be used to determine magnetization direction for a sample in remanence or changes in the sample magnetization upon varying the applied magnetic field or temperature. Secondly, XMCD can be used to separate the orbital $\langle L_z \rangle$ and spin $\langle S_z \rangle$ contributions to the magnetic moment. This is possible through sum rules (Thole et al., 1992; Carra et al., 1993) that relate these contributions to the integrated dichroic signal. It is this elemental and orbital specificity and the ability to deconvolve the magnetic moment into its orbital and spin components that makes XMCD a unique tool for the study of magnetic materials.

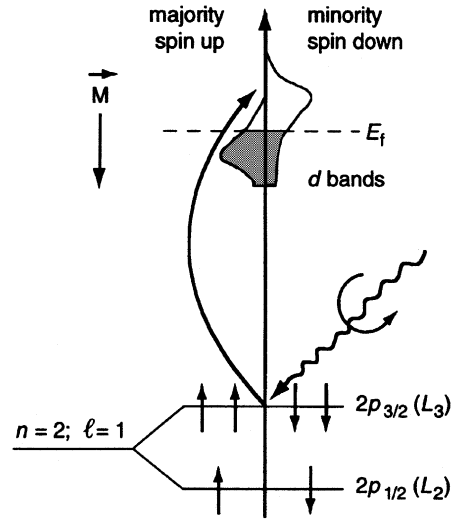


Figure 1. Schematic illustration of the energy bands and transitions involved in XMCD measurements. A circularly polarized photon excites a photoelectron from a spin-orbit-split core state to a spin-polarized valence band.

In discussing x-ray spectra, one must refer quite frequently to the x-ray absorption edges; the terminology for labeling these edges should therefore be made clear from the outset of this discussion. The common convention has been to denote the principal quantum number ($n = 1, 2, 3, 4, \ldots$) of the core state by K, L, M, N,... and the total angular momentum ($s_{1/2}$, $p_{1/2}$, $p_{3/2}$, $d_{3/2}$, etc.) of the core state by 1, 2, 3, 4,.... Therefore, an edge referred to as the L_3 edge would denote an absorption threshold with an initial $2p_{3/2}$ core state. Table 1 lists the elements and edges for which XMCD measurements have been taken along with the corresponding transition, energy range, and reference to a typical spectrum. Most XMCD work in the soft x-ray regime (E < 2000 eV) has centered on the $3d$ transition metal (TM) $L_{2,3}$ edges and rare earth (RE) $M_{4,5}$ edges. These edges probe the largely localized TM $3d$ and RE $4f$ states, which are the primary magnetic electrons in almost all magnetic materials. The relatively large magnetic moments of these states lead to large dichroic effects in this energy regime. Hard x-ray measurements ($E > 2000$ eV) have focused on the TM K edge and the RE $L_{2,3}$ edges, which probe much more diffuse bands with small magnetic moments leading to smaller dichroic effects. Although the edges probed and thus the final states are different for these regimes, the physics and

Table 1. Representative Sample of the Elements and Edges Measured Using XMCD

Elements	Edges	Orbital Character of Transitions	Energy Range (keV)	Reference(s) Showing Typical Spectra
3*d* (Mn-Cu)	*K*	$s \rightarrow p$	6.5–9.0	Stahler et al. (1993)
3*d* (V-Cu)	$L_{2,3}$	$p \rightarrow d$	0.5–1.0	Rudolf et al. (1992); Wu et al. (1992)
4*d* (Rd, Pd)	$L_{2,3}$	$p \rightarrow d$	3.0–3.5	Kobayashi et al. (1996)
4*f* (Ce-Ho)	$M_{4,5}$	$d \rightarrow f$	0.9–1.5	Rudolf et al. (1992); Schillé et al. (1993)
4*f* (Ce-Lu)	$L_{2,3}$	$p \rightarrow d$	5.7–10.3	Fischer et al. (1992); Lang et al. (1992)
5*d* (Hf-Au)	$L_{2,3}$	$p \rightarrow d$	9.5–13.7	Schütz et al. (1989a)
5*f* (U)	$M_{4,5}$	$d \rightarrow f$	3.5–3.7	Finazzi et al. (1997)

analysis of the spectra is essentially the same, and they differ only in the techniques involved in spectrum collection.

This unit will concentrate on the techniques employed in the hard x-ray regime, but considerable mention will be made of soft x-ray results and techniques because of the fundamental importance in magnetism of the final states probed by edges in this energy regime. Emphasis will be given to the methods used for the collection of XMCD spectra rather than to detailed analysis of specific features in the XMCD spectra of a particular edge. More detailed information on the features of XMCD spectra at specific edges can be found in other review articles (Krill et al., 1993; Stahler et al., 1993; Schütz et al., 1994; Stöhr and Wu, 1994; Pizzini et al., 1995).

Competitive and Related Techniques

Many techniques provide information similar to XMCD, but typically they tend to give information on the bulk macroscopic magnetic properties of a material rather than microscopic element–specific information. It is apparent that the basic criteria for choosing XMCD measurements over other techniques are the need for this element-specific information or for a separate determination of the orbital and spin moments. The need to obtain this information must be balanced against the requirements that XMCD measurements must be performed at a synchrotron radiation facility, and that collection of spectra can take several hours. Therefore, the amount of parameter space that can be explored via XMCD experiments is restricted by the amount of time available at the beamline.

The methods most obviously comparable to XMCD are the directly analogous laser-based methods such as measurement of Faraday rotation and magneto-optical Kerr effect. These techniques, while qualitatively similar to XMCD, are considerably more difficult to interpret. The transitions involved in these effects occur between occupied and unoccupied states near the Fermi level, and therefore the initial states are extended bands that are much more difficult to model than the atomic-like core states in XMCD measurements. In fact, recent efforts have focused on using XMCD spectra as a basis for understanding the features observed in laser-based magnetic measurements. Another disadvantage of these measurements is that they lack the element specificity provided by XMCD spectra, because they involve transitions between all the conduction electron states. A final potential drawback is the surface sensitivity of the measurements.

Because the size of the dichroic signal scales with sample magnetization, XMCD spectra can also be used to measure relative changes in the magnetization as a function of temperature or applied field. This information is similar to that provided by vibrating sample or SQUID magnetometer measurements. Unlike a magnetometer, which measures the bulk magnetization for all constituents, dichroism measures the magnetization of a particular orbital of a specific element. Therefore, XMCD measurements can be used to analyze complicated hysteresis behavior, such as that encountered in magnetic multilayers, or to measure the different temperature dependencies of each component in a complicated magnetic compound. That these measurements are generally complementary to those obtained with a magnetometer should nevertheless be stressed, since it is frequently necessary to use magnetization measurements to correctly normalize the XMCD spectra taken at fixed field.

The size of the moments and their temperature behavior can also be ascertained from magnetic neutron or nonresonant magnetic x-ray diffraction. These two techniques are the only other methods that can deconvolve the orbital and spin contributions to the magnetic moment. For neutron diffraction, separation of the moments can be accomplished by measuring the magnetic form factor for several different momentum transfer values and fitting with an appropriate model, while for non-resonant magnetic x-ray scattering the polarization dependence of several magnetic reflections must be measured (Blume and Gibbs, 1988; Gibbs et al., 1988). The sensitivity of both these techniques, however, is very limited and thus measurements have been restricted to demonstration experiments involving the heavy rare earth metals with large orbital moments ($\mu_{orb} \sim 6\,\mu_B$). XMCD on the other hand provides orbital moment sensitivity down to $\sim 0.005\,\mu_B$ (Samant et al., 1994). Furthermore, for compound materials these measurements will not be element specific, unless a reflection is used for which only one atomic species produces constructive interference. This is usually only possible

for simple compounds and generally not possible for multilayered materials. As diffractive measurements, these techniques have an advantage over XMCD in that they are able to examine antiferromagnetic as well as ferromagnetic materials. But, the weakness of magnetic scattering relative to charge scattering for nonresonant x-ray scattering makes the observation of difference signals in ferromagnetic compounds difficult.

A diffractive technique that does provide the element specificity of XMCD is x-ray resonant exchange scattering (XRES; Gibbs et al., 1988). XRES measures the intensity enhancement of a magnetic x-ray diffraction peak as the energy is scanned through a core hole resonance. This technique is the diffraction analog of dichroism phenomena and involves the same matrix elements as XMCD. In fact, theoretical XRES calculations reduce to XMCD at the limit of zero momentum transfer. At first glance, this would seem to put XMCD measurements at a disadvantage to XRES measurements, because the former is just a simplification of the later. It is this very simplification, however, that makes possible the derivation of the sum rules relating the XMCD spectrum to the orbital and spin moments. Most XRES measurements in the hard x-ray regime have been limited to incommensurate magnetic Bragg reflections, which do not require circularly polarized photons. In these cases, XRES measurements are more analogous to linear dichroism, for which a simple correlation between the size of the moments and the dichroic signal has not been demonstrated. Lastly, even though the XRES signal is enhanced over that obtained by nonresonant magnetic scattering, it is still typically lower by a factor of 10^4 or more than the charge scattering contributions. Separating signals this small from the incoherent background is difficult for powder diffraction methods, and therefore XRES, and magnetic x-ray scattering in general, have been restricted to single-crystal samples.

Spin-polarized photoemission (SPPE) measures the difference between the emitted spin-up and spin-down electrons upon the absorption of light (McIlroy et al., 1996). SPPE spectra can be related to the spin polarization of the occupied electron states, thereby providing complimentary information to XMCD measurements, which probe the unoccupied states. The particular occupied states measured depend upon the incident beam energy, UV or soft x-ray, and the energy of the photoelectron. Soft x-ray measurements probe the core levels, thereby retaining the element specificity of XMCD but only indirectly measuring the magnetic properties, through the effect of the exchange interaction between the valence band and the core-level electrons. UV measurements, on the other hand, directly probe the partially occupied conduction electron bands responsible for the magnetism, but lose element specificity, which makes a clear interpretation of the spectra much more difficult. The efficiency of spin-resolving electron detectors for these measurements, however, loses 3 to 4 orders of magnitude in the analysis of the electron spin. This inefficiency in resolving the spin of the electron has prevented routine implementation of this technique for magnetic measurements. A further disadvantage (or advantage) of photoemission techniques is their surface sensitivity. The emitted photoelectrons typically come only from the topmost 1 to 10 monolayers of the sample. XMCD, on the other hand, provides a true bulk characterization because of the penetrating power of the x-rays. Magnetic sensitivity in photoemission spectra has also been demonstrated using circularly polarized light (Baumgarten et al., 1990). This effect, generally referred to as photoemission MCD, should not be confused with XMCD, because the states probed and the physics involved are quite different. While this technique eliminates the need for a spin-resolving electron detector, it does not provide a clear relationship between measured spectra and size of magnetic moments.

PRINCIPLES OF THE METHOD

XMCD measures the difference in the absorption coefficient over an absorption edge upon reversing the sample magnetization or helicity of the incident circularly polarized photons. This provides magnetic information that is specific to a particular element and orbital in the sample. Using a simple one-electron theory it is easy to demonstrate that XMCD measurements are proportional to the net moment along the incident beam direction and thus can be used to measure the variations in the magnetization of a particular orbital upon application of external fields or change in the sample temperature. Of particular interest has been the recent development of sum rules which can be used to separate the orbital and spin contributions to the magnetic moment. XMCD measurements are performed at synchrotron radiation facilities, with the specific beamline optics determined by the energy regime of the edge of interest.

Basic Theory of X-Ray Circular Dichroism

Although a number of efforts had been made over the years to characterize an interplay between x-ray absorption and magnetism, it was only in 1986 that the first unambiguous evidence of magnetization-sensitive absorption was observed in a linear dichroism measurement at the $M_{4,5}$ edges of Tb metal (Laan et al., 1986). This measurement was quickly followed by the first observation of XMCD at the K edge of Fe and, a year later, of an order-of-magnitude-larger effect at the L edges of Gd and Tb (Schütz et al., 1987, 1988). Large XMCD signals directly probing the magnetic $3d$ electrons were then discovered in the soft x-ray regime at the $L_{2,3}$ edge of Ni (Chen et al., 1990).

The primary reason for the lack of previous success in searching for magnetic contributions to the absorption signal is that the count rates, energy tunability, and polarization properties required by these experiments make the use of synchrotron radiation sources absolutely essential. Facilities for the production of synchrotron radiation, however, were not routinely available until the late 1970s. The mere availability of these sources was clearly not sufficient, because attempts to observe XMCD had been made prior to 1988 without success (Keller, 1985). It was only as the stability of these sources increased that the systematic errors which can easily creep into difference measurements were reduced sufficiently to allow the XMCD effect to be observed.

The basic cause of the enhancement of the dichroic signal that occurs near an absorption edge can be easily understood using a simple one-electron model for x-ray absorption (see Appendix A). While this model fails to explain all of the features in most spectra, it does describe the fundamental interaction that leads to the dichroic signal, and more complicated analyses of XMCD spectra are typically just perturbations of this basic model. In this crude treatment of x-ray absorption, the transition probabilities between spin-orbit-split core states and the spin-polarized final states are calculated starting from Fermi's golden rule. The most significant concept in this analysis is that the fundamental cause of the XMCD signal is the spin-orbit splitting of the core state. This splitting causes the difference between the matrix elements for the left or right circularly polarized photons. In fact, the presence of a spin-orbit term is a fundamental requirement for the observation of any magneto-optical phenomena regardless of the energy range. Without it no dichroic effects would be seen.

Specifically, the example illustrated in Appendix A demonstrates that, for an initial $2p_{1/2}$ (L_2) core state, left (right) circularly polarized photons will make preferential transitions to spin-down (-up) states. The magnitude of this preference is predicted to be proportional to the difference between the spin-up and spin-down occupation of the probed shell and as such the net moment on that orbital. Therefore the measured signal scales with the local moment and can be used to determine relative changes in the magnetization upon applying a magnetic field or increasing the sample temperature. Furthermore, the XMCD signal for an L_3 edge is also predicted to be equal and opposite that observed at an L_2 edge. This 1:−1 dichroic ratio is a general rule predicted for any pair of spin-orbit-split core states [i.e., $2p_{1/2,\ 3/2}$ ($L_{2,3}$), $3d_{3/2,\ 5/2}$ ($M_{4,5}$), etc.].

Although this simple model describes the basic underlying cause of the XMCD effect, it proves far too simplistic to completely explain most dichroic spectra. In particular, this model predicts no dichroic signal from K-edge measurements, because the initial state at this edge is an s level with no spin-orbit coupling. Yet it is at this edge that the first (admittedly weak) XMCD effects were observed. Also, while experimental spectra for spin-orbit-split pairs have been found to be roughly in the predicted 1:−1 ratio, for most materials they deviate considerably. To account for these discrepancies, more sophisticated models need to incorporate the following factors, which are neglected in the simple one-electron picture presented.

1. The effects of the spin-orbit interaction in the final states.
2. Changes in the band configuration due to the presence of the final state core hole.
3. Spin-dependent factors in the radial part of the matrix elements.
4. Contributions from higher-order electric quadrupole terms, which can contribute significantly to the XMCD signal for certain spectra.

The first two of these factors are significant for all dichroic spectra, while the later two are particularly relevant to RE L-edge spectra.

First consider the effects of the spin-orbit interaction in the final states. The inclusion of this factor can quickly explain the observed dichroic signal at TM K-edge spectra (i.e., repeat the example in Appendix A with initial states and final states inverted). This observation of a K-edge dichroic signal illustrates the extreme sensitivity of the XMCD technique, because not only do $4p$ states probed by this edge posses a relatively small spin moment ($\sim 0.2\ \mu_B$), but the substantial overlap with neighboring states results in a nearly complete quenching of the orbital moment. This orbital contribution ($\sim 0.01\ \mu_B$), however, is nonzero, and thus a nonzero dichroic signal is observed.

The influence of the final state spin-orbit interaction also explains the deviations from the predicted 1:−1 ratio observed at L and M edges. The spin-orbit term tends to enhance the L_3- (M_5-) edge dichroic signal at the expense of the L_2 (M_4) edge. In terms of the simple one-electron model presented, the spin-orbit interaction effectively breaks the degeneracy of the m_l states, resulting in an enhancement of the XMCD signal at the L_3 edge. An example of this enhancement is shown in Figure 2, which plots XMCD spectra obtained at the Co L edges for metallic Co and for a Co/Pd multilayer (Wu et al., 1992). The large enhancement of the L_3-edge dichroic signal indicates that the multilayer sample possesses an enhanced Co $3d$ orbital moment compared to that of bulk Co.

A quantitative relationship between the degree of this enhancement and the strength of the spin-orbit coupling is expressed in terms of sum rules that relate the integrated dichroic (μ_c) and total (μ_o) absorptions over a

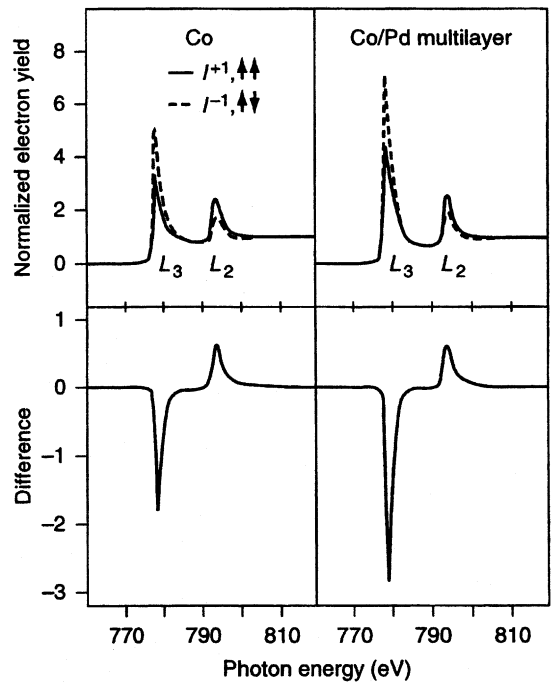


Figure 2. Normal and dichroic absorption at the Co $L_{2,3}$ edges for Co metal and Co/Pd multilayers. Courtesy of Wu et al. (1992).

spin-orbit-split pair of core states to the orbital and spin moments of that particular final state orbital (Thole et al., 1992; Carra et al., 1993). The integrated signal rather than the value at a specific energy is used in order to include all the states of a particular band. This has the added benefit of eliminating the final state core-hole effects, thereby yielding information on the ground state of the system. Although the general expressions for the sum rules can appear complicated, at any particular edge they reduce to extremely simple expressions. For instance, the sum rules for the $L_{2,3}$ edges simplify to the following expressions:

$$\frac{\int_{L_3+L_2} \mu_c dE}{\int_{L_3+L_2} 3\mu_o dE} = \frac{\langle L_z \rangle / 2}{(10 - n_{occ})}, \quad (1)$$

$$\frac{\int_{L_3} \mu_c dE - 2\int_{L_2} \mu_c dE}{\int_{L_3+L_2} 3\mu_o dE} = \frac{\langle S_z \rangle + 7\langle T_z \rangle}{(10 - n_{occ})} \quad (2)$$

Here $\langle L_Z \rangle$, $\langle S_Z \rangle$, and $\langle T_Z \rangle$ are the expectation values for the orbital, spin, and spin magnetic dipole operators in Bohr magnetons [$\boldsymbol{T} = \frac{1}{2}(\boldsymbol{S} - 3\hat{\boldsymbol{r}}(\hat{\boldsymbol{r}} \cdot \boldsymbol{S})$] and n_{occ} is the occupancy of the final-state band (i.e., $10 - n_{occ}$ corresponds to the number of final states available for the transition). The value of n_{occ} must usually be obtained via other experimental methods or from theoretical models of the band occupancy, which can lead to uncertainties in the values of the moments obtained via the sum rules. To circumvent this, it is sometimes useful to express the two equations above as a ratio:

$$\frac{\int_{L_3+L_2} \mu_c dE}{\int_{L_3} \mu_c dE - 2\int_{L_2} \mu_c dE} = \frac{\langle L_z \rangle}{2\langle S_z \rangle + 14\langle T_z \rangle} \quad (3)$$

This yields an expression that is independent of the shell occupancy and also eliminates the need to integrate over the total absorption, which can also be a source of systematic error in the measurement. Comparison with magnetization measurements can then be used to obtain the absolute value of the orbital moment. Applying the sum rules to the spectra shown in Figure 2 yields values of 0.17 and 0.24 μ_B for the orbital moments of the bulk Co and Co/Pd samples, respectively. This example again illustrates the sensitivity of the XMCD technique, because the clearly resolvable differences in the spectra correspond to relatively small changes in the size of the orbital component.

Several assumptions go into the derivation of the sum rules, which can restrict their applicability to certain spectra; the most important of these is that the two spin-orbit split states must be well resolved in energy. This criterion means that sum rule analysis is restricted to measurements involving deep core states and is generally not applicable to spectra taken at energies below ~250 eV. Moreover, the radial part of the matrix elements is assumed to be independent of the electron spin and identical at both edges of a spin-orbit-split pair. This is typically not the case for the RE L edges, where the presence of the open $4f$ shell can introduce some spin dependence (Konig et al., 1994).

Also, the presence of the magnetic dipole term $\langle T_Z \rangle$ in the these expressions poses problems for determining the exact size of the spin moment or the orbital-to-spin ratio. The importance of the size of this term has been a matter of considerable debate (Wu et al., 1993; Wu and Freeman, 1994). Specifically, it has been found that, for TM $3d$ states in a cubic or near-cubic symmetry, the $\langle T_Z \rangle$ term is small and can be safely neglected. For highly anisotropic conditions, however, such as those encountered at interfaces or in thin films, $\langle T_Z \rangle$ can become appreciable and therefore would distort measurements of the spin moment obtained from the XMCD sum rules. For RE materials, $\langle T_Z \rangle$ can be calculated analytically for the $4f$ states, but an exact measure of the relative size of this term for the $5d$ states is difficult to obtain.

PRACTICAL ASPECTS OF THE METHOD

The instrumentation required to collect dichroic spectra naturally divides XMCD measurements into two general categories based on the energy ranges involved: those in the soft x-ray regime, which require a windowless UHV-compatible beamline, and those in the hard x-ray regime, which can be performed in nonevacuated environments. The basic elements required for an XMCD experiment, however, are similar for both regimes: i.e., a source of circularly polarized x-rays (CPX), an optical arrangement for monochromatizing the x-ray beam, a magnetized sample, and a method for detecting the absorption signal. The main differences between the two regimes, other than the sample environment, lie in the methods used to detect the XMCD effect.

Circularly Polarized X-Ray Sources

In evaluating the possible sources of CPX, the following quantities are desirable; high circular polarization rate (P_c), high flux (I), and the ability to reverse the helicity. For a given source, though, the experimenter must sometimes sacrifice flux in order to obtain a high rate of polarization, or vice versa. Under these circumstances, it should be kept in mind that the figure of merit for circularly polarized sources is $P_c^2 I$, because it determines the amount of time required to obtain measurements of comparable statistical accuracy on different sources (see Appendix B). Laboratory x-ray sources have proved to be impractical for XMCD measurements because they offer limited flux and emit unpolarized x-rays. The use of a synchrotron radiation source is therefore essential for performing XMCD experiments. Three different approaches can be used to obtain circularly polarized x-rays from a synchrotron source: using (1) off-axis bending magnet radiation, (2) a specialized insertion device, such as an elliptical multipole wiggler, or (3) phase retarders based on perfect crystal optics.

In a bending magnet source, a uniform circular motion by the relativistic electrons or positrons in the storage ring is used to generate synchrotron radiation. This radiation, when observed on the axis of the particle orbit, is purely linearly polarized in the orbital plane (σ polarization).

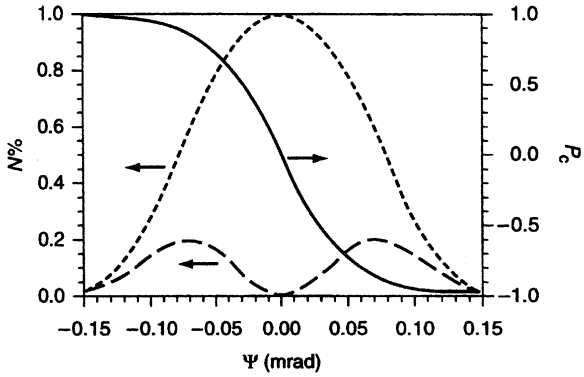


Figure 3. Horizontally (short dashes) and vertically (long dashes) polarized flux along with circular polarization rate (solid) from an APS bending magnet at 8.0 keV.

Appreciable amounts of photons polarized out of the orbital plane (π polarization) are only observed off-axis (Jackson, 1975). These combine to produce a CPX source because the σ- and π-polarized photons are exactly $\delta = \pm\pi/2$ radians out of phase. The sign of this phase difference depends on whether the radiation is viewed from above or below the synchrotron orbital plane. Therefore, the off-axis synchrotron radiation will be elliptically polarized, with a helicity dependent on the viewing angle and a degree of circular polarization given by

$$P_c = \frac{2E_\sigma E_\pi}{|E_\sigma|^2 + |E_\pi|^2} \sin\delta, \qquad (4)$$

where E_σ and E_π are the electric field amplitudes of the radiation. An example of this is shown in Figure 3, which plots the σ and π intensities along with the circular polarization rate as a function of viewing angle for 8.0 keV x-rays at an Advanced Photon Source bending magnet (Shenoy et al., 1988).

Although simple in concept, obtaining CPX in this fashion does have some drawbacks. The primary one is that the off-axis position required to get appreciable circular polarization reduces the incident flux by a factor of 5 to 10. Furthermore, measurements taken at the sides of the synchrotron beam where the intensity changes more rapidly are particularly sensitive to any motions of the particle beam. Moreover, the photon helicity cannot be changed easily because this requires moving the whole experimental setup vertically. This movement can result in slightly different Bragg angles incident on the monochromator, thereby causing energy shifts that would have to be compensated for. Attempts have been made to overcome this by using slits, which define beams both above and below the orbital plane simultaneously in order to make XMCD measurements. These efforts have had limited success, however, since they are particularly sensitive to sample inhomogeneity (Schütz et al., 1989b).

Standard planar insertion devices are magnetic arrays placed in the straight sections of synchrotron storage rings to make the particle beam oscillate in the orbital plane. Because each oscillation produces synchrotron radiation, a series of oscillations greatly enhances the emitted flux over that produced by a single bending magnet source. These devices produce linearly polarized light on axis, but the off-axis radiation of a planar device, unlike that of a bending magnet, is not circularly polarized, because equal numbers of right- and left-handed bends in the particle orbit produce equal amounts of left- and right-handed circular polarization, yielding a net helicity of zero. Specialized insertion devices for producing circularly polarized x-rays give the orbit of the particle beam an additional oscillation component out of the orbital plane (Yamamoto et al., 1988). In this manner, the particle orbit is made to traverse a helical or pseudo-helical path to yield a net circular polarization on axis of these devices. When coupled with a low-emittance synchrotron storage ring, these devices can provide a high flux with a high degree of polarization. The main disadvantage of these devices is that their high cost has made the beamlines dedicated to them rather rare. Also, to preserve the circular polarization through monochromatization, specialized optics are required, particularly for lower energies (Malgrange et al., 1991).

Another alternative for producing CPX is using x-ray phase retarders. Phase retarders employ perfect crystal optics to transform linear to circular polarization by inducing a $\pm\pi/2$ radian phase shift between equal amounts of incoming σ- and π-polarized radiation. (Here σ and π refer to the intensities of the radiation polarized in and out of the scattering plane of the phase retarder, and should not be confused with the radiation emitted from a synchrotron source. Unfortunately, papers on both subjects tend to use the same notation; it is kept the same for this paper in order to conform with standard practice.) Equal σ and π intensities incident on the phase retarder are obtained by orienting its plane of diffraction at a 45° angle with respect to the synchrotron orbital plane. As the final optical element before the experiment, it offers the greatest degree of circular polarization incident on the sample. Thus far phase-retarding optics have only been developed for harder x-rays, with soft x-ray measurements restricted to bending-magnet and specialized-insertion-device CPX sources.

Typically materials in the optical regime exhibit birefringence over a wide angular range; in the x-ray regime, however, materials are typically birefringent only at or near a Bragg reflection. For the hard x-ray energies of interest in XMCD measurements, a transmission-phase retarder, shown in Figure 4, has proved to be the most suitable type of phase retarder (Hirano et al., 1993). In this

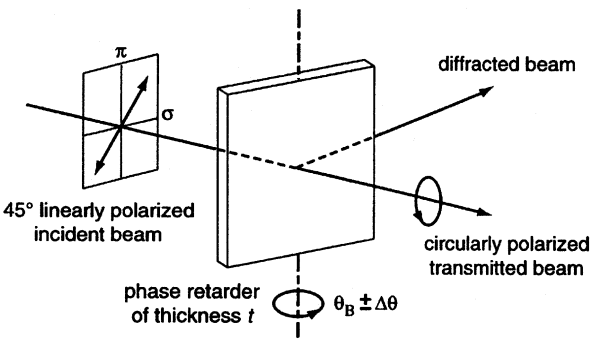


Figure 4. Schematic of a transmission phase retarder.

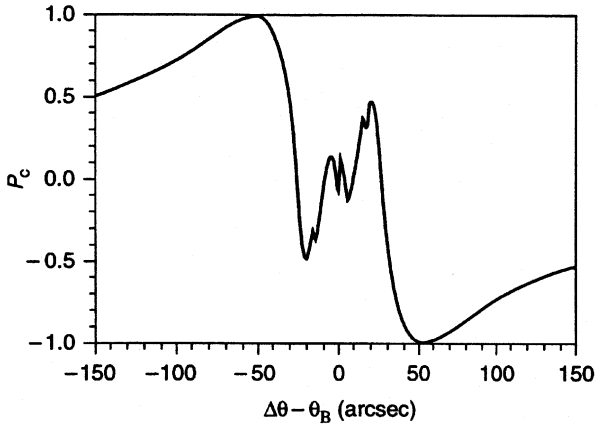


Figure 5. Calculated degree of circular polarization of the transmitted beam for a 375-μm-thick diamond (111) Bragg reflection at 8.0 keV.

phase retarder, a thin crystal, preferably of a low-Z material to minimize absorption, is deviated ($\Delta\theta \sim 10$ to 100 arcsec) from the exact Bragg condition, and the transmitted beam is used as the circularly polarized x-ray source. Figure 5 plots the predicted degree of circular polarization in the transmitted beam as a function of the off-Bragg position for a 375-μm-thick diamond (111) crystal and 8.0 keV incoming x-ray beam. Note that, for a particular crystal thickness and photon energy, either helicity can be obtained by simply reversing $\Delta\theta$. Further, when used with a low-emittance source, a high degree of polarization ($P_c > \sim 0.95$) can be achieved with a transmitted beam intensity of ~10% to 20% of the incident flux. This flux loss is comparable to that encountered using the off-axis bending magnet radiation. The phase retarder, however, provides easy helicity reversal and can be used with a standard insertion device to obtain fluxes comparable to or greater than those obtained from a specialized device (Lang et al., 1996). The main drawback of using a phase retarder as a CPX source is that the increased complexity of the optics can introduce a source of systematic errors. For instance, if the phase retarder is misaligned and does not precisely track with the energy of the monochromator, the $\pm\Delta\theta$ movements of the phase retarder will not be symmetric about the Bragg reflection. This leads to measurements with different P_c values and can introduce small energy shifts in the transmitted beam.

Detection of XMCD

The most common method used to obtain XMCD spectra is to measure the absorption by monitoring the incoming (I_o) and transmitted (I) fluxes. The absorption in the sample is then given by:

$$\mu_o t = \ln\left(\frac{I_o}{I}\right) \tag{5}$$

The dichroism measurement is obtained by reversing the helicity or magnetization of the sample and taking the difference of the two measurements ($\mu_c = \mu^+ - \mu^-$). Typically the thickness of the sample, t, is left in as a proportionality factor, because its removal requires an absolute determination of the transmitted flux at a particular energy. Some experimenters choose to remove this thickness factor by expressing the dichroism as a ratio between the dichroic and normal absorption, μ_c/μ_o. Absorption not due to resonant excitation of the measured edge, however, is usually removed by normalizing the pre-edge of the μ_o spectra to zero; therefore, expressing the XMCD signal in terms of this ratio tends to accentuate features closer to the absorption onset.

Rather than taking the difference of Eq. 5 for two magnetizations, the dichroic signal is also frequently expressed in terms of the asymmetry ratio of the transmitted fluxes only. This asymmetry ratio is related to the dichroic signal by Eq. 6.

$$\begin{aligned}\frac{I^+ - I^-}{I^+ + I^-} &= \frac{e^{-\mu^+ t} - e^{-\mu^- t}}{e^{-\mu^+ t} + e^{-\mu^- t}} = \frac{e^{0.5\mu_c t} - e^{-0.5\mu_c t}}{e^{0.5\mu_c t} + e^{-0.5\mu_c t}} \\ &= -\tanh(0.5\mu_c t) \cong -f0.5\mu_c t \end{aligned} \tag{6}$$

The last approximation can be made, because reasonable attenuations require $\mu_o t \sim 1$ to 2 and $\mu_c < \sim 0.05\,\mu_o t$. The factor f is introduced in an *ad hoc* fashion to account for the incomplete circular polarization in the incident beam and the finite angle (θ) between the magnetization and beam directions, $f = P_c \cos\theta$. In some cases, the factor f also includes the incomplete sample magnetization (M') relative to the saturated moment at $T = 0$ K. If the experimenter is interested in obtaining element specific magnetizations as a function of field or temperature, however, the inclusion of this factor would defeat the purpose of the measurement.

This type of measurement is illustrated in Figure 6, which shows a plot of a hysteresis measurement taken using the dichroic signal at the Co and Fe L_3 edges of a Co/Cu/Fe multilayer (Chen et al., 1993; Lin et al., 1993). In this compound the magnetic anisotropy of the Co layer is much greater than that of the Fe layer. Thus, in an applied magnetic field, the directions of Fe moments reverse before those of Co moments. By monitoring the strength of the dichroic signal as a function of applied field for the Fe and Co L_3 edges, the hysteresis curves of each constituent were traced out. Similarly, XMCD can also be used to measure the temperature variation of the magnetization of a specific constituent in a sample. This is demonstrated in Figure 7 (Rueff et al., 1997), which shows the variation of the dichroic signal at the Gd L edges as a function of temperature for a series of amorphous GdCo alloys. These XMCD measurements provide orbital- and element-specific information complementary to that obtained from magnetometer measurements.

Measurement Optics

Two basic optical setups, shown in Figure 8, are used for the measurement of x-ray attenuations. The scanning monochromator technique (Fig. 8a) selects a small energy range ($\Delta E \sim 1$ eV) to impinge on the sample, and the full

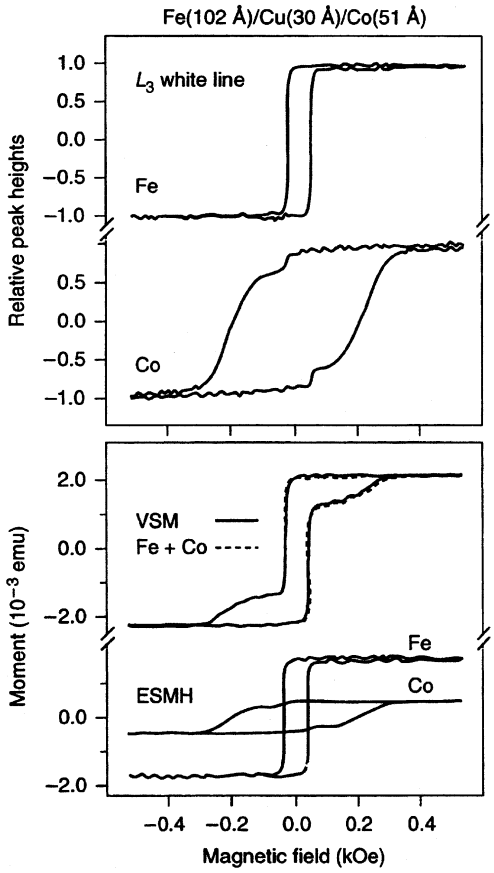


Figure 6. Hysteresis measurements of the individual magnetizations of Fe and Co in a Fe/Cu/Co multilayer obtained by measuring the L_3 edge XMCD signal (top), along with data obtained from a vibrating sample magnetometer (VSM) and a least-squares fit to the measured Fe and Co hysteresis curves (bottom). Courtesy of Chen et al. (1993).

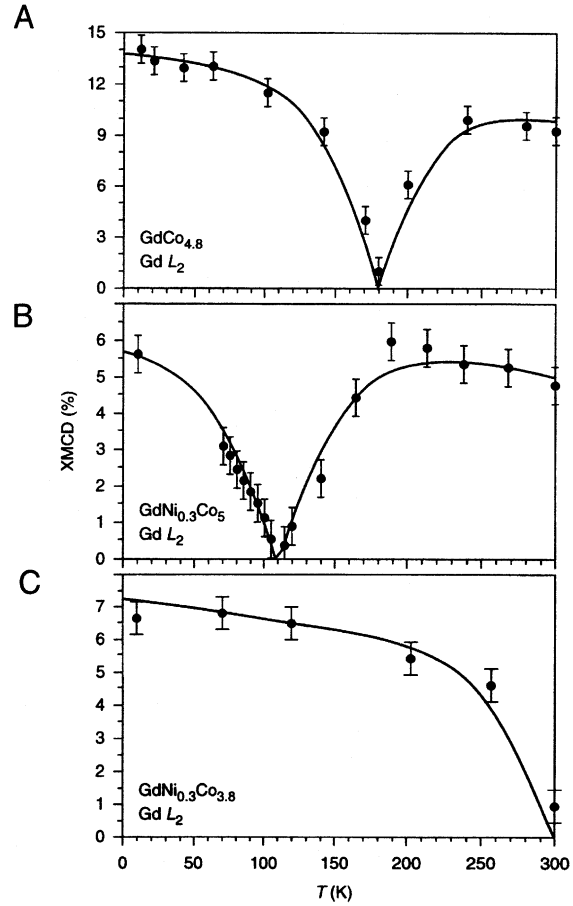


Figure 7. Variation of the Gd L_2-edge XMCD signal as a function of temperature for various GdNiCo amorphous alloys. Courtesy of Rueff et al. (1997).

spectrum is acquired by sequentially stepping the monochromator through the required energies. In the dispersive monochromator technique (Fig. 8b), a broad range of energies ($\Delta E \sim 200$ to 1000 eV) is incident on the sample simultaneously and the full spectrum is collected by a position-sensitive detector. Both of these setups are common at synchrotron radiation facilities where they have been widely used for EXAFS measurements. Their conversion to perform XMCD experiments usually requires minimal effort. The added aspects are simply use of CPX, using the methods discussed above, and a magnetized sample.

Although the scanning monochromator technique can require substantial time to step through the required beam energies, the simplicity of data interpretation and compatibility with most synchrotron beamline optics have made it by far the most common method utilized for XMCD measurements. Typically, these measurements are taken by reversing the magnetization (helicity) at each energy step, rather than taking a whole spectrum, then changing the magnetization (helicity) and repeating the measurement. The primary reason for this is to minimize systematic errors that can be introduced as a result of the decay or movements of the particle beam.

In the hard x-ray regime, because of the high intensity of the synchrotron radiation, the incident (I_o) and

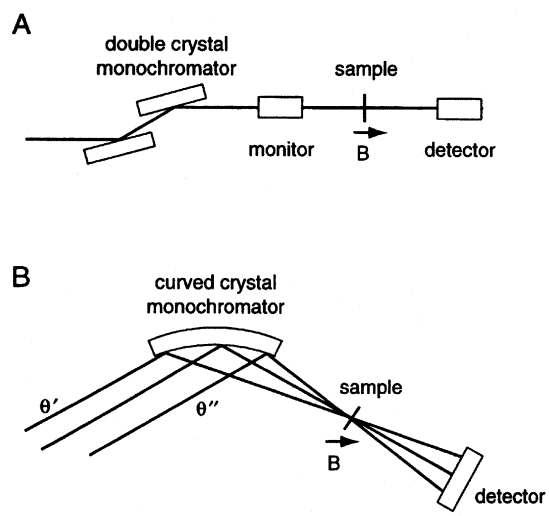


Figure 8. Experimental setups for the measurement of the absorption. (**A**) Scanning monochromator (double crystal); (**B**) dispersive monochromator (curved crystal).

attenuated (I) beam intensities are usually measured using gas ionization chambers. For these detectors, a bias (500 to 2000 V) is applied between two metallic plates and the x-ray beam passes though a gas (generally nitrogen) flowing between them. Absorption of photons by the gas creates ion pairs that are collected at the plates, and the induced current can then be used to determine the x-ray beam intensity. Typical signal currents encountered in these measurements range from $\sim 10^{-9}$ A for bending magnet sources to $\sim 10^{-6}$ A for insertion-device sources. While these signal levels are well above the electronic noise levels, $\sim 10^{-14}$ A, produced by a state-of-the-art current amplifier, other noise mechanisms contribute significantly to the error. The most common source of noise in these detectors arises from vibrations in the collection plates and the signal cables. Vibrations change the capacitance of the detection system, inducing currents that can obscure the signal of interest—particularly for measurements made on a bending magnet source, where the signal strength is smaller. This error can be minimized by rigidly mounting the ion chamber and minimizing the length of cable between the ion chamber and the current amplifier. Even when these effects are minimized, however, they still tend to be the dominant noise mechanism in XMCD measurements. This makes it difficult to get an *a priori* estimate of the error based on the signal strength. Therefore, the absolute error in an XMCD measurement is frequently obtained by taking multiple spectra and computing the average and variance in the data.

Using a scanning monochromator, one can also indirectly measure the effective absorption of a sample by monitoring the fluorescence signal. Fluorescence radiation results from the filling of the core hole generated by the absorption of an x-ray photon. The strength of the fluorescence will thus be proportional to the number of vacancies in the initial core state and therefore to the absorption. An advantage of this technique over transmittance measurements is that fluorescence arises only from the element at resonance, thereby effectively isolating the signal of interest. In a transmittance measurement, on the other hand, the resonance signal from the absorption edge sits atop the background absorption due to the rest of the constituents in the sample.

Fluorescence measurements require the use of an energy-sensitive detector to isolate the characteristic lines from the background radiation of elastically and Compton scattered radiation. Generally Si or Ge solid-state detectors have been used for fluorescence detection, since they offer acceptable energy resolution ($\Delta E \sim 200$ eV) and high efficiencies. The count rate offered by these solid-state detectors, however, is typically restricted to <500 kHz. Fluorescence signals this strong are easily obtained even using bending magnet sources; thus, the detector becomes the rate-limiting factor for data acquisition. Considering that a typical XMCD difference signal for a RE L edge is $\sim 0.5\%$, obtaining a 10% accuracy for this difference requires a total count of $\sim 4 \times 10^6$. Thus, acquiring this accuracy over the entire scanned range, which may contain ~ 100 points, can require data acquisition times on the order of several hours, irrespective of the CPX source utilized.

A further complication encountered with fluorescence data is that the signal must be corrected for the self absorption of the sample. For a thick sample, this correction is given by:

$$\mu_o(E) = \frac{I_f}{I_o} \sim \frac{\mu_o(E)\sin\theta}{\mu_T(E)/\sin\theta + \mu_T(E_f)/\sin\varphi} \qquad (7)$$

where $\mu_T(E)$ is the total absorption at energy E, E_f is the fluorescence energy, θ is the entrance angle of the incident x-rays, and ϕ is the exit angle of the fluorescent x-rays. From the count rates involved and the process needed to deconvolve the true absorption from the measured fluorescence signal, it is clear that this technique is at a disadvantage compared to a straightforward transmission measurement. Fluorescence measurements, however, are the preferred technique when examining dilute systems or thin magnetic films on thick substrates. For these samples, $\mu_T(E)$ is nearly constant, so the correction in Eq. 7 becomes negligible. Further, the difference signals in transmission arising from dilute constituents are minimal; thus, they benefit greatly from isolation of absorption only due to the atomic species of interest achieved with the fluorescence technique.

Another indirect measure of the absorption spectra used for XMCD is the measurement of the total electron yield from a sample. In this method, an electrical lead is attached directly to the sample, and a metallic enclosure with an entrance hole for the beam surrounds it. The enclosure is kept at a small bias (~ 10 V) with respect to the sample, effectively turning the sample itself into an ionizing detector. This technique is quite similar to fluorescence measurements, in that, instead of probing the radiative decay of the core hole, it probes the decay of the core hole via Auger transitions. The main contribution to the total yield signal, however, does not arise directly from the Auger electrons themselves or even from the ejected photoelectrons, but rather from the secondary inelastically scattered electrons. Despite the complications encountered with secondary electron processes, it has been shown that the total yield signal obtained is proportional to the x-ray absorption (Gudat and Kunz, 1972). To minimize possible systematic errors, the incident beam intensity is monitored using a similar detection method. A partially transmitting wire mesh is placed in the incident beam, again surrounded by a metal enclosure with an applied bias, and the induced current measured to determine I_o.

Electron yield measurements are the detection method most often utilized in the soft x-ray regime, because the photon attenuation lengths at these energies (~ 100 Å) make transmittance measurements difficult. Furthermore, the probability of fluorescence decay relative to Auger decay of the core hole becomes small below ~ 2000 eV. The main drawback (or advantage) of this technique is its high surface sensitivity, since nearly all the Auger electrons originate within ~ 50 Å of the surface.

A particularly novel use of electron yield detection in XMCD has been imaging of magnetic domains (Stöhr et al., 1993). In this technique, a photoelectron microscope is used to detect the secondary electrons from a particular

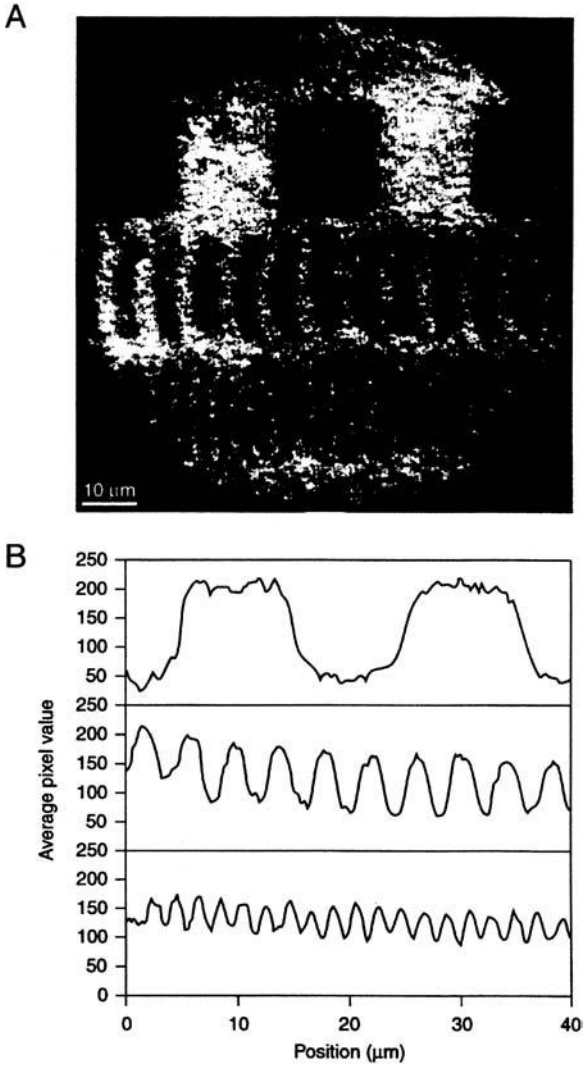


Figure 9. Image of the magnetic domains in Co/Pt multilayer obtained from XMCD signal detected using a photoelectron microscope (**A**). Also shown are integrated traces of the image (**B**). Reprinted with permission from Stöhr et al. (1993). Copyright 1993 American Association for the Advancement of Science.

spot on the sample, thereby obtaining a two-dimensional absorption profile (Tonner et al., 1994). By tuning the energy to an absorption edge with a strong XMCD signal and utilizing CPX, magnetization-sensitive contrast in the resultant image is obtained. This is demonstrated in Figure 9, which shows just such an image taken at the Co L edges of a CoPtCr film. In this particular material, the magnetic domains align themselves perpendicular to the plane of the sample. Therefore, taking the difference in the signal measured with either helicity produces a picture of the magnetic domains of the sample. Figure 9 is a combination of the images taken at both the L_2 and L_3 edges. Domain sizes down to $10\,\mu m \times 0.5\,\mu m$ are clearly visible. The uniqueness of this measurement over the same information available from other techniques is again the element specificity of XMCD. Thus the degree of magnetic alignment of individual constituents can be measured.

An alternative to the scanning monochromator is the use of energy dispersive optics (Baudelet et al., 1991). In a dispersive arrangement a curved crystal is used to monochromatize the beam. The angle that the incident beam makes with the diffraction planes ($\theta' \rightarrow \theta''$) varies across the crystal, and thus a range of energies ($E' \rightarrow E''$) are diffracted (Fig. 8b). The dispersion of the monochromator crystal is such that the photon energies vary approximately linearly with angle, focusing to a point ($\sim$100 μm) $\sim$3 to 4 m from the monochromator, where the sample is generally placed. A position-sensitive detector placed behind the sample can then be used to collect the full spectrum. There are several advantages to using dispersive optics for collecting XMCD data. The foremost is that acquiring the spectra requires no motorized motions, eliminating the time required to scan the monochromator and thereby greatly enhancing data collection rates. Another advantage is that the small focal spot reduces errors caused by sample nonuniformity. The disadvantage of the technique is that the data analysis is in general more complicated than for data obtained using a scanning monochromator. The charge-coupled device detectors generally used for these measurements possess an inherent readout noise that must be carefully subtracted. This subtraction can be further complicated by structure that may be introduced into this noise over time by the x-ray beam; thus, the noise must be monitored periodically throughout the data collection process by collecting background spectra. Furthermore, the I_o spectrum may not be uniform with energy, as a result of variations in the reflectivity of the monochromator crystal or post-monochromator mirror optics; therefore, an image of the incident beam must also be taken. This technique has almost exclusively been used with off-axis bending-magnet CPX sources, although recently the applicability of phase retarders with dispersive optics has been demonstrated (Giles et al., 1994).

Sample Magnetization

Once the CPX source and optics have been chosen, the sample must be magnetized. The analysis presented in Appendix A assumed that the magnetization and photon helicity were collinear. A more detailed calculation demonstrates that for dipolar transitions the magnitude of the XMCD signal should scale with the cosine of the angle between the magnetization and beam. Therefore, in order to maximize the possible signal, the magnetic field should be as close to the beam direction as possible. Problems arise, however, in applying a field exactly parallel to the beam direction. Demagnetization effects make thin foils extremely difficult to magnetize normal to the foil plane and thus they are usually magnetized in plane. In a transmittance measurement the x-ray beam must be incident at some angle to the plane of the sample, which results in a finite angle between the beam and the field directions ($\sim 25^\circ \rightarrow 45^\circ$). Even when a solenoid coaxial with the incident x-ray beam is used to provide the magnetic field, the sample plane should not be perpendicular to the solenoid axis but rather at some angle. In this arrangement,

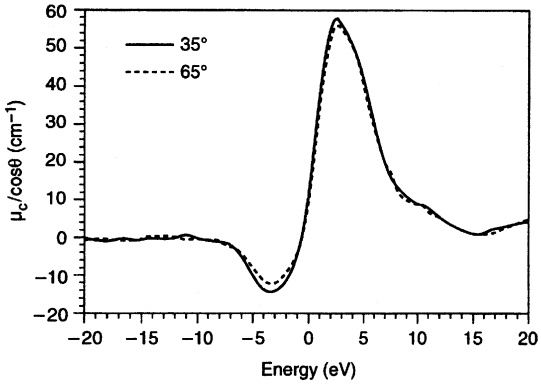


Figure 10. XMCD spectra at the Dy L_3 edge of a $Dy_{0.4}Tb_{0.6}$ alloy for different angles between the incident beam and magnetization directions, showing the contributions arising from quadrupole transitions. Courtesy of Lang et al. (1995).

unless the solenoid generates a very high magnetic field ($|B| > 1$ T), the direction of the local magnetization direction will generally lie along the sample plane rather than along the field direction.

Varying the angle that the beam makes with the sample magnetization can also be used as a means of identifying features in the dichroic spectra that do not arise from dipolar transitions which scale with the cosine. Figure 10 shows the XMCD signal taken for two different angles at the Dy L_3 edge in a $Dy_{0.4}Tb_{0.6}$ alloy (Lang et al., 1995). These spectra are normalized to the cosine of the angle between the beam and the magnetization directions; therefore, the deviation in the lower feature indicates nondipolar-like behavior. This feature has been shown to arise from quadrupolar transitions to the 4*f* states and is a common feature in all RE $L_{2,3}$-edge XMCD spectra (Wang et al., 1993). Although quadrupolar effects in the absorption spectrum are smaller than dipolar effects by a factor of $\sim$100, their size becomes comparable in the difference spectrum due to the strong spin-polarization of the RE 4*f* states. The presence of quadrupolar transitions can complicate the application of sum rule to the analysis of these spectra but also opens up the possibility of obtaining magnetic information on two orbitals simultaneously from measurements at only one edge.

METHOD AUTOMATION

Acquisition of XMCD spectra requires a rather high degree of automation, because the high intensities of x-ray radiation encountered at synchrotron sources prevent any manual sample manipulation during data acquisition. The experiment is normally set up inside a shielded room and all experimental equipment operated remotely through a computer interface. These interfaces are typically already installed on existing beamlines and the experimenter simply must ensure that the equipment that they are considering for the experiment is compatible with the existing hardware. Note that the time required to enter and exit the experimental enclosure is not minimal. Therefore, computerized motion control of as many of the experimental components (i.e., magnet, sample, slits, etc.) as possible is desirable.

The typical beamline equipment used in XMCD data acquisition includes motorized stages, various NIM (nuclear instrumentation module electronic standard) electronic components, power supplies, and cryostats with associated temperature controllers. The data acquisition system for the collection of the XMCD spectra should be fully automated. The computer should step through the required energies if a scanning monochromator is used, alternate the magnet field or helicity, and record the detector counts. In addition, since obtaining enough statistical accuracy in full XMCD spectra frequently requires repeated scanning for a substantial time, the software should also collect multiple scans and save the data without manual intervention.

DATA ANALYSIS AND INITIAL INTERPRETATION

The first thing to realize when taking an initial qualitative look at an XMCD spectrum is that these measurements probe the unoccupied states, whose net polarization is opposite that of the occupied states. Thus, the net difference in the XMCD signal reflects the direction of the so-called minority spin band, because it contains the greater number of available states. It is therefore easy to make a sign error for the dichroic spectra, and care should be taken when determining the actual directions of the moments in a complicated compound. To prevent confusion, it is essential to measure the direction of the magnetic field and the helicity of the photons used from the outset of the experiment.

If XMCD is being used to measure changes in the magnetization of the sample, the integrated dichroic signal should be used rather than the size of the signal at one particular energy for two reasons in particular. The integrated signal is less sensitive to long-term drifts in the experimental apparatus than are measurements at one nominal energy. And, systematic errors are much more apparent in the entire spectrum than at any particular energy. For instance, the failure of the spectrum to go to zero below or far above the absorption edge would immediately indicate a problem with the data collection, but would not be detected by measurement at a single energy.

When analyzing spectra taken from different compounds, one should note that a change in the magnitude of the dichroic signal at just one edge does not necessarily reflect a change in the size or direction of the total magnetic moment. Several factors complicate this, some of which have already been mentioned. For example, the development of an orbital moment for the spectra shown in Figure 2 greatly enhanced the signal at the L_3 edge but diminished the signal at the L_2 edge. Therefore, although the size of the total moment remains relatively constant between the two samples, the dichroic signal at just one edge shows substantial changes that do not necessarily reflect this. The total moment in this case must be obtained through the careful application of both sum rules or comparison with magnetometer measurements. The strength of the TM *K*-edge dichroism also tends not to

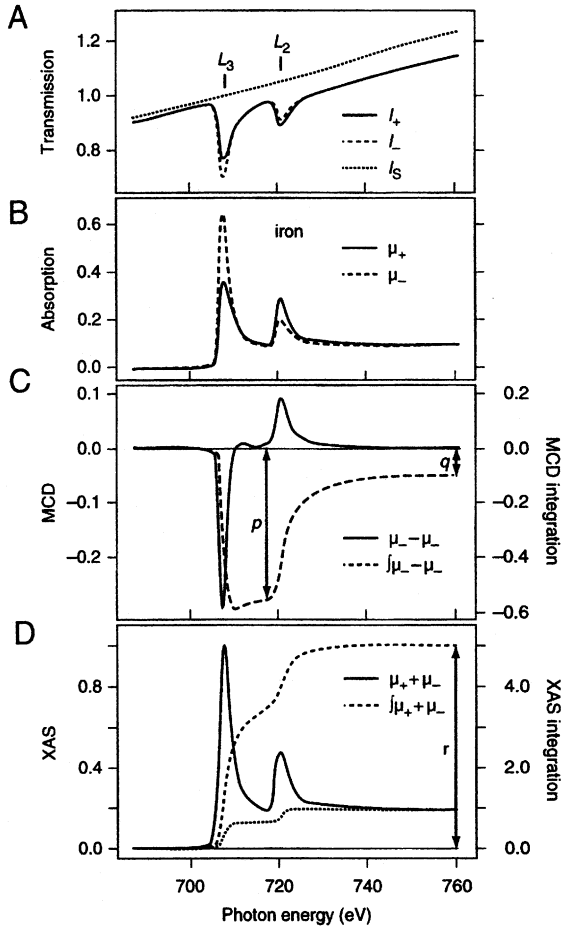


Figure 11. Steps involved in the sum rule analysis of the XMCD spectra demonstrated for a thin layer of metallic Fe. (**A**) Transmitted intensity of parylene substrate and substrate with deposited sample. (**B**) Calculated absorption with background subtracted. (**C**) Dichroic spectra and integrated dichroic signal. (**D**) Absorption spectra and edge-jump model curve along with integrated signal. Courtesy of Chen et al. (1995).

reflect the size of the moment for different materials. This is particularly true for the Fe *K*-edge spectra which demonstrate a complete lack of correlation between the size of the total Fe moment and the strength of the dichroic signal, although better agreement has been found for Co and Ni (Stahler et al., 1993). In addition, RE $L_{2,3}$-edge spectra show a strong spin dependence in the radial matrix elements that can in some cases invert the relationship between the dichroic signal and the size of the moment (Lang et al., 1994). These complications notwithstanding, the dichroic spectra at one edge and within one compound should still scale with changes in the magnetization due to applied external fields or temperature changes.

If the XMCD signal is not complicated by the above effects and both edges of a spin-orbit-split pair are measured, the sum rules may be applied. The steps involved in the implementation of the sum rules are illustrated in Figure 11 for a thin Fe metal film deposited on a parylene substrate (Chen et al., 1995). The first step in sum rules analysis is removal from the transmission signal of all absorption not due to the particular element of interest. In this example, this consists of separately measuring and subtracting the substrate absorption (Fig. 11a), leaving only the absorption from the element of interest (Fig. 11b). In compound materials, however, there are generally more sources of absorption than just the substrate material, and another method of background removal is necessary. The common method has been subtraction from the signal of a λ^3 function fit to the pre-edge region of the absorption spectrum.

The next step in the application of the sum rules is determining the integrated intensities of the dichroic signal. The easiest method is to plot the integral of the dichroic spectrum as a function of the scanning energy (Fig. 11c). The integrated values over each edge can then be directly read at the positions where the integrated signal becomes constant. In the figure, p corresponds to the signal from the L_3 edge alone and q corresponds to the integrated value over both edges. Although there is some uncertainty in the appropriate value of the p integral, since the dichroic signal of the L_3 edge extends well past the edge, choosing the position closest to the onset of the L_2 edge signal minimizes possible errors. This uncertainty diminishes for edges at higher energies where the spin-orbit splitting of the core levels becomes more pronounced.

The final step in the sum rule analysis is obtaining the integral of the absorption spectrum (Fig. 11d). At first glance, it seems that this integral would be nonfinite because the absorption edges do not return to zero at the high-energy end of the spectrum. The sum rule derivation, however, implicitly assumes that the integral includes only absorption contributions from the final states of interest (i.e., the 3*d* states in this example). It is well known that the large peaks in the absorption spectrum just above the edge, sometimes referred to as the "white line," are due to the strong density of states of *d* bands near the Fermi energy, while the residual step-like behavior is due to transitions to unbound free electron states. To deconvolve these two contributions, a two-step function is introduced to model the absorption arising from the transitions to the free electron states. The absorption-edge jumps for this function are put in a 2:1 ratio, expected because of the 2:1 ratio in the number of initial-state electrons. The steps are also broadened to account for the lifetime of the state and the experimental resolution. The integral r shown in Fig 11d is then obtained by determining the area between the absorption curve and the two-step curve. Once the integrals r, p, and q are determined, they can easily be put into Eqs. 1, 2, and 3 to obtain the values of the magnetic moments.

SAMPLE PREPARATION

XMCD measurements in the hard x-ray regime require little sample preparation other than obtaining a sample thin enough for transmission measurements. Measurements in the soft x-ray regime, on the other hand, tend to require more complicated sample preparation to account for UHV conditions and small attenuation lengths. Hard x-ray XMCD measurements therefore tend to be favored if

the same type of information (i.e., relative magnetization) can be obtained using edges at the higher energies. The incident photon fluxes typically do not affect the sample in any way, so repeated measurements can be made on the same sample without concern for property changes or degradation.

Attenuation lengths (i.e., the thickness at which $I/I_o \approx 1/e$) for hard x-rays are generally on the order of $\sim$2 to 10 μm. While it is sometimes possible to roll pure metals to these thicknesses, the rare earth and transition metal compounds generally studied in XMCD are typically very brittle, so that rolling is impractical. The sample is therefore often ground to a fine powder in which the grain size is significantly less than the x-ray attenuation length ($\sim$1 μm). This powder is then spread on an adhesive tape, and several layers of tape combined to make a sample of the proper thickness. Most of the absorbance in these tapes is usually due to the adhesive rather than the tape itself; thus, a commercial tape brand is as effective as a specialized tape such as Kapton that minimizes x-ray absorbance. An alternative to tape is to mix the powder with powder of a light-Z material such as graphite and make a disk of appropriate thickness using a press. A note of warning: the fine metallic powders used for these measurements, especially those containing RE elements, can spontaneously ignite. Contact with air should therefore be kept to a minimum by using an oxygen-free atmosphere when applying the powder and by capping the sample with a tape overlayer during the measurements.

The techniques used to detect the XMCD signal in the soft x-ray energy regime are generally much more surface sensitive. Therefore, the surfaces must be kept as clean as possible and arrangements made to clean them in situ. Alternatively, a capping layer of a nonmagnetic material can be applied to the sample to prevent degradation. Surface cleanliness is also a requirement for the total electron-yield measurements; that is, the contacts used to measure the current should make the best possible contact to the material without affecting its properties. For transmittance measurements in this regime, the thinness of the sample attenuation lengths ($\sim$100 Å) makes supporting the sample impractical. Thus, samples are typically deposited on a semitransparent substrate such as parylene.

PROBLEMS

There are many factors that can lead to erroneous XMCD measurements. These generally fall into two categories, those encountered in the data collection and those that arise in the data analysis. Problems during data collection tend to introduce distortions in the shape or magnitude of the spectra that can be difficult to correct in the data analysis; thus, great care must be taken to minimize their causes and be aware of their effects.

Many amplitude-distorting effects encountered in XMCD measurements, and in x-ray absorption spectroscopy in general, come under the common heading of "thickness effects." Basically these are systematic errors that diminish the relative size of the observed signal due to leakage of noninteracting x-rays into the measurement. Leakage x-rays can come from pinholes in the sample, higher-energy harmonics in the incoming beam, or the monochromator resolution function containing significant intensity below the absorption-edge energy. While these effects are significant for all x-ray spectroscopies, they are particularly important in the near-edge region probed by XMCD.

First, to minimize pinhole effects, the sample should be made as uniform as possible. If the sample is a powder, it should be finely ground and spread uniformly on tape, with several layers combined to obtain a more uniform thickness. All samples should be viewed against a bright light and checked for any noticeable leakages. The sample uniformity should also be tested by taking absorption spectra at several positions on the sample before starting the XMCD measurements. Besides changes in the transmitted intensity, the strength and definition of sharp features just above the absorption edge in each spectrum provide a good qualitative indication of the degree of sample nonuniformity.

Another source of leakage x-rays is the harmonic content of the incident beam. The monochromator used for the measurement will pass not only the fundamental wavelength but multiples of it as well. These higher-energy photons are unaffected by the absorption edges and will cause distorting effects similar to those caused by pinholes. These harmonics are an especially bothersome problem with third-generation synchrotron sources because the higher particle-beam energies found at these sources result in substantially larger high-energy photon fluxes. Harmonics in the beam are generally reduced by use of an x-ray mirror. By adjusting the reflection angle of the mirror, near-unit reflectivity can be achieved at the fundamental energy, while the harmonic will be reduced by several orders of magnitude. Another common way of eliminating harmonics has been to slightly deviate the angle of the second crystal of a scanning monochromator (Fig. 8a) so that it is not exactly parallel to the first. The monochromator (said to be "detuned") then reflects only a portion of the beam reflected by the first crystal, typically $\sim$50%. The narrower reflection width of the higher-energy photons makes harmonics much more sensitive to this detuning. Thus while the fundamental flux is reduced by a factor of $\sim$2, the harmonic content of the beam will be down by several orders of magnitude. Eliminating harmonics in this manner has a distinct disadvantage for XMCD measurements, however, because it also reduces the degree of circular polarization of the monochromatized beam. Therefore, monochromator detuning should not be used with off-axis bending-magnet or specialized-insertion-device CPX sources.

A third important consideration in minimizing thickness effects is the resolution of the monochromator. After passing through the monochromator, the beam tends to have a long-tailed Lorentzian energy distribution, meaning that a beam nominally centered around E_o with a resolution of ΔE will contain a substantial number of photons well outside this range. When E_o is above the edge, the bulk of the photons are strongly attenuated, and thus those on the lower end of the energy distribution will constitute a greater percentage of the transmitted intensity.

Typically a monochromator with a resolution well below the natural width of the edge is used to minimize this effect. This natural width is determined by the lifetime of the corresponding core state under study. For instance, a typical RE L-edge spectrum with a core-hole broadening of $\sim$4 eV is normally scanned by a monochromator with $\sim$1 eV resolution. Increasing the resolution, however, typically increases the Bragg angle of the monochromator, which reduces the circular polarization of the transmitted beam.

While all the methods for reducing the thickness effects mentioned above should be used, the primary and most efficient way to minimize these effects is to adjust the sample thickness so as to minimize the percentage of leakage photons. This is accomplished by optimizing the change in the transmitted intensity across the edge. It should be small enough to reduce the above thickness effect errors yet large enough to permit observation of the signal of interest. It has been found in practice that sample thicknesses (t) of $\sim$1 attenuation length ($\mu t \sim 1$) below the absorption edge and absorption increases across the edge of $<$1.5 attenuation lengths ($\Delta\mu t < 1.5$) minimize the amplitude distortions.

Another source error in the XMCD measurement is sharp structure arising from spurious reflections in the monochromator or phase retarder, if used. These effects are usually called "glitches" because they have a small energy width and typically affect one data point. Generally the monochromator or phase plate must be adjusted by rotation about its diffraction vectors to move the offending glitch out of the energy range of interest. If this is not possible, an algorithm must be incorporated into the data reduction to omit the offending point. Similar glitch effects occur during data collection if data acquisition at one point started before a scanning monochromator had had time to settle into the relaxed position, causing the data point for one magnetization (helicity) to be taken under different conditions than the other. This can become particularly bothersome if the monochromator has some sort of feedback mechanism to maintain stability. These type of glitches can be minimized by introducing a 0.5- to 1.0-s delay between motor motions and triggering of the detectors.

The electromagnets used to magnetize samples are also a frequent cause of systematic errors in XMCD spectra. The efficiency of a gas ionization detector can be affected by the presence of a magnetic field. The resultant induced nonlinearities in these detectors can be as great as or greater than the dichroic signal itself. Therefore, for measurements taken by switching the magnetization of the sample, the detectors should be placed well away from the magnets. If this is not possible, the detectors or the magnet should be shielded with a high-permeability metal foil to minimize stray magnetic fields. Similarly, other detectors not directly sensitive to magnetic fields can nevertheless be indirectly influenced by the effects that the electromagnet power supplies can have on their electronic components. Strong electromagnets draw substantial currents, which can induce voltage drops in the circuit that powers them. To minimize cross-talk between the detector electronics and the power supply, they should be connected to different power sources. It is important to note that this does not simply mean lines on different circuit breakers, but rather entirely different, isolated lines. This problem can be particularly bothersome for solid-state detectors in which the switching of a magnet power supply can induce a change in the gain of a spectroscopy amplifier, thereby shifting the position of a fluorescence line or other measured signal.

Besides the errors encountered in the acquisition of the data, several errors can be introduced in the analysis of the spectra, especially in implementing the sum rules. Foremost among these are uncertainties in the values of the beam polarization and the sample magnetization. Both these factors scale directly with the size of the dichroic signal, and any uncertainty in them directly correlates with the size of the error of a magnetic moment obtained from XMCD. The rate of circular polarization is generally calculated from known parameters since it is difficult to measure directly. Because these calculations make assumptions about storage-ring parameters and the degree of perfection of the beamline optics, the error in the values for the polarization rate tend to be quite large and frequently dominate the error in a measurement of the absolute size of the magnetic moment. For measurements of the relative changes in the XMCD spectra, however, this uncertainty is eliminated since the polarization does not change. The magnetization of the sample is usually obtained from magnetometer measurements, but for a multiconstituent sample with different temperature dependencies estimation of each sublattice magnetization can prove difficult.

Systematic errors can also occur at each step in the implementation of the sum rules. As noted in the discussion of Figure 11, the values of the integrals p, q, and r may vary slightly depending on the energy chosen to extract the integrated value. The greatest uncertainty in the sum rule analysis, however, arises in the integration of the absorption signal, which involves two steps, each of which can introduce uncertainties into the measurement. The first step is subtraction of the background absorption and the second (which causes a greater uncertainty) is modeling of the absorption from transitions to free electron states with a two-step function. Although the two-step function itself should be an adequate model of the transitions involved, the error arises in the placement of the inflection point of the two steps. This has generally been chosen to be at the maximum in the absorption spectrum, but this choice is rather arbitrary. A way around introducing this error is to use Eq. 3 to determine an orbital to spin ratio rather than the absolute moments themselves.

ACKNOWLEDGMENTS

The author would like to thank Professor A.I. Goldman for offering suggestions on and proofreading this work, as well as Dr. C.T. Chen and Dr. J. Stöhr for providing copies of their figures for publication herein. This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, under contract No. W-31-109-ENG-38.

LITERATURE CITED

Baudelet, F., Dartyge, E., Fontaine, A., Brouder, C., Krill, G., Kappler, J. P., and Piecuch, M. 1991. Magnetic properties of neodymium atoms in Nd-Fe multilayers studied by magnetic x-ray dichroism on Nd L_{II} and Fe K edges. *Phys. Rev. B* 43:5857–5866.

Baumgarten, L., Schneider, C. M., Petersen, H., Schäfers, F., and Kirschner, J. 1990. Magnetic x-ray dichroism in core-level photoemission from ferromagnets. *Phys. Rev. Lett.* 65:492–495.

Blume, M. and Gibbs, D. 1988. Polarization dependence of magnetic x-ray scattering. *Phys. Rev. B* 37:1779-1789.

Brouder, C. 1990. Angular dependence of x-ray absorption spectra. *J. Phys. Condens. Matter* 2:701–738.

Brouder, C. and Hikam, M. 1991. Multiple-scattering theory of magnetic x-ray circular dichroism. *Phys. Rev. B* 43:3809–3820.

Carra, P., Thole, B. T., Altarelli, M., and Wang, X. 1993. X-ray circular dichroism and local magnetic fields. *Phys. Rev. Lett.* 70:694–697.

Chen, C. T., Sette, F., Ma, Y., and Modesti, S. 1990. Soft-x-ray magnetic circular dichroism at the $L_{q2,3}$ edges of nickel. *Phys. Rev. B* 42:7262–7265.

Chen, C. T., Idzerda, Y. U., Lin, H.-J., Meigs, G., Chaiken, A., Prinz, G. A., and Ho, G. H. 1993. Element-specific magnetic hysteresis as a means for studying heteromagnetic multilayers. *Phys. Rev. B* 48:642–645.

Chen, C. T., Idzerda, Y. U., Lin, H.-J., Smith, N. V., Meigs, G., Chaban, E., Ho, G. H., Pellegrin, E., and Sette, F. 1995. Experimental confirmation of the x-ray magnetic circular dichroism sum rules for iron and cobalt. *Phys. Rev. Lett.* 75:152–155.

Finazzi, M., Sainctvit, P., Dias, A.-M., Kappler, J.-P., Krill, G., Sanchez, J.-P., Réotier, P. D.d., Yaouanc, A., Rogalev, A., and Goulon, J. 1997. X-ray magnetic circular dichroism at the U $M_{4,5}$ absorption edges of UFe_2. *Phys. Rev. B* 55:3010–3014.

Fischer, P., Schütz, G., Scherle, S., Knülle, M., Stahler, S., and Wiesinger, G. 1992. Experimental study of the circular magnetic x-ray dichroism in Ho-metal, $Ho_2Fe_5O_{12}$, Ho_2Fe_{23} and Ho_2Co_{17}. *Solid State Commun.* 82:857–861.

Gibbs, D., Harshman, D. R., Isaacs, E. D., McWhan, D. B., Mills, D., and Vettier, C. 1988. Polarization and resonance properties of magnetic x-ray scattering in Ho. *Phys. Rev. Lett.* 61:1241–1244.

Giles, C., Malgrange, C., Goulon, J., Bergevin, F.d., Vettier, C., Dartyge, E., Fontaine, A., Giorgetti, C., and Pizzini, S. 1994. Energy-dispersive phase plate for magnetic circular dichroism experiments in the x-ray range. *J. Appl. Cryst.* 27:232–240.

Gudat, W. and Kunz, C. 1972. Close similarity between photoelectric yield and photoabsorption spectra in the soft-x-ray range. *Phys. Rev. Lett.* 29:169–172.

Hirano, K., Ishikawa, T., and Kikuta, S. 1993. Perfect crystal x-ray phase retarders. *Nucl. Instr.Meth.* A336:343–353.

Jackson, J. D. 1975. Classical Electrodynamics. John Wiley & Sons, New York.

Keller, E. N. 1985. Magneto X-ray Study of a Gadolinium-Iron Amorphous Alloy. Ph.D. thesis, University of Washington.

Kobayashi, K., Maruyama, H., Iwazumi, T., Kawamura, N., and Yamazaki, H. 1996. Magnetic circular x-ray dichroism at Pd $L_{2,3}$ edges in Fe-Pd alloys. *Solid State Commun.* 97:491–496.

Konig, H., Wang, X., Harmon, B. N., and Carra, P. 1994. Circular magnetic x-ray dichroism for rare earths. *J. Appl. Phys.* 76:6474–6476.

Krill, G., Schillé, J. P., Sainctavit, J., Brouder, C., Giorgetti, C., Dartyge, E., and Kappler, J. P. 1993. Magnetic dichroism with synchrotron radiation. *Phys. Scripta* T49:295–301.

Laan, G. v. d., Thole, B. T., Sawatzky, G. A., Goedkoop, J. B., Fluggle, J. C., Esteva, J.-M., Karnatak, R., Remeika, J. P., and Dabkowska, H. A. 1986. Experimental proof of magnetic x-ray dichroism. *Phys. Rev. B* 34:6529–6531.

Lang, J. C., Kycia, S. W., Wang, X., Harmon, B. N., Goldman, A. I., McCallum, R. W., Branagan, D. J., and Finkelstein, K. D. 1992. Circular magnetic x-ray dichroism at the erbium L_3 edge. *Phys. Rev. B* 46:5298–5302.

Lang, J. C., Wang, X., Antropov, V. P., Harmon, B. N., Goldman, A. I., Wan, H., Hadjipanayis, G. C., and Finkelstein, K. D. 1994. Circular magnetic x-ray dichroism in crystalline and amorphous $GdFe_2$. *Phys. Rev. B* 49:5993–5998.

Lang, J. C., Srajer, G., Detlefs, C., Goldman, A. I., Konig, H., Wang, X., Harmon, B. N., and McCallum, R. W. 1995. Confirmation of quadrupolar transitions in circular magnetic x-ray dichroism at the dysprosium L_{III} edge. *Phys. Rev. Lett.* 74:4935–4938.

Lang, J. C., Srajer, G., and Dejus, R. J. 1996. A comparison of an elliptical multipole wiggler and crystal optics for the production of circularly polarized x-rays. *Rev. Sci. Instrum.* 67:62–67.

Lin, H.-J., Chen, C. T., Meigs, G., Idzerda, Y. U., Chaiken, A., Prinz, G. A., and Ho, G. H. 1993. Element-specific magnetic hysteresis measurements, a new application of circularly polarized soft x-rays. *SPIE Proc.* 2010:174–180.

Malgrange, C., Carvalho, C., Braicovich, L., and Goulon, J. 1991. Transfer of circular polarization in Bragg crystal x-ray monochromators. *Nucl. Instr. Meth.* A308:390–396.

McIlroy, D. N., Waldfried, C., Li, D., Pearson, J., Bader, S. D., Huang, D.-J., Johnson, P. D., Sabiryanov, R. F., Jaswal, S. S., and Dowben, P. A. 1996. Oxygen induced suppression of the surface magnetization of Gd(0001). *Phys. Rev. Lett.* 76:2802–2805.

Pizzini, S., Fontaine, A., Fiorgetti, C., Dartyge, E., Bobo, J.-F., Piecuch, M., and Baudelet, F. 1995. Evidence for the spin polarization of copper in Co/Cu and Fe/Cu multilayers. *Phys. Rev. Lett.* 74:1470–1473.

Rudolf, P., Sette, F., Tjeng, L. H., Meigs, G., and Chen, C. T. 1992. Magnetic moments in a gadolinium iron garnet studied by soft-x-ray magnetic circular dichroism. *J. Mag. Mag. Matr.* 109:109–112.

Rueff, J. P., Galéra, F. M., Pizzini, S., Fontaine, A., Garcia, L. M., Giorgetti, C., Dartyge, E., and Baudelet, F. 1997. X-ray magnetic circular dichroism at the Gd L edges in Gd-Ni-Co amorphous systems. *Phys. Rev. B* 55:3063–3070.

Samant, M. G., Stöhr, J., Parkin, S. S. P., Held, G. A., Hermsmeier, B. D., Herman, F., Schilfgaarde, M. v., Duda, L.-C., Mancini, D. C., Wassdahl, N., and Nakajima, R. 1994. Induced spin polarization in Cu spacer layers in Co/Cu multilayers. *Phys. Rev. Lett.* 72:1112–1115.

Schillé, J. P., Sainctavit, P., Cartier, C., Lefebvre, D., Brouder, C., Kappler, J. P., and Krill, G. 1993. Magnetic circular x-ray dichroism at high magnetic field and low temperature in ferrimagnetic $HoCo_2$ and paramagnetic Ho_2O_3. *Solid State Commun.* 85:787–791.

Schütz, G., Wagner, W., Wilhelm, W., Kienle, P., Zeller, R., Frahm, R., and Materlik, G. 1987. Absorption of circularly polarized x-rays in iron. *Phys. Rev. Lett.* 58:737–740.

Schütz, G., Knulle, M., Wienke, R., Wilhelm, W., Wagner, W., Kienle, P., and Frahm, R. 1988. Spin-dependent photoabsorption at the L-edges of ferromagnetic Gd and Tb metal. *Z. Phys. B* 73:67–75.

Schütz, G., Wienke, R., Wilhelm, W., Wagner, W., Kienle, P., Zeller, R., and Frahm, R. 1989a. Strong spin-dependent absorption at the $L_{2,3}$-edges of 5*d*-impurities in iron. *Z. Phys. B* 75:495–500.

Schütz, G., Frahm, R., Wienke, R., Wilhelm, W., Wagner, W., and Kienle, P. 1989b. Spin-dependent *K*- and *L*-absorption measurements. *Rev. Sci. Instrum.* 60:1661–1665.

Schütz, G., Fischer, P., Atternkofer, K., Knulle, M., Ahlers, D., Stahler, S., Detlefs, C., Ebert, H., and Groot, F. M. F. d. 1994. X-ray magnetic circular dichroism in the near and extended absorption edge structure. *J. Appl. Phys.* 76:6453–6458.

Shenoy, G. K., Viccaro, P. J., and Mills, D. M. 1988. Characteristics of the 7-GeV Advanced Photon Source: A Guide for Users. Argonne National Laboratory ANL-88–9.

Smith, N. V., Chen, C. T., Sette, F., and Mattheiss, L. F. 1992. Relativistic tight-binding calculations of x-ray absorption and magnetic circular dichroism at the L_2 and L_3 edges of nickel and iron. *Phys. Rev. B* 46:1023–1032.

Stahler, S., Schutz, G., and Ebert, H. 1993. Magnetic *K* edge absorption in the 3*d* elements and its relation to local magnetic structure. *Phys. Rev. B* 47:818–826.

Stöhr, J., Wu, Y., Hermsmeier, B. D., Samant, M. G., Harp, G. R., Koranda, S., Dunham, O., and Tonner, B. P. 1993. Element-specific magnetic microscopy with circularly polarized x-rays. *Science* 259:658–661.

Stöhr, J. and Wu, Y. 1994. X-ray magnetic circular dichroism: Basic concepts and theory for 3*d* transition metal atoms. *In* New Directions in Research with 3rd Generation Soft X-ray Synchrotron Radiation Sources (A. S. Schlachter and F. J. Wuilleumier, eds.).. pp. 221–251. Kluwer Academic Publishers, Dordrecht, Netherlands.

Thole, B. T., Carra, P., Sette, F., and Laan, G. v. d. 1992. X-ray circular dichroism as a probe of orbital magnetization. *Phys. Rev. Lett.* 68:1943–1946.

Tonner, B. P., Dunham, D., Shang, J., O'Brien, W. L., Samant, M., Weller, D., Hermsmeier, B. D., and Stöhr, J. 1994. Imaging magnetic domains with the x-ray dichroism photoemission microscope. *Nucl. Instr. Meth.* A347:142–147.

Wang, X., Leung, T. C., Harmon, B. N., and Carra, P. 1993. Circular magnetic x-ray dichroism in the heavy rare-earth metals. *Phys. Rev. B* 47:9087–9090.

Wu, Y., Stöhr, J., Hermsmeier, B. D., Samant, M. G., and Weller, D. 1992. Enhanced orbital magnetic moment on Co atoms in Co/Pd multilayers: A magnetic circular x-ray dichroism study. *Phys. Rev. Lett.* 69:2307–2310.

Wu, R., Wang, D., and Freeman, A. J. 1993. First principles investigation of the validity and range of applicability of the x-ray magnetic circular dichroism sum rule. *Phys. Rev. Lett.* 71:3581–3584.

Wu, R. and Freeman, A. J. 1994. Limitation of the magnetic-circular-dichroism spin sum rule for transition metals and importance of the magnetic dipole term. *Phys. Rev. Lett.* 73:1994–1997.

Yamamoto, S., Kawata, H., Kitamura, H., Ando, M., Saki, N., and Shiotani, N. 1988. First production of intense circularly polarized hard x-rays from a novel multipole wiggler in an accumulation ring. *Phys. Rev. Lett.* 62:2672–2675.

KEY REFERENCES

Lovesey, S. W. and Collins, S. P. 1996. X-ray Scattering and Absorption by Magnetic Materials. Oxford University Press, New York.

Provides an overall theoretical picture of the interaction of x-rays with magnetic materials. Chapter 5 is devoted to XMCD measurements and includes several excellent examples of recent experimental work.

Koningsberger, D. C. and Prins, R. 1988. X-ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS, and XANES. John Wiley & Sons, New York.

Although this is primarily devoted to EXAFS measurements, the methods for and problems encountered with the acquisition of data are also applicable to XMCD measurements. Also useful is the theoretical work presented on XANES spectra, which cover the same energy range as XMCD, although without polarization analysis.

Chen et al., 1995. See above.

Idzerda, Y. U., Chen, C. T., Lin, H.-J., Meigs, G., Ho, G. H., and Kao, C.-C. 1994. Soft x-ray magnetic circular dichroism and magnetic films. *Nucl. Motr. Meth.* 347:134–141.

These two references provide excellent discussion of the various experimental limitations encountered when applying the XMCD sum rules.

Stöhr and Wu, 1994. See above.

A good overall theoretical treatment of XMCD, with emphasis on the interpretation of 3d L edge spectra.

APPENDIX A: ONE-ELECTRON PICTURE OF DICHROISM

This appendix provides a simple theory to explain the origin of dichroic signals. In this theory, the x-ray absorption is treated within the one-electron picture and the electric dipole approximation is assumed to be valid. The explanation below follows the same reasoning as put forth by Brouder and Hikam (1991) and Smith et al. (1992), and the reader should refer to these papers for a more detailed explanation of all the edges.

Using the first-order term in Fermi's golden rule, the x-ray absorption cross-section due to dipole transitions may be written as follows (Brouder, 1990):

$$\sigma(\varepsilon,\omega) = 4\pi^2\alpha\hbar\omega \sum_{f,m_j,\beta} |\langle f|\varepsilon^\beta \cdot \boldsymbol{r}|jm_j\rangle|^2 \delta(E_f - E_i - h\omega) \quad (8)$$

where the α is the fine-structure constant.

Here the common bracket notation is followed in which the expression above represents an integral over the spacial and spin dimensions of the system and $|jm_j\rangle$ and $|f\rangle$ represent the initial- and final-state wave functions involved in the transition. Because the initial atomic core states are strongly spin-orbit coupled, the wave functions are expressed in terms of the total angular momentum j. As an example of an absorption calculation, consider the above expression for an L_2 edge with initial $2p_{1/2}$ core state. By the dipole selection rules, the final states accessible from this edge would be either the s or d states. As a general rule, however, the transitions to the $\ell - 1$ states are suppressed by 1 to 2 orders of magnitude compared to the $\ell + 1$ states due to the negligible overlap of the former with the initial level. Thus to a very good approximation only transitions to the d states need be considered. Neglecting any spin-orbit coupling in the final states, the

final d-band states can be written as the product of spatial and spin wave functions:

$$\langle f| = \varphi_{\ell m}(r)\langle s| \sim 2m_\ell \langle\uparrow(\downarrow)| \tag{9}$$

where $m = -2, -\ell, \ldots, 2$ are the magnetic quantum numbers and $\uparrow$ and $\downarrow$ refer to the spin-up and spin-down states. The electric dipole operator $(\varepsilon \cdot r)$ for left (right) circularly polarized photons can be expressed in terms of spherical harmonics by:

$$\boldsymbol{\varepsilon}^{\pm} \cdot \mathbf{r} = \pm\left[\frac{4\pi}{3}\right]^{\frac{1}{2}} rY_{1\pm1} \tag{10}$$

Here the sign $(+,-)$ in the polarization vector (spherical harmonic) refers to (left-, right-) handed circular polarization. To obtain matrix elements involving three spherical harmonics, the initial $|jm_j\rangle$ states ($m_j = \pm 1/2$ for an L_2 edge) can be expanded in terms $|lm_l\rangle$ of wave functions using Clebsch-Gordan coefficients:

$$\left|\frac{1}{2}-\frac{1}{2}\right\rangle = \phi_{1\frac{1}{2}}(r)\left[\sqrt{\frac{1}{3}}Y_{10}|\downarrow\rangle - \sqrt{\frac{2}{3}}Y_{1-1}|\uparrow\rangle\right] \tag{11}$$

$$\left|\frac{1}{2}\frac{1}{2}\right\rangle = \phi_{1\frac{1}{2}}(r)\left[-\sqrt{\frac{1}{3}}Y_{10}|\uparrow\rangle - \sqrt{\frac{2}{3}}Y_{11}|\downarrow\rangle\right] \tag{12}$$

If we assume that the spin functions are orthogonal and the radial parts of the wave functions are independent of spin, we can define quantities $\rho^{\uparrow}$ and $\rho^{\downarrow}$ that incorporate all the constants and the radial matrix elements. Here $\rho^{\uparrow}(\rho^{\downarrow})$ represents a transition to a spin-up (-down) state and, to a first-order approximation, can be considered proportional to the spin-up (-down) spin density of unoccupied states. Taking a sum on initial and final states, Eq. 8 for left circularly polarized light (ε^+) reduces to:

$$\begin{aligned}\sigma^{+}_{L_2} \sim{}& \frac{1}{3}|\langle 21|11|10\rangle|^2\rho^{\downarrow} + \frac{2}{3}|\langle 20|11|1-1\rangle|^2\rho^{\uparrow} \\ &+\frac{1}{3}|\langle 21|11|10\rangle|^2\rho^{\uparrow} + \frac{2}{3}|\langle 22|11|11\rangle|^2\rho^{\downarrow}\end{aligned} \tag{13}$$

The integrals denote the product of three spherical harmonics of index $|lm_l\rangle$. Here all the null matrix elements of terms with spin opposite to the factors in Eqs. 11 and 12 above have been omitted, and the selection rule $m_f = m_i + 1$ for the integration of three spherical harmonics has been used. Upon performing the integrations and normalizing, Eq. 13 reduces to:

$$\sigma^{+}_{L_2} \sim \frac{3}{5}\rho^{\downarrow} + \frac{12}{5}\rho^{\uparrow} + \frac{3}{5}\rho^{\uparrow} + \frac{2}{5}\rho^{\downarrow} = \rho^{\uparrow} + 3\rho^{\downarrow} \tag{14}$$

Thus left circularly polarized x-rays will be three times more likely to make a transition to a spin-down state than a spin-up state. Repeating the above steps for right-handed circular polarization yields exactly the opposite result,

$$\sigma^{-}_{L_2} \sim 3\rho^{\uparrow} + \rho^{\downarrow} \tag{15}$$

Therefore, right circularly polarized light will make preferential absorption to spin-up states. The same procedure can be repeated for the L_3 edge with an initial $2p_{3/2}$ state to obtain the following results for left and right circular polarization, respectively:

$$\sigma^{+}_{L_3} \sim 5\rho^{\uparrow} + 3\rho^{\downarrow}, \quad \sigma^{-}_{L_3} \sim 3\rho^{\uparrow} + 5\rho^{\downarrow} \tag{16}$$

Here the assumption has been made that the $2p_{1/2}$ and $2p_{3/2}$ radial matrix elements are essentially equal.

$$\rho^{\uparrow(\downarrow)}_{1/2} = \rho^{\uparrow(\downarrow)}_{3/2} = \rho^{\uparrow(\downarrow)} \tag{17}$$

Using Eqs. 14, 15, and 16, the following expressions for the XMCD at $L_{2,3}$ edges can be written as shown in Eq. 18.

$$\begin{aligned}\mu_c(L_2) &\sim \sigma^{+} - \sigma^{-} \sim -2\left(\rho^{\uparrow} - \rho^{\downarrow}\right), \\ \mu_c(L_3) &\sim \sigma^{+} - \sigma^{-} \sim 2\left(\rho^{\uparrow} - \rho^{\downarrow}\right)\end{aligned} \tag{18}$$

Therefore XMCD is, to a first-order approximation, proportional to the spin density of unoccupied states, and the ratio of the signal from the $L_{2,3}$ spin-orbit split states is 1:−1.

APPENDIX B: FIGURE OF MERIT FOR CPX

The figure of merit for CPX sources ($P_c^2 I$) can be obtained by expressing the measured signal as a sum of terms arising from charge and magnetic effects,

$$I^{\pm} \sim \sigma_c I + \sigma_m I, \quad \sigma_c \gg \sigma_m \tag{19}$$

Here $I^{\pm}$ indicates the measured intensities taken with opposite helicities (or magnetizations), I is the incoming beam intensity, and σ_c and σ_m are the charge and magnetic cross-sections. This expression applies not only to XMCD measurements but more generally to other techniques that utilize CPX for magnetic measurements (i.e., magnetic Compton scattering and nonresonant magnetic x-ray scattering). For ferromagnetic materials, the magnetic cross-section depends linearly on P_c (or the magnetization), and thus can be separated out, $\sigma_m \sim P_c\sigma'_m$ making the difference-to-sum ratio

$$\frac{I^{+} - I^{-}}{I^{+} + I^{-}} \sim \frac{2P_c\sigma'_m I}{2\sigma_c I} \sim P_c\frac{\sigma'_m}{\sigma_c} \tag{20}$$

Therefore the measured signal depends linearly on the degree of circular polarization. The percentage error in this quantity, which is the quantity to be minimized in any experimental measurement, is obtained by adding the errors in the numerator and denominator in quadrature:

$$\left(\Delta\%\left(\frac{I^{+} - I^{-}}{I^{+} + I^{-}}\right)\right)^2 = \left(\frac{\sqrt{I^{+} - I^{-}}}{I^{+} + I^{-}}\right)^2 + \left(\frac{\sqrt{I^{+} - I^{-}}}{I^{+} - I^{-}}\right)^2 \tag{21}$$

The first term above will always be much smaller than the second, and thus can be neglected, yielding:

$$\Delta\%\left(\frac{I^+ - I^-}{I^+ + I^-}\right) \cong \frac{\sqrt{I^+ + I^-}}{I^+ - I^-} \sim \frac{\sqrt{2\sigma_c \bar{I}}}{2P_c\sigma_m I} \sim \frac{1}{P_c\sqrt{\bar{I}}} \qquad (22)$$

Therefore, the minimum error in the measurement is achieved by maximizing $P_c \cdot \sqrt{\bar{I}}$. Although this quantity is sometimes referred to as a figure of merit, the experimenter is generally more interested in the amount of time required to perform a measurement on one source compared to another. The time in this quantity is implicitly included in the flux, and therefore to obtain a source comparison for comparable times the quantity above must be squared, yielding a figure of merit of $P_c^2 I$.

J. C. LANG
Advanced Photon Source
Argonne National Laboratory
Argonne, Illinois

X-RAY PHOTOELECTRON SPECTROSCOPY

INTRODUCTION

X-ray photoelectron spectroscopy (XPS) uses x rays of a characteristic energy (wavelength) to excite electrons from orbitals in atoms. The photoelectrons emitted from the material are collected as a function of their kinetic energy, and the number of photoelectrons collected in a defined time interval is plotted versus kinetic energy. This results in a spectrum of electron counts (number per second) versus electron kinetic energy (eV). Peaks appear in the spectrum at discrete energies due to emission of electrons from states of specific binding energies (orbitals) in the material. The positions of the peaks identify the chemical elements in the material. Peak areas are proportional to the number of orbitals in the analysis volume and are used to quantify elemental composition. The positions and shapes of the peaks in an XPS spectrum can also be analyzed in greater detail to determine the chemical state of the constituent elements in the material, including oxidation state, partial charge, and hybridization.

X-ray photoelectron spectroscopy is widely applied to all types of solids, including metals, ceramics, semiconductors, and polymers, in many forms, including foils, fibers, and powders. It has also been used to obtain spectra of gas-phase compounds. In general, it is a nondestructive method of analysis. When applied to solids, XPS is a surface-sensitive technique. The nominal analysis depth is on the order of 1 to 10 nm (10 to 100 monolayers). Surface sensitivity can be increased by collecting the emitted photoelectrons at to glancing angles to the surface. Typical spatial resolutions on a sample surface are on the order of 1 to 5 mm, with so-called small-spot systems having spatial resolutions as low as 25 μm. Analysis times range from under 5 min to identify the chemistry and composition of a sample to 1 hr or more for characterizing chemical states of trace elements.

The best sensitivity of XPS for quantifying elemental composition in solids is on the order of 0.1 at. %. All elements with atomic number greater than three can be detected. The detection sensitivity varies for each element, with some elements requiring greater concentrations to reach a nominal detection threshold in a reasonable analysis time. Relative uncertainties in atomic compositions are typically $\sim\pm$ 10% for samples with homogeneous compositions throughout the sampling volume. This error can be reduced by analyzing standards to calibrate the sensitivity factors of the instrument for each element. Materials with heterogeneous distributions of elements represent a greater challenge to obtaining accurate compositions with XPS. Samples with contaminant or other thin film overlayers will always be found to have attenuated concentration values for the supposed bulk material, and samples exposed to air before analysis typically show accentuated carbon and oxygen concentrations. Atomic concentrations reported from XPS analysis should be understood to be indicative of those from a hypothetical sample of homogeneous composition throughout the sampling volume, unless otherwise stated.

Thin film samples can be analyzed with XPS to determine film thickness. Typical methods of analysis involve measuring composition either as a function of sputter depth into the sample (sputter profiling) or as a function of electron emission angle (angle-resolved XPS). The former method requires the use of a noble-gas ion-etching process to sputter away successive layers of the sample, and is therefore destructive. Sputter profiling is limited in film thickness only by the patience of the user and the robustness of the instrumentation. Angle-resolved XPS is limited to films with a thickness that is on the order of the sampling depth of XPS (1 to 10 nm). Both techniques have other limitations, primarily involving surface roughness for angle-resolved XPS and both surface roughness and film thickness for sputter profiling.

The chemical state information obtained with XPS includes oxidation states or hybridization states for chemical bonds of the elements. Typical examples include determining whether a constituent element is or is not oxidized in a metal alloy or characterizing whether carbon is present as a carbonyl, ether, or ester linkage in a polymer. Application of peak fitting routines is generally required to distinguish peaks from overlapping oxidation or hybridization states. Assignment of chemical state using XPS is not without ambiguities, because other factors besides chemical environment shift the peaks in an XPS spectrum. The use of references with known composition and chemistry can usually help resolve these ambiguities.

The primary limitation of XPS is the need for ultrahigh vacuum conditions during analysis. This generally limits the type of material to those with a low vapor pressure ($<10^{-8}$ mbar) at room temperature and limits the sample size to that which will fit through the introduction ports on the vacuum chamber. Some compounds, such as polymers, can also degrade under the x-ray flux. Another limitation in interpretation of spectra from insulator and semiconductor samples arises from sample charging. This is an

artifact that can be corrected either during or after data acquisition, although the extent of the correction is often not straightforward to determine.

For those wishing to find further information on XPS, a discussion of the available literature is provided at the end of this unit (see Key References).

Competitive and Related Techniques

A common technique that is sometimes considered to compete with XPS for quantitative chemical analysis of materials surfaces is Auger electron spectroscopy (AES; AUGER ELECTRON SPECTROSCOPY). By comparison to XPS, AES is more surface sensitive, with sampling depths typically below 1 nm. Sub-micron spatial resolutions are also possible with AES. In this regard, AES is often the preferred technique for compositional analysis of surfaces in metallurgy. The disadvantage of AES in comparison to XPS is that AES must be done on conductive materials because it uses electrons as the excitation source; it cannot analyze ceramic or polymeric materials as well as XPS can, if at all. In terms of sample damage, AES is also more destructive of ceramic or polymeric materials and molecular layers adsorbed on surfaces. Both XPS and AES have comparable accuracies for quantitative analysis of elemental composition. The sensitivity of AES is better for light elements. Finally, XPS is generally considered as being able to provide much richer information about the chemical state of the elements in the material, although reports have shown exceptions to this (Sekine et al., 1996). This often makes XPS a better technique than AES for analysis of the chemistry of fundamental processes occurring at or on surfaces, such as chemical bonding, adhesion, or corrosion. The two techniques actually complement each other very closely, especially for analytical determination of elemental compositions at the surface of a material. Although the terminology ESCA was introduced for and has become synonymous primarily with XPS because of the resolution of the elemental chemical state information it provides, AES can also be considered as an electron spectroscopy for chemical analysis (ESCA), especially in cases were it is more widely used as an analytical tool for quantitative surface analysis than is XPS.

Another common technique that competes with XPS for analysis of elemental composition is energy-dispersive spectroscopy (EDS; see ENERGY-DISPERSIVE SPECTROMETRY) or energy-dispersive x-ray analysis (EDXA). The EDS technique is commonly available with scanning electron microscopes. By comparison to XPS, EDS has better spatial resolution on the surface but probes deeper into the bulk of the material, on the order of microns. No chemical state (oxidation state) information about the elements can be obtained with EDS. Finally, XPS has a better (elemental) resolution of the lighter elements.

A technique that is closely related to XPS is ultraviolet photoelectron spectroscopy (UPS; see ULTRAVIOLET PHOTOELECTRON SPECTROSCOPY). In UPS, the source is ultraviolet light rather than x rays. The principles of the two techniques are practically the same; both are based on photoemission. Because the UPS source has a lower energy than typical XPS sources, UPS can only probe electrons in the valence band energy region of a material. The sensitivity of UPS to changes in the valence band structure of an element is greater than that of XPS. Therefore, UPS is often a used for determination of band structures of materials, especially semiconductors.

Other techniques of interest in relation to XPS are x-ray adsorption fine structure (XAFS; see XAFS SPECTROSCOPY), extended XAFS (EXAFS), and near-edge XAFS (NEXAFS). These techniques use x-rays of variable energy from synchrotron sources, whereas in XPS, the source energy (wavelength) remains constant. The information obtained from XAFS and NEXAFS pertains more to the fundamental nature of the chemical bond for a specific element in a material rather than the composition of the material. The average coordination number (number of bonds) surrounding a specific atom (element) in a material can be determined with EXAFS.

PRINCIPLES OF THE METHOD

X-ray photoelectron spectroscopy is based on the principle of photoemission. The process is illustrated in Figure 1 for electrons in atomic orbitals. The principles are nearly the same for molecular orbitals and free electron or valence bands. A photon (x ray) of frequency $\nu = c/\lambda$, where λ is its wavelength, has a kinetic energy of $h\nu$. As this photon passes through a material, it can interact with the electrons in the material. Absorption of the x ray by an atom can lead to an electronic excitation of the atom. The photon absorption process conserves energy, leaving an electron in the atom to a first approximation with a kinetic energy (KE or E_K) that is equal to the energy of the incident photon less the initial binding (electronic ground state) energy (BE or E_B) of the electron, $E_K = h\nu - E_B$. Under proper circumstances, when the electron does not scatter

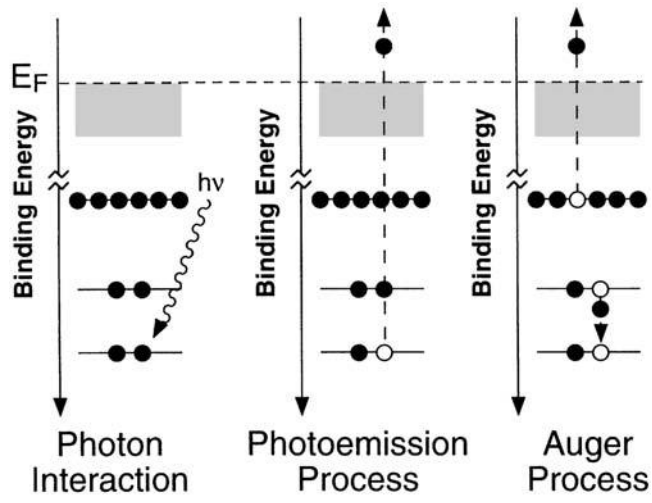


Figure 1. Schematics of the excitation process leading to photoemission of a core electron and the subsequent Auger relaxation process that can occur. Core-level electrons are solid circles, holes are open circles, and the valence band is a shaded box. The position of the Fermi energy is shown by the dotted line at E_F.

back, it acquires a sufficient velocity to escape from the material. In most common XPS instruments, electrostatic or electromagnetic analyzers focus the escaping electrons (photoelectrons) at a specific KE from the sample onto an electron detector. The number of electrons that arrive at the detector is counted for a certain time interval, and the electron counting is subsequently repeated at a new KE value. Alternatively, the number of electrons at the detector is counted concurrently while sweeping the energy selection range of the analyzer. Either method produces a spectrum of number of electrons emitted from the sample per unit time (electron counts per time, or I_i^e) versus the KE of the electron. The KE scale of the spectrum is converted to a BE scale using energy conservation. Peaks appear in the spectrum at energies where electronic excitation has occurred.

The electronic excitation by the photon leaves behind a hole (a site of positive charge). How this hole is subsequently filled is important in the overall energy balance, as discussed in the section on high-resolution scans. One important relaxation process that can occur is the Auger process. This is shown in the third panel in Figure 1. An upper lying (lower BE) electron drops into the hole, perhaps from the valence band as shown. The energy released by this process causes ejection of a second electron called the Auger electron. The Auger process is independent of how the initial hole was created. When electrons rather than x-rays are used to excite the initial core electron, the method is called Auger electron spectroscopy (AES), as detailed in AUGER ELECTRON SPECTROSCOPY. Auger electron peaks will also appear in XPS spectra, and their analysis in this case is considered under the acronym, x-ray induced AES (XAES). The position of an XAES peak relative to a corresponding photoemission peak is used to define a characteristic parameter called the Auger parameter. The Auger parameter is affected by the chemistry of the material. Further consideration of the use of the Auger parameter is given elsewhere (Briggs and Seah, 1994).

Not all photons that enter a material are successful at causing a photoemission event. The probability of such an event occurring is represented by the photoelectron cross-section, σ. Photoelectron cross-sections provide a start in determining the flux, J_i^e, of emitted electrons from a given orbital (or band) for a given x-ray flux F and given number of orbitals n present in the material. The relationship is $J_i^e \propto F\sigma n$, where the electron flux J_i^e has units of electrons/m^2/sec, F has units of photons/m^2/sec, and n is the number of orbitals per unit area viewed by the analyzer (orbitals/m^2). Although the cross-section has explicit units of area (m^2), it contains implicit units of (number of photoelectrons)/(number of photons)(number of orbitals, as can be derived from the flux equation).

The XPS Survey Spectrum

Using the above information, we can prepare a schematic of what might be expected for an XPS spectrum from a pure material, Au, under ideal conditions. A survey, or wide-range scan covers the majority or all of the levels that can be excited by the x-ray energy being used. In what follows, we will use a scale of BE rather than KE. Further consideration of how the x-ray source energy influences the BE scan range and spectrum shape is presented in the section on practical aspects of XPS.

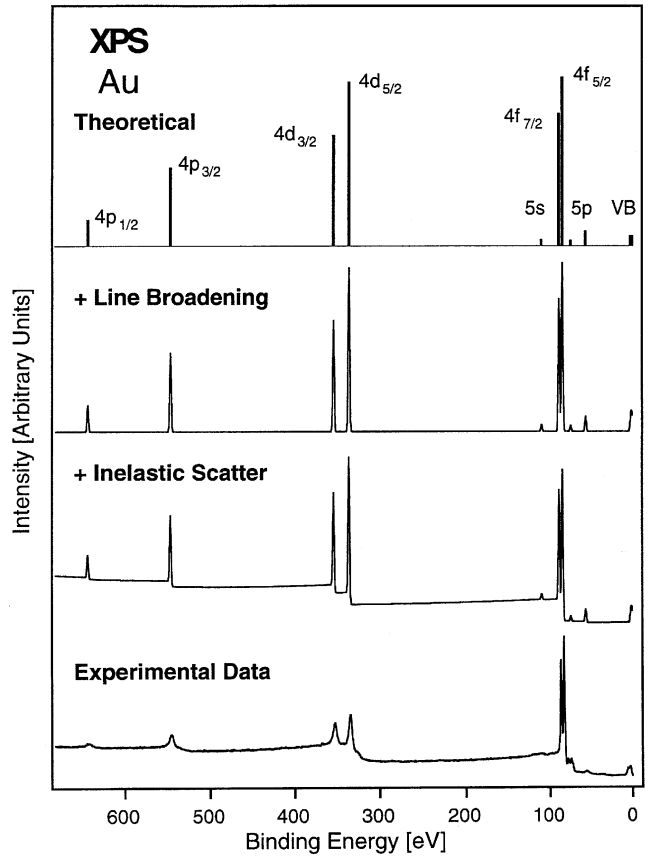


Figure 2. A series of XPS spectra for Au, generated from a theoretical standpoint and incorporating two factors associated with the overall photoemission process from a solid. An experimental XPS spectrum of Au is included for comparison.

Binding energies for electrons in Au can be obtained from standard handbooks (Lide, 1993; Moulder et al., 1995), and values of σ have also been calculated (Scofield, 1976). Using this information, we can generate a theoretical XPS survey spectrum in the BE range 0 to 650 eV for an assumed ideal, monochromatic x-ray source (with an as yet unspecified energy). The result is presented in the uppermost plot in Figure 2. It is essentially a series of delta function lines appearing at the respective BE values for each orbital. The heights of the lines are scaled in proportion to σ for each orbital.

One noticeable feature of the theoretical spectrum is a splitting of p through f orbitals into two peaks. This is due to coupling between the electron spin and orbital angular momenta for these orbitals. The s orbitals do not split because they have no orbital angular momentum. One of two methods can be applied to determine the magnitude of this coupling, j-j coupling or L-S (Russell-Sanders) coupling. The former predominates for atomic numbers below ~20. In j-j coupling, the magnitude of total angular momentum for a given electron is the vector sum of its spin and orbital angular momenta. Since s orbitals have

no angular momentum, electrons in spin-up or spin-down configurations have the same magnitude of total angular momentum. In p orbitals, vector coupling of spin-up or spin-down states with the ± 1 orbital angular momentum vectors gives total angular momentum of either 1/2 (from $|1 - 1/2|$ or $|-1 + 1/2|$) or 3/2 (from $|1 + 1/2|$ or $|-1 - 1/2|$). These two spin-orbit couplings have different binding energies in the magnetic field of the nucleus, leading to the splitting of the levels. In L-S coupling, the total angular momentum of is determined by summing (vectorially) the total spin angular momentum (S) and the total orbital angular momentum (L) after the electron is removed. Similar splitting of orbitals arises. The nomenclature used in XPS to designate the spin-orbit splitting is applied as though j-j coupling predominates and is adopted from spectroscopy (rather than from x-ray notation). It uses principal quantum number, orbital designation (s, p, d, f), and $|1 \pm s|$. For example, the $3p_{3/2}$ peak designates electrons from p orbital of the $n = 3$ level with a total angular momentum of either $|1 + 1/2|$ (spin-up in an "upwardly" aligned p orbital) or $|-1 - 1/2|$ (spin down in a "downwardly" aligned p orbital).

The width of the spin-orbit splitting depends on the magnitude of the coupling between electron spins and orbital angular momentum. This coupling is proportional to $\langle 1/r^3 \rangle$, the expectation value of $1/r^3$, where r is the radius of the orbital. In practical terms, for a given element, the width of the spin-orbit splitting decreases with increasing principal quantum number ($4p$ levels are split more than $5p$ levels), and with increasing angular momentum quantum number ($4p$ levels are split more than $4f$ levels). The strength of the spin-orbit splitting also increases with increasing atomic number because the nuclear charge increases. For example, the experimentally determined $4f$ level splittings of the series Ir–Pt–Au (increasing atomic sizes) are ~3.0 eV, 3.4 eV, and 3.7 eV. The splitting of valence band levels increases within a homologous series going down a period—for example Cu $2p$ (20 eV), Ag $3p$ (30 eV), and Au $4p$ (100 eV)—because the increase due to an increase in nuclear charge overrides the decrease due to an increase in principal quantum number. The spin-orbit split levels have different relative occupancies, roughly in proportion to the respective degeneracies of the levels. Thus, p 1:2, d 2:3, and f 3:4. These relative occupancies are included in the theoretical spectrum in Figure 2 by means of the relative cross-sections.

When the energy of the excitation source $h\nu$ is greater than the work function of the sample, electrons from valence bands (VB) will also be emitted from the sample. This is the principle behind the photoelectric effect. The valence bands are the topmost bonding levels with energies just above the Fermi energy (E_F). In the representative model spectrum for Au, the VB has been modeled as a band of finite width, just above 0 eV in BE, the nominal position of E_F. Excitation from the VB levels is typically better examined with UPS because the excitation cross-sections are greater. Both XPS and UPS provide useful information about the chemical state of a material, with XPS is the only appropriate method of the two for investigations of changes in core levels and UPS applicable to the valence-band region. Studies of valence bands during gas adsorption and reaction phenomena on single crystal surfaces benefit by using both XPS and UPS. The two techniques are also useful for studying the valence band chemistry of polymeric materials.

Photoemission of electrons from solids will not produce the delta function responses shown in Figure 2 for the theoretical spectrum. Even under the assumption of ideal sources and analyzers, we must account for finite lifetime broadening effects. These effects are tied to the time needed for the ion created by the XPS process to relax back to its ground state, neutral atom condition. Heisenberg's uncertainty principle specifies $\Delta E_t \geq \hbar/2\Delta t_l$, where ΔE_l is the expected broadening of the emission feature due to a lifetime uncertainty of Δt_l. Shorter lifetimes lead to larger energy spreads in the photoemission distribution for the respective level. Core and valence band hole lifetimes are on the order of 10^{-14} to 10^{-16} sec, leading to lifetime broadenings on the order of 0.05 to as much as 5 eV. In general, lifetime broadening will be be greater for photoemission from lower-lying (higher BE) levels than from upper-level (valence) levels, because the lower lying holes relax in a shorter time. As discussed in the section on high-resolution spectra, ions created in solids by the photoemission process will also relax differently from those created in gas-phase compounds, leading to a shifting of the apparent BE and to wider photoemission peaks from solids as compared to those from gases. Gas phase photoelectron peaks can have peak widths that reach the theoretical thermal broadening limit of kT (Boltzmann's constant times temperature), or ~0.025 eV at room temperature.

Lifetime broadening was simulated by convolving a Gaussian with a 2 eV half-width with the theoretical spectrum in Figure 2. This does not cover the full range of broadening effects, primarily because the true broadening is not constant for all levels. Lifetime broadened photoemission lineshapes may also contain some Lorentzian character. This is discussed further below (see Data Analysis and Initial Interpretation).

One remaining factor can be illustrated for photoemission from solids using the survey scan range shown in Figure 2. As a photoelectron leaves a solid, it may lose energy to its surroundings. At this point, we will only consider inelastic scattering of the exciting photoelectron, as shown in Figure 3. The scattering occurs through interaction of the outgoing electron with phonon (vibrational) modes of the material. Those electrons that lose too

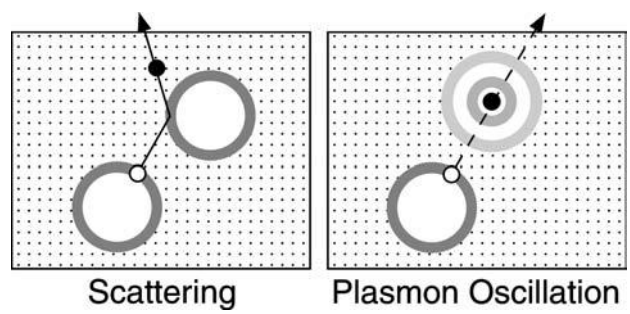


Figure 3. A pictorial representation of electron scatter and the bulk plasmon loss processes that can occur as an electron travels through a solid.

much energy through inelastic collisions will not leave the solid and are of no concern. Those electrons that lose energy yet still can leave the solid are collected at lower KE values than those at which they started. They contribute to an increase in the background on the higher BE side of the main photoemission peak and are not considered part of the primary photoemission process.

In general, the background due to inelastically scattered electrons extends over a range as much as 500 eV in BE above a primary photoelectron peak (which is up to 500 eV in KE below the E_K of the photoelectron). The shape of the loss signal over this range of energy depends on the material being analyzed and the initial E_K of the photoelectron. Each photoemission peak represents an additional source of electrons that can scatter as they leave the solid. Therefore, inelastic scattering leads to an overall increase in the background signal with increasing BE in an XPS spectrum.

Electrons will also scatter inelastically in multiple collisions as they travel through the solid. Because this is a cumulative process, the number of inelastically scattered electrons begins to increase again at low KE values. A peak in number of multiply scattered electrons is reached ~20 to 50 eV in E_K. This peak is called the secondary electron peak. Secondary electron emission decreases to zero at a value of zero E_K. The consequence of the secondary electron peak is that the background intensity in an XPS spectrum will begin to increase dramatically as BE approaches within ~100 eV of the source energy ($h\nu$). This will be shown in spectra presented in the section on practical aspects of the technique.

A simulated electron loss curve was convolved with the lifetime-broadened spectrum in Figure 2. The shape of the energy loss curve was simulated to give a resulting spectrum with a comparable background to an experimental spectrum around the Au 4*f* peaks. The actual shape of the inelastic loss spectrum and the secondary background can be measured with electron energy loss spectroscopy (EELS; see SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). The hypothetical loss spectrum did not account for the secondary electron peak. The result, and the experimental spectrum for Au, are included in Figure 2.

Lifetime broadening and inelastic electron scattering, when convolved with a theoretically generated spectrum of delta functions, can been seen to account for a significant amount of the overall shape of a survey scan in XPS. Some of the remaining differences between the generated and experimental spectra in Figure 2 arise from factors that depend on the source and analyzer. These factors are discussed below (see Practical Aspects of the Method), where the utility of survey scans is also considered. Further differences between generated and experimental spectra become apparent when we focus on the exact shape of a specific photoelectron peak, the background immediately near this peak, and the peak position. These factors are considered below.

The XPS High-Resolution Spectrum

A high-resolution XPS spectrum is one that defines a narrow energy region, nominally 10 to 20 eV, around a photoemission peak. High-resolution spectra are taken to obtain improved quantitative analysis for elements in the material as well as chemical state information about the elements. In the following, we concentrate on the photoemission principles that define the shape of the peak in a high-resolution scan. How high-resolution spectra are analyzed to determine the composition of a material will be discussed below (see Data Analysis and Initial Interpretation).

The chemical state of an element affects the position and shape of its photoemission peak. The chemical state is an initial-state effect, in that it is defined before the photoemission process occurs. Unfortunately, the chemical state of an atom is not the only factor involved in shifting or broadening an XPS peak. Factors that occur after the photoemission process can also influence the apparent BE of the photoelectron. These factors are called final-state effects. Final-state effects will now be considered.

Final-State Effects. Final state effects can be classified into five types, plasmon loss, Auger excitation, relaxation, shake-up or shake-off, and multiplet splitting. All but the first one are concerned with how the core hole, that is left immediately after the photoemission process, is filled. They all shift, and in some cases broaden, the XPS peak.

The plasmon loss process is illustrated in Figure 3. It is a loss process that is similar to inelastic scattering discussed above, and it also leads to an increase in background intensity in survey scans. It occurs when the outgoing photoelectron interacts with the free electrons in a solid material (often called the free electron gas). Unlike what is traditionally considered under the term inelastic scattering, however, the plasmon loss process is quantized and does not extend over a broad energy range below the principal photoemission peak.

Two types of plasmon loss processes are recognized, bulk and surface. In bulk plasmon loss, the outgoing photoelectron loses energy by interacting via collective oscillations with the free electron gas in the bulk of material. These oscillations will be at resonance at the so-called bulk plasmon frequency, ω_p, for the free electron gas. Therefore, the probability that an exiting photoelectron loses energy will be greatest for loss energies defined approximately by $\Delta E_{loss} = h\omega_p$. Values of ω_p can be readily calculated for simple free electron gas models for various materials (Ibach and Lüth, 1990). They range from 3 to 20 eV and decrease with decreasing free electron density (in correspondence with the decrease in ω_p). The increasing magnitude of ω_p is what causes ceramics (and polymers) to be transparent just above or in the infrared range (the lowest energy), semiconductors to become transparent near (lower energy) ultraviolet frequencies, and metals to be opaque to high frequencies (energies) of radiation (up to x-ray frequencies). Surface plasmon losses occur when the outgoing photoelectron interacts with surface plasmon states in the material. The corresponding energies are generally lower than those for the respective bulk plasmon loss.

The net effect of the (bulk or surface) plasmon-loss process is the creation of a new peak or peaks at higher apparent BE (lower E_K) than the main XPS peak. The loss peak

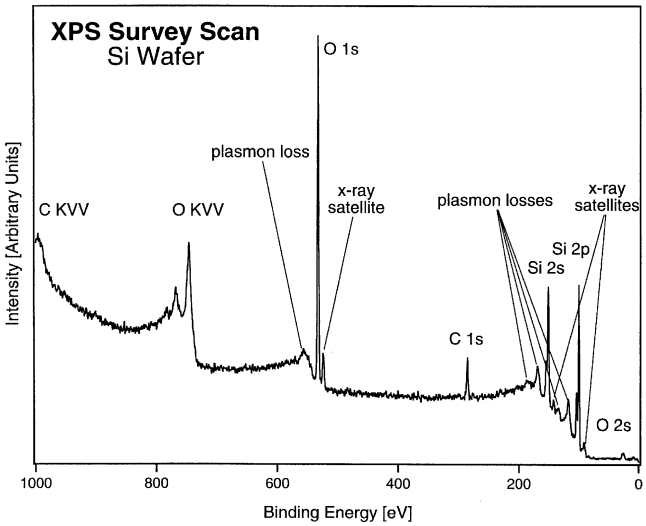


Figure 4. A survey scan from a Si wafer showing plasmon loss peaks that appear above the main Si peaks and O 1*s* peak. X-ray satellites are also labeled.

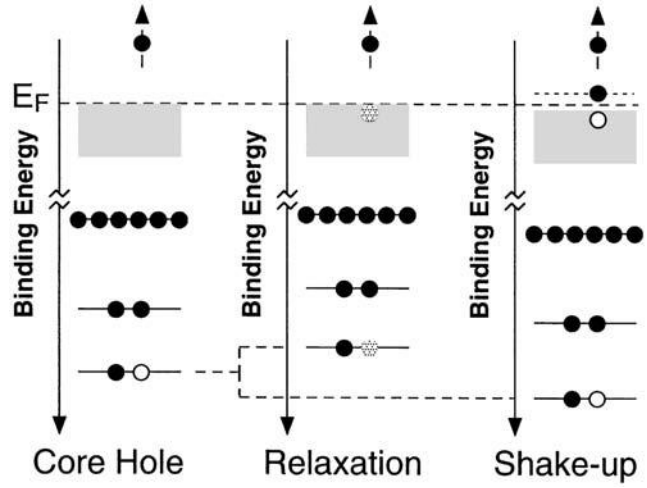


Figure 5. Schematics of the relaxation and shake-up processes in solids and their effects on the position of the core level created by the photoemission event.

will be closer to the primary photoelectron peak in ceramics than in metals. Because the plasmon-loss process is quantized, peaks can also appear at integer multiples of ω_p. This will give rise to a series of peaks at equally spaced increments in BE above the primary XPS peak. The intensity of these multiple plasmon-loss peaks decreases with distance from the primary peak. An example is shown for Si in Figure 4. The Si 2*p* peak has multiple plasmon loss peaks. The position of a plasmon loss peak relative to a primary photoemission peak will depend on the type of material. The loss peak will be furthest away (toward higher BE) from the primary peak in materials with high free-electron densities, such as metals. It will be closest to the main peak in ceramics and polymers. Plasmon loss peaks are broader and generally lower in intensity than the primary XPS peak. They can be distinct in survey scans as well as in high-resolution spectra, as seen in Figure 4. As with inelastically scattered (background) electrons, bulk and surface plasmon loss peaks are not included in the calculations of the total number of electrons emitted from a given orbital for a given element in the sample. In other words, they are not considered in calculating atomic concentrations of elements.

The Auger excitation process is shown in the third panel of Figure 1. It is a means of removing the initial photoemission core hole. The Auger excitation process leaves the original atom in a doubly charged state. This energy reconfiguration will affect the measured BE of the outgoing photoelectron. The consequences are such that the energy gained by filling the initial core hole with an electron is subsequently lost by creation of the second core hole, for a balance that is nearly zero.

The relaxation process is a partial filling of the positive charge in the initial hole. This is represented in Figure 5 by a partial shading of the initial core hole after the photoemission event. The relaxation process can occur via a rehybridization of the atom, in which case it is an intraatomic process. This is illustrated in the figure by a removal of partial charge from the valence band. Relaxation can also occur by extraatomic (also called interatomic) processes, in particular through a process equivalent to diffusion of the localized core hole throughout the valence band of the material.

Relaxation processes typically occur instantaneously in comparison to the lifetime of the core hole and always lead to a decrease in apparent BE for the outgoing photoelectron. This is shown in Figure 5 by the shift of the orbital positions toward the Fermi energy due to the relaxation process. A useful if somewhat naive analogy is to consider relaxation as leaving the atom less positively charged than immediately after the electron is ejected. The outgoing electron senses this as an increase in negative charge, and is correspondingly repelled to higher E_K. Higher E_K translates to lower apparent BE in the XPS spectrum.

The magnitude of the decrease in BE due to relaxation can be on the order of several electron volts. It is distinctly apparent in comparison of gas-phase and solid-phase XPS spectra from compounds. Relaxation is also one reason why work functions of bulk materials are lower than the first ionization energies of the corresponding isolated (gas phase) atoms. Relaxation shifts in photoemission peaks also manifest themselves in XPS spectra from small particles versus bulk solids, although only at particle sizes approaching the scale of nanoclusters. This can occur with samples such as metal particles on heterogeneous catalysts or isolated clusters embedded in a different matrix material. Finally, the magnitude of the relaxation shift can be different for atoms in different chemical environments.

Relaxation occurs to the same extent for all atoms in a material to the extent that the atoms are in the same chemical and structural state. It therefore does not typically result in a new feature (peak) in an XPS spectrum. It may only be apparent in comparing of the position of a photoemission peak from one sample to another. As noted above, it is most prominent when comparing spectra from gas phase and bulk materials or from particles of different

sizes or in different matrices. In materials having clusters of different sizes or particles in different chemical states, variations in the extent of relaxation due to these chemical or structural inhomogeneities can lead to relatively minor peak-broadening effects.

The fourth final-state effect is either the shake-up or shake-off process. This is shown in the third panel in Figure 5. The process involves rehybridization of the ion in such a way as to excite an upper-lying (valence band) electron. In the shake-up process, illustrated in Figure 5, the electronic excitation leads to promotion of a valence-band electron, for example, from the highest occupied molecular orbital (HOMO), to an unoccupied state above the Fermi energy (the lowest unoccupied molecular orbital or LUMO). In the shake-off process, the electron is promoted to a higher virtual state, although it does not necessarily leave the vicinity of the material (in which case, it could be considered an Auger electron).

Both shake-up and shake-off processes shift the BE of the initial hole to higher values. One view of why this is so is shown in Figure 5. The creation of a localized, occupied state above the initial E_F of the material causes an upward shift of E_F. This is equivalent to shifts in E_F that occur when semiconductors are doped. The upward shift in E_F causes the increase in E_B of all electrons in states below E_F. Other factors, such as changes in orbital energies during the rehybridization process, are also involved.

Unlike relaxation, which leads to an overall increase in the background of a spectrum, the shake-up or shake-off processes lead to the appearance of a new peak, called a satellite peak (not to be confused with satellite lines due to the x-ray source). The position of the satellite peak can be up to a few electron volts in BE above the primary photoelectron peak. In some cases, the intensity of the satellite peak can be comparable to that of the main peak. The satellites can also be turned off or on by the chemical state of the element. An example of this is seen in high-resolution spectra of the Cu $2p_{3/2}$ peaks for unoxidized and oxidized foils shown in Figure 6. Satellites are only weakly visible above the main peak for the sputtered foil. The spectrum from the oxidized foil clearly shows two shake-off satellites at ~7 and 10 eV above the main Cu $2p_{3/2}$ peak.

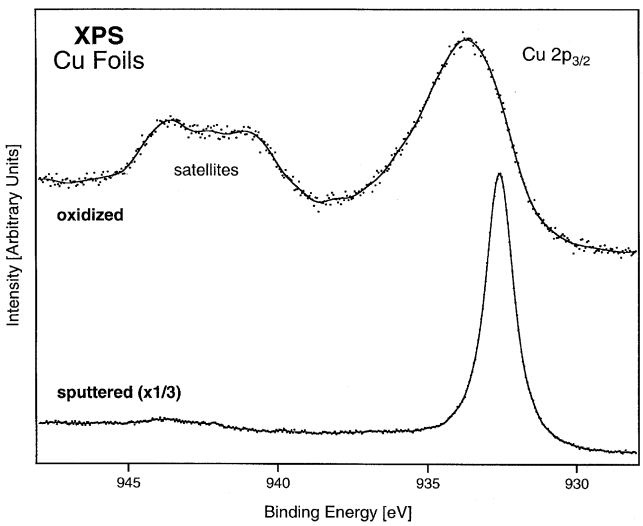


Figure 6. High-resolution spectra from sputtered and oxidized Cu foils showing how the intensity of the Cu $2p_{3/2}$ satellite peaks can be influenced by chemical state.

Because the shake-up and shake-off processes occur after the photoemission process, the electrons that appear as satellite peaks in the XPS spectra should actually be considered as an integral part of the photoemission event. In particular, with reference to the equation $I_i^e \propto F\sigma n$, this means the electrons contributing to the satellite peaks should also be counted to determine the total number of orbitals, n, that produced the photoelectrons. In simpler terms, satellite peaks must generally be considered in all calculations of atomic concentrations. We will return to this point in considering data analysis of high-resolution spectra for atomic concentrations. Looking ahead, it suggests that acquisition of high-resolution spectra should always favor a wider range on the high BE side of a principal photoemission peak in order to cover potential changes in the appearance of satellites that may be present.

The concluding final state effect for consideration is multiplet splitting. Multiplet splitting occurs when the hole created by the photoemission process leaves behind an unpaired electron. The basis for the process is similar to that leading to the spin-orbit splitting of p, d, and f levels discussed in conjunction with survey scans (see discussion of The XPS Survey Spectrum). Multiplet spitting is especially important for photoemission from filled s levels, since photoemission always leaves behind an unpaired electron. The unpaired electron can couple with the nuclear and remaining electron spins as either a spin-up or a spin-down state. These two couplings will have different energies, and thereby shift the position of the orbital. The magnitude of the splitting is on the order of a few electron volts. It can be turned off and on by the chemical state of the atom, as with satellites. The width of the multiplet splitting will be largest for valence s levels, and it decreases in magnitude with increasing BE.

Initial-State Effects. Initial-state effects influence the BE of the orbital before the photoemission process occurs. Three types of initial state effects are recognized: chemical state shifts, inhomogeneous matrix broadening, and charging shifts. The first two are considered below. Charging is an artifact that will be dealt with in the section on problems associated with XPS (see Problems).

Chemical state information is typically what is sought from a high-resolution XPS scan. The spectra in Figure 6 show that we can clearly distinguish between the chemical states of Cu and CuO, for example. The shapes of the Cu $2p$ peaks and their positions are distinctly different for Cu in the neutral (0) and charged (+2) oxidation states. The enhancement of the shake-up satellite adds a distinct signature to the peak shape. Differentiation of chemical state information is unfortunately not this straightforward for all elements, as the background information on final state effects already suggests.

The chemical state of an atom depends on the type of bond it forms: metallic, covalent, or ionic. Dispersive or hydrogen bonding also plays a role, as does the

hybridization in covalently bonded systems. These factors affect the partial charge on the atom. In metals, atoms are expected to be neutral (have an oxidation state of zero). In covalent materials, atoms can have partial charges due to electronegativity differences as well as differences in hybridization. At extremes of electronegativity difference, ionic bonding leads to consideration of the formal oxidation state (charge) of the atom.

Although bonding occurs almost exclusively through valence electrons, the E_B of core levels are also affected, as seen clearly in Figure 6, where Cu 2*p* peaks are shown for Cu^0 and Cu^{2+} states. This can be simply illustrated based on the picture of the electrons in orbitals as charges contained on hollow, conductive shells. The core electrons are contained spatially on shells within the outer valence shells. The work needed to remove the inner shell (core) electrons is affected proportionally by changes in the potentials on the valence shell. In atoms that acquire a partial negative valence charge, the core levels will shift to lower E_B because less work is needed. The core levels in atoms that become more positively charged will correspondingly shift to higher E_B. The analogy presented to explain peak shifts due to relaxation effects is also useful to consider here. An outgoing photoelectron will sense an increase in negative valence charge on an atom as it leaves the atom. This causes the outgoing photoelectron to gain KE in comparison to one that has been ejected from a neutral atom. This leads to a decrease in the measured BE of a photoelectron from a negatively charged ion relative to one from the same orbital in a neutral atom. The opposite shift (toward higher BE values) occurs for electrons leaving ions of greater positive charge.

The chemical shift due to a change in valence charge can be expressed as

$$\Delta E_B = k\Delta\left(\frac{q}{r}\right) + \sum_{j\neq i} k\frac{\Delta q_i}{r_{ij}} \qquad (1)$$

where ΔE_B is the chemical state shift, $k\Delta q$ is the change in valence charge on the atom, and r is the radius of the valence orbital. An increase in negative charge is considered as a positive value of Δq. The summation accounts for interaction with charges from surrounding atoms. This model considers the core level as being confined within a uniformly charged shell of radius r. The radius of the valence orbital will also change as the valence charge changes. Neglecting this effect for small changes in charge, we can replace the valence radius by its expectation (average) value, $\langle r_2\rangle$. We can also neglect the second term when Δq is small, since the r_{ij} values will be larger than $\langle r_2\rangle$. This leads to

$$\Delta E_B \approx k\left(\frac{\Delta q}{\langle r_2\rangle}\right) \qquad (2)$$

A consequence of this is that for the same amount of valence charge shift the chemical state shift in photoelectron E_B will be larger for atoms of smaller radius. Alternatively, less valence charge perturbation is needed to give the same shift in E_B in smaller atoms. We expect therefore to be better able to distinguish subtle changes in chemical state for light elements (with smaller atomic radii) as opposed to heavy elements (with larger atomic radii). This has practical applications in analysis of polymer materials. Although changes in valence charge due to rehybridization can be relatively small, shifts in the peak positions of core levels for carbon, oxygen, and nitrogen often readily observable for different hybridizations. An example of this is shown in Figure 7. After deconvolving the various overlapping peaks, we can clearly see contributions from carbon in aliphatic, ester, and carbonyl hybridization states in the C 1*s* peak from the polymer material.

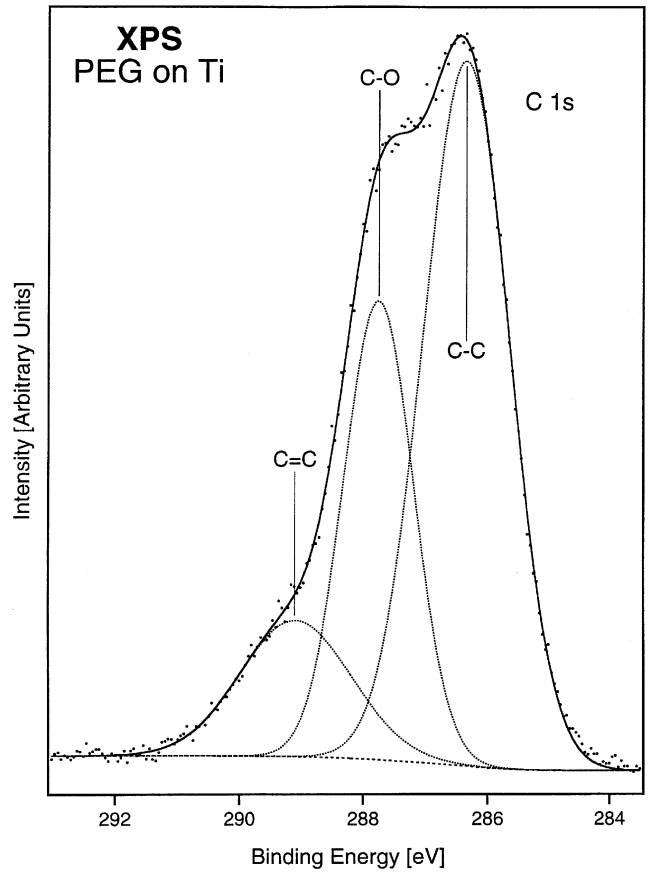


Figure 7. A C 1*s* peak from poly(ethylene glycol) (PEG) adsorbed on a Ti surface. The peak has been resolved into components for the various hybridization states.

We may want to compare shifts in E_B values for an element in two different materials. In this case, Equation 1 becomes

$$E_{B_2} - E_{B_1} = k\left(\frac{\Delta q_2}{\langle r_2\rangle} - \frac{\Delta q_1}{\langle r_1\rangle}\right) + (V_2 - V_1) \qquad (3)$$

where E_{B_i} is the binding energy of a given orbital in chemical state i. The first difference term accounts for differences in valence charge between the two atoms, and the V_i terms are abbreviations of the summation term in Equation 1. They essentially account for differences in Madelung potentials for the atom in the two different states.

Tabel 1. Peak Shifts Measured Going from a Pure Metal to an Oxide[a]

Compounds	Orbital	ΔBE (eV)	R (nm)
Fe/FeO	Fe $2p_{3/2}$	2.9	0.126
Cu/CuO	Cu $2p_{3/2}$	1.8	0.128
Os/OsO_2	Os $4f_{7/2}$	1.3	0.135
Pt/PtO	Pt $4f_{7/2}$	3.2	0.139
$Pt/Pt(OH)_2$	Pt $4f_{7/2}$	1.6	0.139

[a] The corresponding atomic radius of the metal atom is included.

Madelung potentials at a localized atom will change if the matrix surrounding the atom contains ionic or charged constituents and the crystalline symmetry of the matrix changes as viewed from the atom.

The effects of both valence potential and Madelung potential terms may cancel each other. This can lead to counterintuitive peak shifts based solely on considerations of chemical bonding. An example is found by comparison of the position for the Cu $2p_{3/2}$ peak in Cu (932.7 eV), Cu_2O (932.4 eV), and CuO (933.8 eV). The Cu $2p$ peak shifts first to slightly lower BE values going from Cu to Cu_2O, then shifts to higher BE values with the bonding of one more oxygen atom. The initial reversal in E_B of the peak shift going from Cu to Cu_2O may be due to differences in the Madelung potential term in Equation 3 or to differences in relaxation effects for the metal and metal oxide matrices.

Further examples of peak shifts for elements in different chemical states are given in Table 1 (Krause and Ferreira, 1975). The shift in BE going through the sequence $Fe/FeO \rightarrow Cu/CuO \rightarrow Os/OsO$ is increasing with decreasing atomic (or ionic) size ($Fe \rightarrow Cu \rightarrow Os$), suggesting an overriding role due to differences in the valence charges. By comparison, in the $Fe/FeO \rightarrow Pt/PtO$ sequence, BE increases with decreasing atomic size ($Fe \rightarrow Pt$), suggesting a greater role for the Madelung terms. Finally, the BE values in the $Fe/FeO \rightarrow Pt/Pt(OH)_2$ sequence are again more in line with a trend of decreasing shift with increasing atomic size ($Fe \rightarrow Pt$).

The second prominent initial-state effect, inhomogeneous matrix broadening, arises because not all atoms are in exactly equivalent bonding configurations in a material. The BE of an orbital will differ from atom to atom because of localized differences either in composition or structure surrounding each atom. This causes variations in the chemical potential or Madelung potential at each site. A specific orbital will therefore have a value of E_B specific to its atom.

In practice, the shifts in E_B of atomic orbitals due to micro-scale inhomogeneities in otherwise homogeneous materials are well below the detection limits of standard XPS instruments. Since the number of micro-inhomogeneities in a material is expected to be distributed as a Gaussian about a mean type, we expect to see peak broadening in the XPS high-resolution spectra rather than peak shifts. An example of this might be expected in comparisons of XPS spectra from a material with a glassy structure versus a spectrum from the same material in its single-crystal or polycrystalline state.

PRACTICAL ASPECTS OF THE METHOD

The practical aspects of XPS will be covered by first considering the instrumentation, including the source, analyzer, and detector. We will then consider unique aspects of the x-ray interaction with the sample for applications to obtain quantitative composition, chemical state, and film thickness information. Limitations of the method with regard to determining these parameters are discussed below (see Problems). The final part of this section covers important practical aspects related to the use of XPS for analysis of the four basic types of materials: metals, semiconductors, ceramics, and polymers.

Instrumentation

Facilities for XPS analysis consist primarily of an x-ray source, electron analyzer, and electron detector. This is all housed in an ultrahigh vacuum (UHV) chamber with nominal operating pressures below $\sim 10^{-8}$ mbar. One reason for UHV conditions is that mean free paths of the photoelectrons are inversely proportional to gas pressure. At room temperature and atmospheric pressure, the mean free path of an electron in air is only hundreds of nanometers. It is tens of kilometers at 10^{-8} mbar. Significant gas ionization can also occur for electrons traveling in gases above $\sim 10^{-4}$ mbar pressure. Another reason UHV conditions are often important during XPS analysis is to keep the sample surface clean. The impact rate of gas molecules with a surface is also inversely proportional to gas pressure. A common method of estimating coverage is to consider the general rule of thumb that 1 cm^2 of an initially pristine surface can be covered in 1 sec by a monolayer of absorbed gas when put under a pressure of 10^{-6} mbar of gas. Note that this rule of thumb is obtained from the historically important unit of dose called the Langmuir, defined using the pressure unit Torr as 1 liter $= 10^{-6}$ Torr sec/cm^2, where 1 Torr = 1.33 mbar. Suppose that analysis of 1 cm^2 of surface is to take 1 hr. To avoid having the analysis area completely covered by a monolayer of absorbed gas in this time requires that we work at gas pressures below $3 - 4 \times 10^{-10}$ mbar. Although typical survey scans usually take only a few minutes with modern instrumentation, high-resolution spectra can require an hour or more in some cases, and XPS can be sensitive enough to detect the growth of the monolayer in this time. The final reason for using UHV conditions is that laboratory x-ray sources will be damaged at higher pressures, especially when the background gas contains significant concentrations of water vapor or oxygen (see GENERAL VACCUM TECHNIQUES for details on achieving and maintaining UHV conditions).

Sources. Laboratory x-ray sources consist of an electron filament in close proximity to a metal surface. The filament is typically a tungsten wire that is heated by passing a current through it. A high voltage, on the order of 10 to 20 kV, is applied between the filament and the metal to accelerate electrons from the filament to the metal. The current that flows between the filament and the target metal (the anode) is the emission current. Electrons strike the metal

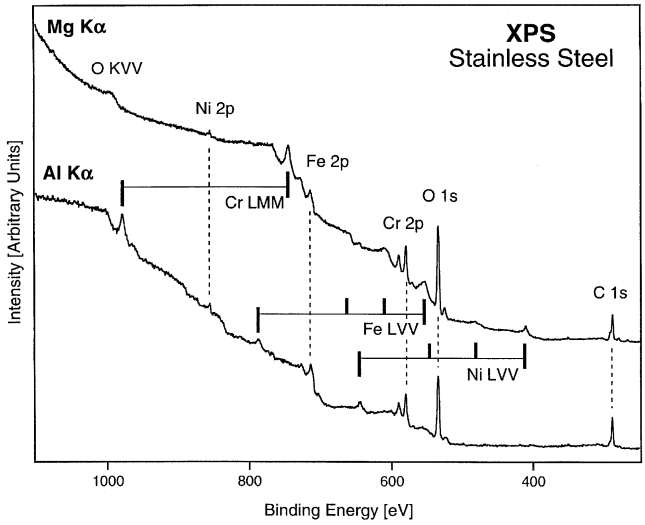


Figure 8. Spectra measured on a stainless steel alloy using either Mg Kα or Al Kα radiation. The shifts of the Auger lines are noted, while the photoemission peaks are seen to stay in one place on the binding energy scale. The background energy between the two spectra is also different at high binding energy due to the onset of the secondary loss peak.

anode and, in deaccelerating, produce x rays. Two types of x rays are produced, Bremsstrahlung ("braking radiation") and emission lines. Bremsstrahlung is a broad background underlying the emission lines and is independent of the choice of metal. Because it is low in intensity and broad in energy, Bremsstrahlung has no utility for XPS. The frequency (energy) of the principal emission line from a laboratory x-ray source is characteristic of the anode metal. The two most frequently used laboratory anodes in XPS are Al and Mg, with primary Kα emission lines at 1486.6 eV and 1253.6 eV, respectively. Other metals are sometimes used to have lower or higher excitation energies. Many laboratory XPS systems have a dual-anode arrangement that permits selection of one or another type of anode metal. The advantages of this arrangement are shown in Figure 8.

The peaks from AES electrons will always appear at constant E_K regardless of the excitation source. Their peak positions will therefore shift on the BE scale as a function of excitation source energy. This is clearly seen in Figure 8, where the AES peaks have shifted by 233 eV to higher BE when Al Kα is used as the excitation source rather than Mg Kα. The true photoemission peaks remain at the same position in BE. In cases where Auger peaks overlap with principal photoemission peaks, having a second x-ray source of a different excitation energy can sometimes help resolve any ambiguities in peak assignments.

The background intensity also increases more sharply at higher BE values for the XPS spectrum taken with an Mg x-ray source compared to that using an Al source. This is due to the onset of the secondary electron peak. For the Mg source, the increase in secondary electron background should start at ~1150 eV (1253 to 100), whereas for the Al source, it should start at ~1380 eV. This means that an Al source can be used to probe core levels at higher values of BE than a Mg source can. This is the same principle that limits UPS to valence band spectra.

Dual x-ray sources (Mg and Al) are typically used in conjunction with one another to shift Auger peaks out of the way of primary photoemission peaks. X-ray anodes with lower excitation energy, such as Ti at 395.3 eV, are also useful. The cross-section for photoemission of electrons from low E_B orbitals is sometimes greater at lower excitation energy, leading to greater sensitivity. In other words, a lower amount (concentration) of an element is needed in a material to give a specific signal intensity in an XPS spectrum when its cross-section is higher. The tradeoff to this is, as the excitation energy of the x-ray source decreases, fewer of the higher atomic number elements can be detected.

Principal emission lines from laboratory x-ray anodes are not monochromatic. A lineshape analysis shows they have a characteristic spreading about the principal emission energy. The x-ray half-width of Al anodes is ~0.85 eV and of Mg is ~0.7 eV, for example. This leads to further broadening of the photoemission peaks in the XPS spectra beyond lifetime broadening effects. The total peak broadening due to lifetime and x-ray source effects will be $\Delta E_T^2 = \Delta E_1^2 + \Delta E_s^2$, where ΔE_s is the source broadening. For a state with an intrinsic width of 0.10 eV, its XPS peak measured with MgKα will have a width of 0.71 eV. Laboratory x-ray sources also have x-ray satellite emission lines, typically at higher energy from the main emission line. Both Al and Mg have a number of such secondary emission lines (Krause and Ferreira, 1975). Their intensities are generally ~10% or lower than that from the primary x-ray emission line from the source. The secondary emission lines from an x-ray source lead to additional photoemission peaks at lower BE in the XPS spectra. These peaks are called x-ray satellites. For example, Kβ lines appear in the XPS spectra in Figure 4 at BE values ~7 eV lower than each of the principal peaks. X-ray satellite peaks should not be confused with the satellite peaks that are characteristic of final state shake-up and shake-off processes. The former appear at lower BE values than the principal photoemission peak and the latter at higher BE values.

In operating laboratory x-ray sources, one characteristically defines any two of the parameters of applied voltage (V), filament emission current (I_F), or x-ray power (P) to define the operating characteristics of the source. They are related as $P = I_F V$. For a given x-ray source geometry, setting any two of these parameters defines the flux of x rays to the sample.

Most XPS systems in use today are equipped with an x-ray monochromator between the x-ray source and the sample. A typical x-ray monochromator is essentially an array of x-ray diffracting elements (quartz crystal strips for example) positioned to diffract the principal emission line from a conventional x-ray source onto the sample. Systems using a monochromator have a number of distinct advantages over conventional laboratory x-ray sources. First, the monochromator removes the x-ray satellite lines and significantly reduces Bremsstrahlung background radiation, leading to cleaner spectra, especially in survey

scan mode. Secondly, an x-ray monochromator reduces the broadening of the x-ray source line significantly, making the resolution of overlapping features in high-resolution XPS spectra an easier task. Finally, a monochromator reduces the potential for sample damage, in one case because of lower overall x-ray flux and in another because inadvertent bombardment by electrons through damaged x-ray windows is no longer possible. Although use of a monochromator decreases the net x-ray flux to the sample, the output flux of modern x-ray sources and the sensitivity of electron detectors have improved to the point where this loss is overridden by the improvement in signal quality that is obtained. For these reasons, monochromatic x-ray sources are to be preferred for XPS analysis in all cases where a choice is offered.

Synchrotron radiation is also a viable source of x-ray radiation. Because the use of synchrotron radiation involves a movable monochromator between the beam line and the analysis system, the advantages are its tunability through a range of x-ray energies, high flux, narrow linewidth, low background intensity, and absence of satellites. The disadvantages are that one must schedule time on a synchrotron facility and be adept at the use of such facilities. Collaboration with research groups doing comparable studies is encouraged at the outset. The tunability of the x-ray energy from synchrotron sources is often exploited for x-ray absorption fine structure (XAFS; see XAFS SPECTROSCOPY) or near-edge XAFS (NEXAFS) studies. Both techniques are related to XPS. They look at the intensity of photoemitted electrons from a sample at a well-defined BE as a function of the energy of the incident x-ray. Further information is provided in XAFS SPECTROSCOPY.

Analyzers. Two types of analyzers are widely used for XPS analysis, the cylindrical mirror analyzer (CMA) and the hemispherical sector analyzer (HSA), shown in the schematic in Figure 9. Historically, the HSA pre-dates the CMA in design (Palmberg et al., 1969; Siegbahn et al., 1969a,b; Palmberg, 1975). The CMA is popular for AES systems, and further details of its configuration and operation can be found elsewhere (Palmberg et al., 1969; Palmberg, 1975). The HSA has become the norm for XPS analysis and will be considered exclusively in the following discussion.

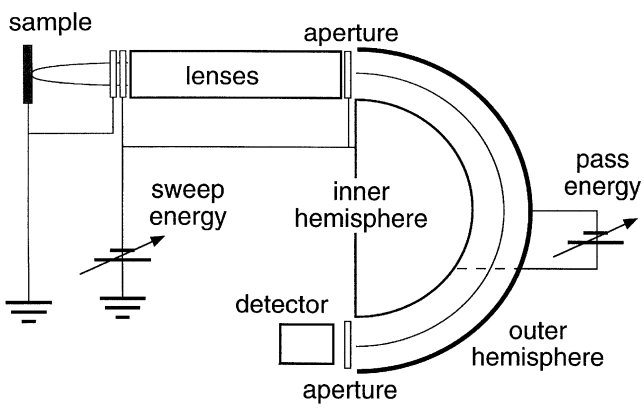


Figure 9. A schematic representation of the key parts of a hemispherical sector analyzer (HSA).

The HSA is recognized on XPS systems by its distinctive clam-shell appearance. This hemispherical portion of the analyzer serves as an electron monochromator. It is typically preceded in the electron path by a series of electrostatic lenses and is followed by an electron detector.

Photoelectrons leaving the sample are collected by the lens system. The frontmost lens is at sample potential (ground) to provide a field-free region between the analyzer and sample. The lenses focus the electrons spatially to the front aperture of the HSA and may provide image magnification. In principle, an image of the aperture is focused by the lenses onto the sample, and the bounds of this image define the analysis area. Retarding fields in front and back of the lenses serve as energy selectors. These retarding fields are swept in potential (voltage) to determine the cutoff of electron E_K values that reach the analyzer. In doing this, the lenses retard the electrons in KE; they do not perform any further energy discrimination. Electrons reach the analyzer with a full range of E_K values above that selected by the sweep energy on the lenses.

The function of the HSA is to select electrons of a particular E_K to pass to the detector. The optimal E_K for an electron to pass through an HSA (or CMA) is called the pass energy, E_p. It is defined for an HSA of inner and outer radii R_i and R_o, respectively, according to

$$E_p = K_s \Delta V$$
$$K_s = \left(\frac{R_o}{R_i} - \frac{R_i}{R_o}\right)^{-1} \quad (4)$$

where K_s is a spectrometer constant and ΔV is the potential difference between the inner and outer hemispheres (outer minus inner potential). For a given set of voltages (potentials) on the inner and outer shells of an ideal HSA of a specific size, only electrons at E_p in KE will pass through to the detector. Those electrons with KE higher than E_p are lost to the outer hemisphere of the HSA, and those that travel too slowly are lost to the inner hemisphere.

All hemispherical sector analyzers allow electrons with E_K values that are not exactly at E_p to reach the detector. This is determined solely by the geometry of the analyzer. The range of E_K values that are passed, $E_p \pm \Delta E_A$, describes the absolute energy resolution, ΔE_A, of the HSA. The primary geometric parameter of concern in establishing ΔE_A is the acceptance angle, α, of electrons at the entrance slit of the HSA. Electrons with E_K equal to E_p that enter the HSA normal to its entrance slit ($\alpha = 0$) will always be transmitted; this defines the ideal case. To increase the number of electrons that can be transmitted through the HSA, we can also allow it to transmit electrons with E_K equal to E_p that enter at an angle α from the normal to the entrance slit. This is determined in practice by specifying the lens and aperture designs according to electromagnetic principles, which typically requires expansion of the governing equations for electron travel in a Taylor series about α. The optimal design for an HSA is to be focusing to second order in α.

When this second-order focusing constraint is applied, a limit on ΔE_A is obtained. It is found to be a function of the entrance slit width W and geometric radius of the optimal path through the analyzer R_p in the form

$$\frac{\Delta E_A}{E_p} \propto \sqrt{W/R_p} \tag{5}$$

where $\Delta E_A/E_p$ is the relative resolution. An optimally designed HSA is therefore not perfectly monochromatic; it introduces peak broadening of the order of ΔE_A. The total peak broadening in an experimental XPS spectrum due to lifetime, source, and analyzer effects is therefore $\Delta E_T^2 = \Delta E_l^2 + \Delta E_s^2 + \Delta E_A^2$. Analyzers with larger values of R_p or smaller values of E_p will give smaller values of ΔE_A, meaning better absolute energy resolution. In practice, typical XPS systems with an HSA are designed with $\Delta E_A/E_p \sim 0.1$. Specialty XPS instruments exist with far better relative resolutions, in which case the diameters of the hemispheres are significantly larger.

The transmission function of the HSA is an important factor to consider for quantitative analysis with XPS (Nöller et al., 1974). It defines how well the lenses and analyzer transmit an electron that leaves the sample at a given E_K. Three cases have been recognized: (a) where the lenses define the transmission, (b) where the lenses define the transmission only along one plane perpendicular to the sample along the lens axis, and (c) where the analyzer defines the transmission. In general, for XPS analysis using an HSA, the HSA defines the transmission except at low E_K (high E_B).

The HSA can typically be operated in one of two modes. In one, the analyzer transmission function is held constant. The transmission function of the HSA varies as $1/E_K$ and as $1/\sqrt{E_p}$. In fixed analyzer transmission mode, the value of E_p is varied to give a constant transmission function over the entire sweep range of KE in the spectrum. The pass energy increases with increasing KE in the spectrum. Because $\Delta E_A/E_p$ depends only on the analyzer geometry, the absolute resolution ΔE_A correspondingly varies with KE in the scan. In practice, this leads to broader peaks at low BE (high KE, where ΔE_A is large) compared to high BE. Constant transmission function mode can be used for AES analysis with an HSA. Peak widths in AES are much broader than those in XPS, and the variation in ΔE_A as a function of E_p is correspondingly not as important. In addition, having a low transmission function at low KE in the scan significantly reduces the overall size of the secondary background signal (below ~50 eV in KE). While constant analyzer transmission mode may have some utility for XPS survey scans, it is to be avoided for high-resolution XPS scans and in all cases where accurate elemental compositional information is desired.

The more common mode of operation of an HSA for XPS analysis is to hold E_p constant while the input sweep energy is varied. In practice, this is done by retarding the electrons through the lenses so that electrons leaving the sample at the desired EK arrive at the entrance slit to the analyzer with EK equal to E_p. Because E_p is constant throughout a scan, ΔE_A is constant over the entire BE range of the XPS sweep. Constant pass energy mode is the one to use for high-resolution scans where constant absolute energy resolution is essential in order to obtain reliable chemical state information. In addition, for automated analysis, elemental compositions must be determined from scans taken with the XPS instrument in constant pass energy mode.

In constant pass energy mode, the transmission function of an HSA typically varies approximately as $1/E_K$, which is the KE of the electrons at the entrance to the lens. This means that even if the flux of electrons emitted from a sample is constant over the entire KE range, the apparent intensity of electrons measured with an HSA in constant E_p mode will decrease with decreasing BE. In experimental spectra, peaks at high BE will be accentuated by the HSA relative to those at low BE. This effect will be considered again in later discussions regarding quantitative analysis using XPS, especially for ARXPS.

At a given value of KE in a scan, the transmission also varies as $1/\sqrt{E_p}$, so that lower values of E_p will give better absolute energy resolution and higher analyzer transmission. This latter point can lead to confusion because, in practice, when E_p is decreased, absolute resolution (ΔE_A) increases but the measured signal intensity decreases (suggesting that transmission function decreases). The resolution of this apparent contradiction is to realize that all electrons that exit the HSA are counted as part of the signal at the given KE in the scan, regardless of what their E_K actually was. In other words, the intensity that is measured at the detector as belonging to a signal at E_K, $I^m(E_K)$, is actually the integrated signal between $E_K \pm \Delta E_A$

$$I^m(E_K) \propto \int_{E_K - E_A}^{E_K + E_A} J^e(E)\,dE \tag{6}$$

As analyzer pass energy is decreased, the relative fraction of electrons that are transmitted by the analyzer (the transmission function) increases (in proportion to $\sqrt{E_p}$). However, the integration range in the above equation decreases (in proportion to E_p). Because the latter effect dominates, measured intensity decreases.

Survey scans are typically taken at a higher pass energy where the signal intensity $I^m(E_K)$ is large but the absolute energy resolution ΔE_A is poor. The objective is to obtain a high signal-to-noise ratio in a relatively short time; the objective is not primarily to resolve overlapping peaks. This was the case for the experimental spectrum in Figure 2. High-resolution scans are generally run at lower E_p to obtain peak resolutions approaching the inherent lifetime and source broadening limit. This means that high-resolution scans typically require longer acquisition times to get comparable signal-to-noise ratios.

Hemispherical analyzers for XPS analysis are more frequently being designed to be spatially focusing. This is because x-ray flux cannot be easily focused, and we often want to analyze only specific spatial regions on a sample. One way to achieve spatial focusing with an HSA is to insert mechanical apertures with various diameter entrance slits into the electron path through the lens,

typically at a point just prior to the analyzer. The image of this aperture is focused back to the sample, and smaller apertures select smaller spots on the sample. While this method has, in principle, no lower limit in object size at the sample, it is limited by the sensitivity of the detector systems. The nominal spot sizes that can be obtained using such mechanical aperture systems are on the order of millimeters, with hundreds of microns being possible only with extremely long (tens of hours) signal acquisition time.

Small-spot, or imaging XPS refers to systems that have electrostatic or electromagnetic lenses that spatially focus the analyzer entrance aperture to a defined spot at the sample (Seah and Smith, 1988; Drummond et al., 1991). No mechanical apertures are used, and the (mechanical) entrance slit width is always a constant size. Small-spot systems in use today can obtain spatial resolutions at the sample on the order of a hundred microns routinely, with acquisition times well within an hour for high-resolution scans. Lower limits in analysis spot size at the sample are currently $\sim$25 μm.

Another advance in analyzer and electrostatic lens design has led to XPS systems that can scan the analysis spot over the sample while acquiring data. Scanning XPS systems are typically coupled to small-spot analysis capabilities. This allows acquisition of chemical state maps over a sample. In this case, the intensity of photoelectrons at specific BE values is plotted versus spatial position. Within the spatial resolution limit of the instrument, scanning XPS can be used to determine such information as the uniformity of coatings on surfaces.

Detectors. Three types of electron detectors are common on XPS systems: electron multipliers, channeltrons, and channel plate arrays. The electron multiplier functions in a manner similar to an amplifier tube, often using multiple amplifier stages connected in series. The channeltron is a smaller, single-component system. The channel plate array is a circular, wafer-shaped detector with a multitude of narrow channels that each function, in a sense, as a separate channeltron. All three detectors have gain factors on the order of 10^3 to 10^6, meaning that one electron in will produce 10^3 to 10^6 electrons out. All of the detectors run at high voltage, often as much as 5 kV. They all require secondary amplification stages external to the detector to boost the output signal further. The external amplifiers are typically connected as closely as possible to the output stage of the detector to reduce concurrent amplification of circuit noise.

An increasingly common arrangement in XPS systems is to use multiple electron detectors arranged for position-sensitive detection. This makes use of the fact that electrons leave the exit aperture of the analyzer at different angles depending on their E_K. Multiple detectors can be configured to collect and process these electrons simultaneously, thereby increasing the signal-to-noise ratios obtained for comparable acquisition times with single-detector schemes.

For the electron multiplier or channeltron, position sensitivity is realized by positioning multiple detectors at key positions clustered around the exit aperture. At any instant, as one detector positioned off-center from the aperture is collecting electrons at $E_p + \Delta E_A$, the on-axis detector(s) collect electrons at E_p, and an opposing detector to the first collects electrons at $E_p - \Delta E_A$. With channel plate arrays, the entire channel plate detector can be hard-wired to become a position-sensitive detector. The positions of electrons arriving on diametrically opposing sides of the circular array are typically sensed and recorded by a four-element resistive network attached in a four-fold symmetrical pattern to the outer edge of the channel plate.

Electron Detection. The final consideration with regard to instrumentation in XPS is how the E_K of the photoelectron is sensed. This is not the same as how the analyzer selects for a particular KE of incoming photoelectrons. The latter is done by applying a retarding field to the lens system, as discussed above. Photoelectrons are counted when they successfully travel through the lens and analyzer systems because they have the appropriate E_K. However, photoelectrons have a different KE as seen from the reference frame of the sample versus the reference frame of the analyzer. The reason for this is shown schematically in Figure 10.

Photoelectrons leave the sample with a E_K determined by their initial E_B, the excitation energy $h\nu$, and the work function of the sample, ϕ_s. The work function of the sample is the difference between the Fermi energy of the sample E_F and the vacuum level. The vacuum level at the sample is where the photoelectrons have zero E_K as seen by the sample.

When the photoelectrons reach the detector, they will have traveled through an analyzer with an overall work function of ϕ_A. The analyzer work function is a constant that is independent of all intervening potentials (voltages) between the sample and the detector. A constraint on a properly configured electron analysis

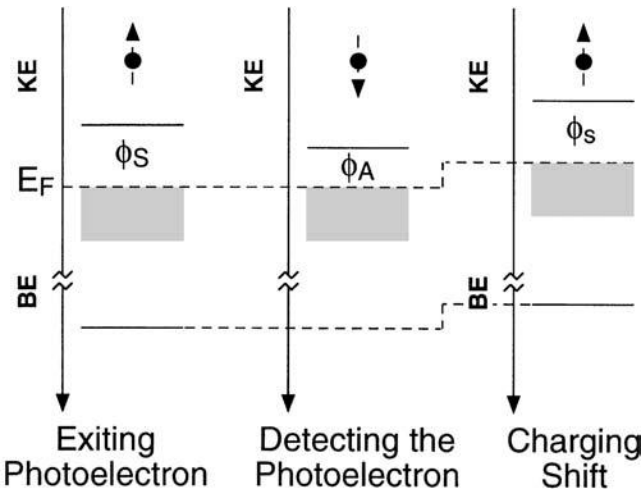


Figure 10. A schematic representation of the translation of electron kinetic energy (KE) to binding energy (BE), with consideration of the work functions of the sample and analyzer. How sample charging affects the considerations is represented by the panel on the far right.

instrument is that the Fermi energy of the sample and the system are equal, and electrons are "detected" (counted) when they reach the Fermi energy of the system. Because the work functions of the sample and analyzer are generally different, the electron will appear to have either a greater KE (as shown in Fig. 10) or lower KE from the perspective of the analyzer. The perspective of the analyzer, not the sample, is the one that counts in sensing the E_K of the electron as it passes through the analyzer.

Combining all of the above considerations, we can write a series of equations that lead to the formal relationship between the E_B of the electron in the sample and the E_K of the electron as it is sensed by the analyzer, where the S and A subscripts denote sample and analyzer, respectively.

$$E_{KS} = h\nu - (E_B + \phi_s)$$
$$E_{KA} = E_{KS} - (\phi_A - \phi_S)$$
$$E_{KA} = h\nu - E_B - \phi_A$$
$$E_B = h\nu - E_{KA} - \phi_A$$

The final equation above states that we must know the work function of the analyzer to know the absolute values of E_B for electrons in samples. In practice, the analyzer work function is a calibration constant for the instrument. It is determined by aligning peak positions from a sample with those of a reference. The typical calibration sample is Au, with the position of the Au $4f_{7/2}$ peaks being set at 84.0 eV. The issue of instrument calibration will be addressed more thoroughly in the section on problems with the technique.

As an example of the above equations in practice, consider a C 1*s* electron with $E_B = 284.6$ eV excited by Mg Kα radiation (1253.6 eV) from a sample with a work function of 5.0 eV and measured through an analyzer with a work function of 4.0 eV. The electron will appear to have $E_{KS} = 1253.6 - (284.6 + 5.0) = 964.0$ eV from the sample. It will appear to have E_{KA} 965.0 eV to the analyzer and will be recorded at this value if KE is used as the scale on the XPS spectrum. In a properly calibrated spectrometer, the E_B of this electron will be determined to be $1253.6 - 965.0 - 4.0 = 284.6$ eV.

Maintenance. The importance of maintaining a well-calibrated instrument for any analytical technique cannot be overemphasized. The sections that follow will highlight how instrument parameters affect quantitative results in XPS analysis. Both compositional (peak areas) and chemical (peak positions) information can be unknowingly distorted by improper or inattentive maintenance and calibration of an XPS instrument. A number of articles are available that specifically discuss calibration procedures in greater detail. In particular, the series of articles by Seah and others is highly recommended for consideration of instrument parameters that affect determination of compositional information with XPS (Anthony and Seah, 1984a,b; Seah et al., 1984; Seah, 1993). A corresponding set from Anthony and Seah is recommended for discussion of calibration of energy scales (which affect chemical information; Anthony and Seah, 1984a,b). A summary article pertaining to both AES and XPS was also published (Seah, 1985). Those who are involved in daily operation of XPS systems are strongly encouraged to apply the procedures set forth in these articles as part of their instrument maintenance and calibration routines. Laboratory experiments can also be designed around these articles as a method of introducing graduate students to the rigor needed for accurate and reproducible quantitative analysis of peak parameters with XPS.

Applications

In analysis of solid materials, the two primary uses of XPS are to determine elemental compositions and to characterize the chemical states of elements. Other applications include determining film thickness or composition profiles at the surface of a material, either through nondestructive or destructive means, and characterizing molecular orientation for adsorbates on surfaces.

The discussion below considers the practical factors that are involved in determining compositions and chemistry using XPS. The use of XPS for determining film thickness and molecular orientation in absorbed layers is also outlined. Finally, a brief review of materials-specific considerations is presented.

Elemental Compositions. The basis behind the use of XPS to determine elemental compositions is provided above (see Principles of the Method). A relatively more rigorous development is given in what follows, using what has been learned from both the principles of the technique and practical aspects of the instrumentation. Those interested in further details and alternate derivations are referred to the reference list (see Key References) and to Jablonski and Powell (1993). The formulation below starts from the perspective of an ideal analyzer viewing an ideal sample at an angle that is fixed relative to the surface plane. Nonideal factors introduced by practical analyzers will then be introduced. Considerations of nonideal samples will be covered at the end.

The flux of electrons emitted from a given orbital for a given element, i, in a sample, J_i^e (electrons/m^2-sec), for a given x-ray flux F (photons/m^2-sec) can be written as

$$J_i^e = F\sigma n_i \qquad (7)$$

where σ (in units of m^2) is the cross-section and n_i (in units of 1/m^2) is the number of orbitals per unit area over which the x-ray beam impinges. We can recast this expression in terms of elemental (atomic) concentrations in a bulk solid by accounting for the volume sampled. For the electron KE values of interest in XPS, x rays penetrate much further into a sample (on the order of microns) than the electrons can escape. Therefore, the total flux of electrons from a bulk sample is limited by the probability of the electron escaping. As discussed above (see Principles of the Method), inelastic scattering and plasmon loss processes are the two main factors that remove photoelectrons from consideration as part of a primary photoelectron peak, where the former dominates. In analogy with scattering processes in ideal gases, the escape probability can be

represented by a mean free path, λ_i (in units of m), for the electrons to travel before scattering. The escape probability decays exponentially with distance traveled relative to λ_i. Therefore

$$J_i^e = F\sigma \int_0^\infty C_i(z)\, \exp(-z/\lambda_i) dz \quad (8)$$

Assuming a homogeneous composition C_i (in units of atoms/m^3) of element i throughout the material, this equation becomes

$$J_i^e = F\sigma\lambda_i C_i \quad (9)$$

This represents the total photoemission flux from orbital i in a homogeneous sample.

A wide range of references are available on the subject of electron mean free paths in materials (Tanuma et al., 1988, 1991a,b, 1996; Marcu et al., 1993). Electron mean free paths increase sharply with decreasing KE at low E_K because the quantum-mechanical wavelength of the electron increases, and this reduces the inelastic scattering efficiency. They increase as $\sqrt{E_K}$ at high KE because the electrons essentially travel further between scattering events. The minimum mean free path is ~0.2 to 0.3 nm in the range of 30 to 100 eV. Mean free paths are also dependent on the type of material, its compositional homogeneity, crystal structure, and microstructure. Variations in electron mean free paths by factors of 2 to 3 are not uncommon from material to material for the same photoelectron E_K from a specific orbital in a given element.

Once the electrons are emitted from the sample, they must pass through the analyzer and be detected. This gives rise to considerations of nonidealities in the instrument, in particular of the analysis system. The relationship between emitted flux J_i^e and measured intensity I_i^m (number per unit time) can be expressed succinctly as

$$I_i^m = f_s f_n \theta_s TDAJ_i^e \quad (10)$$

where A is the sampling area on the surface, D is the transmission coefficient of the detector, T is the transmission coefficient of the lens and analyzer, θ_s is a collection efficiency factor for the frontmost lens of the spectrometer, f_n is a response function for the lens, analyzer, and detector, and f_s is a response function for the system electronics. The five coefficients other than sampling area generally depend on E_K of the incoming photoelectron (as viewed by the analyzer). They range between zero and unity. How they depend on E_K cannot always be exactly determined. This introduces errors in the consideration of atomic concentration, as discussed below (see Data Analysis and Initial Interpretation).

The transmission function of the analyzer has been discussed in the previous section and is typically the most prominent factor considered in the above equation. The detector response function D defines the gain of the detector as a function of the E_K of the electron at the detector. It is typically important to consider only when the pass energy varies, which should not occur during quantitative XPS analysis. In most practical cases, it is not possible or even really necessary to determine D separately from the value of T.

The angular efficiency factor θ_s is an abbreviated form of an integral over the solid angle of view of the analyzer, often called the analyzer acceptance angle. It accounts for any potential dependence of photoemission intensity from a solid on the analysis angle, often called the electron take-off angle. In some cases, this factor is also abbreviated as $d\sigma/d\Omega$ and termed the differential cross-section as a function of measurement angle. Even for pure elements as solids, photoemission intensity does indeed depend on emission angle, and therefore on take-off angle, especially for single crystal substrates. This is due both to the angular dependence of photoemission from orbitals of nonspherical shape as well as to diffraction phenomena. The angular efficiency factor is for the most part only important in analysis using XPS when considering spectra from highly ordered (single crystal) materials as a function of take-off angle or when considering variations in intensity as a function of analyzer acceptance angle. The former is typically a far more widely performed measurement than the latter.

The two response-function terms, f_n and f_s, account for non-idealities in the instrument and the electronics respectively. Basically, these terms account for temporal variations in such things as stray magnetic field strengths within the analyzer or fluctuations in voltage precision and stability as a function of sweep energy. Typically, they are also not readily separable from the value of T.

The overall transmission function of an analyzer, $T' = f_n f_s \theta TD$, can be determined by using ideal electron sources with known emission behavior as a function of E_K. A typical methodology might involve the use of an electron gun and analysis of the intensity of elastically scattered electrons from a sample reaching the detector as a function of electron E_K. An alternate method is to compare a measured spectrum $I^m(E_K)$ from one instrument to a spectrum that is known to represent the expected emission spectrum $J^e(E_K)$ from a well-defined sample. Overall instrument response functions are useful for fundamental studies into the behavior of an XPS instrument. They are typically not recorded as part of the everyday operation of the system. As will be seen below, the behavior is instead lumped into one factor, the sensitivity factor S_i, that is provided in tabular form by the instrument manufacturer.

Substituting the electron flux equation for a homogeneous sample into the above equations leads to a relationship between measured electron intensity and the concentration of an element in a homogeneous material.

$$I_i^m = F\sigma f_a AT\lambda_i C_i \quad (11)$$

where the angular efficiency as well as the detector and instrument response functions have been lumped into the factor f_a. The factor AT, the product of the viewing area of the analyzer and its transmission function, is called the analyzer étendue, G. As mentioned previously, for an HSA operating in constant pass energy mode under conditions where the analyzer defines the transmission function, G varies approximately as $1/E_K$.

The practical applications of XPS for determining atomic concentrations start from the above equation. It is typically presented in a far more abbreviated form, namely

$$I_i^m = FS_iC_i \tag{12}$$

where S_i is the instrument-dependent sensitivity factor for orbital i in an element. In practice, sensitivity factors should be provided by instrument vendors for each orbital in all elements. In most cases, sensitivity factors are only provided for the most prominent orbitals. As seen from the above discussion, they are not directly transferable from one instrument to another, since they contain instrument response and transmission functions. Also, sensitivity factors are confined to a particular sample-analyzer geometry. Finally, they do not take into account nonidealities in the sample, especially in regard to composition and structure. Despite all these apparent limitations, the use of tabulated sensitivity factors for an instrument has become the accepted practice for determining atomic concentrations with XPS, as discussed below (also see Data Analysis and Initial Interpretation).

Equation 12 can be used in three ways. The first is to determine absolute atomic (mole) fractions, x_i for an element in a material. In this regard, we can write

$$x_i = \frac{I_i{}^m/S_i}{\sum I_i^m/S_i} \tag{13}$$

showing in principle that only sensitivity factors and measured intensities are required to determine atomic concentrations of a material. This equation is based on analysis of a material with a homogeneous composition throughout. This restriction will be relaxed below in the discussion on other applications of XPS.

To obtain absolute accuracy in atomic compositions using the above equation, we must have values of S_i that were obtained (a) from a sample of similar if not identical atomic composition and structure to the one we are analyzing and (b) from a sample-analyzer configuration equivalent to that used to determine S_i. While the second condition can generally be met without much difficulty, meeting the first condition can be difficult if not impossible. Therefore, the above equation can be considered an approximation for almost all real-case systems. Given the extent of variations that are possible in λ_i and θ_i from one material to another, relative inaccuracies in x_i can be as much as $\pm 20\%$ using the above equation. A worthwhile note from this discussion is that absolute atomic fractions determined from XPS will be accurate to better than $\sim \pm 1\%$ only if they have been derived after a painstakingly thorough calibration of the XPS instrument concerned. This issue is separate from the precision of the measurements and will be highlighted again below (see Data Analysis and Initial Interpretation).

A far more useful form of Equation 12 (Briggs and Seah, 1994) can be derived by considering analysis to obtain "relative" atomic concentrations. Then, we can write

$$\frac{x_i}{x_j} = \frac{I_i^m/S_i}{I_j^m/S_j} = \frac{S_j}{S_i}\left(\frac{I_i^m}{I_j^m}\right) = R_{ji}\left(\frac{I_i^m}{I_j^m}\right) \tag{14}$$

where R_{ji} is the relative sensitivity factor of element (orbital) j to element i. Be aware in applications of this form whether you are using R_{ji} or R_{ij}, which are related as $R_{ji} = 1/R_{ij}$. Relative sensitivity factors are far less dependent on variations in sample composition and structure, as well as instrument configuration, than absolute concentrations. Changes in these factors tend to cancel each other in taking the ratio. Therefore, while uncertainties in absolute concentrations determined using XPS can be high, the uncertainties in relative atomic concentrations are in many cases bounded primarily by errors in measurement precision (repeatability). This often means that relative atomic concentrations determined with XPS can be compared more reliably from sample to sample and from system to system than absolute atomic concentrations.

The final form of Equation 12 (Briggs and Seah, 1994) that is of practical use arises when taking samples of known (and uniform) atomic composition for comparison to samples of unknown (and uniform) composition with similar micro- and molecular structure. In this case, the formulation can be recast to give

$$\frac{x_i}{x_{io}} = \left(\frac{I_i^m}{I_{io}^m}\right)\left(\frac{S_{io}}{S_i}\right) \tag{15}$$

where the "o" subscript indicates the sample with known composition. A particularly useful form of this equation for illustration purposes and applications arises when the known material is pure and we are comparing with binary (two-element) materials. The equation can then be rewritten as

$$\begin{gathered}\frac{x_A}{x_B}\left(\frac{x_{Ao}}{x_{Bo}}\right) = \left(\frac{S_B/S_{Bo}}{S_A/S_{Ao}}\right)\left(\frac{I_A^m/I_{Ao}^m}{I_B^m/I_{Bo}^m}\right) \\ \frac{x_a}{x_B} = f_{BA}^m\left(\frac{I_A^m/I_{Ao}^m}{I_B^m/I_{Bo}^m}\right)\end{gathered} \tag{16}$$

where F_{BA}^m is known as the matrix factor for the binary system. The matrix factor accounts primarily for the changes produced by putting element B in a matrix of element A. It is not always unity. To a first approximation, it can be taken as a ratio of the square root of the molar specific volumes of the elements

$$F_{BA}^m \approx \sqrt{\frac{\hat{V}_B}{\hat{V}_A}} \approx \left(\frac{R_B}{R_A}\right)^{3/2} \tag{17}$$

or the corresponding atomic radii for similar crystal systems. To use this equation, we would first measure intensities for the principal XPS peaks from pure materials, I_{Ao}^m and I_{Bo}^m. Then, intensities from samples of known composition, I_A^m and I_B^m would be measured to calculate the matrix factor for the binary system. Thereafter, alloy composition can be calculated directly from measured peak intensities of unknown binary alloys. This equation is limited primarily to homogeneous alloy systems.

Care must be used in reporting atomic compositions determined with XPS. They must be understood to be accurate only for a sample that has a uniform composition throughout the entire sampling volume. This is rarely

the case with real-world samples. Concentrations can vary dramatically across an analysis area, for example when analysis is done across grain boundaries in metals or across copolymer regions in polymers. One must keep in mind that even the best spatial resolution of ~25 μm obtainable with small-spot XPS is often larger than spatial variations in composition across the surfaces of most real-world samples. Conventional XPS systems can have analysis areas at the sample as large as a few mm^2. In cases where the sample is known to have lateral variations in composition, the atomic concentrations determined with XPS will generally be spatially averaged values.

Concentrations also vary with sample depth. In this regard, a key parameter is the average sampling depth in comparison to the depth over which the composition varies. The average sampling depth depends on the mean free path λ_i of the electrons. For example, most samples quickly become covered by an overlayer of carbonaceous and water contamination upon exposure to air. This overlayer, if thicker than λ_i, will mask the true composition of the underlying material, and anomalously high values of atomic concentration for oxygen and carbon will result. Other examples of potential problems that can arise in determining compositions when using XPS spectra from samples with nonuniform composition profiles have been pointed out elsewhere (Tougaard, 1996).

Further discussion on how to extract information about compositional variations with sample depth is given in the following section titled Other Applications. Analysis of XPS spectra to obtain values for I_i^m is also considered (see Data Analysis and Initial Interpretation). Methods for preparing samples to reduce the contaminant overlayer are given in the section on sample preparation.

Chemical State. Obtaining chemical state information from an XPS spectrum can often be as important as obtaining elemental composition. Any application that must consider the relative amounts of different oxide, valence, or hybridization states of an element will be interested in using XPS to determine chemical state. Specific examples of this for different families of materials (metals, ceramics, semiconductors, polymers, and composites) are found in almost any journal that publishes studies from XPS (see Key References).

The BE of a photoemission peak is affected by the chemical state of the element (see Principles of the Method). When an atom acquires a partial positive charge due to bonding or rehybridization, we expect a shift to higher BE values. Correspondingly, a shift to lower BE values is expected when an atom becomes negatively charged. Variations in the chemical state of atoms throughout the material can also affect the width of the photoemission peak.

As also pointed out (see Principles of the Method), other factors besides the chemical state of the element influence both the BE and width of a photoemission peak. The Madelung potential term arises when the lattice structure about the atom changes. The extent of intra- or extra-atomic relaxation may also change. This means, a peak shift to higher or lower BE values cannot always be interpreted as arising solely from a change in chemical state.

In practice, one should therefore not view XPS as an unambiguous means of identifying unknown chemical states of elements. The better approach is to have initial information regarding the bonding of the element in the material. This information should then serve as a guide to select the main factors that could cause peak shifts or peak broadening for photoemission from the sample. Other factors that lead to peak shifts and broadening are considered below (see Problems). Finally, an enormous number of reference sources are available that contain XPS peak positions and even reference spectra from compounds (see Key References). The reader is also encouraged in this regard to search the Internet. These reference texts and pertinent literature on the sample being analyzed should be consulted for fuller insight into peak assignments.

For further discussion on resolving overlapping peaks via peak fitting procedures and on deconvolving instrument response functions and source broadening effects from high-resolution XPS spectra (see Data Analysis).

Other Applications. Two other applications of XPS deserve mention. Both are extensions of the use of XPS to determine atomic concentrations. They apply in cases where composition varies with sample depth, $C_i(z)$. The first is the use of what is commonly termed sputter profiling. The second is called angle-resolved XPS (ARXPS). Both methods rely on the limitation that sampling depth in XPS is only on the order of the electron mean free path, λ_i.

Sputter profiling is a destructive technique. Details are provided in the reference from Briggs and Seah (1994). It is useful for mapping any form of concentration profile as a function of depth. In sputter profiling, the sample is bombarded with high-energy ions, typically Ar^+. This etches away layers of the sample. Sputtering rates can be controlled to remove on the order of sub-monolayers of material with careful calibration of the sputter gun and its configuration relative to the sample. Sputter depth profiling consists of performing a series of analysis-sputter cycles on a sample. Sputter rates and times during each cycle are set by the user. Typical results from sputter profiling are plots of I_i^m, x_i, or peak positions versus sample depth. Sample depth is obtained from sputter time by calibrating the sputter rate. This is done by using a layered sample with a known thickness. Oxide layers on Si wafers and oxide layers on Ta are most common and can be readily obtained through commercial vendors.

A number of factors limit the utility of sputter profiling with XPS (Hofmann, 1993; Wetzig et al., 1997), including sample surface roughness (Gunter et al., 1997) and compositional heterogeneity. Both contribute to nonuniform sputtering rates across a sample surface, especially in systems where sputtering is done at an incident angle that is not normal to the sample surface. A common method of overcoming this limitation, at least for rough sample surfaces, is to rotate the sample under the sputter gun. Another limit on sputter profiling is the inconsistency between sampling depth and sputter depth per cycle. Generally, the sputter depth per analysis cycle is set to be lower than the sampling depth (mean free path of the escaping

electrons). This leads to the effect that I_i^m will start to decrease or increase before the depth is reached where concentration actually increases or decreases. This problem is essentially due to a convolution of the true concentration profile $C_i(z)$ with functions representing the effective sampling depth during analysis $f_d(z)$ as well as any non-uniformities in sputtering depth $f_n(z)$, giving a measured concentration profile $C_i^m = C_i(z) \oplus f_d(z) \oplus f_n(z)$. A number of methods exist for handling the deconvolution of $f_d(z)$ and $f_n(z)$ from C_i^m in a simple or rigorous manner (Blaschuk, 2001; Hofmann, 2001; Lesch et al., 2001). Ion bombardment can also cause remixing at the sputter crater of components in a heterogeneous sample, and segregation of specific bulk elements to surfaces and interfaces can occur (Hofmann, 1992). These factors will lead to false composition readings during sputter profiling. Finally, care must be taken during sputter profiling to make sure that sputtering is done uniformly over an area that is larger than the analysis area.

With the exception of nonuniform sputtering rates, the above limitations are less important in sputter profiling during AES analysis. The analysis area and sampling depth are both much smaller in AES than they are in XPS. Therefore, in cases where it can be used, AES is generally preferred for sputter profiling, especially when only composition (not chemical state information) is desired. The only other concern prior to sputter profiling is to determine whether any of the AES or XPS peaks of interest overlap. This can sometimes rule out or favor one technique over the other. For example, N peaks overlap with Ti peaks in AES but not in XPS, so nitrided Ti samples would be better analyzed with depth profiling using XPS.

In ARXPS, compositions are measured as a function of sampling angle relative to the surface plane, often called the analyzer take-off angle. For samples with layered compositions, changing the analyzer take-off angle results in changes in measured intensity, I_i^m, because emitted flux, J_i^e, varies from elements within the different layers. Measurements of I_i^m as a function of take-off angle can be used to determine values for layer thickness. Other applications include the determination of molecular orientation in ordered adlayers and crystal packing symmetry in lattices. Two extensive reviews by C. Fadley of the principles and applications of ARXPS are recommended as a starting point for further details (Fadley, 1976, 1984).

One case is useful to present as an illustration of the methodology. This is an example for a bilayer material, such as occurs when a coating, oxide, or contaminant layer covers a surface. An illustration of the set up is shown in Figure 11. The overlayer has a thickness, d; the outmost surface and boundary interface are assumed to be flat on the microscale, and the boundary between the overlayer and the underlying substrate is abrupt (a step function).

Consideration of this situation starts with modification of Equation 8 to account for a change in the effective path length the electrons must travel to escape. The path length is dependent on the take-off angle, θ, of the photoelectrons. We define take-off angle relative to the surface plane, so that glancing angle analysis occurs as θ approaches zero (some texts define take-off angle with respect to the surface normal). Under these conditions, for a fixed value of

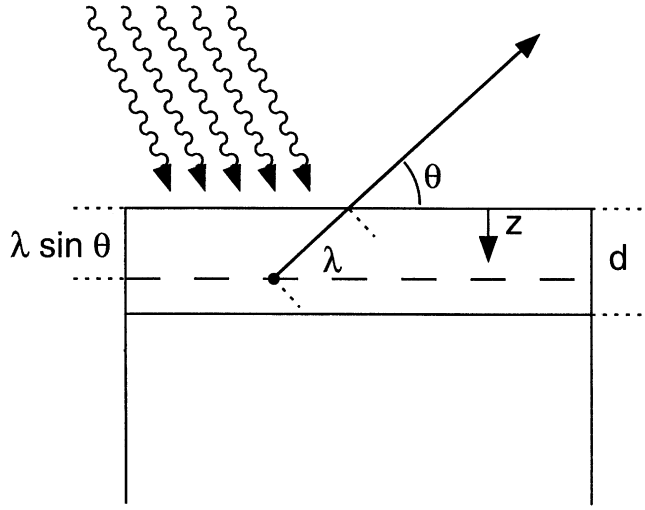


Figure 11. An example of a simple system that illustrates the application of angle resolved XPS. An overlayer of thickness d is being analyzed at a take-off angle θ with respect to the surface plane.

θ and a semi-infinite sample, we can write Equation 8 for the overlayer o and underlying substrate s to be

$$J_o^e = F\sigma_o \int_0^d C_o(z)\, \exp(-z/\lambda_o \sin\theta) dz \qquad (18)$$

$$J_s^e = F\sigma_s \int_d^\infty c_s(z)\, \exp(-z/\lambda_s \sin\theta) dz \qquad (19)$$

We assume the overlayer and substrate have uniform composition throughout. We can then obtain representative equations for the measured intensity of photoelectrons from an orbital in the overlayer or in the substrate as

$$I_o^m = F\sigma_0 f_{a,o} G_o C_o \lambda_o (1 - \exp(-d/\lambda_o \sin\theta) \qquad (20)$$

$$I_s^m = F\sigma_s f_{a,s} G_s C_s \lambda_s \exp(-d/\lambda_s \sin\theta) \qquad (21)$$

Note that, factors dealing with the behavior of the instrument, f_a and G, as well as the cross-section σ depend on E_K of the electron and the orbital being considered, respectively.

A ratio of measured intensity for a photoemission peak in the overlayer I_o^m to that from the underlying substrate I_s^m has the form

$$\frac{I_o^m}{I_s^m} = \left(\frac{\sigma_o}{\sigma_s}\right)\left(\frac{G_o}{G_s}\right)\left(\frac{\lambda_o(1 - \exp(-d/\lambda_o \sin\theta))}{\lambda_s \exp(-d/\lambda_s \sin\theta)}\right)\left(\frac{C_o}{C_s}\right) \qquad (22)$$

In this equation, the parameter sought is d, the overlayer thickness. This can be determined if we obtain I_o^m/I_s^m as a function of θ, and know all the other parameters. In practice, the ratios of cross-sections and transmission functions are usually taken as unity to simplify the expression. The concentration ratios are either set equal to unity as well or are estimated from materials properties. An example of the use of this equation is provided below (see Data Analysis).

Photoemission intensity for one peak from an overlayer can also be compared relative to the intensity from same peak for the pure sample of the element concerned using the above equations. This leads to

$$\frac{I_o^m}{I_p^m} = \frac{\lambda_o(1 - \exp(-d/\lambda_o \sin\theta))}{\lambda_p \sin\theta}\left(\frac{\bar{V}_p}{\bar{V}_o}\right)x_o \qquad (23)$$

where x_o is the mole fraction of the element in the overlayer and the $V]_i$ values are the molar specific volumes of the element in the overlayer or pure material.

Angle-resolved XPS can also be used to determine the bonding angle of a molecular adsorbate on a surface. Consider a linear-chain molecule that can bond in either an up or down configuration on a surface, and imagine that one end is tagged with a different element than the other. If we form a monolayer on a microscopically flat surface, analysis with XPS at glancing take-off angle will accentuate the signal from the element that is at the outermost position in the absorbed layer. This has applications for molecular coatings and polymers where information about the chemistry of the exposed surface relative to that for the underlying surface layer is of utmost importance. On microscopically flat surfaces, ARXPS can often be used to distinguish between perpendicular and horizontal configurations of molecules. It can also be used to determine the bond angle of molecular adsorbates in certain cases where the adsorbate has well-defined long-range order. These types of angle-resolved XPS measurements should be confirmed using other techniques.

Surface roughness complicates the analysis with ARXPS. In addition, the angular resolution of the data can never be any better than the acceptance angle of the front lens on the analyzer. Most commercial systems have an acceptance angle of 15° to 20°. This may not present problems for film thickness measurements, but is too wide for sophisticated ARXPS measurements of bond angles for molecular adsorbates. Some HSA-equipped XPS instruments have variable magnifications for the lenses. The acceptance angle of the analyzer will always be larger at higher magnifications. Both surface roughness and an overly wide analyzer acceptance angle lead to a fall off of intensity from theoretically expected values as take-off approaches glancing angles.

Materials-Specific Considerations

Almost all of the previously mentioned journals (see Introduction) present articles on XPS as applied to all classes of materials. The scope ranges from fundamental (*Surf. Sci.* or *Surf. Interface Sci.*) to more applied (*Appl. Surf. Sci.* or *J. Vac. Sci. Technol.*). The use of XPS for analysis of biomaterials surfaces is also becoming more widely represented in *Colloid Interface Sci.* and *J. Biomed. Mater. Research.*

With regard to metals, the ASM Handbook (ASM Handbook Committee, 1986) is a reasonable starting point for practical applications of XPS in metallurgy. Analysis of metals that are used in routine applications is often not straightforward with XPS. Most metals will have oxide or contaminant overlayers on their surfaces, and surface segregation is also common in alloys. Structural variations, such as grains, precipitates, or other microconstituents can also be present, especially in multi-component metal alloys. These factors will all affect the measured intensity of a photoelectron peak I_i^m by introducing nonidealities into the equations that were derived primarily for analysis of homogeneous samples. In addition, as seen in Figure 8, the more elements that are in a metal sample, the greater is the chance that Auger peaks will interfere with XPS peaks in the spectrum. This can make the determination of composition and chemical state difficult. Finally, on metals with oxide layers, the oxide layer can act as an electrical insulator, leading to peak shifts due to differential charging (see Problems).

As mentioned above (see Competitive and Related Techniques), better spatial and depth resolution are typically obtained using AES. This may be preferred for analysis of metals, especially for compositional analysis across grains or within two-phase regions. Because of the larger sampling depth and spot size of XPS, compositions will be averaged in these cases. The distinct advantage of XPS over AES for analysis of polycrystalline metals is that the oxidation states of metals are generally much easier to resolve with XPS, though this is not always the case (Sekine et al., 1996). Assignment of peak shifts to changes in metal oxidation state can be complicated when multiple oxidation states exist. Peak fitting is generally required to resolve the overlapping features. The use of sputtering to remove the outermost contamination layer, as well as glancing angle analysis to become more sensitive to surface compositions, can often help.

Analysis of semiconductors is relatively routine with XPS. The overriding concern is generally to remove the outermost contamination layer. With extrinsic semiconductors, sensitivity factors for the dopant concerned should be compared to those for the intrinsic material, keeping in mind that lower sensitivity factors mean greater amounts (concentrations) of dopant will be needed to give the same signal-to-noise ratio in an XPS spectrum. The mean free paths of the relevant photoelectrons in the material should be considered when analyzing for dopant compositions that may vary with depth. Sample charging can occur in XPS of semiconductors. In this regard, though, XPS presents fewer problems than AES.

With ceramics, XPS is generally the only spectroscopic laboratory method that can be used to determine (surface) composition and chemical state information. Analysis with AES is typically not feasible because of sample charging. Even with XPS, ceramic samples charge considerably, leading to peak shifts.

In regard to analysis of polymers with XPS, the articles by D.T. Clark are good, though somewhat dated, starting points (Clark and Feast, 1978; Clark, 1979). Analysis of polymers with XPS is a common method of determining both concentrations and hybridization states of elements in the material. As shown in Figure 7, XPS can easily resolve certain hybridization states of carbon. It can also resolve variations in functional groups attached to the carbon. In most cases, shifts of the C 1*s* peak can be assigned solely to changes in the electronegativity of the surrounding atoms. The primary difficulty with polymer samples is charging. Some polymer samples are also damaged by high x-ray flux.

Metal and semiconductor single-crystal samples are often used as model systems for analysis with XPS. Quantification of the rates of chemical reaction processes and determination of the fundamental nature of chemical bonds formed by adsorbates to surfaces are two areas that benefit by the use of XPS on such model systems. Single-crystal metals and semiconductors have also been used to develop much of the theoretical basis behind the shape of the XPS spectrum. Angle-resolved XPS also benefits by using microscopically if not atomically flat surfaces.

METHOD AUTOMATION

All XPS systems in operation today are run by computers, with many modern systems being entirely automated after introduction of the sample. The user generally has to define a number of parameters to control the instrument. On older-generation instruments, some of the parameters may be input manually. In most cases, those parameters that are input manually are not supposed to vary from run to run.

Starting with the source, after selecting an anode type (in systems with more than one anode), one typically defines any two of three operating parameters of x-ray source voltage, emission current, or power. This defines the type of x ray (its primary energy $h\nu$) and the flux F. We are typically not interested in knowing the absolute value of F for the x-ray source, only in knowing that it is the same value from analysis to analysis. Most XPS instruments also allow the user to select an analysis area on the sample, either using manual apertures or through computer settings to control focus voltages to the lenses of the analyzer. Modern systems also allow the user to position the sample in different locations and possibly different orientations under the analyzer using computer control.

The scan parameters are set next. Among these parameters are: the sweep starting and/or ending energy, $\mathrm{BE}_{\mathrm{start}}$ and $\mathrm{BE}_{\mathrm{end}}$; sweep width, $\Delta\mathrm{BE}_{\mathrm{sweep}}$; step size, $\delta\mathrm{BE}_{\mathrm{step}}$, or number of points per sweep n_{pps}; dwell time at each step, t_{dwell}, or the total time per sweep, t_{sweep},; and finally the number of scans, n_{scans}, or total analysis time t_{total}. These parameters are related as listed below.

$$\Delta\mathrm{BE}_{\mathrm{sweep}} = |\mathrm{BE}_{\mathrm{end}} - \mathrm{BE}_{\mathrm{start}}|$$
$$\Delta\mathrm{BE}_{\mathrm{step}} = \Delta\mathrm{BE}_{\mathrm{sweep}}/n_{\mathrm{pps}}$$
$$t_{\mathrm{sweep}} = n_{\mathrm{pps}}\, t_{\mathrm{dwell}}$$
$$t_{\mathrm{total}} = t_{\mathrm{sweep}}\, n_{\mathrm{scans}}\, n_{\mathrm{sweeps}}$$

Not all of these parameters are used by all systems. Some are set automatically by the system depending on the type of scan (survey or high-resolution, for example). In this regard, the remaining input parameter is the pass energy E_{p}.

In addition to automated data acquisition, most instruments with computer control can be programmed to cycle through a series of spectra and determine peak parameters. This is generally done during sputter profiling or angle-resolved XPS. The instrument automatically shows profiles of composition as a function of sputter time (or depth) or take-off angle.

Automation of the data acquisition sequence leads to problems when the results deviate from those expected or when the system has problems during operation. The former case arises, for example, when samples have different degrees of charging, shifting the peaks out of preprogrammed sweep ranges. In some cases, trying to abort or break an established acquisition routine can lead to the loss of the settings, or worse, loss of data, especially on older-generation instruments. The user is therefore cautioned to be well aware of the potential variations that could arise before automating data acquisition.

Automation of data analysis does not typically present as many problems. The user can always repeat the analysis using different parameters. Unfortunately, some older XPS systems with automated data analysis may not allow this luxury, so caution is again warranted. Also, one should always do spot checks on the analysis results after an automated analysis. The computer is not an "all-knowing" instrument, and exceptions to the standard methods of data analysis can always arise.

DATA ANALYSIS AND INITIAL INTERPRETATION

Analysis of an XPS spectrum starts by determining parameters for the photoemission peaks. The three primary parameters that define a photoemission peak in XPS are its area A_{pi}, position BE_{i}, and full-width-at-half-maximum (FWHM) Δw_{i}. Additional consideration may be given to peak shape, as discussed below (see Peak Integration and Fitting). To obtain accurate values for these peak parameters, we have to process the XPS spectrum correctly. This means applying methodologies that are well grounded in the principles of the method. Data analysis may be preceded or followed by application of methodologies to improve the appearance of a spectrum. Finally, interpretation of the results should keep in mind the uncertainties that arise during data analysis.

The following discussion first presents the methods used to determine peak parameters from XPS spectra. This is followed by a discussion of methods that can be used to improve the appearance of an XPS spectrum. We then consider how to interpret the results, with particular emphasis on uncertainty analysis. The concluding presentation in this section offers an example of the analysis of ARXPS data to determine the film thickness of an oxide layer on a Si wafer.

Analysis

The most routine analysis of an XPS spectrum involves identifying the peaks with their corresponding elements. Tables of peak positions are available to help with this task. Most XPS systems also have software libraries of peak positions that will automatically mark elements in a spectrum. This automation is only as good as the software routines for locating peaks; it should not be taken as a definitive indication of the presence of an element, and it may miss small peaks on noisy backgrounds. The best method of determining unambiguously that an element is present is to cross-correlate for all peaks that

should appear for that particular element. For example, if a peak at 100 eV is identified as the Si 2*p*, a corresponding 2*s* peak should be found at ~150 eV with the appropriately scaled height. Auger peaks should also be present if the scan range covers them.

We generally also want to obtain elemental compositions from an XPS spectrum. This requires quantitative analysis of the spectrum. Quantitative analysis for compositions is based on background subtraction and peak integration. In cases where more than one chemical state is present for a given element, we may need to fit the peaks with components. Finally, we may want to quantify the positions and half-widths of peaks in order to compare changes from sample to sample. The practical aspects associated with these procedures are discussed in the sections immediately below.

Background Subtraction. The area of a photoemission peak is determined by subtracting an appropriate background and integrating (numerically) under the resultant curve. The background underlies the true spectral features, as illustrated in Figure 2. The types of background shapes that can be used range from a constant to a complex shape and are based to some degree on theoretical considerations of the inelastic scattering processes that generate a background in XPS spectra. The choice of a background shape can affect the resulting peak parameters. Although this effect is often below the precision (repeatability) of the measurements from sample to sample, proper attention should be paid to the application of appropriate background subtraction methods with XPS spectra. This is especially true in cases where peak fitting is to be done and where subtle changes in peak parameters, especially peak position or FWHM, are important to resolve.

A constant or linear background removes a straight line from under a spectrum, typically bringing the lowest value(s) in the spectrum to a value of zero. The constant background removes a constant offset throughout the spectrum. The higher-BE side of a primary photoemission peak will generally have a greater intensity due to inelastically scattered electrons and potentially to satellite peaks. A linear background rather than a constant one would therefore appear to be a better choice in most cases. This is the case for both spectra in Figure 6. The intensities on the left and right sides of the peaks in Figure 7 are nearly equal, so a constant offset could be removed. In both of these cases though, better choices for background shapes exist. In practice, a constant or linear background should only be applied as an approximation when the background intensity on the left and right sides of the photoemission features under consideration are nearly identical or have identical slopes.

A widely applied background shape with some degree of theoretical support is the integrated or Shirley background (Shirley, 1972). It is based on the principle that background intensity I_i^b (BE) at a given BE is proportional to the total intensity of photoelectrons at higher E_B. This is similar to the nature of the inelastic scattering process that leads to a significant portion of the background in XPS. The intensity of scattered (background) electrons increases at lower KE (higher BE) away from a photoemission peak in proportion to the intensity of photoelectrons at higher E_K (lower E_B). The Shirley background is derived analytically from a formulation similar to

$$I_i^b(\mathrm{KE}) = C\int_{\mathrm{KE}}^{\infty}(I_i^m(E) - I_i^b(E))dE \qquad (24)$$

The integration is best represented in KE space, since higher E_K electrons contribute to the background at a given KE. The value C is essentially a normalization constant to bring the background into alignment at the endpoints. This function can be readily coded into XPS analysis software and is therefore routinely available with the instrument or as part of off-line data analysis software. The above equation must be solved iteratively using numerical integration, and has no user-defined parameters other than the start and end ranges of the background. Some older XPS instruments may only use a simpler single-pass numerical integration for increased processing speed.

Disagreements exist about the accuracy of the Shirley background routine. It has been shown to distort peak parameters (peak positions) in certain critical cases. In the experience of the author, the potential errors introduced by the use of the iterative Shirley background versus a supposedly more accurate background shape are usually well within the systematic errors of determining peak parameters for most routine XPS experiments. Only when a high degree of accuracy is essential during peak fitting should the user generally have to consider other (more accurate) background shapes.

The Shirley background is only appropriate for fitting as a baseline to high-resolution peaks. The algorithm does not normally converge well over the wide BE range of a survey scan. Application of the Shirley background also leads to problems in cases where the signal intensity increases or decreases dramatically on either side of a photoemission peak. This happens when a photoemission peak of interest sits as a shoulder on a sloping background from a significantly larger peak. The resulting complex background shape is a convolution of the overlapping peak intensity and the normal (inelastic) background. Off-line processing routines exist to handle such complex background shapes. Whenever possible, data acquisition should encompass a full range of all overlapping peaks to avoid this problem.

The Tougaard background shape for XPS has the strongest basis in theory (Tougaard, 1988). It is derived on first-principles consideration of the intensity of inelastically scattered electrons as a function of decreasing KE away from the E_K of the primary photoelectron. The resultant analytical expression in its general form is an integral over peak intensity that has two materials dependent fitting parameters, B and C.

$$I_i^o(\mathrm{KE}) \approx C\int_{\mathrm{KE}}^{\infty}\frac{E-\mathrm{KE}}{\left[B+(E-\mathrm{KE})^2\right]^2}I_i^m(E)dE \qquad (25)$$

The fitting parameters relate to the nature of the inelastic scattering in the material of concern and are usually left to be determined by measurements with techniques

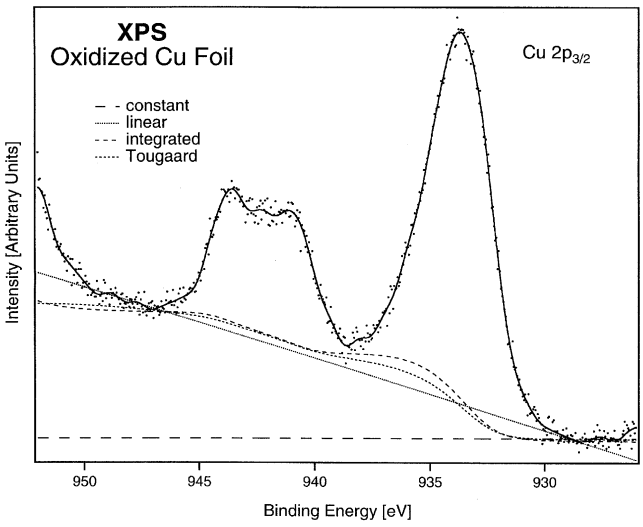


Figure 12. Examples of constant, linear, and integrated (Shirley) types of background shapes fit to a Cu $2p_{3/2}$ peak and the companion satellite for CuO.

such as EELS (see SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). Since this degree of rigor is beyond the scope of interest for typical XPS users, the Tougaard background is usually used with predefined "universal" values of fitting parameters, in particular $B \approx 1643$ eV^2. The value of C is primarily a normalization constant to fit the background at the ends of the spectrum under consideration. The resulting background shape is similar to the Shirley background.

Examples of a constant, linear, iterative integrated (Shirley), and Tougaard backgrounds applied to the CuO peak shown earlier are given in Figure 12. This figure clearly shows the potential for error that can arise when a proper background is not used prior to further processing of a spectrum. The constant background is inappropriate for determination of peak area. Compare the linear, integrated, and Tougaard backgrounds. The resulting integrated area under the entire Cu $2p_{3/2}$ peak may be different in the three cases; however the extent of the differences is likely to be within the reproducibility of the measurements.

Peak Integration and Fitting. Once a suitable background is removed, A_{pi} can be found by numerical integration. The units of A_{pi} are typically counts-eV/sec. This is directly proportional to measured intensity $I_{\mathrm{i}}^{\mathrm{m}}$ in the corresponding equations above (see Practical Aspects of the Method). The additional factor of eV cancels in the calculation of concentration or concentration ratio.

Broad peaks may be deconvolved into (fitted by) more than one component peak to help resolve overlapping peaks. The spectrum in Figure 7 shows resolution of the broad photoemission feature into three component peaks. The following guidelines apply when doing peak fitting: (1) peak fitting should always be done after removing an appropriate background; (2) the shape of the component peak matters; (3) the initial conditions used to start the peak fit should be best approximations to the expected chemistry of the sample; and (4) peak parameters should be appropriately constrained during the fitting process.

With regard to background subtraction, many peak-fitting algorithms in use today for analysis of XPS spectra do not include the capability to automatically fit an appropriate background while optimizing fit parameters for component peaks. In such cases, the appropriate background must be carefully chosen and removed beforehand to obtain consistent, accurate results from peak fitting. Note in Figure 12 that differences are to be expected in the final results for peak fitting when different backgrounds are used, because the initial peak shapes will be different. In particular, small peaks that are shoulders on larger peaks can be accentuated or lost merely by shifting the position or changing the shape of the background that is removed. Calibration of the combined baseline subtraction and peak-fitting methodologies with samples of known chemistry and composition is recommended. Examples where simultaneous fitting of baselines and peaks was done specifically to obtain detailed information about sample chemistry have been reported (Castle et al., 1990; Castle et al., 2000; Salvi et al., 2001).

A large number of peak shapes can be used to fit to a feature in an XPS spectrum. All peak shapes have at least three parameters: peak height, peak FWHM (or half-width), and peak position. Peak half-width is half of the FWHM value. Additional parameters are used to mix fractional amounts of the different peak shapes and to define such factors as asymmetry or tail height and extent. The doublet is a peak shape that is particular to XPS spectra. It is used to fit to the multiplet split p, d, or f peaks, and it is essentially defined as two peaks separated by a certain distance.

The Appendix provides further information about peak shapes commonly used to fit XPS spectra. This information should clarify how the parameters are used to define the peak shape. In confronting the wide range of peak shapes, the first-time XPS user is cautioned to proceed conservatively. The Gaussian and Gaussian-Lorentzian mixed peak shapes (and their doublets) provide adequate representation of XPS peak shapes for almost all routine analysis work. The peak shapes with greater degrees of freedom should be used only with a clear understanding of the physical significance behind their application. In other words, do not choose a peak shape with more free parameters than can be reasonably justified within the confines of the chemistry and physics of the system with the goal of covering all potential contingencies. Rather, expand the degrees of freedom (number of peak-fit parameters) from a minimum to a larger set when no other options appear adequate to cover the case being examined.

During peak fitting, any or all of the parameters that define a specific component peak are allowed to vary until an optimal resultant peak shape, compared to the input raw spectrum, is obtained. The computational methods involved in optimizing the parameters are beyond the scope of this unit. Suffice it to say that local and global optima exist in the space of the fit parameters. Therefore, do not use more peaks than absolutely justified, and always start with peak parameters that provide the best representation of the chemistry as known from the sample. Starting

with too many peaks or with parameters that are too far from the expected values could lead to an end result that is far removed from physical reality. This problem can be avoided in almost all cases in a straightforward manner. At the start, use a minimum number of peaks and peak parameters in order to obtain a resultant "fit" spectrum that represents the raw data well at the outset. This may require iterative, manual adjustments of the initial peak parameters. With practice, this step will become a matter of good practice during peak fitting.

In correspondence with setting the initial peak parameters correctly, allowing too many parameters to vary during the optimization can lead the optimization routine astray. The fitting may fall into an unrealistic local minimum or, in some cases, fail to converge properly. Most peak-fitting routines allow the user to constrain the peak parameters either to stay within certain bounds or to stay constant during peak fitting. As the number of potential peak-fitting parameters increases, the number of parameters that are constrained should also increase.

In conclusion, with regard to peak fitting to XPS spectra, let the expected chemistry of the material be the guide in defining and accepting the peak parameters, instead of letting the fitting results determine the chemistry. Further information about the steps involved during peak fitting of XPS spectra is available in a flowchart and comparative study (Crist, 1998). In addition, reports of methods that have been used to validate algorithms for peak fitting to XPS spectra (Seah and Brown, 1998) should be consulted when user-defined routines are developed or used. Finally, as a way to define the level of confidence in the methods being used during peak fitting, standardized XPS test data have been developed for peak fitting (Conny and Powell, 2000).

Peak Position and Half-Width. The remaining peak parameters, BE_i and Δw_i, are determined by the position of the peak maximum and the width at half maximum height, respectively. Unlike peak area, they can routinely be determined by eye from XPS spectra. Be careful in such cases. A common mistake made in setting the positions of components in an overlapping feature by eye is to set the peak positions too close together. Underlying component peaks, when properly resolved using peak fitting, will end up being further apart than the result obtained "by hand." This is especially true in cases where one or more of the peaks is a shoulder on a larger peak. Consider in particular the spectrum in Figure 7. Look carefully at where the component peak positions actually are, according to the peak fit components, in comparison to where an otherwise untrained observer might set them "by eye" on the spectrum of the raw data.

In conjunction with the above, before calculating the FWHM of a peak "by hand," remember to consider the peak height after removing a background. This is done by drawing a baseline by hand, typically a sloping, linear background (not just a constant baseline). Someone doing an analysis of an XPS spectrum by hand for the first time is apt to make a common mistake with overlapping peaks by drawing two separate, linear baselines, one under each overlapping peak. While this may be an appropriate method for such techniques as infrared spectroscopy, where the baseline shape can be dramatically curved throughout a narrow spectral range, it is entirely inappropriate in XPS spectra, where the baseline extends smoothly under all peaks associated with the spectral feature. Review again the sloping linear baseline shown in Figure 6 as a reference; the slope is not disjoint at the junction of the two satellite peaks.

Depending on the computuation package used for peak fitting, peak parameters may be selected through autodetection routines. One method to locate a peak is by the position of maxima in a residual spectrum. The residual spectrum is determined as the difference between the original spectrum and the summation envelope of peaks already fit to the original spectrum. The FWHM for the peak in the original data can also be estimated by the peak width at half the height of the residual spectrum peak. Another method is by scanning derivatives of a spectrum for minima and zero crossings. Assymetries in the shapes of the derivatives can help determine whether closely overlapping peaks exist in a spectrum that otherwise appears to be a single peak. Numerical differentiation is inherently a smoothing process that requires good S/N in the original spectrum to avoid distoration of the derivatives (smoothing) or loss of signal (increase in noise). Either method may have difficulties resolving peaks that closely overlap. New methods that involve wavelet transformation may also prove useful in autodetection of peaks in a spectrum (Ehrentreich et al., 1998) and for improving the fitting process (Zhang et al., 2000).

Principal Component Analysis. Principal component analysis (PCA) is a strong multivariate statistical data analysis tool applied to a variety of fields from science to engineering (Rummel, 1990). It is a potent technique in the arsenal of chemometrics (Malinowski and Howery, 1980) and surface spectroscopies (Solomon, 1987). It has been applied to resolve problems in XPS (Sastry, 1997; Artyushkova and Fulghum, 2001; Richie et al., 2001; Gilbert et al., 1982; Simmons et al., 1999; Balcerowska and Siuda, 1999; Clumpson, 1997; Holgado et al., 2000) and the related surface spectroscopic technique AES (Balcerowska et al., 1999; Passeggi, 1998). Applications of PCA to XPS include reduction of noise (Sastry, 1997), resolution (deconvolution) of overlapped spectra (Gilbert et al., 1982), resolution of XPS spectra into surface and bulk components (Simmons et al., 1999), subtraction of inelastic background signals (Balcerowska and Siuda, 1999), and detection of principal components (Clumpson, 1997; Holgado et al., 2000).

In mathematical terminology, PCA transforms the space of data from a set of possibly correlated variables into a smaller number of uncorrelated variables. This is expressed simply in matrix format as

$$|S| + |W|\Lambda|P| \tag{26}$$

where $|S|$ is data (as a matrix), $|\Lambda|$ is a weightings matrix, $|L|$ is an eigenvalue matrix, and $|P|$ is the matrix of principal components. The eigenvalue matrix is a square diagonal matrix that is obtained from the data through singular

value decomposition, the principal component matrix has the same dimensions as the data matrix, and the weightings determine how each principal component (scaled by its eigenvalue) is factored to obtain each part of the data. The eigenvalues λ_i in $|\Lambda|$ are arranged in order from largest to smallest across the diagonal. This means, components in $|P|$ are arranged in order from most to least significant in contributing to $|S|$. The statistical significance of each component in $|P|$ can also be determined through indicator functions.

In terms of spectroscopy, PCA transforms a set of spectra from a system into a set of independent, representative spectra for the system (the principal spectra). The principal spectra can serve to regenerate all of the features in any of the original spectra. The data matrix $|S|$ contains the experimental spectra (m spectra of n data points each), all from the same region (BE range in XPS). The principal component matrix $|P|$ contains spectra (m spectra of n data points each) in order starting from one that contributes the most to $|S|$. Peaks that appear consistently throughout spectra in $|S|$ will appear in the first principal spectrum in $|P|$. The least significant principal spectrum in $|P|$ generally will only be noise because the magnitude of noise contributes stochastically throughout all spectra in $|S|$.

The utility of PCA in XPS is primarily to decompose high-resolution spectra. Recall that high-resolution spectra are typically taken to determine oxidation or hybridation states of an element, and the desired information is traditionally obtained using peak fitting. Principal component analysis can be considered to have three advantages over peak fitting. Peak fitting often requires that certain parameters of component peaks be well-defined and constrained to obtain realistic results, whereas PCA requires no *a priori* or constrained information about the principal spectra. Peak fitting typically requires subjective input from the user during the analysis, whereas PCA is entirely objective in its analysis. Finally, peak fitting requires significantly more time to analyze one spectrum than PCA takes to return results from a set of spectra. To its disadvantage for applications in XPS, PCA does not return results that can be directly and unambiguously related to specific information about the chemistry of the system. For example, PCA often returns principal spectra that show negative going peaks and therefore would otherwise appear to be entirely unrealistic.

The number of spectra m provided in $|S|$ is important in order to be able to perform proper PCA. It should be significantly (from a statistical sense) larger than the number of significant principal spectra expected in $|P|$. If m is too small, then all spectra in $|P|$ will appear to be equally significant in contributing to $|S|$. As a guide, m should usually be at least a factor of three or more times larger than the number of significant spectra expected from $|P|$. For XPS, the number of significant spectra expected from $|P|$ can be taken as the number of component peaks that will otherwise be fit to a spectrum in $|S|$ with peak fitting. For peak fitting, the minimum value of m is the number of samples with different types of chemistry. To increase this minimum m to a more reliable value for PCA, spectra can be acquired from different samples of the same type, from different regions of the same sample, or even from different orientations (analyzer acceptance angles) of the same region on a sample.

In summary, peak fitting and PCA are complementary and supplementary tools to analyze XPS spectra. Peak fitting requires subjectivity and time to analyze the data (fit peaks to a spectrum), whereas PCA requires subjectivity to interpret the results (determine the physical significance of the principal spectra) and time to obtain the data. The results from PCA are obtained objectively and can guide the otherwise subjective analysis done with peak fitting. For peak fitting results to be statistical reliable, the number of component peaks fit to any spectrum in $|S|$ cannot exceed the number of significant principal spectra determined by PCA to be in $|P|$. In XPS, when more component peaks are used to peak fit than can be supported by PCA, the chemistry of the samples has been unjustifiably represented. Alternatively, when fewer components are used to peak fit than number of significant components found by PCA, certain chemical states of the element being considered have been neglected.

Appearance. When signal quality is poor, as can be the case in spectra from elements at low concentrations, improving the appearance of an XPS spectrum is sometimes considered an important part of data analysis and interpretation. Methods that increase the signal-to-noise ratio, S/N, are applied with the goal of enhancing the results. The use of signal smoothing procedures to improve the precision of quantitative results, particularly elemental compositions as well peak positions and half-widths, must be approached with care. While signal smoothing can be a useful tool, over-reliance on smoothing is a mistake.

The first concern in data acquisition should be to obtain a spectrum with sufficient S/N to obtain reliable quantitative information. The S/N ratio is, in practical terms, the height of a photoemission peak after subtracting a background divided by the root mean square of the intensity of the random noise at a location somewhere near the photoemission peak. The random noise is a stochastic process proportional to a deviation about a mean signal. Increasing the acquisition time in a data channel proportionally decreases the variance of a signal about the mean. Noise therefore increases as the square root of acquisition time. Signal increases proportionally with acquisition time. Therefore S/N increases as the square root of total acquisition time. The formulations provided in the previous section can be used to determine the corresponding relationships to scan, sweep, and dwell time as well as number of scans or sweeps. The value of S/N increases as the square root of any one of these when all others are held constant. In practice, S/N ratios of $\sim$2 to 3 are necessary lower bounds for determining peak parameters. Spectra with such low values of S/N are generally not suitable, however, to determine peak parameters reliably. A more reasonable S/N is above $\sim$10. The raw data in Figure 7 has a value of S/N value of $\sim$75 to 80, while that for CuO in Figure 6 has a value of $\sim$20 to 25. To obtain a comparable value of S/N for these two spectra, we would have to acquire the latter spectrum (Figure 6) for a 16- to 25-fold longer acquisition time.

Smoothing routines are used when data acquisition has failed to obtain sufficient *S*/*N*. Smoothing algorithms are, however, not without their problems. Those that depend on *n*-point averaging methods are notorious for changing peak parameters. Although this change may be below the level of uncertainty in measuring the peak parameters, a change will occur. The FWHM of a peak will be affected in all cases. For overlapping peaks, peak positions will shift, and asymmetrical peaks will change shape. A theoretically more rigorous smoothing algorithm is performed in the time (Fourier) domain using appropriately defined Fourier filter functions. This has been shown to distort peak shapes less. Finally, the advent of algorithms to perform noise reduction in chemical spectroscopy using PCA (Sastry, 1997) or wavelet analysis (Jetter et al., 2000; Shunsuke, 2000; Harrington, 1998) may offer a choice of a better post-processing method to improve the S/N in XPS spectra without distorting peak parameters.

Smoothing is to be avoided in cases where rigorous and exacting quantitative analysis of XPS spectra is desired. Nothing substitutes in this case for acquiring more data to increase *S*/*N* in the raw spectrum. Peak fitting is also best done to raw (noisy) data. To a first order, the optimization algorithms essentially "do not care" about the level of random noise in a spectrum. Smoothing of noisy spectra is appropriate to provide the viewer with a guide to follow the otherwise noisy signal. This is what has been done in Figure 6. The author strongly recommends that, when smoothing is used as an integral method of the data analysis and interpretation, both smoothed and raw (noisy) spectra always be shown simultaneously in the published XPS spectra.

One primary objective of high-resolution analysis with XPS is to obtain spectra that are not distorted by line shape broading due to the x-ray source or analyzer. On systems without a monochromator, computer software methods exist to deconvolve the line shape of the x-ray source. They depend traditionally on iterative routines or operate in the Fourier domain (Jansson, 1984). They all require accurate representations of the line shapes of the x-ray source, as published elsewhere (Krause and Ferreira, 1975). They are not without their problems. Spectra with low S/N or with inadvertent "spikes" will not deconvolve well. Methods to deconvolve x-ray lineshapes are falling out of favor, especially as monochromatic x-ray sources become more popular. Alternative (neither Fourier or iterative) deconvolution techniques however are being used successfully to improve peak resolution. They include the maximum entropy method (Pia and McIntyre, 2001; Pia and McIntyre, 1999; Pratt et al., 1998; Splinter and McIntyre, 1998; Schnydera et al., 2001) and PCA (Gilbert et al., 1982; Simmons et al., 1999). These methods appear to avoid some of the problems with the traditional Fourier or interative deconvolution methods and can be used to remove broadening effects from the analyzer.

Finally, some older XPS systems without a monochromator offer a software option to remove x-ray satellite lines. While this can be useful in survey scans, it has little utility in high-resolution scans. The scan range can be set to avoid the x-ray satellite. If an x-ray satellite interferes with a principal photoemission peak and accurate quantitative analysis is desired, the better option is to use an XPS system with a monochromatic x-ray source.

Interpretation

The results obtained from XPS are a list of the elements present, the compositions, and the positions and halfwidths of their associated peaks. How this information is to be interpreted is an important aspect of the proper use of XPS. The first section below deals with interpretation of the compositions from XPS. This is followed by a discussion about defining chemical states.

Compositions. As pointed out in the section on practical aspects of the technique, compositions determined by XPS are to be interpreted as arising from homogeneous samples. This leads to problems when the sample composition is not uniform. In particular, most samples analyzed immediately after insertion into the vacuum system will contain a layer of carbonaceous material, adsorbed water, and oxides depending on the type of material. The contaminant layer can be a few monolayers thick in some cases. This approaches the mean free path of most low KE electrons. Attenuation of peak intensities from the underlying material is therefore to be expected. The consequence of this is that atomic concentrations will have relative uncertainties approaching ±10% or more in most practical cases.

Consider a rather simple example. A monolayer of (graphitic) carbon on an iron surface will have thickness of ~0.1 nm. The prominent Fe 2*p* line appears at ~706 eV in BE. For Mg Kα radiation, this corresponds to ~550 eV in KE, for a mean free path of ~1 nm. The intensity of electrons from in the Fe 2*p* peak will be attenuated by a monolayer of carbon to ~91% at 90° take-off angle, 87% at 45°, and 56% at 10°. Based on Equation 15 and assuming sensitivity factors do not change, the same sample of otherwise pure Fe with only a uniform monolayer of carbon on its surface would therefore appear to have compositions of 9%, 13%, and 44% carbon at the respective take-off angles.

Most contaminant layers are thicker than a monolayer. Doubling the thickness of the layer nearly doubles the respective carbon compositions. For this reason, unexpectedly high values of atomic concentrations for carbon and oxygen on samples just inserted into the vacuum system are not unusual. Values as high as 50% to 60% carbon due to a contaminant overlayer are not uncommon, even on supposedly well cleaned surfaces. Metals (other than Au) will also always be covered by an oxide layer. Methods that can be used to reduce this contaminant overlayer are discussed in the section on sample preparation.

The C 1*s* peak appears at ~284 to 286 eV in BE. All photoemission peaks at higher BE than the C 1*s* line will likely have lower mean free paths. The concentrations of the corresponding elements will therefore be attenuated relative to carbon. The peak intensities for those elements below the C 1*s* peak in BE will be attenuated less.

An important objective of XPS analysis for sample composition should also be to estimate the uncertainties in composition values. A thorough guide to uncertainty analysis of data can be found elsewhere (Taylor, 1997).

Uncertainties in atomic concentrations determined from XPS generally contain two contributions, random (statistical) uncertainties and measurement errors. Statistical uncertainties are reduced by repeating the measurements multiple times. Certainly, more than one sample should be analyzed to obtain valid concentrations for the sample. A general guide in numerical data analysis is that at least ten measurements of a value are needed to begin to have a sample size that statistically approaches the behavior of a true representative set for the system. Small-sample statistics must otherwise be applied. In this regard, methods to report on the confidence level for values and to propagate uncertainties in values through calculations should be followed. For any composition value reported, one should, at the minimum, always report an average, the calculated standard deviation, and the number of samples used to obtain the value.

The extent of uncertainty in composition that arises due to measurement errors in an XPS spectrum is considered below for illustration. First, assume an ideal sample, where the composition is homogeneous throughout. For any one given measurement made to determine concentration with XPS, the relative uncertainty in absolute concentration scales in proportion to the relative uncertainty in measuring the peak area and the sensitivity factor. The peak area A_i and element sensitivity factor S_i enter in the calculation of atomic composition x_i (atom or mole percent) as a ratio $K_i = A_i/S_i$. The errors in measuring peak areas and element sensitivity factors may be correlated because they are measured using the same devices in both cases. For such situations, an upper boundary on the relative uncertainty in composition due to measure errors can be established as

$$\frac{\Delta x_i}{x_i} \leq \frac{\Delta K_i}{K_i} + \frac{\sqrt{\Sigma \Delta K_j^2}}{x_i K_i} \tag{27}$$

where $\Delta K_i^2 = K_i^2[(\Delta A_i/A_i)^2 + (\Delta S_i/S_i)]^2$. Note that the terms in the above equation would be added in quadrature if the measurement errors for area and for sensitivity factor respectively were uncorrelated from all other values in their sets. The above formulation is still valid to establish a conservative upper boundary for the influence of measurement errors on composition. Relative errors in S_i values can be as much as $\pm 10\%$, especially when the instrument has not been calibrated. The greater the number of elements considered, the greater the summation over all values of relative uncertainty in sensitivity factor will become. Therefore, the relative uncertainty in any one composition will increase as more components are considered. Relative errors in peak areas depend on how the measurement was made (by hand or by computer, for example) and on the choice of background. They can be reduced by making sure that the absolute peak area, A_i, is as large as possible. The effects of excessive noise on a spectrum complicate matters and will generally act to increase the value of $\Delta A_i/ = A_i$.

With an assumption that sensitivity factors are accurate, $(\Delta S_i/S_i = 0)$, $\Delta K_i/K_i = \Delta A_i/A_i$, and assuming all peaks areas are measured to the same degree of relative uncertainty, $f_a = \Delta A/A$, and recalling for any two elements that $K_A/K_B = x_A/x_B$, the above equation can be recast to

$$\frac{\Delta x_i}{x_i} \leq f_a \left[1 + \frac{\sqrt{\Sigma x_j^2}}{x_i^2}\right] \tag{28}$$

This shows that, all else being equal, components with the lowest composition x_i will have the largest relative uncertainty in their composition $\Delta x_i/x_i$. It also shows that, as more components are incorporated in a calculation of composition, the relative uncertainty in the absolute composition of any one component $\Delta x_i/x_i$ increases. The above equation also can be applied for the case where peak areas are measured with infinite accuracy and only sensitivity factors have measurement errors, in which case f_a is the relative error in sensitivity factor. Relative uncertainties in concentration ratios for any two components A and B are expressed as

$$R_{\mathrm{AB}} = \left(\frac{x_{\mathrm{A}}}{x_{\mathrm{B}}}\right) \leq \frac{\Delta K_{\mathrm{A}}}{K_{\mathrm{A}}} + \frac{\Delta K_{\mathrm{B}}}{K_{\mathrm{B}}} \tag{29}$$

When either sensitivity factors or peak areas are considered infinitely accurate, this expression becomes

$$R_{\mathrm{AB}} \leq f_{\mathrm{a,A}} + f_{\mathrm{a,B}} \tag{30}$$

where $f_{a,i}$ is the relative error in determining the peak area or sensitivity factor for component i. An important point of this equation is that, to a first approximation, uncertainties in the ratios of component compositions are not affected by the number of components being considered or the absolute compositions of the components. For this reason, composition ratios from XPS can be considered more significant (less uncertain) than absolute concentrations.

Order of magnitude numerical calculations can serve to illustrate the above formulations. First consider how increasing the number of components in a system increases the relative uncertainty in the composition of each component. When either sensitivity factors or areas are considered accurate, for systems with n components of equal composition (ideal stoichiometric compounds), the relative uncertainty on any one composition becomes $f_a(1 + n\sqrt{n})$, where f_a is the relative uncertainty in either sensitivity factor or peak area. The multiplier to f_a is 3.8, 6, 9, and 12 for systems of two to five components of equal composition, respectively. Specifically, for a 1:1 stoichiometric binary material such as CuO or NaCl, when the relative error in measuring peak areas or elemental sensitivity factors is 5% ($f_a = 0.05$), the relative uncertainty in measuring the composition of either component will be $\pm 19\%$. Therefore, the composition for either component in such a binary could only be measured to an accuracy $x = 0.50 \pm 0.10$ (atom fraction) using XPS. For this case, XPS measurements of compositions between 40 atomic percent and 60 atomic percent could not be distinguished

as statistically different. By further example, for a five-component material of equal compositions measured with a 5% relative error on peak areas or sensitivity factors, the absolute composition of any one component could only be measured to a precision of 0.20 ± 0.12 atomic percent. Incorporation of the errors for both area and sensitivity factors in the above formulation only serves to increase the overall inaccuracy in measuring the true value of composition. These examples illustrate clearly why absolute atomic concentrations reported from XPS should typically be considered to have a relative uncertainty (imprecision) of at least $\pm 10\%$ due to measurement errors. Calibration or otherwise clear validation of the sensitivity factors for the spectrometer being used and affirmation of stringent care in measuring peak areas should typically be sought before accepting reports of atomic concentrations with relative uncertainties lower than $\pm 10\%$.

A corresponding example for the concentration ratio R_{AB} can be developed. When either peak areas or sensitivity factors are accurate, the relative error in any value of R_{AB} is the sum of relative errors in the individual terms. For the above systems of two to five components, in cases where $f_a = 0.05$ for all components, the value of R_{AB} will be 0.10, or 10%. For the binary CuO, the ratio of copper to oxygen can be measured at best to an accuracy of 1.0 ± 0.1. For a five-component system of equal compositions throughout, the ratio of any two compositions can be measured at best to an accuracy of 1.0 ± 0.1. The relative error in measuring a concentration ratio does not depend on the number of components included in the analysis. This illustration can serve to help save analysis time. Suppose that changes in concentrations of only two components are of overriding importance in the analysis of a multicomponent system. Only XPS spectra from those two components will need to be obtained. Although absolute concentrations will not be possible to determine, the ratio of concentrations will be a reliable indicator of the behavior of the system.

Chemical State Information. In regard to elemental and chemical state identification, reference spectra have been published. One standard for reference spectra from elements are the handbook from Perkin-Elmer Corporation (Moulder et al., 1995). A reference book has also recently been made available on polymers (Kratos Analytical, 1997). Further reference libraries of XPS spectra are available from commercial vendors such as XPS International (*http://www.xpsdata.com/*); the journal *Surface Science Spectra* publishes experimental XPS spectra taken under well-defined conditions (sample type and treatment and instrument parameters) as references for other researchers.

Unambiguous interpretation of shifts and broadening of primary photoemission peaks in XPS spectra should be based on systematic elimination of the various factors that contribute to these effects. The first factor to consider and eliminate is the potential for sample charging. This artifact has been mentioned already and will be discussed in detail in the section on problems with the technique. The following discussion is based on the assumption that sample charging does not occur or that peak shifting or broadening effects due to sample charging have been eliminated or, more likely, appropriately accounted for. Consideration is based first on comparison of XPS spectra from a "bulk" material for a before and after situation—before a treatment and after the treatment. Overlayers are considered separately. Comparison of XPS spectra from one experimental system to another, in order to find similarities in the chemistry of the materials, is then considered.

For the same instrument conditions, relatively small amounts of peak broadening without a commensurate peak shift or appearance of a new peak is indicative of a movement toward greater heterogeneity in the sample chemistry or structure. The effects may go hand in hand. A change in structure may also be associated with subtle changes in the type of bonding between constituent elements in the material. These effects are often only worth pursuing in detail for samples with otherwise well defined chemistry and structure.

Peak broadening to such an extent that a new peak appears in close conjunction with the initial peak usually indicates a change in the chemistry in a portion of the material. The new peak may overlap with the initial peak so that peak fitting is required to resolve it. As long as some indication of the initial peak is present, assignment of the new peak to chemical state effects can usually be done with confidence. Note that changes in structure without changes in chemistry of a bulk material can also give rise to a new peak. Such changes are not as prevalent in materials.

When a peak shifts or significantly changes shape such that the initial peak disappears, the chemistry or structure (or both chemistry and structure) of a bulk material has changed. In some cases, these factors cannot be separated. Reference XPS spectra should be consulted. The changes may be obvious, as shown in the comparison of the Cu 2*p* peaks in Figure 6. In cases where reference spectra are not available, the chemistry and structure of the material are best probed with other techniques that provide supplemental information (e.g., x-ray diffraction for structure).

Thin film overlayers have unique properties that make interpretation of changes in their XPS spectra more difficult. Oxide layers on metals and other comparable insulator-like overlayers are prone to charging, especially as their thickness increases. Molecular overlayers on surfaces can show peak shifts due to changes in the degree of extra-atomic relaxation in the layer, and changes in the shapes of satellite peaks are often associated with changes in chemistry. Angle-resolved XPS is one method that can be used to accentuate "bulk" versus near surface behavior. Where possible, thin film adsorbates should also be analyzed with techniques such as attenuated total reflectance Fourier transform infrared (ATR-IR) spectroscopy, glancing angle FTIR, EELS, AES (AUGER ELECTRON SPECTROSCOPY), surface Raman (RAMAN SPECTROSCOPY OF SOLIDS), or thin film x-ray diffraction (XRD; X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ALALYSIS) to confirm proposed changes in chemistry or structure.

In conclusion, analysis of chemical state information using XPS is rarely without ambiguities in complex

Table 2. Peak Areas, Element Sensitivity Factors, and Atomic Compositions from Analysis of Si Wafer Using XPS

Orbital	Peak Area (Counts-eV/sec)	Sensitivity Factor	Atomic Concentration (Atomic %)
O 1*s*	94960	0.711	42
C 1*s*	11964	0.296	13
Si 2*p*	48366	0.339	45

systems. The technique should not be viewed as a definitive one. Analysis of the material with as many techniques as appropriate to the system at hand, and comparison to published reference spectra or spectra from similar materials, is often highly recommended.

Example. The example presented here concerns analysis of an Si wafer for composition, chemical state information, and oxide film thickness. A survey scan, obtained with a PHI 5400 system using Mg radiation at 15 kV and 325 W at a take-off angle of 45° and pass energy of 89.5 eV, is shown in Figure 4. The main photoemission peaks and Auger KVV transitions are labeled. Plasmon loss peaks predominate for both the O 1*s* and Si peaks. The x-ray satellite peaks are also visible for these lines. Peak areas, sensitivity factors, and compositions determined from the survey scan are given in Table 2.

High-resolution scans of the Si 2*p* peak taken with a pass energy of 8.95 eV at three different take-off angles are shown in Figure 13. Two distinct states of Si are apparent. The peak at higher BE is from SiO_2, and the lower BE state is from Si. The SiO_2 peak is from the oxide layer on top of the wafer. The relative intensities of the two peaks change as a function of take-off angle. At glancing angle ($\Theta = 10°$), the peak from SiO_2 predominates because analysis is more surface sensitive. The asymmetry in the Si peak for the Si state, clearly visible at 45° and 90° take-off angles, is due to the close overlap of the Si $2p_{3/2}$ and $2p_{1/2}$ peaks.

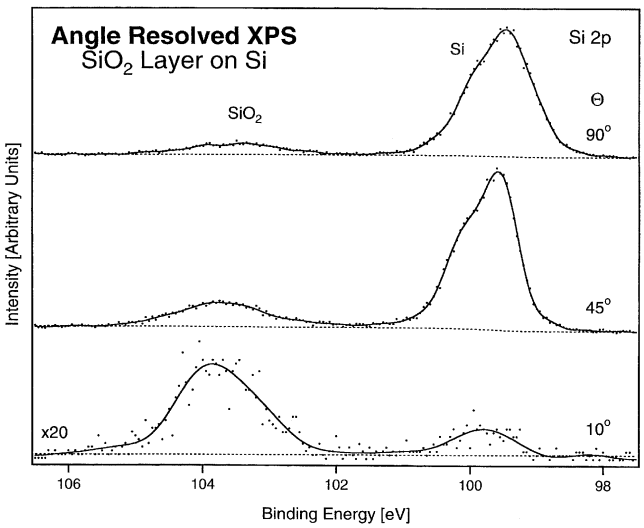


Figure 13. High-resolution scans of the Si 2*p* peak from a Si wafer at three different take-off angles. The oxide and neutral states of Si are clearly visible.

The areas under the two peaks were determined at each take-off angle by numerical integration (trapeziod rule) with computer software. The ratios of the SiO_2/Si peak areas were 4.71 (10°), 0.248 (45°), and 0.125 (90°). This data can be used in Equation 22 to determine the thickness of the oxide film. In this case, the cross-sections and mean free paths for the overlayer and substrate peaks can be taken to be the same since they are both the same orbital. Because an HSA was used for analysis, the ratio of étendue values is an inverse ratio of the KE of the peaks, which in this case is

$$\frac{G_o}{G_s} = \frac{E_{Ks}}{E_{Ko}} \approx \frac{(1253.6 - 99.4)}{(1253.6 - 103.4)} \approx 1.0 \tag{31}$$

Equation 22 therefore reduces to

$$\frac{I_o^m}{I_s^m} \approx \frac{(1 - \exp(-d/\lambda\sin\theta))}{\exp(-d/\lambda\sin\theta)} \left(\frac{C_o}{C_o}\right) \tag{32}$$

This can be rewritten as

$$\frac{d}{\lambda} = \ln\left[\frac{(A_o/A_s)}{(C_o/C_s)} + 1\right]\sin\theta \tag{33}$$

We can estimate the ratio of atomic concentrations for Si in the oxide layer and the underlying substrate to a first approximation. This can be done by using the densities and atomic concentrations in the layers

$$C_o/C_s = (x_{si}\rho_o M_s/\rho_s M_o) \tag{34}$$

where x_{Si} is the mole fraction of Si in the oxide (0.33 for SiO_2), ρ_i is the mass density of i, and M_i is the molar mass. If the mass densities of Si and SiO_2 are taken to be equal, the value of C_o/C_s is ~0.16. Using the area ratios from above, the ratio of film thickness to mean free path is determined to be 0.59 (10°), 0.66 (45°), and 0.58 (90°). The average ratio is 0.61 ± 0.04. Escape depths for the Si 2*p* peak are ~2.0 to 3.0 nm. Using an average of 2.5 nm, the oxide film thickness is found to be 1.5 ± 0.2 nm.

Before reviewing a more rigorous approach to analysis of the above ARXPS data, we can consider where mistakes are typically made in the above simple analysis. The first common mistake is to proceed without confirming that the ratios of cross-sections, mean free paths, analyzer étendues, and concentrations are unity. In the above calculation, since closely spaced peaks from the same orbital were used, all of the ratios but that for the atomic concentration could be safely approximated as unity. Had the ratio of concentrations also been taken as unity, the film thickness to mean free path ratios would have been determined to be 0.30 (10°), 0.16 (45°), and 0.12 (90°). This would have led to a reported oxide film thickness of 0.5 ± 0.2 nm. The significant difference between this value and the above, due to a failure to account for concentration differences, should be noted. When ARXPS data from two different peaks are used, the other ratios will not

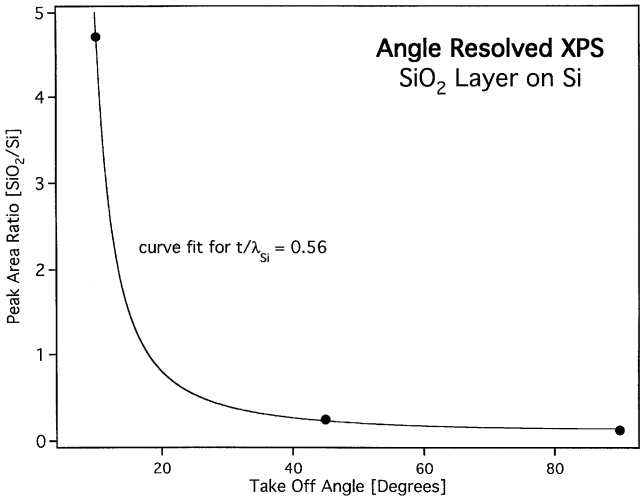


Figure 14. Results from nonlinear curve fitting of the peak area ratios to the equation for angle-resolved XPS. The solid line shows the fit for the given ratio of film thickness to electron mean free path.

necessarily be unity either. A second common mistake is to use ARXPS data from one take-off angle and apply Equation 22 in a significantly reduced form

$$d/\lambda = \ln[(A_o/A_s) + 1] \tag{35}$$

This equation can only be used for data taken at 90° take-off angle under situations where the above mentioned ratios can confidently be set to unity. As seen by the above sample calculations, gross inaccuracies can be expected in all other cases.

A better approach to the determination of film thickness from ARXPS data is to use the full compliment of data for non-linear curve fitting to Equation 22. The ratios of terms can either be determined or left lumped as a single fitting parameter. An example of the results using this method is shown in Figure 14. The three area ratios were fit to Equation 22 using a non-linear χ^2 optimization routine. The ratios of cross-sections, étendues, and concentrations were lumped into a single parameter, the mean free paths from the oxide and substrate were taken to be equal, and the other fitting parameter was the ratio of film thickness to electron mean free path. The results show good agreement with a fit to d/λ of 0.56. The lumped parameter was close to unity. As the figure shows, ARXPS data should be taken near glancing take-off angles for the best sensitivity to film thickness. The data above ~45° contribute far less to the shape of the fitting curve. The confidence in the fitting would be improved by having data at 10° increments from 10° to 90°. Deviations that arise in fitting ARXPS data will also be more apparent at low take-off angles, as discussed in the section on problems with XPS.

SAMPLE PREPARATION

Materials surfaces exposed to atmospheric conditions will contain a layer of contaminants, primarily adsorbed water and hydrocarbons. One objective of sample preparation prior to analysis with XPS could be to remove this layer without affecting the chemistry of the underlying material. For samples with intentionally applied molecular adsorbate layers, treatments that remove the contaminant layer may also remove the adsorbed layer, so that special precautions may be needed. Sample preparation should also focus on obtaining a representative cross-section of the material for analysis. When the extent of spatial variations in composition across a sample are expected to be larger than the analysis area at the sample surface, the recommended practice is to analyze over multiple regions on the sample surface. Finally, the optimal place for preparing a sample is often in the same vacuum chamber as the XPS analysis.

Because of constraints on the sample introduction ports of typical XPS vacuum chambers, samples are usually restricted to be a certain size. Typical sample sizes are ~3 to 5 mm thick and small enough to fit in a circle of 1 to 3 cm diameter (depending on the XPS instrument). Some newer XPS systems have been designed to handle larger size samples, up to 10 to 15 cm diameter (for example, for the Si wafer industry). The geometry of the sample is sometimes also restricted. Flat samples (like a hockey puck) are ideal in all systems. Curved surfaces and crevices present problems because either the x-ray source or the analyzer can be "shadowed" by the sample. Analysis cannot be done on hidden inner surfaces of samples.

Samples must be low out-gassing materials so that the pumps on the XPS chamber can maintain a suitable vacuum during sample analysis. This generally means they should have vapor pressures well below 10^{-8} mbar. Some XPS systems include a cold stage that cools the sample during analysis, potentially to as low as liquid N_2 temperatures. This can be beneficial in reducing the vapor pressure of otherwise high-vapor-pressure materials. Polymer samples are notorious for their outgassing problems. Small samples and thin films are best to use in this case, to minimize the total amount of material that outgasses. Porous materials and powders often present the same problems with extensive outgassing. They can often be analyzed if they have been well outgassed beforehand. This can often be done using a sample-preparation chamber that is attached to the main vacuum chamber containing the XPS instrumentation. In certain cases where ultraclean vacuum conditions in the main XPS analysis chamber are essential (with pressures below 10^{-9} mbar), samples such as polymers and powders that can potentially outgas a great deal are not permitted into the main chamber for XPS analysis. The policies on this vary from laboratory to laboratory.

For bulk-like sample analysis with XPS, the first step in sample preparation is to select a representative sample of the material being considered. Homogeneity of sample composition and structure are important factors. Samples with heterogeneities in composition or structure, such as grains or phase microstructures in metals or phase domains in copolymers, can be analyzed successfully either by analyzing over a larger area relative to the size of the grains or domains or by selectively analyzing over a statistically meaningful number of well-defined smaller

areas across the sample surface. Samples with thin film overlayers should be considered as candidates for sputter profiling or angle-resolved XPS.

Cleaning samples for bulk-like analysis does not typically require a significant effort unless the material is corroded or dirty. Sample surfaces can generally be washed and degreased by standard means using solvents such as water, methanol, or acetone. The solvents should be low in their content of residual contaminants—distilled, deionized water and research grade solvents are recommended. Cleaning the sample with an aqueous detergent wash and distilled water rinse in an ultrasonic cleaner has more recently become a recommended final step. In addition, the use of an ozone-generating UV lamp to reactively remove hydrocarbons is becoming more widely adopted. Polishing is not always necessary; however rough or pitted surfaces tend to entrap contaminants more readily than smooth surfaces. If polishing is done, extra care should be given to making sure to clean the polishing compound from the sample surface.

Prior to XPS analysis for bulk-like information, the thin outermost contaminant layer on a sample can often be removed by reactive ion etching. Sputter cleaning to remove the first few surface layers on a sample is often a routine part of the XPS analysis procedure. Excessively rough or pitted samples will not sputter-clean adequately. Insulator and polymer samples also present problems. These materials generally charge while being sputtered. This can cause uneven sputter rates across the surface. Polymers can also degrade during sputtering.

Preparation of samples with thin-film, adsorbed layers is generally done using well-defined recipes specific to each system. When ultraclean surfaces are a requirement and sputter cleaning is not a viable cleaning method, well-defined sample-preparation methods using ultraclean chemicals can be essential. Even the simple task of rinsing a sample in a beaker of solvent subjects its surface to the potential for further contamination. Under normal conditions, the air-liquid interface of solvents in beakers collects a layer of adsorbed contaminants, and these contaminants can transfer to a sample surface when it is passed through the air-liquid interface. As illustrated above (see Practical Aspects of the Method), even a monolayer of contaminants attenuates the underlying signal intensity, leading to inaccurate results.

Some XPS facilities are equipped with a stage where the sample can be heated while it is in the vacuum. This can be used to desorb chemically adsorbed overlayers. Some XPS systems also have sample-preparation chambers where the samples can be exposed to reactive gases (oxygen). This can be used to react off contaminant layers. Finally, XPS systems can also be equipped with all manner of tools within the vacuum system to cleave, cut, notch, or slice a sample. This is the best method to obtain a pristine surface.

Preparation of powder samples for XPS analysis presents unique problems. In almost all cases, free-standing powders are not put into the main vacuum system for XPS analysis. Should the powders spill, the potential for damage to vacuum components inside the chamber, especially the pumping units, is far too high. Therefore, powders must be mounted securely to a holder in some manner. Common techniques for mounting powders include the use of double-sided tape, mounting wax, and soft metal foils. Double-sided tape is usually of a special quality that is used extensively in vacuum systems because of its low outgassing properties. Powders can also be pressed or sprinkled into thermoset or thermoplastic mounting waxes that are subsequently allowed to cure or harden. Finally, soft metal foils, in particular indium or gold, are often used to mount powders. The powder is pressed firmly into the metal. Typically, only a single layer or two of particles that are uniformly distributed across the sample surface is necessary to get a reasonable signal. In cases where the powder is somewhat self-adherent when compacted, a thicker layer can often be formed on the tape, wax, or metal foil. In all of these cases, the resulting mounted powder sample should be tapped firmly while its surface is in a vertical orientation in order to remove all loosely adhering particles.

Particles may also be mounted as free-standing samples after being tightly compacted into free-standing pellets. The pellets should be stable to the routine stresses involved in handling and mounting. They should not crumble easily. When pellets are used, they can also be mounted in sample holders that have wells to hold the pellet. Finally, powders have also been mounted by compacting them somewhat and then containing them in a specially designed cup-like holder under a wire gauze. This requires the same care to contain loose particles from spilling as all the other methods.

One disadvantage of mounting particles in any form other than pellets is that the mounting material will inevitably appear in the XPS spectrum. Double-sided tape contains carbon, silicon (adhesives), and oxygen. So do most mounting waxes. The indium foil will have an oxide coating. This problem must be handled on a case-by-case basis. Another problem is that powders in any form are notorious for having a carbonaceous overlayer, and sputter cleaning is not usually a viable option. Finally, porous powders can have problems with outgassing, and XPS will not provide a representative analysis of materials within pores due to limitations on electron escape depth.

SPECIMEN MODIFICATION

Under proper conditions, XPS typically does not modify the sample. Polymers are, however, sensitive to photo-induced reactions during XPS analysis. In cases where the photoemission yield or secondary electron background is high, damage to polymers may occur by electron-induced chemical degradation. Photostimulated desorption of gaseous adlayers on surfaces can also occur for all materials. Finally, if electron beam neutralizers are used to reduce sample charging, the electron current can cause the sample to degrade.

PROBLEMS

The most significant problem in XPS analysis is sample charging. As photoelectrons are emitted from the sample,

a continual current (sample current) must flow from ground. Insulating and even semiconducting samples also have measurable capacitance in addition to resistance properties. Oxide layers on metals behave in the same way. This leads to a build-up of charge on these samples, typically on the surface being analyzed. The charge build-up is always positive, leading to a shift of the photoelectrons to lower KE and therefore higher BE. The shift is uniform across the entire XPS spectrum by as much as 2 to 3 eV. Inhomogeneous samples can also have differential charging across the sample surface. This leads to peak broadening, since the extent of charging is not uniform across the sampled areas.

Three methods are available to take care of sample charging. One method draws on experiences in SEM and AES. The sample surface is coated with a thin (sub-monolayer) film of conductive metal, typically Au. This method requires a metal-deposition system that can coat samples. Such systems are commonly found at scanning electron microscopy (SEM) facilities.

A second method is to supply additional electrons to the sample from a secondary source. This is usually done using an electron gun or an electron-emitting filament. The electrons that are supplied are flooded at a very low E_K, below $\sim$50 eV. There are afew problems with this method. First, even at low E_K, electrons are damaging to some classes of materials, particularly polymers and sometimes ceramics. Secondly, the exact amount of electron flood current to supply is dependent on the sample type and surface condition. Trial and error experiments can be required to find the flood current that exactly compensates for the positive charge build-up on a surface. Changing the sample or sample orientation relative to the flood gun will change the current needed. Finally, when sample charging is inhomogeneous, an electron flood gun will not reduce peak broadening. Despite these limitations, electron flood guns are commonly used to correct for sample charging in many XPS systems.

The third method is to post-process the spectrum. Since charging shifts the spectrum uniformly, post-correction requires shifting all of the peaks back (to lower BE values) by a uniform amount. The problem with this method is that a reference is needed to align the peak positions. Without prior knowledge of the exact oxidation states of all principal peaks in the spectrum, the amount of charging correction shift to apply is entirely arbitrary. The most common standard internal reference to compensate for charging is to align the main C 1*s* peak to a value for graphitic carbon, 284.6 eV. Two problems arise with this. First, not all carbon is graphitic in nature. The C 1*s* peak position is particularly sensitive to changes in hybridization state, as seen in Figure 7. This means that the internal reference state may not be correct for the sample at hand. The second problem is that even when graphitic carbon is present, differential charging and the presence of other states of carbon will broaden the C 1*s* peak. If this broadening is asymmetric, the C 1*s* peak position will only be exact if it is determined by rigorous peak fitting procedures.

None of the above methods to correct for sample charging are reliable in all situations. When reporting peak positions from samples that may have charging shifts, one potentially better method of comparing chemical differences from sample to sample is to report relative peak positions. Define one peak that does not change position from sample to sample as an E_B standard for the set of experiments. Report all other E_B values relative to this standard peak. Changes in offsets from the reference peak should be a better indication of any possible changes in chemical state or final state effects. This is comparable to what is done in determining Auger parameters, where photoemission peak positions are calculated relative to respective XAES peaks.

As reported above (see Practical Aspects of the Method), the accuracy of the energy calibration of the XPS instrument should always be confirmed. Three common errors arise: offset, sloping, and nonlinear scaling. Offset errors are typically due to an incorrect setting of the analyzer work function. This will cause all of the peak positions to be off by a constant amount, toward either lower or higher BE. Offset errors are corrected by calibrating the work function of the instrument using the peak position of a well-established standard, typically the Au $4f_{7/2}$ line. Sloping errors typically arise when the electronics for the instrument have linear deviations from their required output voltages. This will cause peaks on one side of the calibration peak to shift to higher BE values and those on the other to shift to lower values, for example. This error is corrected by calibrating the instrument electronics for linearity across a wide BE range using at least two appropriately spaced reference peak positions. Finally, nonlinear scaling errors arise due to nonlinearities in the electronics. They will cause the BE of peaks to fall on a curved rather than straight line in a plot of measured BE versus expected BE.

One caution in operating XPS systems with conventional x-ray sources is to always confirm the integrity of the x-ray source window. The window is a thin metal foil at the end of the x-ray source between the electron filament and the sample. This x-ray window is thin enough to be nearly transparent to x rays, yet opaque to electrons. A typical x-ray window is Al. Small pinholes in the window will allow electrons from the x-ray filament to bombard the sample. This leads to an increase in the overall background of the spectrum without increasing the overall signal. The flood of high-KE electrons can also damage the sample, particularly polymers. The x-ray window evaporates over time and should be replaced as needed.

A second problem with laboratory x-ray sources arises as the anode degrades with time. The active metal coating (Mg or Al) on the anode (which is typically a block of Cu) evaporates with time or will oxidize if high water vapor or oxygen partial pressures are continually present. This problem leads to a significant decrease in S/N and signal relative to background intensity in the XPS spectrum. In systems with dual anodes (Mg/Al being most common), the metals from the two separate anodes can diffuse across the boundary separating them. This will cause the appearance of ghost peaks in survey scans. For example, when a Mg anode becomes contaminated with Al, every photoemission peak in a spectrum taken with the Mg anode will have a "ghost" peak at a value that is 233 eV higher in BE. The relative height of the ghost to the main peak

will depend on the extent of contamination. To check for ghost peaks when using a Mg/Al dual anode system, look for the most intense photoemission peak, then look either 233 eV higher or lower in BE, depending on the original anode, to determine whether cross-contamination has occurred.

Finally, a concern that must be kept in mind when analyzing ARXPS data is that film roughness and analyzer acceptance angle will both act to degrade the quality of the results. The analyzer acceptance angle limits the angular resolution and lowest take-off angle that can be measured. Commercial instruments typically have half angles of acceptance at the input lens on the order of 10°; therefore measurements below ~10° in take-off angle will show a significant decrease in signal intensity, and angular steps less than ~10° will not add any significantly new information. Instruments are available with half angles of acceptance as low as 2°. Film roughness will also cause a deviation from Equation 22. Measured peak intensity ratios will be lower than the expected values at low take-off angles, and can be higher than expected at near normal take-off angles. Film thickness values estimated using Equation 22 that are less than the anticipated mean surface roughness of the film should be suspect.

SUMMARY

The overall protocol for XPS analysis is generally to define the sample, determine the preparation and cleaning methods, mount and pre-pump the sample, load and align it, set up the analysis parameters, and obtain the XPS spectrum. After the sample has been prepared and cleaned outside of the vacuum system, the rest of the procedure can take anywhere from 15 min to a day to complete. Outgassing of the sample in a prep-vacuum preparation chamber is usually essential for porous or low-vapor-pressure materials.

A typical sequence of automated sample analysis on a modern XPS instrument proceeds after inserting a sample or set of samples. The user then defines, for each sample, the number of spatial regions and their locations, the number of acquisition sets, and the parameters for each set. High-resolution scans can be taken in sequence within a data acquisition set. This mode is called multiplex analysis on some instruments. Sputter profiling or angle-resolved analysis adds additional parameters for each acquisition set, such as the duty cycle of the sputter gun or the take-off angles for analysis. On modern instruments and with a well-established routine, a user can in principle load a set of samples, program the XPS instrument for data acquisition and analysis, and walk away.

ACKNOWLEDGMENTS

The bulk of this unit was written against a background of questions that arose during a graduate lecture course on surface spectroscopic techniques taught at the University of Alabama in Huntsville in the spring of 1998. The input of the class was beneficial in defining the scope of this tutorial. The ARXPS data were obtained by a group of students as part of the laboratory assignment for the course. Critical reviews of the manuscript by M. A. George and J. C. Gregory are greatly appreciated.

LITERATURE CITED

Anthony, M. T. and Seah, M. P. 1984a. XPS: Energy calibration of electron spectrometers. 1. An absolute, traceable energy calibration and the provision of atomic reference line energies. *Surf. Interface Anal.* 6:95–106.

Anthony, M. T. and Seah. M. P. 1984b. XPS: Energy calibration of electron spectrometers. 2. Results of an interlaboratory comparison. *Surf. Interface Anal.* 6:107–115.

Artyushkova, K. and Fulghum, J. 2001. Identification of chemical components in xps spectra and images using multivariate statistical analysis methods. *Journal of Electron Spectroscopy and Related Phenomena* 121:33–55.

ASM Handbook Committee (ed.) 1986. Metals Handbook, Vol. 10: Materials Characterization, 9th ed. American Society for Metals, Metals Park, Ohio.

Balcerowska, G. and Siuda, R. 1999. Inelastic background subtraction from a set of angle-dependent XPS spectra using PCA and polynomial approximation. *Vacuum Surface Engineering, Surface Instrumentation and Vacuum Technology* 54:195–199.

Balcerowska, G., Bukaluk, A., Seweryn, J. and Rozwadowski, M. 1999. Grain boundary diffusion of Pd through Ag thin layers evaporated on polycrystalline Pd in high vacuum studied by means of AES, PCA, and FA. *Vacuum Surface Engineering, Surface Instrumentation and Vacuum Technology* 54:93–97.

Blaschuk, A. G. 2001. Comparison of concentration profiles obtained by ion sputtering and nondestructive layer-by-layer analysis. *Metallofiz. Noveishie Tekhnol.* 23:255–271.

Briggs, D. and Seah, M. P. (eds.). 1994. Practical Surface Analysis: Auger and X-ray Photoelectron Spectroscopy, Vol. 1, 2nd ed. John Wiley & Sons, New York.

Bruhwiler, P. A. and Schnatterly, S. E. 1988. Core threshold and charge-density waves in alkali metals. *Phys. Rev. Lett.* 61:357–360.

Castle, J. E., Ruoru, K., and Watts, J. F. 1990. Additional in-depth information obtainable from the energy loss features of photoelectron peaks. *Corrosion Sci.* 30:771–798.

Castle, J. E., Chapman-Kpodo, H., Proctor, A., and Salvi, A. M. 2000. Curve-fitting in xps using extrinsic and intrinsic background structure. *Journal of Electron Spectroscopy and Related Phenomena* 106:65–80.

Clark, D. T. 1979. Structure, bonding, and reactivity of polymer surfaces studies by means of ESCA. *In* Chemistry and Physics of Solid Surfaces, Vol. 2 (R. Vanselow, ed.) pp. 1–51. CRC Press, Boca Raton, Fla.

Clark, D. T. and Feast, W. J. (eds.). 1978. Polymer Surfaces. John Wiley & Sons, New York.

Clumpson, P. 1997. Using principal components with chemical discrimination in surface analysis. *Spectra in Quantitative Analysis II 13.* VAM Bulletin (*http://www.vam.org.uk/news/ news_bulletin.asp*).

Conny, J. M. and Powell, C. J. 2000. Standard test datafor estimating peak parameter errors in xps: I, II, and III. *Surface and Interface Analysis* 29.

Crist, V. 1998. Advanced peak fitting of monochromatic xps spectra. *J. Surf. Anal.* 4:428–434.

Doniach, S. and Sunjic, M. 1970. Many electron singularity in x-ray photoemission and x-ray line spectra from metals. *J. Phys.* C 3.

Drummond, I. W., Ogden, L. P., Street, F. J., and Surman, D. J. 1991. Methodology, performance, and application of an imaging x-ray photoelectron spectrometer. *J. Vacuum Sci. Technol.* A9:1434–1440.

Ehrentreich, F., S. Nikolov, G., Wolkenstein, M., and Hutter, H. 1998. The wavelet trans-form: A new preprocessign method for peak recognition of infrared spectra. *Mikrochimica Acta* 128:241–250.

Ertl, G. and Küppers, J. 1985. Low Energy Electrons and Surface Chemistry. Springer-Verlag, Weinheim, Germany.

Fadley, C. S. 1976. Solid state and surface analysis by means of angular-dependent x-ray photoelectron spectroscopy. *Prog. Surf. Sci.* 11:265–343.

Fadley, C. S. 1984. Angle-resolved x-ray photoelectron spectroscopy. *Prog. Surf. Sci.* 16:275–388.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. North-Holland Publishing, New York.

Gilbert, R., Llewellyn, J., Swartz, W., and Palmer, J. 1982. Application of factor analysis to the resolution of overlapping XPS spectra. *Appl. Spectrosc.* 36:428–430.

Gunter, P. L. J., Gijzeman, O. L. J., and Niemantsverdriet, J. W. 1997. Surface roughness effects in quantitative XPS: Magic angle for determining overlayer thickness. *Appl. Surf. Sci.* 115:342–346.

Harrington, P. 1998. Different discrete wavelet transforms applied to denoising analytical data. *Journal of Chemical Information in Computer Science* 36:1161–1170.

Hofmann, S. 1992. Cascade mixing limitations in sputter profiling. *J. Vacuum Sci. Technol.* B10:316–322.

Hofmann, S. 1993. Approaching the limits of high resolution depth profiling. *Applied Surf. Sci.* 70-71:9-19

Hofmann, S. 2001. Profile reconstruction in sputter profiling. *Thin Solid Films* 398-399:336–342

Holgado, J., Rafael, A., and Guillermo, M. 2000. Study of CeO 2 XPS spectra by factor analysis: Reduction of CeO_2. *Applied Surface Science* 161:301.

Ibach, H. and Lüth, H. 1990. Solid State Physics: An Introduction to Theory and Experiment. Springer-Verlag, New York.

Jablonski, A. and Powell, C. J. 1993. Formalism and parameters for quantitative surface analysis by Auger electron spectroscopy and x-ray photoelectron spectroscopy. *Surf. Interface Anal.* 20:771–786.

Jansson, P. A. 1984. Deconvolution with Applications in Spectroscopy. Academic Press, Orlando, Fla.

Jetter, K., Depczynski, U., Molt, K., and Niemoller, A. 2000. Principles and applications of wavelet transformation to chemometrics. *Anal. Chim. Acta* 420:169–180.

Kratos Analytical. 1997. High-Resolution XPS Spectra of Polymers. Kratos Analytical, Chestnut Ridge, N.Y.

Krause, M. O. and Ferreira, J. G. 1975. K x-ray emission spectra of Mg and Al. *J. Phys. B: Atomic Mol. Phys.* 8:2007–2014.

Lesch, N., Aretz, A., Pidun, M. Richter, S., and Karduck, P. 2001. Application of sputter-assisted EPMA to depth profiling analysis. *Mikrochimica Acta* 132:377–382.

Lide, D. R. (ed.) 1993. Handbook of Chemistry and Physics, 73rd ed. CRC Press, Ann Arbor, Mich.

Malinowski, E. and Howery, D. 1980. Factor Analysis in Chemistry. John Wiley & Sons, New York.

Marcu, P., Hinnen, C., and Olefjord, I. 1993. Determination of attenuation lengths of photoelectrons in aluminum and aluminum oxide by angle-dependent x-ray photoelectron spectroscopy. *Surf. Interface Anal.* 20:923–929.

Moulder, J. F., Stickle, W. F., Sobol, P. E., and Bomben, K. D. 1995. Handbook of X-ray Photoelectron Spectroscopy. Perkin-Elmer, Eden Prairie, Minn.

Nöller, H. G., Polaschegg, H. D., and Schillalies, H. 1974. A step towards quantitative electron spectroscopy measurements. *J. Electron Spectrosc. Related Phenomena* 5:705–723.

Palmberg, P. W. 1975. A combined ESCA and Auger spectrometer. *J. Vacuum Sci. Technol.* 12:379–384.

Palmberg, P. W., Bohn, G. K., and Tracy, J. C. 1969. High sensitivity auger electron spectrometer. *Appl. Phys. Lett.* 15:254–255.

Passeggi, M. 1998. Auger electron spectroscopy analysis of the first stages of thermally stimulated oxidation of GaAs(100). *Appl. Surf. Sci.* 133:65–72.

Pia, H. and McIntyre, N. S. 1999. High resolution xps studies of thin film gold-aluminum alloy structures. *Surface Science* 421:L171-176.

Pia, H. and McIntyre, N. S. 2001. High-resolution valence band xps studies of thin film au-al alloys. *Journal of Electron Spectroscopy and Related Phenomena* 119:9–33.

Pratt, A. R., McIntyre, N. S., and Splinter, S. J. 1998. Deconvolution of pyrite, marcasite and arsenopyrite xps spectra using the maximum entropy method. *Surface Sci.* 396:266.

Richie, R., Oswald, S., and Wetzig, K. 2001. Xps and factor analysis for investigation of sputter-cleaned surface of metal (Re, Ir, Cr)-silicon thin films. *Appl. Surf. Sci.* 179:316–323.

Rummel, J. 1990. Applied Factor Analysis. Northwestern University Press, Evanston, Ill.

Salvi, A. M., Decker, F., Varsano, F, and Speranza, G. 2001. Use of xps for the study of cerium-vanadium (electrochromic) mixed oxides. *Surface and Interface Analysis* 31:255–264.

Sastry, M. 1997. Application of principal component analysis to X-ray photoelectron spectroscopy: The role of noise in the spectra. *Journal of Electron Spectroscopy and Related Phenomena* 83:143–150.

Schnydera, B., Alliata, D., Kotza, R., and Siegenthaler, H. 2001. Electrochemical intercalation of perchlorate ions in hopg: An sfm/lfm and xps study. *Applied Surface Science* 173:221–232.

Scofield, J. H. 1976. Hartree-Slater subshell photoionization cross-sections at 1254 and 1487 eV. *J. Electron Spectrosc. Related Phenomena* 8:129–137.

Seah, M. P. 1985. Measurement: AES and XPS. *J. Vacuum Sci. Technol.* A3:1330–1337.

Seah, M. P. 1993. XPS reference procedure for the accurate intensity calibration of electron spectrometers—results of a BCR intercomparison co-sponsored by the VAMAS SCA TWP. *J. Electron Spectrosc. Related Phenomena* 20:243–266.

Seah, M. P and Brown, M. T. 1998. Validation and accuracy of software for peak synthesis in xps. *Journal of Electron Spectroscopy and Related Phenomena* 95:71–93.

Seah, M. P. and Smith, G. C. 1988. Concept of an imaging XPS system. *Surf. Interface Anal.* 11:69–79.

Seah, M. P., Jones, M. E., and Anthony, M. T. 1984. Quantitative XPS: The calibration of spectrometer intensity-energy response functions. 2. Results of interlaboratory measurements for commercial instruments. *Surf. Interface Anal.* 6:242–254.

Sekine, T., Ikeo, N., and Nagasawa, Y. 1996. Comparison of AES chemical shifts with XPS chemical shifts. *Appl. Surf. Sci.* 100:30–35.

Shirley, D. A. 1972. High-resolution x-ray photoemission spectrum of the valence band of gold. *Phys. Rev. B* 5:4709–4714.

Shunsuke, M. 2000. Application of spline wavelet transformation to the analysis of extended energy-loss fine structure. *J. Electron Microsc.* 49:525–529.

Siegbahn, K. 1985. Photoelectron Spectroscopy: Retrospects and Prospects. Royal Society of London.

Siegbahn, K., Nordling, C., Fahlman, A., Nordberg, R., Hamrin, K., Hedman, J., Hohansson, G., Bergmark, T., Karlsson, S. E., Lindgren, I., and Lindberg, B. 1969a. ESCA: Atomic, molecular, and solid structure studied by means of electron spectroscopy. *Nova Acta Regiae Soc. Uppsala* 4:20.

Siegbahn, K., Nordling, C., Johansson, G., Hedman, J., Heden, P. F., Hamrin, R., Gelius, U., Bergmark, T., Werme, L. O., Manne, R., and Baer, Y. 1969b. ESCA Applied to Free Molecules. North-Holland Publishing, Amsterdam.

Simmons, G., Angst, D., and Klier, K. 1999. A self modeling approach to the resolution of XPS spectra into surface and bulk components. *Journal of Electron Spectroscopy and Related Phenomena* 105 :197–210.

Solomon, J. 1987. Factor analysis in surface spectroscopies. *Thin Solid Films* 154:11–20.

Splinter, S. J. and McIntyre, N. S. 1998. Resolution enhancement of x-ray photoelectron spectra by maximum entropy deconvolution. *Surface and Interface Analysis* 26:195–203.'

Tanuma, S., Powell, C. J., and Penn, D. R. 1988. Calculations of electron inelastic mean free paths I. Data for 27 elements and four compounds over the 200-2000 eV range. *Surf. Interface Anal.* 11:577.

Tanuma, S., Powell, C. J., and Penn, D. R. 1991a. Calculations of electron inelastic mean free paths. II. Data for 27 elements over the 50-2000 eV range. *Surf. Interface Anal.* 17:911–926.

Tanuma, S., Powell, C. J., and Penn, D. R. 1991b. Calculations of electron inelastic mean free paths. III. Data for 15 inorganic compounds over the 50-2000 eV range. *Surf. Interface Anal.* 17:927–939.

Tanuma, S., Ichimura, S., and Yoshihara, K. 1996. Calculations of effective inelastic mean free paths in solids. *Surf. Interface Anal.* 100:47–50.

Taylor, J. R. 1997. An Introduction to Error Analysis,2nd ed. University Science Books, Sausalito, Calif.

Tougaard, S. 1988. Quantitative analysis of the inelastic background in surface electron spectroscopy. *Surf. Interface Anal.* 11:453–472.

Tougaard, S. 1996. Quantitative XPS: Non-destructive analysis of surface nanostructures. *Appl. Surf. Sci.* 100:1–10.

Wetzig, K., Baunack, S., Hoffmann, V., Oswald, S, and Prassler, F. 1997. Quantitative depth profiling of thin layers. *Fresenius J. Anal. Chem.* 358:25–31.

Zhang, X. Q., Zheng, J. B., and Gao, H. 2000. Comparison of wavelet transform and fourier self-deconvolution (FSD) and wavelet FSC for curve fitting. *Analyst* 125:915–919.

KEY REFERENCES

ASM Handbook Committee, 1986. See above.

Covers use of XPS for metallurgical applications.

Barr, T. L. 1994. Modern ESCA. CRC Press, New York.

A good overview of XPS, with descriptions of important photoemission phenomena from a quantum mechanical perspective. gives a more rigorous quantum mechanical presentation of certain photoemission processes in a manner that complements the information given in Briggs and Seah (1994). Also provides an important historical perspective, especially regarding the contributions of Siegbahn to the development of the experimental aspects of the technique that has become known as electron spectroscopy for chemical analysis (ESCA; Siegbahn et al., 1969a,b; Siegbahn, 1985). Recommended tor reader who is approaching XPS for the first time.

Briggs, D. 1998. Surface analysis of polymers by xps and static sims. *In* Cambridge Solid State Science Series (D. R. Clarke, S. Suresh, and I. M. Ward, eds.).. Cambridge University Press, Cambridge, U.K.

An in-depth treatment of the instrumentation, physical basis, and applications of XPS (and static secondary ion mass spectroscopy) with a specific focus on polymers. The book includes details from five case studies, including analysis of copolymer, biopolymer, and electropolymer surfaces.

Briggs, D. and Beamson, G. 1992. High Resolution XPS of Organic Polymers: The Scientia ESCA300 Database. John Wiley & Sons, New York.

This reference includes high resolution XPS spectra of over 100 organic polymers recorded at nearly intrinsic signal line widths. It contains a full spread of complete data, describes the database and the procedures used, and lists other relevant information.

Briggs and Seah, 1994. See above.

Perhaps the most comprehensive discussion of all significant practical aspects of XPS. Also covers Auger electron spectroscopy (treated separately in AUGER ELECTRON SPECTROSCOPY). Unattributed information in this unit is drawn primarily from this reference.

Crist, B. V. 1991. Handbook of Monochromatic XPS Spectra: Polymers and Polymers Damaged by X-Rays. John Wiley & Sons, New York.

Crist, B. V. 2000a. Handbook of Monochromatic XPS Spectra: Semiconductors. John Wiley & Sons, New York.

Crist, B. V. 2000b. Handbook of Monochromatic XPS Spectra: The Elements and Native Oxides. John Wiley & Sons, New York.

This recent three volume set by B. Vincent Crist includes spectra from native oxides, polymers, and semiconductors. Each volume contains an introductory section giving extensive details of the instrument, sample, experiment, and data processing. All spectra in the volumes are collected in a self consistent manner, all peaks in the survey spectra are fully annotated, and each high-energy resolution spectrum is peak-fitted with detailed tables containing binding energies, FWHM values, and relative percentages of the component.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. North-Holland Publishing, New York.

Covers use of XPS for thin film applications.

Ghosh, P. K. 1983. Introduction to Photoelectron Spectroscopy, Chemical Analysis, Vol. 67. John Wiley & Sons, New York.

Though dated, covers the theoretical aspects of XPS with important and illustrative examples from gas phase systems. Provides a good treatment of XPS and the associated photoemission phenomena. Recommended tor reader who is approaching XPS for the first time.

Kratos Analytical, 1997. See above.

Handbook which is an essential source for reference XPS spectra of polymers.

Moulder et al., 1995. See above.

Handbook which is an essential source for reference XPS spectra of the elements

Walls, J. M. 1989. Methods of Surface Analysis: Techniques and Applications. Cambridge University Press, Cambridge, U.K..

Recommended for reader who is approaching XPS for the first time. Provides a well-rounded overview of the technique as a complement to the detail offered in Briggs and Seah (1994).

Brundle, C. R. and Baker, A. D. 1978. Electron Spectroscopy—Theory, Technique, and Applications. Academic Press, London.

Carlson, T. A. 1975. Photoelectron and Auger Spectroscopy. Plenum Press, New York.

Watts, J. F. 1990. An Introduction to Surface Analysis by Electron Spectroscopy. Oxford Science Publications, Oxford.

Woodruff, D. P. and Delchar, T. A. (eds.). 1988. Modern Techniques in Surface Science. Cambridge University Press, Cambridge, U.K.

The above four references include text and serial chapters that have been written over the past 25 years above XPS; some are no longer in print.

Brundle, C. R., Evans, J. C. A., and Wilson, S. (eds.). 1992. Encyclopedia of Materials Characterization: Surfaces, Interfaces, Thin Films. Butterworth-Heinemann, Boston.

Cahn, R. W. and Lifshin, E. (eds.). 1993. Concise Encyclopedia of Materials Characterization, 1st ed. Pergamon Press, New York.

Sibilia, J. P. (ed.) 1988. A Guide to Materials Characterization and Chemical Analysis. VCH Publishers, New York.

Vickerman, J. C. 1997. Surface Analysis: The Principal Techniques. John Wiley & Sons, New York.

Wachtman, J. B. and Kalman, Z. H. 1993. Characterization of Materials. Butterworth-Heinemann, Stoneham, Mass.

Walls, J. M. and Smith, R. (eds.). 1994. Surface Science Techniques. Elsevier Science Publishing, New York.

The above six publications are encyclopedic references that generally provide less depth on the specific topic of XPS but may provide a better perspective on its relation to other spectroscopic and analytical techniques. Walls and Smith (1994), Vickerman (1997), Brundle et al. (1992), and Cahn and Lifshin (1993) provide, potentially, the best overview. Sibilia (1988) is somewhat more limited in its coverage, and Wachtman and Kalman (1993) is relatively less practical (and correspondingly more abstract) in its style.

Christmann, K. 1991. Introduction to Surface Physical Chemistry, Topics in Physical Chemistry. Springer-Verlag, New York.

Drummond et al., 1991. See above

Gunter et al., 1997. See above.

Prutton, M. 1994. Introduction to Surface Physics. Oxford University Press, New York.

Somorjai, G. A. 1994. Introduction to Surface Chemistry and Catalysis. John Wiley & Sons, New York.

The above five publications include a discussion section on XPS while providing a comprehensive background on surface chemistry or surface physics.

Articles pertaining to the principles and applications of XPS are frequently presented in the journals Applied Surface Science, Colloid and Surface Science, Journal of Vacuum Science & Technology, Surface and Interface Science, and Surface Science.

INTERNET RESOURCES

http://srdata.nist.gov/xps/

This Web page at the National Institute of Standards and Technology (NIST) is particularly worth citing. It contains positions of XPS and XAES peaks for elements in various compounds as well as plot (Wagner plots) of the Auger parameter for elements in various compounds.

APPENDIX: PEAK SHAPES

In general, a peak shape is named for the analytical function that defines it. Common peak shapes include the Gaussian, Lorentzian, and Voigt. Mixtures of these, such as a Gauss + Lorentz, and variations of these, such as an assymetric Lorentz or assymetric Voigt, are also widely used. For fitting to XPS spectra, special peak shapes such as a Gauss + tails (Briggs and Seah, 1994) or the Doniach-Sunjic (Bruhwiler and Schnatterly, 1988; Doniach and Sunjic, 1970) have also been developed. Some peak shape functions can be derived directly from a first-principles analysis of the response expected from the spectroscopic event (excitation of a photoelectron). These include the Gaussian and Lorentzian functions. Alternatively, analytical functions for peak shapes such as the Gaussiann + tails include formulations from semi-empirical analysis of experimental data.

Peak shapes can be classified in many ways. One method is according to the number of parameters needed to define the shape analytically. In this respect, at least four classes of peak shapes are commonly used to fit peaks in chemical spectroscopies, starting from those that require three parameters to those that must use six. Each class may contain more than one peak shape, but all peak shapes within a given class use the same number of parameters to define them.

The principal parameters used in an analytical function to define a peak shape in spectroscopy are the peak height, half-width or full-width at half-maximum, position, area, assymetry or shape factor, and mixing ratio. Certain types of XPS peaks are also defined by relative parameters from another peak, such as relative height or offset (relative position). This happens for spin-split doublets (for example, $p_{3/2}$ and $p_{1/2}$ peaks) and for satellites (shake-up and loss peaks). Finally, peak fitting can be done using an experimental spectrum or set of spectra rather than analytical peak shapes. Such reference style peaks require only two parameters to fit to an experimental spectrum: their height and their offset relative to the spectrum.

The above discussion suggests that XPS peak shapes can be categorized globally into three styles, single peaks, relative peaks (such as spin-split doublets), and reference peaks (using a full experimental spectrum). The first two styles are obtained from analytical functions and will be called singlets and ntuplets for the number of peaks they define in a spectrum. Reference-style peaks are derived from experimental data and will not be considered in any further detail in the discussion below. In principle, an ntuplet-style peak uses n times as many parameters as its comparably shaped singlet, where n is the number of ntuplets. Rather than expand the number of peak classes beyond four to accommodate this (potentially unlimited) expansion of parameters, classification of a peak shape is restricted to the number of principal parameters it

requires. Singlet and ntuplet peak styles correspondingly then each contain only four different peak classes. As an example, a Gaussian singlet style peak and a Gaussian doublet-style peak (an ntuplet style with two peaks) are in the same peak class because they both use the same number of principal parameters (three as shown below). The doublet will require an additional set of three relative parameters to define shape of the second (relative) peak.

The four peak classes of XPS peak shapes are labeled A through D below. For each class, examples of analytical functions are given for singlet style peak shapes. A representative doublet style is given for a Gaussian peak shape (the first peak shape in class A). Where possible, the formula are given in a format suitable for input into a symbolic math or graphing package for programming purposes. In all of the functions, S_{ps} is the analytical signal for the peak shape with a name that is abbreviated as ps, x is the energy scale (assumed to be in eV binding energy in XPS), w is the full width of the peak at exactly half of its height (full-width at half-maximum or FWHM in eV), p is the peak position (eV), and a is the peak area (counts-eV/sec). Other parameters are defined as needed.

The classification scheme above and the functions provided below are by no means unique or unequivocal.

Class A (3 Parameters)

Two of the peak shapes in this class, Gaussian and Lorentzian are the most commonly used in spectroscopy.

Gaussian. Many processes that lead to a signal in chemical spectroscopy are random in nature and result statistically in a Gaussian signature. In statistical analysis, a normal probability curve is a Gaussian function defined by its mean value or position μand its standard deviation, σ, which is measured as the half-width of the Gaussian peak at a distance of about 61% of the peak height. The normal probability curve also has an integrated area of unity.

The normal probability function is not useful for fitting peaks to spectroscopic data. Almost all spectroscopic peaks have areas that are not identically equal to unity, and w (the FWHM or the width at 50% of the height) of a spectroscopic peak an experimental spectrum is more easily measured than σ (the width of the corresponding Gaussian function at about 61% of its height). The Gaussian functions given below are therefore defined in terms of spectroscopic parameters, especially w.As a point of reference for a Gaussian function, the FWHM is $w = 2\sigma\sqrt{2\ln(2)}$. Also, for any form of the Gaussian, the relationship between h, σ, w, and a is $a = 2\sqrt{2\pi}h\sigma = \sqrt{\pi\ln(2)}hw$.

Three parameters define a normal probability Gaussian curve, μ, σ, and an area of unity. Therefore p and any two parameters from the set of h, w,and a are needed to define a spectroscopic Gaussian peak. Correspondingly, three different Gaussian shape peak functions exist in Class A depending on which two parameters are chosen in addition to p.

Position/height/full-width-half-maximum. A singlet style peak is represented as

$$S_{GPHF} = h\exp\left(-4\frac{\ln(2)(x-p)^2}{w^2}\right) \tag{36}$$

For this peak, the area is

$$a = hw\sqrt{\pi}/2\sqrt{\ln(2)} \tag{37}$$

An example of a doublet style where both peaks have the same Gaussian shape is given below. In the function, D_{DGHF} is the signal, r_h is the relative height, o is the offset, and r_w is the relative fwhm of the doublet peak

$$\begin{aligned} D_{DPHE} = {} & h\exp\left(-4\frac{\ln(2)(x-p)^2}{w^2}\right) \\ & + r_h h\exp\left(-4\frac{\ln(2)(x-p-o)^2}{(r_w w)^2}\right) \end{aligned} \tag{38}$$

Position/area/full-width-half-maximum

$$S_{GPAF} = \left(\frac{ea\sqrt{\ln(2)}}{w\sqrt{\pi}}\right)\exp\left(-4\frac{\ln(2)(x-p)^2}{w^2}\right) \tag{39}$$

The height of this peak is

$$h = 2a\sqrt{\ln(2)}/w\sqrt{\pi} \tag{40}$$

Position/height/area

$$S_{GPHA} = h\exp\left(-2\frac{h^2\pi(x-p)^2}{a^2}\right) \tag{41}$$

The FWHM of this peak is

$$w = 2a\sqrt{\ln(2)}/h\sqrt{\pi} \tag{42}$$

Lorentzian. The Lorentzian peak shape is also fundamental to chemical spectroscopy. It is often considered a more natural line shape for chemical spectroscopy than the Gaussian function because the broadening it displays is characteristic of the lifetime broadening found in quantized energy states. The Lorentzian function gives a broader peak at the base than does the Gaussian. Two variations of this peak shape exist:

Position/height/full-width-half-maximum

$$S_{LPHF} = h\Big/\left(4\frac{(x-p)^2}{w^2}+1\right) \tag{43}$$

The area of this peak is $a = \pi w/2$.

Position/height/area

$$S_{\mathrm{LPHA}} = h \Big/ \left(\frac{(x-p)^2 \pi^2}{a^2} + 1 \right) \tag{44}$$

The FWHM of this peak is $w = 2a/\pi$.

Class B (4 Parameters)

These peak shapes include a fourth parameter that is either f_a—*a peak shape assymetry factor—or* f_g, a fractional Gaussian. In these peak shapes, h and w do not always define the actual height and half-width of the peak as measured in the resultant curve.

Gaussian + Lorentzian. This is a combination of the two Class A peak shapes. The f_g value ranges from 0 (pure Lorentzian) to 1 (pure Gaussian). Two different variations can be formed

$$S_{\mathrm{G+L(A)}} = f_g S_{\mathrm{GPHF}} + (1 - f_g) S_{\mathrm{LPHF}} \tag{45}$$

$$S_{\mathrm{G+L(B)}} = f_g S_{\mathrm{GPHA}} + (1 - f_g) S_{\mathrm{LPHA}} \tag{46}$$

Voigt. The Voigt function has the form

$$S_{\mathrm{V}} = \left(\frac{1}{\alpha} \right) \int_{-\infty}^{\infty} \frac{h \exp(-y^2)}{w_{\mathrm{L}}^2 + ((x-p)/w_{\mathrm{G}}) - y)^2} dy \tag{47}$$

$$\alpha = \int_{-\infty}^{\infty} \frac{\exp(-y^2)}{w_{\mathrm{L}}^2 + y^2} dy \tag{48}$$

where w_{L} and w_{G} are Lorentzian and Gaussian components of the FWHM of the peak. While an analytical solution for α exists, the remaining integral cannot be solved analytically. The Voigt peak shape is therefore typically obtained numerically. One method is to use a lookup function fVoigt(p_1, f_a), where $f_a$is a shape function and the value of p_1 is defined below

$$S_{\mathrm{v}} = \mathrm{f\,Voigt}(p_1, f_{\mathrm{a}}) \tag{49}$$

$$p_1 = 2 \sqrt{\ln(2) \frac{(x-p)}{w}} \tag{50}$$

The function f Voigt returns a value after numerical interpolation from a table. The shape factor f_a varies from 0 to infinity. Larger values give flatter, broader peaks (less Gaussian in shape).

Assymetric Lorentzian. This form of the Lorentzian gives a skew to the higher x side of the peak as the assymetry factor f_a varies from 0 to 1. A value of 0 gives a symmetrical Lorentzian, as can be shown by analytical geometry. As f_a increases from 0, the location of the maximum of the peak shifts toward greater x and the value of h no longer denotes the true peak height. For f_a above 0.5, the function returns to a symmetrical, flat, broad peak shape centered at p

$$S_{\mathrm{AL}} = \left(\frac{h}{\alpha} \right) \cos\left(-f_a + (1 - f_a) \arctan\left(2 \frac{x-p}{w} \right) \right) \tag{51}$$

$$\alpha = 2 \sqrt{\frac{(x-p)^2}{w^2}} + 1 \tag{52}$$

The area of the peak when $f_{\mathrm{a}} = 0$ is $a = \pi w/2$ and otherwise goes to infinity as f_{a} goes to unity.

Doniach-Sunjic. This peak shape has been used to fit the d and f levels of transition metals (Bruhwiler and Schnatterly 1988; Doniach and Sunjic 1970). The actual height, FWHM, and position of the resultant principle peak are only equal to the input values when the assymetry factor $f_{\mathrm{a}} = 0$. As f_{a} increases, the peak becomes assymetrical at higher x values above p and the maximum in the peak shifts toward higher values of x away from p

$$S_{\mathrm{DS}} = \frac{h}{\alpha} \cos\left(\frac{\pi f_{\mathrm{a}}}{2} + (1 - f_{\mathrm{a}}) \arctan(p - x, w) \right) w^{1 - f_{\mathrm{a}}} \tag{53}$$

$$\alpha = ((p-x)^2 + w^2)^{(1-f_{\mathrm{a}})/2} \times \cos\left(\frac{\pi f_{\mathrm{a}}}{2} + (1 - f_{\mathrm{a}}) \arctan(0, w) \right) \tag{54}$$

In the above expression, arctan (y, x) computes the principal value of the argument of the complex number $x + iy$ so that $-\pi < arctan(y,x) \leq \pi$. When f_{a} is unity, the Doniach-Sunjic peak shape is undefined (the function returns a horizontal line at a value of h). The area of this peak is $a = \pi h w$ for $f_{\mathrm{a}} = 0$ and goes to infinity as f_{a} goes to unity.

Class C (5 Parameters)

The simplest peaks in this class are just combinations of the Class A peak shapes and the assymetric Class B peak shapes. One additional parameter f_g defines the mixing ratio.

Gaussian + assymetric Lorentzian

$$S_{\mathrm{G+AL}} = f_g S_{\mathrm{GPHF}} + (1 - f_g) S_{\mathrm{AL}} \tag{55}$$

Class D (6 Parameters)

These peak shapes include parameters to define tails. Tails are assymetric extensions on one side of the peak. In XPS spectroscopy, tails are used to model the increase in intensity on the high binding energy side of the peak due to inelastic scattering.

Gaussian + Tails. Three additional parameters define the shape of the tail. They are c_t—constant tail height fraction—e_t—exponential tail height—and *fc*—fraction of constant tail (versus exponential tail). The tail function signal

S_{TF} is set to zero on the low x side of the peak position p using the logical test $(x \geq p)$.

$$S_{TF} = F_c c_t + (1 - f_c) \exp\left(-\frac{|x-p|}{e_t}\right)(x \geq p) \quad (56)$$

$$S_{GPHF+T} = S_{GPHF}(1 - S_{TF}) + hS_{TF} \quad (57)$$

JEFFREY J. WEIMER
University of Alabama at Huntsville
Huntsville, Alabama

SURFACE X-RAY DIFFRACTION

INTRODUCTION

Surface X-ray Diffraction

Surface x-ray diffraction allows the determination of atomic structures of ordered crystal surfaces in an analogous way to x-ray crystallography for the determination of three-dimensional crystal structures. In doing this, crystallographic parameters such as lattice constants and Debye-Waller factors (thermal vibration amplitudes) are also accessed, so surface x-ray diffraction should be considered as a valid method for measuring these quantities too. Surface x-ray diffraction also probes surface morphological properties, such as roughness and facet formation, and allows the investigation of their thermodynamic and kinetic aspects, such as the study of phase transitions in surfaces.

Because x rays probe deep inside matter, buried interfaces can be treated in exactly the same manner as surfaces, so throughout this unit the words *surface* and *interface* will be used synonymously. In fact, some of the best applications of the method are for buried interfaces, where complementary electron-based probes such as low-energy electron diffraction (LEED; LOW-ENERGY ELECTRON DIFFRACTION and reflection high-energy electron diffraction cannot reach.

Surface x-ray diffraction is not a laboratory technique. While there are one or two exceptions of installations in individual laboratories, the majority are associated with national or international facilities. Even the very early experiments carried out by Eisenberger and Marra (1981) used synchrotron radiation. By now, facility-based science is already rather common and will become more so in the future, as it has considerable economies of scale and advantages of centralized operational safety. The practice of surface x-ray diffraction should necessarily be considered as a facility-based operation, and it will be assumed that the reader will have access as a "user" to one of these facilities. This is usually handled by the submission of a proposal to be judged on the importance of the science to be carried out. Access is then usually free of charge.

Because surface x-ray diffraction is facility based, it is not a technique for which one can purchase the equipment piecemeal from a vendor. Every surface x-ray diffraction installation is slightly different, but most were derived from and are based on the description that follows, which is the X16A facility of the National Synchrotron Light Source (NSLS). Others, such as the Advanced Photon Source (APS, Argonne), the European Synchrotron Radiation Facility (ESRF, Grenoble), Hamburger Synchrotronstrahlungslabor (HASYLAB, Hamburg), Daresbury Laboratory, Photon Factory (PF, Tsukuba), and Laboratoire pour l'Utilisation de Rayonnement Synchrotron (LURE, Orsay), use basically the same control and analysis programs and procedures as those described here.

Competitive and Related Techniques

The need to travel to a facility to make surface x-ray diffraction measurements can be inconvenient, to say the least, so it is important to discuss its strengths and weaknesses compared with other techniques. Scanning probe methods, such as scanning tunneling microscopy (STM, SCANNING TUNNELING MICROSCOPY and atomic force microscopy (AFM), are extensively used in surface science. These imaging methods are unbeatable for detecting singular events on surfaces, and the "microscopist's eye" will often detect a wide variety of local behavior. Thus STM has been very important in understanding instabilities of surfaces, step properties, and island morphologies during growth. When "average" quantities are needed for understanding general behavior, such as thermodynamics, the averaging must be carried out explicitly; diffraction methods will access ensemble-average quantities directly and will filter out the exceptions. This is particularly important when studying the dynamical behavior associated with phase transitions, where the correlations within a moving arrangement of atoms can still be detected by diffraction. At their limit, scanning probe microscopies can reach atomic resolution, but they always detect the outermost features on a surface and will not see below the surface. Diffraction methods can attain atomic positions to an accuracy of 0.01 Å routinely, while scanning microscopies can only achieve that in the surface-normal direction under certain circumstances.

The direct competitors with surface x-ray diffraction are thus the various forms of electron diffraction (Electron Techniques), which provide information of an analogous form: LEED, reflection high-energy electron diffraction (RHEED), and transmission electron diffraction (TED). Since electrons interact more strongly with matter than do x rays, there are some important differences. First, electron diffraction is dynamical and requires a nonlinear theory to explain the intensity of the diffraction seen. This has been accomplished by deriving a careful theoretical description of the electron as it interacts with the sample in the case of LEED and to a lesser extent for RHEED. After 20 years of development, the method has reached the point where structures can be refined as well as just calculated, and thermal vibrations can now be included. With x rays, the simpler kinematical description can be used, which makes the calculations very easy, so that powerful refinement procedures can be followed. Transmission electron diffraction is near to this kinematical

limit also, and the linear approximation has been successfully used to obtain surface structures. The potential user can make a choice between a hard experiment and an easy analysis in the case of x rays or between an easy experiment and a hard analysis in the case of electron diffraction.

Second, for the same reason, the penetration of x rays is significantly larger than electrons. The sensitivity of LEED and RHEED is limited to about four atomic layers, and the accuracy becomes reduced in the lower layers. To examine interfaces, LEED has been used, but it requires physical removal of half of the sample. X rays are not limited in this way and have many applications for interfaces: an important example is the electrochemical interface, which can be studied in situ. Transmission electron diffraction falls in between: samples must be thinned to 100 to 1000 Å but can still contain interfaces. The small differences in accuracy that can be obtained in surface structure determination by the different diffraction techniques can be attributed to the accessible range of reciprocal space. X-ray diffraction and TED tend to have higher accuracy parallel to the surface, because the largest component of momentum transfer is usually in that direction. Low-energy electron diffraction is a backscattering technique, which tends to give more accurate coordinates in the normal direction.

It used to be said that the momentum resolution of x rays was an unmatchable asset, but this has become less important in recent years. First, the spot profile analysis LEED (SPA-LEED) method was developed by Sheithauer et al. (1986) as well as the spot profile analysis RHEED (SPA-RHEED) method by Müller and Henzler (1995). These offer momentum resolution, which is comparable to x-ray diffraction, and are commercially available from Omicron. Surface-phase transitions can now be studied as well with LEED or RHEED as with x-ray diffraction. Second, the merit of probing correlations extending over large distances in real space has diminished with the widespread availability of STM and AFM, which routinely scan the same range.

PRINCIPLES OF THE METHOD

Surface is a generic word that is often used as an antonym to bulk. Deviations from physical properties of bulk matter are often attributed to (i.e., blamed on) "surface effects." The study of surfaces is therefore a productive way of understanding these deviations. The study of surfaces always starts with their structure, and that is what is probed with surface x-ray diffraction. Typical examples of surfaces, generalized to include internal interfaces, are shown in Figure 1.

The general subject of diffraction will not be reviewed here since it appears elsewhere in this volume and also is covered in several excellent textbooks (James, 1950; Warren, 1969; Guinier, 1963; Lipson and Cochran, 1966; Woolfson, 1997). It is assumed that the reader is comfortable with the construction of a three-dimensional (3D) reciprocal space and the idea that the diffraction from a crystal is localized to a set of specific "Bragg" points forming a reciprocal lattice. Reciprocal space is spanned by the vectorial

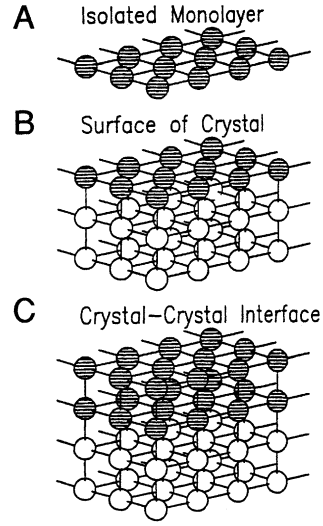


Figure 1. Entities that can be studied with surface x-ray diffraction. (**A**) isolated monolayer (2D crystal). (**B**) Same as A but connected to a crystalline substrate. (**C**) Same as A but at the interface between two different crystals. From Robinson and Tweet (1992).

quantity **q**, which is the momentum transfer of the diffraction experiment. The axes of reciprocal space are chosen to lie along simple directions within the reciprocal lattice. The coordinates are then normalized so that the Bragg points appear at simple integers, called Miller indices, *hkl*. This construction of reciprocal space was introduced in SYMMETRY IN CRYSTALLOGRAPHY and will be assumed henceforth.

In surface x-ray diffraction, it is conventional to choose a coordinate system in reciprocal space that places the third index of the momentum transfer, q_z, perpendicular to the surface. This is usually compatible with the Miller index notation because the naturally occurring surfaces are usually close-packed planes of atoms, which have small-integer Miller indices according to the usual construction. This convention often requires use of an unusual setting of the bulk crystal unit cell, e.g., hexagonal for a cubic (111) surface or tetragonal for a cubic (110) surface. The convention is exactly the same as that used in LEED; see LOW-ENERGY ELECTRON DIFFRACTION. Another common terminology is to separate the components of momentum transfer, **q**, parallel and perpendicular to the surface by writing $\mathbf{q} = (q_P, q_z)$. Here q_P strictly represents both in-plane components. The reason for this separation is that a surface is a highly anisotropic object with completely different properties in the two directions.

Diffraction from Surfaces

The most important characteristic of diffraction from surfaces is illustrated as a 3D sketch in Figure 2, which corresponds roughly to Figure 1 transformed into reciprocal space. Bulk crystal diffraction is concentrated at points of reciprocal space because of the 3D periodicity of the crystal. A surface or interface has reduced dimensionality, so its diffraction is no longer confined to a point but extends continuously in the direction perpendicular to the plane of

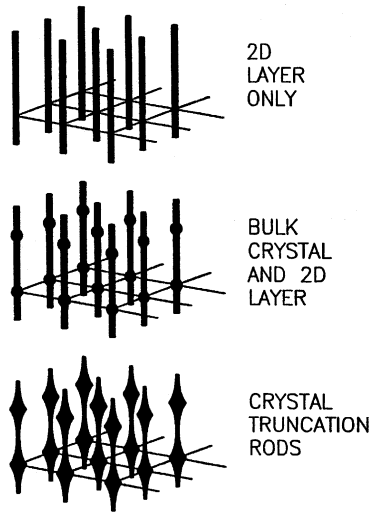


Figure 2. Schematic of the diffraction from the objects in Figure 1. The 3D points of reciprocal space are indicated as circles; 2D rods of reciprocal space are indicated as bars. From Robinson and Tweet (1992).

the surface or interface. These lines of diffraction, sharp in two in-plane directions and diffuse in the perpendicular direction, are called rods. The diffraction pattern of a two-dimensional (2D) crystal, as illustrated in Figure 2, is a 2D lattice of rods in reciprocal space.

A surface rarely exists as an isolated entity (Fig. 1A) but is normally intimately connected to a substrate of bulk crystalline material (Fig. 1B) or materials (Fig. 1C). The diffraction is then a superposition of 3D and 2D features, points and rods, as shown. However, the periodicities of the two lattices are the same (by construction), and so their diffraction patterns will be intimately related. This causes the lateral positions of the features to align so that the intensity distributions merge together. The 2D rods line up between the 3D points, and the two features simply merge together. The resulting object in reciprocal space, illustrated in the bottom panel of Figure 2, is called a crystal truncation rod (CTR). An example of an experimental study of a CTR is given below (see Practical Aspects of the Method).

Crystal truncation rods have been treated theoretically as extensions of bulk diffraction by Andrews and Cowley (1985), Afanas'ev and Melkonyan (1983), and Robinson (1986), and their connection to the surface structure has been derived. The sensitivity to the surface is found to vary smoothly along the rod. Near the divergence of the intensity at the 3D Bragg points, the intensity of the CTR depends only on the bulk structure, through its bulk structure factor. According to the kinematic theory of diffraction, the intensity follows a $(q_z - G_z)^{-2}$ law near each bulk Bragg peak at position $\mathbf{G} = (G_P, G_z)$. In the more precise dynamical theory discussed by Afanas'ev and Melkonyan (1983), this nonphysical divergence at $q_z = G_z$ becomes finite and follows a Darwin curve instead, but the asymptotic behavior remains the same as in the kinematic theory.

The connection between the intensity distribution along a CTR and the corresponding surface structure is

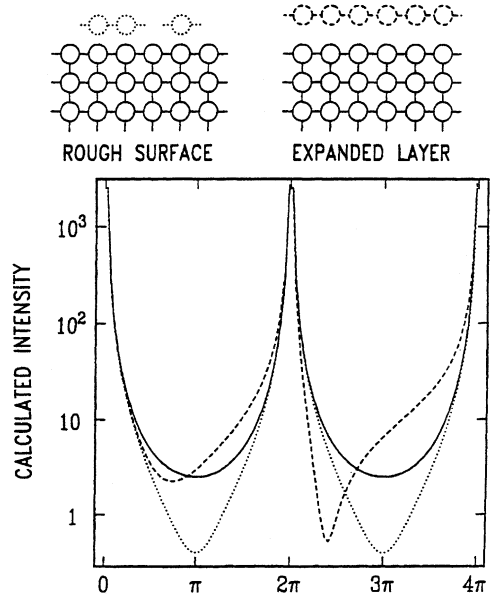


Figure 3. Calculated CTRs for an ideally terminated surface (solid curve), a surface with modified top-layer spacing (dashed), and a rough surface (dotted). The intensity is plotted as a function of perpendicular momentum transfer, q_z, in normalized units.

illustrated by a simple calculation in Figure 3. The full curve is for an ideally terminated simple-cubic lattice of atoms; its functional form is just $1/(\sin q_z)^2$, whose divergence at Bragg points $q = 2n\pi$ is clearly visible. The dashed curve corresponds to an outward displacement of a single layer of atoms at the surface; the intensity curve near the Bragg peaks, where there is little surface sensitivity, is barely changed, but the intensity at the CTR minimum is strongly modified. Finally, the dotted curve is for a rough surface, modeled by random omission of a fraction of the atoms in the top layer; here, again, the biggest effect is at the CTR minimum, this time with a symmetric drop of the intensity curve.

A simple rule of thumb is that the measurement is the most surface sensitive at the position on the CTR that is furthest from the bulk Bragg peaks, where the intensity is also weakest. This rule applies not only to the structure itself but also to the sensitivity to fluctuations in the surface, e.g., its roughness. In performing an experiment, one is frequently faced with a choice between surface sensitivity and signal level, since one quantity trades off directly with the other.

Measurement of Surface X-ray Diffraction

The surface x-ray diffraction signal can be measured with a standard diffractometer. For reasons explained below (see Practical Aspects of the Method), it is common to use a diffractometer with extra degrees of freedom. Synchrotron radiation is highly desirable as it results in convenient signal levels around hundreds to thousands of counts per second. The primary responsibility of the diffractometer is to select $\mathbf{q}$ by suitable choice of its angles while keeping the beam always on the center of the sample.

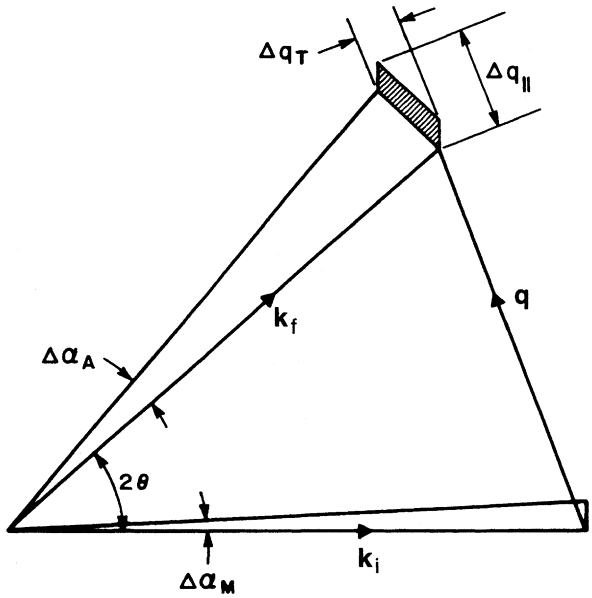


Figure 4. Construction of the resolution function within the scattering plane of a diffractometer. From Robinson (1990b).

Measurements then consist of scans of the **q** vector through the various features of the surface x-ray diffraction.

The shape and size of the resulting scans depend on the nature of the diffraction features being measured, e.g., due to the degree of order in the surface. The results also depend on the resolution function of the instrument, which is always convolved with the true diffraction pattern. The resolution function is controllable to a certain extent by instrumental parameters, particularly slits that define the beam divergence in different directions. The usual construction is shown in Figure 4, which is based on the vector construction of **q** in terms of the incident and exit wave vectors $\mathbf{k}_i$ and $\mathbf{k}_f$: $\mathbf{q} = \mathbf{k}_i - \mathbf{k}_f$. The variation of $\mathbf{k}_i$ is represented by the angular width of directions allowed by the entrance slits and monochromator, $\Delta\alpha_M$, while the variation of $\mathbf{k}_f$ is defined by $\Delta\alpha_A$. The resulting resolution function is the shaded parallelogram at the top of the picture. Simple geometry shows that its dimensions parallel and perpendicular to the vector **q** are given by Moncton and Brown (1983):

$$\Delta q_{||} = k(\Delta\alpha_M + \Delta\alpha_A)\cos\theta$$
$$\Delta q_T = k(\Delta\alpha_M + \Delta\alpha_A)\sin\theta \qquad (1)$$

where $k = 2\pi/\lambda$ is the length of both wave vectors and θ is half the diffraction angle 2θ.

Note that this picture applies to the scattering plane of the diffractometer, spanned by the 2θ angle of the diffractometer, and the notations Δq_P and Δq_T refer to parallel with (or radial) and transverse to the direction of **q**. The scattering plane is not the same as the surface plane of the sample, except in the grazing incidence geometry defined below, which is often used for precisely that reason. The third dimension of the resolution function, denoted $\Delta q_\perp$, is also defined by slits, this time in the direction perpendicular to the scattering plane; the resolution function then becomes a parallelogram-based prism by simple extension of Figure 4 to 3D. This out-of-plane resolution is often the largest of the three because, in the grazing incidence geometry, the diffraction features are extended rods in that direction, and a considerable enhancement of intensity is thereby attained.

Many setups, including the one at X16A, gain additional flexibility by dividing the $\Delta q_\perp$ range into narrower sections by use of a linear position-sensitive detector (PSD). The x-ray counts recorded by the PSD are assigned a "position" along the detector by the height of its output pulse. This is accumulated in a multichannel analyzer (MCA) and then read out in bins, each corresponding to a section of $\Delta q_\perp$. The total counts in each bin are saved in the output file, to allow the data to be subsequently broken down into different q_z positions with a correspondingly narrower $\Delta q_\perp$. Since surface x-ray diffraction in the grazing incidence geometry should give features that are broad in q_z, all the bins should rise and fall together in intensity and give independent structure factors at different q_z values. However, if a powder grain or multiple scattering glitch happens to accidently satisfy the diffraction condition during some surface measurement, it will be recorded in only one MCA bin and can be suppressed.

Surface Crystallographic Measurements

A subset of the general class of surface x-ray diffraction measurements is that pertaining to atomic structure determination alone. Here it is assumed that the surface under investigation is well enough ordered that the widths of all diffraction features are limited by the resolution function alone. This means that, in this case, any disorder that happens to be present is undetectable, and the surface is indistinguishable from ideal. It has been shown by Robinson (1990b) that in this limit the integrated intensity of the diffraction features is independent of the (unseen) disorder and so is representative of the crystallographic structure factor alone. Measurements of structure factors can then be analyzed in terms of the atomic structure of the surface, in a close analogy to bulk x-ray crystallography described by Woolfson (1997) or Lipson and Cochran (1966).

If the disorder is visible in the form of peak broadening that is not too severe, it may still be possible to reach the ideal limit by deliberately worsening the resolution function. While it is often hard to control $\Delta\alpha_M$ since this is a function of the beamline optics, $\Delta\alpha_A$ can be opened up significantly by opening the 2θ detector slits. In all cases, the test for this desirable situation is that all diffraction features are resolution limited in their width. There are other limits in which the structure factor can be reliably measured if the peaks are significantly broader than the resolution function, then their peak intensity is representative of the square of the structure factor, provided a different, but calculable, correction is used.

The measurements needed for crystallography are most easily obtained by rocking scans of the principal diffractometer angle θ. In the detailed protocols below, the name

of the θ angle known to the usual computer software is TH; these names will be cross-referenced where appropriate. This ensures that the narrow direction of the resolution function sweeps roughly perpendicular to the peak. If the peak is slightly misaligned, such a scan will usually catch it either earlier or later and still give the correct structure factor; this will not work if the misalignment is too severe, say more than two peak widths away, because, in general, there would be components of the misalignment in all possible directions. The use of a rocking scan is also prescribed by the definition of integrated intensity, the quantity coupled to the structure factor, as explained by Robinson (1990b) (see Appendix B). This corrects for all kinds of disorder, including mosaic spread, provided this is not too severe. It is also the most reliable way to estimate the background underlying the peak.

Grazing Incidence

It was explained (see Diffraction from Surfaces, above) that the use of grazing incidence orients the resolution function in a favorable direction for surface x-ray diffraction. This means that the incident beam, the exit beam, or both make a very small angle with the surface to be measured. This presents a special challenge in the diffractometry which is discussed below (see Practical Aspects of the Method). It also has other consequences.

The diffraction under grazing incidence conditions is no longer strictly kinematical because such a beam will partly reflect from the surface and also undergo a slight refraction as it crosses to the interior of the crystal. The effect, described by classical optics by Born and Wolf (1975), is small until the incidence angle α_i (or exit angle α_f) becomes comparable with the critical angle for total external reflection, α_C, which is of order a few tenths of a degree for most situations. When $\alpha_i < \alpha_C$, the beam is completely reflected from the sample and only an evanescent wave continues inside. This can be used to considerable advantage when extreme surface sensitivity is desired, since the typical penetration depth is of the order of 100 Å. Fortunately, Vineyard (1982) introduced a sound theoretical basis called the distorted-wave Born approximation for working with the effects of the critical angle. This approximation provides the grazing incidence method with great power.

The distorted-wave Born approximation has an additional consequence for surface x-ray diffraction: the intensity of the diffraction becomes modified under grazing incidence conditions. Born and Wolf (1975) show that the effect can be reduced to a simple transmission function for the intensity, $|T(\alpha)|^2$, as a function of either α_i or α_f, which accounts for the refraction effects. An example of calculations of $|T(\alpha)|^2$ is given in the upper panel of Figure 5. As can be seen, there is a potential factor of 4 enhancement for both the incident and exit beams. The practical drawbacks with the factor of 16 are that it is only attainable when the collimation of both beams is severely restricted and the alignment is very difficult to achieve. The lower panel of Figure 5 shows the effect of the distortion on the α_f profile of an in-plane bulk Bragg peak, which in the absence of the refraction effect would lie at $\alpha_f = 0$. Careful examination of the experimental curve can be used to identify whether the crystal diffracts from all layers up to its surface or whether there are inactive "dead" layers present, as shown.

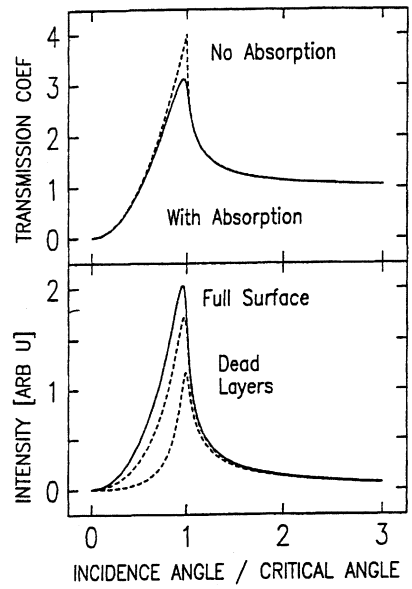


Figure 5. Calculations of $|T(\alpha)|^2$ and the profile of an in-plane diffraction peak for gold at $\lambda = 1.5$ Å. The dashed curves show the effect of 5 Å and 20 Å of inactive dead layers of gold present above the crystalline part of the sample. Adapted from Dosch et al. (1991).

PRACTICAL ASPECTS OF THE METHOD

An overview of the generic vacuum diffractometer configuration is given in Figure 6. This instrument, which operates at beamline X16A at NSLS, is the most widely used design. All of the instrumental functions described in this section are generic to all installations of surface x-ray diffraction, although some specific details of the X16A instrument are given for clarity of illustration. Further machine-dependent details can be found below in the detailed protocols in the Appendices.

The overall size of the instrument in Figure 6 is ~2 m long by 1 m wide and it resides at the beamline endstation. The instrument slides on air bearings, which allows it to be slid quickly into and out of the beamline hutch at the beginning and end of each experimental run. The configuration consists of an ultrahigh-vacuum (UHV) chamber (left side) coupled to a precision diffractometer. These two components will be described separately.

Vacuum System

On the left side, looking towards the x-ray source, lies the main vacuum chamber, which is used for sample preparation. It has turbomolecular (TMP), ion, and sublimation pumps (See GENERAL VACCUM TECHNIQUES) located in a cross below the level of the main chamber. The TMP and ion pumps have gate valves that allow them to be isolated, for protection against power failure for the TMP and to

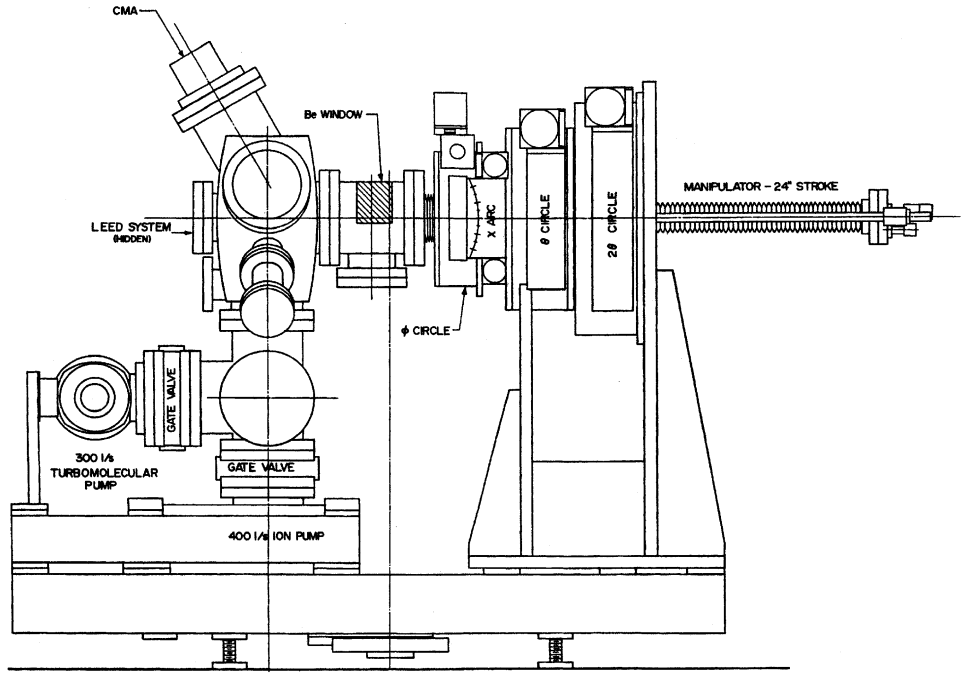


Figure 6. Overview of the vacuum diffractometer configuration used at the X16A facility at NSLS by Fuoss and Robinson (1984). Other installations vary slightly from this design.

reduce exposure to certain gases (e.g., Ar for sputtering and O_2 for dosing) for the ion pump. Numerous flanges are pointing at the sample, allowing the installation of various instruments for sample preparation and diagnosis. Samples are prepared by combinations of sputtering using an ion-bombardment gun, annealing by use of filaments, dosing with gases controlled by leak valves, and evaporation of substances using dosers. Dosers include commercial Knudsen cells and electron-beam evaporators, heated crucibles of molten materials, simple filaments coated with metal films, and getter-type alkali sources. Diagnosis is by means of LEED, Auger electron spectroscopy (AES; see AUGER ELECTRON SPECTROSCOPY), and residual gas analysis (RGA). Most conventional surface science preparations can be carried out using combinations of these tools.

Most importantly, the capability of performing an x-ray diffraction experiment in situ is provided in the chamber. X rays can enter and leave the chamber through a semicylindrical beryllium window, shown in the center of Figure 6. The window was brazed onto the vacuum wall at high temperature, so it is completely UHV compatible. It has 0.5 mm wall thickness so it transmits at least 95% of x rays with energies above 8 keV. In the diffraction position, the sample lies at the extreme right-hand side of the window with its flat face, upon which the surface has been prepared, facing left. The sample is rigidly connected to the diffractometer to the right. A manipulator wand located behind the diffractometer allows the sample to be detached from its rigid diffractometer mount and carried under UHV to the various instruments in the preparation chamber. There is a bellows and rotating seal coupling between the diffractometer and the chamber, which allows the passage of the diffractometer motions. These last essential features were specifically designed for the surface x-ray diffractometer and so will be discussed next.

Because preparation of 10^{-10} Torr pressure is time consuming, a loadlock is used to allow simple manipulations of samples, thermocouples, and heaters to be carried out without venting the main UHV chamber. The basic principle of operation is shown in Figure 7. For example, load locking is very useful for batch processing a series of samples or for introducing samples that cannot be baked at 150°C. Most instruments for surface x-ray diffraction provide some sort of load lock for rapid sample turnaround. Specific details for the operation of the X16A vacuum system are given later in the detailed protocol (see Appendix A), including the load-lock procedure, the venting procedure, and the bakeout procedure.

Sample Manipulator

Once it is retracted, the manipulator moves entirely with the sample on the diffractometer. A spring-loaded kinematical mount comprising three balls resting in 3 V-grooves 120° apart ensures that any flexure of the long arm of the manipulator is decoupled from the sample, which is thereby allowed to rest against the diffractometer ϕ axis. To carry the sample into the preparation chamber, the manipulator is extended by collapsing the 0.8-m-long bellows using a motor and gear on a threaded shaft. Electrical connections to the sample traverse the full length of the manipulator from its accessible rear flange. The last 75 mm of the shaft is illustrated in Figure 7. It is crammed with intricate features that give it great versatility: a 90° elbowlike flexure driven by a push-rod running down the length of the manipulator that allows the sample to face the LEED and AES; a spring-loaded decoupling joint associated with the kinematic mount (not shown); a vacuum sealing flange penetrated by electrical and thermal feedthroughs for the load-lock system described below; and at the end 25 mm of clear space that allows the user to assem-

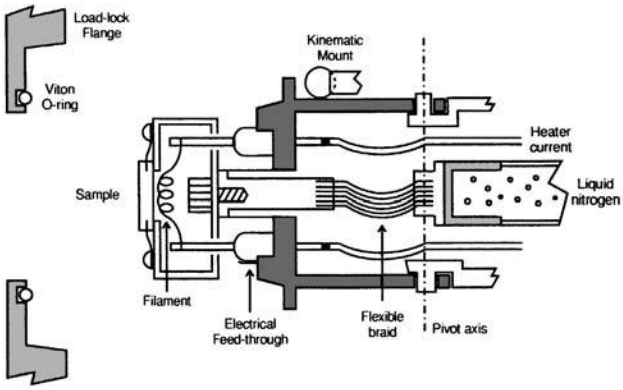


Figure 7. Detail of the manipulator head shown in cross-section. The pivot allows the sample to point at 90° to the main axis to be presented to surface characterization instruments. The viton seal of the load lock, through which the sample is passed, is shown on the left. The mating surface is halfway down the manipulator head and contains vacuum feedthroughs for cooling and electrical connections to the sample. All parts to the left of this are accessible to the user during the load-lock procedure.

ble his or her sample with the appropriate heaters, coolers, and sensors.

The rotating seal design is shown in Figure 8. It allows the passage of rotations across the vacuum wall using a rigid shaft that has neither play nor backlash. It achieves this with three concentric sliding seals made of Teflon between which differential vacua are applied, the first stage at 10^{-3} Torr with a roughing pump, the second at 10^{-6} torr with a TMP pump. If any seal momentarily leaks during rotation, the pressure burst on the UHV side is minimized by the use of these differentials.

Five-Circle Diffractometer

The diffractometer geometry is illustrated with the aid of Figure 9. This corresponds to a "five-circle" geometry, which was invented for surface diffraction by Vlieg et al. (1987). The standard names given to the angle variables describing the diffractometer setting are the same as the names of the axes in the figure. The principal axes are called (confusingly) θ and 2θ due to their association with the two-circle diffractometer setting describing Bragg's law. The symbols ϕ and χ are Euler angles of the sample with respect to the main horizontal (θ) axis. Finally α is the rotation of the entire instrument about the vertical axis. Historically at X16A, α was an afterthought, added to the instrument when it was found that the original four axes were too restrictive in their coverage of reciprocal space in the out-of-plane direction. Later instrument designs have incorporated α axes as a matter of course.

The convention for the zero settings of the axes is defined as follows: $\alpha = 0$ when the incident beam is perpendicular to the main horizontal axis. The remaining definitions all correspond to $\alpha = 0$: $2\theta = 0$ (dial = 270 for the X16A instrument) when the incident beam strikes the center of the detector; $\theta = 0$ (dial = 270) when the χ axis lies exactly along the incident beam direction; $\chi = 0$ when θ and ϕ are collinear; and $\phi = 0$ is arbitrary and usually chosen to be the same as the mechanical dial setting. The names used by the usual computer program for these angular settings are TTH, TH, PHI, CHI, and ALP, as discussed below (see Appendix C).

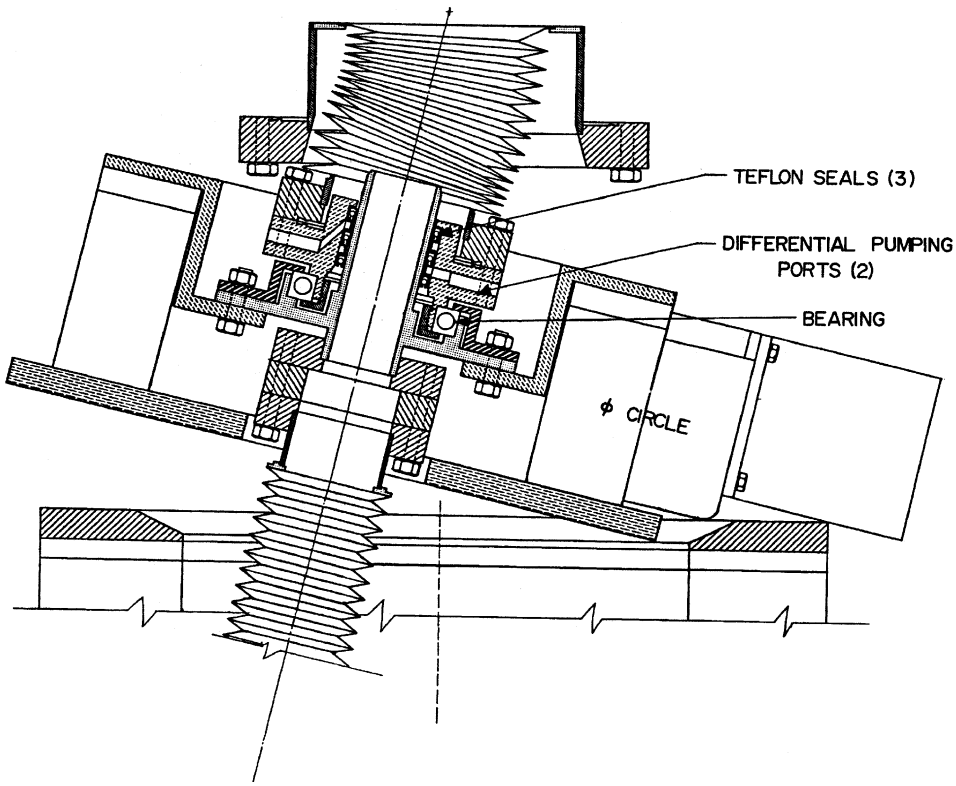


Figure 8. Cross-sectional view of the bellows and rotating vacuum seal that couple the rigid motions of the diffractometer to the sample in UHV. From Fuoss and Robinson (1984).

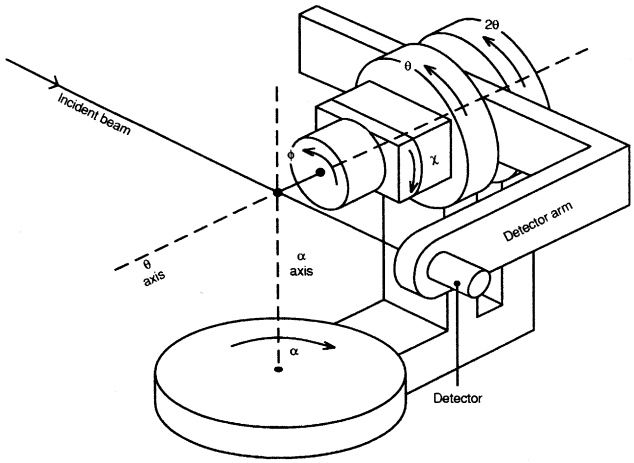


Figure 9. Isometric view of the five-circle diffractometer. The names of the axes and their positive directions of turning are indicated.

The additional degrees of freedom over the four circles of a nonspecialized diffractometer are needed for two reasons: (1) the additional constraints on the diffractometer setting due to grazing incidence or exit conditions and (2) the limited range of the angles due to potential collisions with the vacuum chamber.

The χ axis is particularly affected by the second constraint since it is limited to $\pm 0^\circ$ at X16A by the coupling bellows and also by the manipulator. There is also a "six-circle" geometry due to Abernathy (1993) with an extra angular out-of-plane degree of freedom for the detector, called GAMMA, a "Z-axis" geometry due to Bloch (1985) that is a subset of six-circle geometry without the ϕ and χ motions, and also a "2+2" geometry used for surface diffraction by Evans-Lutterodt and Tang (1995).

Laser Alignment

Many of the typical measurement methods of surface x-ray diffraction employ the grazing angle geometry (see Principles of the Method) on the incident beam, the exit beam, or both. It is therefore crucial that the physical orientation of the sample's surface be well known. To allow for the general case (rather commonly occurring) that the surface orientation is not a simple crystallographic orientation, perhaps due to miscut, it is important that the optical surface orientation be defined independently of the crystallographic orientation. This step is known as "laser alignment."

A laser beam is directed roughly down the main (θ) axis of the instrument and reflected from the polished sample surface and allowed to strike a screen located behind the laser. If the θ axis is rotated, the reflected spot will precess in a circular orbit. The center of this circle is the correct laser reflection angle if the sample face were exactly perpendicular to the axis. The beam can then be directed onto this center by adjustment of the diffractometer ϕ and χ angles until the laser spot remains stationary upon rotating θ. These settings, called fphi and fchi in the detailed protocol below (see Appendix C), are then a unique representation of the sample's optical orientation and are entered into the diffractometer control program so that the incident and exit angles can be precisely controlled.

Sample Crystallographic Alignment

Sample alignment is normally achieved by finding the diffractometer settings for diffraction from a small number of bulk Bragg peaks. If the lattice parameters are known, a complete orientation matrix is specified by the positions of just two Bragg peaks using the method of Busing and Levy (1967). If the lattice parameters are not known, it is necessary to specify the positions of at least three noncoplanar Bragg peaks. The orientation and calculation of arbitrary settings will be made by the control program once this information has been furnished.

The usual procedure for locating the Bragg peaks is to manually vary the diffractometer angles to maximize the diffracted intensity. This is less straightforward with a five- or six-circle diffractometer than with the more standard four-circle instrument, simply because there are more degrees of freedom in which the user can get lost. A reflection has 3 degrees of freedom, so the five-circle diffractometer has two motions too many. In particular, the detector moves on two axes, one of which (α) also moves the sample: there is no longer a simple connection between Bragg's law and the detector motion. Here are some guidelines on the suggested method of finding peaks:

1. Set up a "dummy" orientation matrix (OM) with the best known values of the lattice parameters (see Appendix C). Two mutually perpendicular in-plane reflections can be specified as having the laser orientation angles. From here, the setting of a suitable bulk out-of-plane Bragg peak can be calculated. In-plane peaks are hard to find because both incident and exit beams must be exactly grazing; out-of-plane peaks have a much wider choice of possible settings. Set the diffractometer at that position.
2. The desired reflection can be found by rotating the sample about its surface normal using either ϕ or θ. The peaks are so strong with synchrotron radiation that the motors can be searched at top speed; the reflection will "flash" on the detector oscilloscope or rate meter. There is usually sufficient thermal diffuse scattering (TDS) for a peak to be found in this way, even when it is several degrees away from the line that is searched.
3. Ride up the TDS until the center of the Bragg peak is found. Diffracted beam attenuators will be needed at this stage. Centering can be quickly and automatically carried out with "line-up" (lup) scans of a single motion (see Appendix C), after which the program calculates the center of mass and sends the motion there.
4. When using a PSD with an entrance window that is wide along the q_z direction, the peak may appear at any position along the length (q_z). Since the diffractometer is aligned to the center point of the detector, the contents of the PSD in the MCA must be viewed

(function key F1 or F2) and the peak made to walk to the center. Alternatively, a region of interest should be defined about the center and only the counts in this region used for reflection centering.

5. Once a peak is centered, it is better to rotate the dummy matrix manually to the corresponding orientation. This way both defining reflections are rotated together. If merely one of the reflections is overwritten, the resulting two-peak orientation matrix may be very unrealistic and will make it hard to find the second peak. Alternatively, the second dummy reflection can be chosen to coincide with the surface normal direction.

Once two peaks are found, the orientation matrix is complete and can be used to find any further points in reciprocal space. There will always be small errors of both crystal and diffractometer alignment, which may accumulate and result in poor prediction of further Bragg peaks. There are two ways of avoiding this problem:

a. Use alignment peaks that are most relevant to the features to be measured. For example, a good alignment for measuring a CTR would be to choose two Bragg peaks through which the rod passes.
b. Use multiple-peak alignment. Diffractometer alignment errors can be partially compensated by using the crystallographic unit cell parameters as free variables. An excellent orientation matrix will always be obtained by using a least-squares refinement of the best lattice passing through many Bragg peaks, especially if they are at large diffraction angles.

DATA ANALYSIS AND INITIAL INTERPRETATION

The following summarizes the different ways to present surface x-ray diffraction data to express one or other sensitivities or strengths of the technique. Different material properties of a given surface or interface would require entirely different kinds of measurement to be made. Each subsection lists a different form of presentation of the data that the reader might encounter in the published literature, then describes the typical measurement involved, and finally cross-references the typical diffractometer scans (Appendix C) and typical data analysis steps (Appendix B) that might have been followed. At the end of the section, a worked example is given.

Rod Profiles

Crystallographic measurements referred to above (see Principles of the Method) are used for surface and interface structure determination. The same procedures apply to CTRs or superstructure reflections. The general form of the data is to convey the magnitude of the structure factor along a line of points in reciprocal space, running along a line perpendicular to the surface. Theta-rocking scans (see Appendix C) are taken at a range of q_z values or Miller index L. These are numerically integrated and background subtracted, e.g., using the program PEAK (see Appendix B). For reliability, it is essential to measure multiple reflections related by the surface symmetry. When these symmetry equivalents are merged together, e.g., using the program AVE (Appendix B), their reproducibility is thereby determined and then used to specify the experimental error in the structure factor magnitude. Further details, including the equations defining the error estimate are given by Robinson (1990b).

The data are usually plotted as smooth curves against the perpendicular momentum transfer, L. Since the structure factor is a continuous function, enough L points must be measured to be sure that the profiles are smoothly varying and that all oscillatory features are resolved. This last caution is particularly apt in the case of thin films, which have many closely spaced thickness fringes; in this case, some researchers choose to simply scan along the q_z ridge of intensity in a "direct rod scan" and then to offset the θ angle and repeat the scan to estimate the background.

The rod profiles are then fit to the functional form of the structure factor of an atomic model of the surface, as outlined above (see Principles of the Method). In the kinematical limit, both for CTRs and superstructure rods, this is given by a complex linear sum over atoms representing the superposition of atomic form factors. Closed-form expressions are needed for the form of the CTRs to account for the infinite sum. The surface roughness must also be accounted for at this stage. The program ROD, written by E. Vlieg and described in Coppens et al. (1992) (see Appendix B), is a widely used computer program for this purpose. Atomic models of bulk and surface regions are specified in parametric form in a very general format. The parameters of the model are then refined by a least-squares minimization to give the atomic coordinates on the model.

It should be noted that surfaces are very different from bulk matter in the specification of atomic models. Each layer of the structure is expected to have a different Debye-Waller factor representing its atomic vibration amplitude. These Debye-Waller factors usually increase significantly in magnitude in passing from the bulk outward in the structure. Another common feature of surfaces is partial occupancy of lattice sites. It is also common to find more than one site, each partially occupied, for a structure with disorder. All of these situations are handled by the ROD program (see Appendix B). Finally, surface atoms are prime candidates for anharmonicity, since they have neighbors on one side but not the other. Surface anharmonicity has been analyzed in the pioneering work of Meyerheim et al. (1995).

Reflectivity

The specular reflectivity is the special case of the CTR emanating from the origin of reciprocal space. While it can be measured in the way described in the previous section, this can also be achieved with the simple two-circle method if the sample is mounted perpendicular to the θ axis. It is customary to plot reflectivity data as the square of the reflection coefficient versus total momentum transfer, q_z. This gives a normalization to the data different

from the crystallographic standard described above: an ideally terminated surface would have an intensity dropping as q_z^{-4} in these units, according to the Fresnel law of reflection. To emphasize the differences, the reflectivity divided by the (calculated) Fresnel reflectivity is often what is plotted. There is a correction for the total external reflection region at small angle, which can be understood with a simple dynamical treatment.

Als-Nielsen (1987) has derived a "master formula" for the analysis of reflectivity data in terms of the spatial derivative of the electron density profile of the surface. When q_z is small, the resolution is insufficient for atoms or atomic layers to be located, so the electron density is parameterized as a series of slabs of variable density connected by error function shaped transition regions of variable width. These parameters are then refined by least squares to fit the data and obtain the electron density profile explicitly.

Reflectivity is the method of choice for measuring the sharpness of the layer profiles of artificial multilayer structures that are chemically modulated by varying their composition during deposition. Thick multilayers of electron-dense materials require a full dynamical treatment for correct analysis, in which transmission and reflection coefficients of each layer interact self-consistently. The theory was worked out by Parratt (1954) and has been implemented by Wormington et al. (1992). It is also commercially available as a software product called REFS from Bede Scientific.

Surface Diffuse Scattering

Surface roughness can be measured quantitatively using the CTR profiles, as mentioned above (see Principles of the Method). It is also measured routinely by the reflectivity method above. In either case, the intensity that is missing from the CTR or the reflectivity, resulting in a lower signal, reappears as diffuse scattering nearby. This was clearly explained in terms of the lateral correlations associated with the roughness in an important paper by Sinha et al. (1988).

The surface diffuse scattering can be measured most easily near the origin of reciprocal space by means of wide θ scans (see Appendix C). Such scans have several characteristic features that have simple interpretations: in the center there appears the CTR or the specular reflectivity as a sharp spike; at the two ends of the scan lie the "Yoneda wings," where the diffuse intensity rises by the $|T(\alpha)|^2$ transmission function as $\alpha_i \approx \alpha_C$ or $\alpha_f \approx \alpha_C$; in between lies the surface diffuse scattering. Multilayer samples are found to have rich surface diffuse scattering patterns that provide information about the correlations between the positions of the various layers making up the multilayer.

Reciprocal Lattice Mapping

In some surfaces undergoing phase transitions or morphological changes, there is valuable information in just the diffraction peak positions, widths, and intensities. For example, surface diffraction is extremely sensitive to the appearance of new periodicities in surfaces, and this has been exploited extensively. The measurement consists of recording wide scans, e.g., using the *iscan* and *jscan* described below (see Appendix C), along simple directions of reciprocal space. Such curves would be plotted as a function of temperature, time, or other control variables. Sometimes two-dimensional scans (*kscan*) are also made to map out all the identifying features. A crystal analyzer, which provides much higher **q**-resolution, may be needed to separate finely spaced features.

The materials properties of interest that can be derived from such measurements concern grain shapes and sizes and the degree to which these are strained. The positions of Bragg peaks in any diffraction experiment determine the lattice parameters of the sample crystal; e.g., high-resolution x-ray diffraction is probably the most accurate of all and can achieve 1 part in 10^5 using the Bond method. Such measurements provide information about strain and its distribution within the sample. Similarly, accurate measurements of diffraction line widths, through the Scherrer formula, determine the size of the diffracting grain. In both these situations it is routine to use a careful lineshape anaylsis to extract the most reliable sample parameters. The semi-automatic program ANA described below (see Appendix B) provides a battery of possible lineshapes that apply to different kinds of diffracting objects or distributions of objects.

Grazing Incidence Measurements

Grazing incidence measurements fall into two categories. The first is a simple extension to all of the above techniques to determine materials properties as a function of depth inside a sample. Grazing incidence of exit conditions is then employed to vary the penetration depth into the bulk, which can be calculated reliably from the known refractive index and absorption values of the material at the x-ray energy used. The distribution of depths sampled is always exponential, so it is never possible to select a particular depth of interest; for this purpose it would be necessary to use a nonlinear technique such as ion scattering (see Chapter 12). Dosch (1992) has suggested the use of Laplace transforms to deconvolve the depth information.

Since the incident and exit angles always have to be under control in a surface x-ray diffraction experiment, it is a routine extension of the normal procedure to deliberately control them. For example, with the diffractometer control program "super," the penetration depth can be chosen by setting the target value of the incident or exit angle, α_i or α_f (B or B2 in Appendix C). A useful extension is to make a *vscan* at a fixed reciprocal lattice position as a function of α_i or α_f.

The second category of grazing incidence measurement is to use the PSD to record an entire α_f profile in a single measurement. This method was pioneered for the study of surface-phase transitions in a depth-resolved way by Dosch (1987, 1992). It has the advantage that diffractometer motions are not required during measurement of small changes with temperature. More recently the method has been improved by Salditt et al. (1996) to record and analyze diffuse scattering from multilayer samples as a function of exit angle α_f in a PSD by making wide θ scans at fixed α_i.

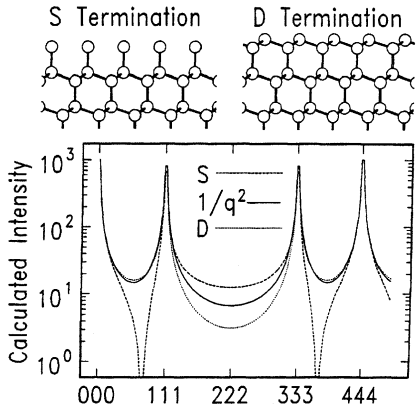


Figure 10. Top: the possible S and D terminations of the Si(111) surface. Bottom: calculated specular reflectivity for the two terminations according to the CTR theory and by superposition of $1/q^2$ intensity distributions.

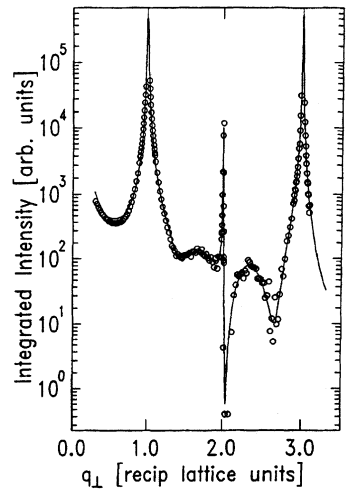


Figure 11. Measured x-ray reflectivity of a UHV-prepared Si(111) sample using a PSD. The solid line is the least-squares fit to a model based on the DAS model for the 7 × 7 reconstructed surface due to Takayanagi et al. (1985). From Robinson et al. (1994).

Worked Example: the Si(111) Surface

When the bulk crystal has a sufficiently simple structure, in the case of all primitive crystal structures, or the NaCl structure, there is only one possible way to cut the lattice to reveal a surface. However, in this case, the CTR presented above (see Principles of the Method) is technically indistinguishable from a simple superposition of $1/q^2$ intensity tails around each Bragg peak, as discussed by Robinson (1990a). Whenever the crystal has a multiple-atom basis that allows for more than one possible termination, the shape of the CTRs depends on which termination exists and is then clearly distinguishable from the superposition of $1/q^2$ intensity tails.

Such is the case of Si(111), where two possible terminations can exist, as illustrated in Figure 10. The double-layer (D) termination occurs when bonds perpendicular to the surface are cut; the single-layer (S) termination occurs when we break the bonds inclined at 19° to the surface, as shown. The density of dangling bonds is three times greater for the latter case, so it might be considered less likely to occur. The calculated CTRs are plotted for the two cases in the lower half of Figure 10. Up to a factor-of-10 difference in the two structure factors is seen to occur.

In reciprocal space, the specular reflectivity line is the direction with the momentum transfer, $\mathbf{q}$, entirely perpendicular to the surface. This corresponds exactly to the zeroth-order CTR diffraction and so can be thought of as a CTR; the analysis of these data is particularly simple because only the vertical components of the atomic coordinates enter, by virtue of the section-projection theorem of Fourier transforms. The question, then, is how is the real Si(111) surface, prepared in UHV, terminated? This is complicated by the fact that it is reconstructed with 7×7 translational symmetry.

Example of Measurements from Si(111)7 × 7

Figure 11 shows the measurements of the specular CTR of Si(111)7×7 using an improved data collection method. An x-ray PSD was placed so that it cuts across the diffracted beam. The background is then measured simultaneously with the integrated intensity of the diffracted beam and can be immediately subtracted. It is clear from the data that neither the S nor D simple termination adequately describes the surface. The figure also shows the result of a least-squares fit to the data, which passes through the points respectably well. The model used was the dimer-adatom-stacking fault (DAS) model for the Si(111)7 × 7 reconstructed surface due to Takayanagi et al. (1985). The heights of the atoms and the vibration amplitudes in the top four layers were adjusted for a good fit using the program ROD described in Appendix B.

Figure 11 also shows an additional sharp feature in the center of the scan at the position of the specular 222 Bragg reflection. This is indeed the specular 222 Bragg reflection, which is forbidden in diffraction from the ideal diamond lattice but is not perfectly extinguished for two reasons suggested by Keating et al. (1971):

1. The vibrations of the ions in the lattice are not spherically symmetric but tend to bulge out along the tetrahedral directions (four of the eight {111} directions) opposite to each bond; the anisotropy is along different directions for the two atoms in the basis, meaning they are crystallographically inequivalent.
2. The electrons forming the sp^3-hybridized bonds between the Si atoms are not spherically symmetric either but have a tetrahedral shape, again meaning the two atoms are crystallographically inequivalent. This affects 4 of the 14 electrons in the Si form factor so it is a relatively large effect.

It was shown many years ago by Keating et al. (1971) that contribution 1, which is temperature dependent, is dominant in neutron diffraction, while contribution 2 is more important in x-ray diffraction experiments. We therefore included extra bond charges in our model to fit this 222 peak in our data. It is orders of magnitude weaker

than the bulk diffraction but still stronger than the surface component of the CTR. A particularly noteworthy feature is the interference between the tails of the 222 and the CTR.

PROBLEMS

This section is a critical appraisal of the weaknesses of the surface x-ray diffraction technique. It might be used in evaluating published work, e.g., in searching for alternative explanations of results obtained with the method.

Surface Crystallographic Refinement

Crystallography has an inherent "phase problem." It is the structure factor amplitudes that are measured, not the phases. The solution of an atomic structure amounts to testing hypothetical models for agreement with the measurements. It is assumed that a good fit will only occur for one model, which therefore finds unique values for the missing phase information.

The difficulty with this logic is that it depends upon the quality and quantity of the data. For good, numerous data, experience has shown that there is usually a unique solution. In fact, many researchers are frustrated by the situation that they cannot find a model that agrees with their data! This is a heuristic support for the notion of uniqueness. "Direct methods" of structure solution, which exist for small-molecule crystallography (Woolfson, 1997), are being developed for surfaces but are not yet ready for widespread use.

The problem arises when there are insufficient data for the solution to be determined. The accepted rule of thumb (Lipson and Cochran, 1966) is for the structure to be overdetermined by a factor of 3:1. This means that for every structural degree of freedom in the model, there must be three independent measurements. It is presumed that nonuniqueness of the solution would be extremely unlikely when three separate measurements must all agree with each other for every parameter determined. For this to work, the counting must be done honestly: we will consider first the question of counting model freedoms and then that of counting the data.

It is possible to tie structural parameters together to reduce the degrees of freedom, but this always requires assumptions or constraints on the possible structures that might exist. Chemical constraints such as bond lengths and angles are usually safe, but sometimes unorthodox chemistry can take place at surfaces. This is often done through a Keating model (Keating, 1966) where a penalty for nonstandard bond lengths and angles is added to the χ^2 functional used for least-squares minimization. Another common constraint is to collect the less-sensitive parameters, such as Debye-Waller factors, into groups applying to several related atoms. Yet another constraint is to enforce exponential decay of relaxation parameters from the surface toward the bulk; this amounts to assuming the form of the solution of the Poisson equation describing continuum elasticity in the material. In each case, if the hypothesis leading to the constraint is questionable, then the conclusions drawn are questionable.

A similar logic concerns counting the data. This relates also to their accuracy, which can be measured in the way described above (see Practical Aspects of the Method). The situation is more complicated for surface than for bulk crystallography because the data are in the form of "rod profiles," which are continuous functions of L. The amount of information in a given rod profile can be counted, but this must be done carefully: every distinct maximum and minimum in the rod counts as one independent data point. Whether or not an extremum is distinct is closely tied to the size of the error bars of the measurement.

In principle, all error analysis should be handled through the definition of the normalized χ^2 functional, which is weighted by a factor $1/(N - P)$, where N and P are the number of data and number of parameters that we have just defined. The danger is that, while P is usually correct, N is frequently overestimated because it counts separately all the data points along each rod, rather than just the number of extrema. Nonspecialists and specialists alike quote a small goodness of fit χ^2 (near 1.0) value as an endorsement of the correctness of a structure. This is nonsense! A small χ^2 simply means that the data have been exhausted and the model is sufficiently detailed; it does not necessarily mean it is right. An easy way to obtain a small χ^2 is to measure very few data!

In examining published work, there are certain indications that the refinement of the structure had been difficult. Debye-Waller factors are sometimes omitted because they refine to negative values when included. Negative Debye-Waller factors are unphysical and so usually indicate some other deficiency in the model. Another bad sign in a structural model is when many atoms with partial occupancy have been included. Partial occupancy is common on surfaces, because of their nature: when a crystal is cut, it must have a boundary, but not necessarily one with an ordered structure.

Other Surface Crystallography Problems

Regarding the question of the quality of structure factor data, some aspects are specific to the case of surfaces. Apart from gross misalignments, there are subtle ways that additional errors can be introduced. When grazing incidence is used, the slightest variation of α_i from one measurement to the next causes a different value of the $|T(\alpha)|^2$ prefactor and hence the observed structure factor. In an extreme case, misalignment can cause α_i to become negative and the beam to disappear below the horizon. Sample curvature can aggravate the problem, leading to a different distribution of α_I values along different incidence azimuths. The only safe way to avoid the problem is to measure symmetry-equivalent reflections and record the data reproducibility.

When CTR data are included, there are several dangers associated with measuring very close to bulk Bragg peaks. Highly structured background is commonly found and must be carefully subtracted. The problems are compounded when the out-of-plane detector resolution is opened up to improve counting rates. There can be a significant distortion of the data from the L variation of the intensity across the detector slit. The safest strategy is to

avoid including data near the Bragg peaks, except where absolutely necessary.

Another problem can arise from unavoidable beam harmonics. The detectors are not always capable of discriminating $\lambda/3$ (and $\lambda/4$, etc.) contributions completely. This means that any measurement at or near an integer fraction of a bulk Bragg peak can give a strong "glitch." For simplicity, it is common to avoid all measurements at small-integer fractions of L.

A different source of possible misconception concerns symmetry. There are many cases of published work where the surface structure has lower symmetry than that of the underlying bulk crystal. For example, all reconstructed surfaces have lower translational symmetry than the bulk, and this is accepted for chemical reasons. The general rule is that a surface structure should have as high a symmetry as possible, compatible with the symmetry of its parent bulk. Only when this is shown to be impossible should the symmetry be broken. Then one symmetry element at a time should be removed from the model, since each symmetry that is lost can greatly increase the number of degrees of freedom.

A common source of confusion is that the symmetry of a surface is not immediately obvious from the symmetry of the diffraction pattern. Domain formation is common on surfaces, where differently oriented regions nucleate in different locations. It is usual to average the intensity contributions from each of the orientational domains, rather than the structure factor. This amounts to the assumption that the domains are far apart, beyond the coherence length of the measurement. The opposite limit of microdomains within the coherence length is usually apparent from a broadening of the peak lineshapes. This can be handled by mixed structures with partial occupancy once the peaks have been integrated correctly to account for their enlarged widths.

Lineshape Analysis

The purpose of analyzing lineshapes is to extract parameters that relate to materials properties so that they can be tracked as the sample is varied in some way. It is usually the peak position and/or width that is of interest. Typically the fitting is done directly to the output data stream in a semi-automatic way, e.g., using the program ANA described in the detailed protocol in Appendix B. If the curve that is used for fitting has the wrong shape, systematic errors will be introduced that may or may not have consequences. For this reason, a battery of lineshapes is usually available, sometimes with several adjustable parameters. A good example of a generally useful lineshape is the Lorentzian function raised to a power:

$$p(x) = A[(x - B)^2 + C^2]^{-D} \qquad (2)$$

where A, B, C, and D are adjustable parameters (B and C are the position and width most often sought). The danger here is that of parameter coupling or nonorthogonality, particularly between parameters C and D. If D is just allowed to vary from one fit to the next, and then later ignored, unwarranted changes in C may occur that are not really present in the data. Here, the solution is to find a value of D that applies to most of the data, then fix the parameter for fitting the data series to look for the trends.

A different situation applies to the handling of background, which is seen as a nonvanishing intensity in the extension of the tails of a peak. The background is due to unwanted additional scattering sources within the sample or the instrument. Usually the background is subtracted by extrapolation. The background can have a profound effect on the stability of fitting lineshapes such as the Lorentzian function raised to a power above, especially if D is the parameter of interest. The best advice here is to measure a lot of background extending far away from the peak; this has a surprisingly beneficial effect in constraining the fits of functions with long tails such as this.

The best procedure for measuring diffuse scattering is again to control and understand the background. By its nature, there is no perfect way to separate diffuse scattering from background. The most difficult situation is that of small-angle scattering near the direction of the incident beam, which applies to the nonspecular reflectivity region for surfaces discussed above (see Data Analysis and Interpretation). Here, the potential sources of background extend all the way back up the beamline to the entrance slits and beyond. A common hazard in diffuse scattering measurements is the tails of the resolution function, which appear to be CTRs pointing in apparently unusual directions. In fact, they are the CTRs of the crystals in the monochromator or analyzer rather than those of the sample!

DETAILED PROTOCOLS

The exact procedure will vary slightly from installation to installation, but the general principles will be the same. To be specific, we discuss here the procedure for the X16A surface diffractometer at NSLS. The generalization to other installations will usually be obvious. Details of the vacuum procedure are given in Appendix A, the data analysis programs in Appendix B, and the diffractometer control program in Appendix C.

Alignment of the Beamline

The X16A beamline, like many others used for diffraction, consists of a grazing incidence toroidal mirror and double-crystal monochromator. It is the responsibility of the NSLS facility staff to help the user get to the point of operating the beamline. This requires verification of all safety-related hardware on a "safety checklist" and that the condition of all cooling and vacuum systems be "green" before the safety shutter will open.

The mirror focusing is 1:1 and so should produce a focal spot on the sample ~1 mm wide by 0.5 mm high. The focal length is fixed and requires the correct incidence angle of 5.65 mrad, which fixes the final height of the focused beam. It is important to ensure that the height is correct by using a focusing screen near the sample position and a TV camera with a close-up lens. The beam image is a U-shaped

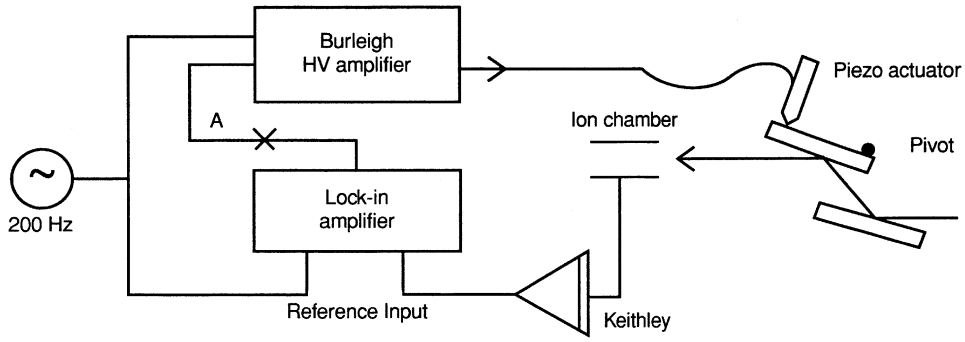


Figure 12. Monochromator feedback circuit. Point A is where the feedback loop should be broken in order to detune the monochromator.

"smile" when the incidence angle is too low and an inverted smile when the incidence angle is too high. The locus of focal images traces out an X shape on the screen upon tilting the mirror, with the optimum focus at the intersection of the X. A tilted smile means the mirror is not steered correctly and requires rotation about the vertical axis. The mirror bend should then be adjusted for the smallest vertical profile, and the beam position should be scanned to verify it is not being cut by any nearby aperture edges.

The double-crystal monochromator is controlled by five manual adjustments. The energy is set by a stepping motor on the back of the tank that is driven by a manual pulser. This can also be controlled by pulses from the beamline computer if needed. The energy reading is a dial counter, which is converted to kilo-electron-volts on a chart in the beamline log book. The θ and χ angles of each crystal are controlled by stepper motors driven by a shaft-encoder knob and selected by one of the four red "engage" buttons. The four angles are also read out by absolute encoders. It is very important not to move these in-vacuum motors far from their starting positions because there are no limit switches. After changing energy, it is necessary to bring the two crystals parallel again by adjustment of θ_2 until a signal is seen in the ion chamber. Side-to-side steering of the beam position is achieved by χ_1, which is generally used to bring the beam onto the sample (see below).

The monochromator feedback circuit is shown in Figure 12. The second crystal is modulated at 200 Hz by a piezoactuator. When its rocking curve is correctly centered on that of the first crystal, the lock-in amplifier reads zero output; if the angle is misset, a positive or negative error signal is generated that corrects the angle. The loop gain and integration time may have to be adjusted for correct performance. To detune the monochromator (e.g., to remove harmonics in the beam), the loop must be broken at point *A* and the crystal angle set manually.

Alignment of the Diffractometer

The intersection of the five axes of the diffractometer is at its center, and this must be positioned accurately on the beam's focal spot. This must be correct to within 100 μm. The beam direction must also be perpendicular to the θ and α axes to within 0.1° and 0.3°, respectively. Because a surface x-ray diffractometer has a sample in a UHV environment also at its center, this presents a special challenge.

The vertical position is set by four screw jacks under the diffractometer table. These are manually or stepping-motor driven in two pairs, front and back, to allow the table to be tilted until the beam is perpendicular to the α axis. The tilt is equal to the beam's known 11.3 mrad angle to the horizontal and preset using a spirit level and rulers. The horizontal position is set coarsely using surveying marks and finely by moving the beam with the χ_2 monochromator adjustment. The perpendicularity of the theta axis is determined by the choice of the zero of α.

The essential alignment tool is an "off-axis pinhole" that is an L-shaped bracket with a 1-mm pinhole drilled through it. This bolts onto the ϕ circle of the diffractometer so that it wraps around the outside of the UHV chamber and presents itself to the incoming or outgoing beam, as shown in Figure 13. The distance from the pinhole to the mounting surface is accurately machined so that it corresponds to the manufactured radius of the χ arcs. Because of its offset from the center, the position of the pinhole with respect to the beam is scanned by moving the ϕ and χ axes of the diffractometer.

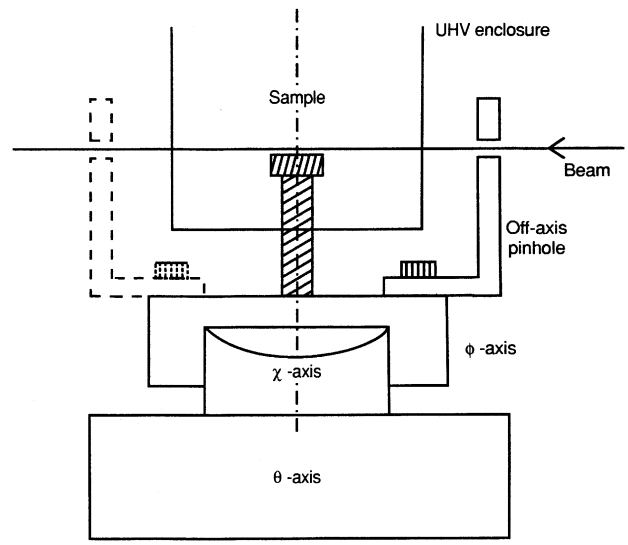


Figure 13. Off-axis pinhole designed to mark the center of the diffractometer without breaking the vacuum. The dashed position is reached by a 180° rotation of the θ axis.

Because the sample is blocking the true center position so that the x-ray beam cannot pass, the alignment must be carried out at a controlled location just above it. This is done by tilting the diffractometer χ by a known amount (usually 1°) from its calibrated zero and moving the beam with the χ_2 (monochromaor) to correspond. The center position of an arbitrarily placed starting beam is determined in (ϕ, χ) space with the pinhole on one side of the sample and then again on the other side. An exact 180° rotation of θ takes the pinhole from front to back, as shown in Figure 13. Half the difference between the two ϕ values multiplied by the lever arm L is the error in the table height, which is thereby corrected. Half the difference between the two χ values is the deviation of the beam from perpendicularity with the θ axis, which defines the zero of α.

Once these corrections have been made, the pinhole is exactly aligned with the diffractometer center, offset horizontally to avoid the sample. The beam, passing through the pinhole, can then be used to set precisely the vertical incident and exit beam collimation slits. If the pinhole is then removed and the beam moved until it grazes across the sample at the center, the horizontal slits can be set too. This is also the opportunity to set up the channel definitions of the PSD (if used), by marking off known distances from the calibrated center line.

The centering of the α axis with respect to the other four is also adjustable, owing to its historical origin as an auxiliary sample motion. The adjustment is by means of a motorized linear translation across the beam and slotted screw-holes in the direction parallel to the beam. Previous settings of these adjustments are recorded in the logbook, and these should be used as starting values. The alignment should not be needed often as the performance does not depend critically upon it; the usual symptom of misalignment is a loss of intensity at large values of perpendicular momentum transfer. It is carried out using the off-axis pinhole once this has been centered exactly on the beam following the procedure above. Then θ is rotated by exactly 90° so that the pinhole is pointing straight up, thereby marking a point exactly above the intersection point of the axes, as can be seen in Figure 9. The center of the α axis is then brought to intersect this marked point by using the translations. This is most easily performed with a high-magnification TV camera hanging down from the hutch roof: when α is rotated by hand, the pinhole position should be invariant.

Angle Calculations

The principal purpose of the diffractometer control program is to make the angle calculations that map the reciprocal crystal coordinates onto the angle setting of the diffractometer. This calculation depends on the mode of operation of the instrument, which is the "five-circle" mode for the installation discussed here. The program also contains the software drivers for the axis motors corresponding to the diffractometer angles so that these can be moved by the program. It also generates the most useful kinds of scans automatically and keeps track of the orientation matrix. Its other major task is to produce the output files that contain the results of the experiments organized in a readable format. Here we make reference to the diffractometer control program "super," but many of the basic principles are common to other programs, e.g., the commercial "spec" program used elsewhere and available from Certified Scientific Software. Details of the command structure of "super" are included in Appendix C.

Angle calculations in the five-circle mode are described in a paper by Vlieg et al. (1987). The diffractometer has 5 degrees of freedom in the most general five-circle mode (alm = 2) and 4 degrees of freedom when alm = 1 (see Appendix C, Basic Super Variables). Three of these are needed to select a point of measurement in (3D) reciprocal space; one more is used to constrain the incidence or exit angle (see bem in Appendix C, Basic Super Variables). The 5th degree of freedom is used to help control the resolution function, bearing in mind that the diffraction features to be measured are highly elongated into "rods" along the direction perpendicular to the surface: the program arranges for the crystallographic c^* axis to lie in a horizontal plane. When the diffraction rods (assumed parallel to c^*) are horizontal, the vertical high-resolution direction lies exactly across them.

The constraints on the five-circle diffractometer will allow inclination of the incident beam (ALP), but not the exit beam (GAMMA), from the sample plane. Because of the limits on CHI, the sample must lie within some degrees from perpendicular to the main TTH/TH axis. A consequence of these constraints is that most reciprocal space is accessible only in the bem = 2 mode (see Appendix C, Basic Super Variables) in which grazing exit conditions are maintained. This has the advantage that the PSD can then record the exit beam transmission profile for all reciprocal lattice settings. The exit angle (set by the variable B2) arises from setting TH near TTH + 90° and tilting CHI to a value near the negative of B2, as can be seen with the aid of Figure 9. The exact values depend on how close the "laser alignment" (see Practical Aspects of the Method) angles are to the ideal flat setting, fchi = 0. For this reason the θ axis should be expected to work roughly in the range $90° < TH < 150°$ and the limits should be set to accept this range, e.g., $-30° < TH < 150°$.

ACKNOWLEDGMENTS

It is a pleasure to thank the various people who have helped develop this technique and the X16A facility over the past 15 years. Paul Fuoss helped develop the beamline and original surface diffractometer hardware. Elias Vlieg invented the five-circle mode and wrote the data analysis programs. Robert Fleming wrote the diffractometer control program. Alastair MacDowell developed the manipulator design. During the design and construction stages, the following people worked very hard: Warren Waskiewicz, Laura Norton, Steve Davey, and Jason Stark. In later years the input and contributions of the following people have been invaluable: Ken Evans-Lutterodt, Rolf Schuster, Peter Eng, Peter Bennett, and Don Walko. Finally, the vision and foresight of the Bell-Labs

management to finance the whole project, at a time when it was unclear that it would work, deserves great commendation.

LITERATURE CITED

Abernathy, D. 1993. An X-ray Scattering Study of the Si(113) Surface: Structure and Phase Behavior. Ph.D. dissertation, Massachusetts Institute of Technology.

Afanas'ev, A. M. and Melkonyan, M. K. 1983. *Acta Crystallogr. A* 39:207.

Als-Nielsen, J. 1987. Solid and liquid surfaces studied by synchrotron X-ray diffraction. *In* Structure and Dynamics of Surfaces II (W. Schommers and P. von Blanckenhagen, eds.). pp. 181-222. Springer-Verlag, Berlin.

Andrews, S. R. and Cowley, R. A. 1985. *J. Phys. C* 18:6247.

Bloch, J. M. 1985. *J. Appl. Crystallogr.* 18:33.

Born, M. and Wolf, E. 1975. Principles of Optics. Pergamon Press, Elmsford, N.Y.

Busing, W. R. and Levy, H. R. 1967. *Acta Crystallogr.* 22:457.

Coppens, P., Cox, D., Vlieg, E., and Robinson, I. K. 1992. Synchrotron Radiation Crystallography. Academic Press, London.

Dosch, H. 1987. *Phys. Rev. B* 35:2137.

Dosch, H. 1992. Critical Phenomena at Surfaces and Interfaces: Evanescent X-ray and Neutron Scattering. Springer-Verlag, Heidelberg.

Dosch, H., Mailander, L., Reichert, H., Peisl, J., and Johnson, R.L. 1991. *Phys. Rev. B* 43:13172.

Eisenberger, P. and Marra, W. C. 1981. *Phys. Rev. Lett.* 46:1081.

Evans-Lutterodt, K. and Tang, M. T. 1995. Angle calculations for a "2 + 2" surface X-ray diffractometer. *J. Appl. Crystallogr.* 28:318–326.

Fuoss, P. H. and Robinson, I. K. 1984. Apparatus for X-ray diffraction in ultra-high vacuum. *Nucl. Instr. Methods A* 222:171.

Guinier, A. 1963. X-ray Diffraction. W.H. Freeman, New York.

James, R. W. 1950. The Optical Principles of the Diffraction of X-rays. G. Bell and Sons, London.

Keating, D., Nunes, A., Batterman, B., and Hastings, J. 1971. *Phys. Rev. B* 4:2472.

Keating, P. N. 1966. *Phys. Rev.* 145:637.

Lipson, H. and Cochran, W. 1966. The Determination of Crystal Structures. Cornell University Press, Ithaca, N.Y.

Meyerheim, H. L., Moritz, W., Schulz, H., Eng, P. J., and Robinson, I. K. 1995. Anharmonic thermal vibrations observed by surface X-ray diffraction for Cs/Cu(001). *Surf. Sci.* 333:1422–1429.

Moncton, D. E. and Brown G. S. 1983. *Nucl. Instr. Methods* 208:579.

Müller, B. and Henzler, M. 1995. SPA-RHEED—A novel method in reflection high-energy electron diffraction with extremely high angular and energy resolution. *Rev. Sci. Instrum.* 66:5232–5235.

Parratt, L. G. 1954. *Phys. Rev.* 95:359.

Robinson, I. K. 1986. Crystal truncation rods and surface roughness. *Phys. Rev. B* 33:3830.

Robinson, I. K. 1990a. *Faraday Discuss. R. Soc. Chem.* 89:208.

Robinson, I. K. 1990b. Surface crystallography. *In* Handbook on Synchrotron Radiation, Vol. III (D.E. Moncton and G.S. Brown, eds.). pp. 221–266. Elsevier/North-Holland, Amsterdam.

Robinson, I. K. and Tweet, D. J. 1992. Surface X-ray diffraction. *Rep. Prog. Phys.* 55:599–651.

Robinson, I. K., Eng, P. J., and Schuster, R. 1994. Origin of the surface sensitivity in surface X-ray diffraction. *Acta Phys. Pol. A* 86:513.

Salditt, T., Lott, D., Metzger, T. H., Peisl, J., Vignaud, G., Legrand, J. F., Grübel, G., Høghøi, P., and Schärpf, O. 1996. Characterization of interface roughness in W/Si multilayers by high resolution diffuse X-ray scattering. *Physica B* 221:13–17.

Sheithauer, U., Meyer, G., and Henzler, M. 1986. *Surf. Sci.* 178:441.

Sinha, S. K., Sirota, E. B., Garoff, S., and Stanley, H. B. 1988. X-ray and neutron scattering from rough surfaces. *Phys. Rev. B* 38:2297.

Takayanagi, K., Tanishiro, Y., Takahashi, S., and Takahashi, M. 1985. *Surf. Sci.* 164:367.

Vineyard, G. H.. 1982. *Phys. Rev. B* 26:4146.

Vlieg, E., van der Veen, J. F., Macdonald, J. E., and Miller, M. 1987. Angle calculations for a five-circle diffractometer used for surface X-ray diffraction. *J. Appl. Crystallogr.* 20:330–337.

Warren, B. E. 1969. X-ray Diffraction. Addison-Wesley, Reading, Mass.

Woolfson, M. M. 1997. An Introduction to X-ray Crystallography. Cambridge, Cambridge University Press.

Wormington, M., Bowen, D. K., and Tanner, B. K. 1992. Principles and performance of a PC-based program for simulation of grazing incidence X-ray reflectivity profiles. *Mater. Res. Soc. Symp. Proc.* 238:119–124.

KEY REFERENCES

Als-Nielsen, 1987. See above.

Describes the principls of x-ray reflectivity from an experimental and theoretical point of view.

Dosch, 1992. See above.

Covers the use of evanescent x-ray waves to probe the depth dependence of phase transitions near surfaces.

Robinson, 1990b. See above.

Describes surface structural analysis from a crystallographic perspective.

Sinha et al., 1988. See above.

Describes phenomenological models of surface roughness and derives their signatures in diffraction.

APPENDIX A: VACUUM PROCEDURE

In order that samples have a chance to remain clean for the duration of an experiment, surface science experiments are usually carried out at pressures of 10^{-10} Torr or below, called the ultrahigh-vacuum (UHV) regime. Under these conditions, according to the kinetic theory of gases, an atom on the surface will be impacted by a residual gas atom from the vacuum about every 30 to 60 min. The vacuum protocols described here apply to the X16A facility at NSLS and are rather specific in this section. At other facilities, the principles will be the same, but the details will be a little different. In the latter case, the information provided here will enable the reader to assess the level of complexity involved in carrying out experiments under UHV.

Load-Lock Procedure

Because preparation of 10^{-10} Torr pressure is time consuming (see below), a load-lock procedure has been developed that allows simple manipulation of samples, thermocouples, and heaters to be carried out without venting the main UHV chamber. The principle of operation is shown in Figure 7. Load locking is very useful for batch processing a series of samples, e.g., or for introducing samples that cannot stand to be baked at 150°C.

To load a sample, first ensure that the loading chamber is under vacuum. Its turbopump should have been at full speed for at least 10 min. Then open the gate valve that separates the loading chamber from the main UHV chamber. The pressure may rise into the 10^{-8} Torr range.

Set the diffractometer to the precalibrated angle settings recorded in the logbook or use the preassigned macro definition, called "load." The manipulator is now exactly in line with the sealing surface and can be extended fully until the motor will not move further. As always, when moving the manipulator, it is imperative to watch the progress so as to avoid unexpected collisions. The centering can be verified by observing the final sealing approach through the load-lock window.

Once sealed, the loading chamber is vented with air by (1) depowering the turbopump, (2) isolating the roughing pump (it is not necessary to turn it off), (3) waiting for the pump to slow down significantly (~2 min), and (4) slowly unscrewing its black venting screw while observing the pressure in the main chamber. If the pressure rises, immediately reverse the procedure and reseat the sample.

If the procedure goes well, the UHV conditions of the main chamber will be retained by the electrical and thermal feedthroughs in the design of the manipulator head. The loading chamber can then be unbolted to allow changing of or modification to the sample.

The reverse procedure is to bolt the loading chamber back over the finished sample, open the valve to the roughing pump, and start up the turbopump. Anywhere from 5 to 30 min after the pump reaches full speed, the manipulator can be retracted into the main chamber and the gate valve closed. The pressure may rise momentarily into the 10^{-7}-torr range before dropping to 10^{-10} again. The fresh sample and filaments will now need to be outgassed by heating.

Venting Procedure

When more serious vacuum interventions are required, the entire chamber will need to be vented. The procedure here is similar to that described above, except there are more precautions.

First, turn off all filaments inside the main chamber and disconnect their power cables to avoid accidental activation. The ion gauge is protected and so may be left on. Close the gate valve to the ion pump if this is not to be vented also; otherwise turn off the power here as well.

Manually override the protection circuit to the turbopump gate valve, so that it does not close when the pump is turned off. This is done by powering the solenoid directly. Then the main chamber is vented with nitrogen gas supplied to its vent valve as follows:

1. Depower the turbopump.
2. Isolate the roughing pump (it is not necessary to turn it off).
3. Wait for the pump to slow down significantly (~2 min).
4. Switch on its venting circuit on the turbo control panel.

The pressure has equilibrated with the atmosphere when the bellows become limp. Overpressurization is avoided by a check valve in the nitrogen feed.

Bakeout Procedure

While the main chamber is open, the following routine maintenance may need to be carried out:

1. Replace the sublimation pump filaments.
2. Check the body of the manipulator for shorts or frayed wiring.
3. Recharge evaporation sources.
4. Clean any evaporated deposits from the interior of the windows.

The vacuum restart procedure is the reverse of the venting procedure given above (see Venting Procedure). The next step is the bakeout, which requires considerable disassembly of the diffraction parts of the experiment in order that they do not obstruct the ovens or get unnecessarily hot themselves. The following settings are advised:

1. The variables TTH, TH, CHI, and ALP should be set to zero (270 on dials of TH and TTH). The motor power can now be turned off.
2. The PSD high voltage should be turned off before moving it for storage.
3. Flight path vacuum should be turned off at its Nupro valve. The entire diffractometer TTH arm can now be dismounted and stored.
4. The turbopump gate valve should be in the protect mode (i.e., connected to its controller).

It is necessary to heat all of the chamber walls to 150°C for ~12 hr. This is achieved with ~1000 W of power distributed as follows:

1. Snap together the toggles that hold together the oven walls, which encase the main chamber. Any gaps can be filled with aluminum foil.
2. Lay the heater tape along the length of the manipulator and encase it with its clamshell oven. Any excess tape can be wrapped around the exposed feedthroughs.
3. Place the heater tape around the bellows/seal (between the diffractometer and the chamber).
4. Apply the following power settings: 110 V to each of two oven heaters and one ion pump heater, 60 V to heater tape around chi-bellows/seal; set the manipulator to 110 V if the tape is small (100 W) or to 70 V if the tape is large (300 W).

To shut down the bakeout, it is adequate to follow these quick steps:

1. Turn off all power.
2. About 30 to 60 min later, partially open roof of oven and unwrap the manipulator.
3. Another 30 to 60 min later, finish dismounting the oven.

The expected performance of the pressure will be in the 10^{-7}-torr range at the start, rising to 10^{-6} when hot and eventually dropping to 10^{-7} at the end of the bake. After the heat is turned off, the pressure will fall steadily and reach 10^{-10} Torr in $\sim$12 h.

APPENDIX B: DATA ANALYSIS PROGRAMS

A suite of data analysis programs have been written to perform the various steps of data analysis. The plot program ANA is the most generally useful, while the full set of four programs are needed for a complete surface crystallographic analysis. They were all written by Elias Vlieg and are much more widely used than just at X16A, so their description here will be generally applicable. The programs are designed to work with the data file structure of "super" (Appendix C) but also accept input in other formats, including "spec."

The programs are interactive with the user. The commands to all programs are either keyed in at a prompt or redirected from a macro (*.mac) file. Each entered line is a continuous string of commands and arguments delimited by spaces. The commands are mostly English or Dutch verbs and may be highly abbreviated, usually to a single letter. This is made possible by the use of a hierarchical "menu" structure that defines the logical sequence of allowed commands. Each command is followed immediately by its arguments (separated by spaces) or by a <CR> to be prompted for the arguments. For example, the setup commands specific to plotting in ANA are unreachable at the main level but are reached by typing *p(lot)*. At any level the command *h(elp)* will provide a list of available commands and *l(ist)* will provide current values and status.

Upon exit from any of the programs (except AVE), an executable macro file is generated that records the current options that have been defined during the last session. This file is automatically executed when the program is restarted so that the previous configuration is restored. In the following descriptions the commands are defined by their full names with the minimum abbreviation indicated using the syntax *p(lot)*.

Plot program ANA

This program maintains five data buffers into which spectra can be read. It uses the same buffers for generated fitting curves, so these can be superimposed. A variety of read formats are supported.

The *s(et)* menu is for setting parameters and modes. The *r(ead)* submenu is used to define filenames and column numbers for input. The *f(it)* submenu sets the mode of fitting, e.g., the weighting scheme.

The *r(ead)* menu is for reading in a spectrum and is followed by a format type, a scan number or filename, and a destination buffer number. Example formats are: *s(uper), c(olumn), x(y),* and *(s)p(ec)*.

The *o(perate)* menu is for manipulating spectra. Mathematical operations are available through the *ad(d), su(m), mu(ltiply),* and *sh(ift)* commands. Another functions is *l(ump)* for combining adjacent groups of data points within a spectrum. The function *me(rge)* combines two spectra, accounting for monitor counting time and removing duplicate data points; *w(rap)* removes 360° cuts in data; and *d(elete)* removes selected data points while *c(ut)* removes data points above or below a certain specified value.

The *p(lot)* menu makes graphs on the screen that can accumulate multiple entries for later printing or conversion to other image formats. A wide selection of *c(urve), s(ymbol),* and *li(nestyle)* options are supported. Curves can be *p(lot)*ed fresh or else *o(verlay)*ed and *ax(es)* and *te(xt)* can be supplied and modified. Logarithmic and linear scales can be chosen with, e.g., *ylog* and *xlin*. The program uses a commercial package called "graphi-c" for its plotting functions, and all file conversions of graphical output (including printing) must be handled through its internal mechanisms: hitting the space bar (a few times) allows access to the print menu of "graphi-c", and the simplest printout is activated by *l* for large or *m* for medium size.

The *f(it)* menu performs least-squares fitting to a library of functions. The command is followed by a code for the fit function required, which is assembled from *g(aussian), l(orentzian), v(oigt), p(owerlaw),* or *k(ummer)* components using commas as delimiters. Finally, the input and output spectrum numbers must be given. For example, the command "f g,l,l 1 2" will fit the sum of a gaussian and two lorentzians to the data in spectrum 1 and put the fit curve in spectrum 2. A two-parameter linear background is always assumed in addition to the chosen functions. Once inside the fit menu, the options are to *g(uess)* starting values of the parameters, specify a *v(alue)* by hand, and then *f(ix)* or *l(oosen)* these values. Finally, *r(un)* executes the least-squares fit and lists the parameter values.

A useful shortcut is the *a(utoplot)* menu with only two commands, *d(ata)* and *f(it)*. This generates a generic format plot with useful text information superimposed, such as the fitted values of the parameters.

Integration Program PEAK

This is a batch program for converting rscans and jscans (see Appendix C, Important Scan Types) into structure factors. It is assumed that the scans cut directly across the desired diffraction features whose intensity is then related to the square of the structure factor; misalignment will, of course, lead to underestimation of the structure factor. It is also assumed that the scan reaches far enough that the background can be estimated by linear interpolation between its two ends. The "stretching" feature of "super"

(see Appendix C, Important Scan Types) is designed to facilitate this.

The program is also responsible for four standard corrections to the data:

1. area correction, due to the changes in the active area of the sample as the diffractometer angles change, introduced by Robinson (1990b);
2. Lorentz factor correction, to convert an angular scan into an integral in reciprocal space, as in Warren (1969);
3. polarization correction, as in James (1950); and
4. monitor and stepsize normalization to account for different counting times and scanning rates that may have been used for different parts of the data.

Under the *c(olumn)* menu, it is necessary that all the diffractometer angles be identified by their column sequence number in the scan file. These must not change from one scan to the next.

The remaining parameters are entered at the command line. The *le(ft)* and *ri(ght)* numbers of points are to be considered as background; all points in between are integrated numerically to a total intensity. The *s(cantype)* can be *r(ocking)* or *i(ndex)* and will affect the Lorentz factor; all other scans found in the batch input will be ignored. For the area correction, the incident beam is taken to have dimensions *wi* vertically by *hi* horizontally, the sample has width *ws*, and the vertical exit beam slit is *we*.

The *r(un)* command followed by starting and ending scan numbers will perform the integration. Two output files are generated. The *.inf file has a line of information about each scan found, such as the date, background, and peak levels. The *.pk file contains the scan number hkl, the structure factor, and the structure factor error as the first three columns.

Averaging Procedure AVE

This procedure uses the *.pk output of PEAK to locate and compare all the symmetry equivalents. It determines an overall estimate of the average systematic error, ε, assumed to be a constant fraction of each individual structure factor value. The value of ε is taken to be a quality factor (R value on structure factors) in assessing the data, as discussed in Robinson (1990b). The program then passes through the data a second time to generate a weighted average value (file *.wgt) of each inequivalent structure factor with an enlarged error bar that combines its statistical errors input with its overall systematic error estimated in the first pass.

The *.ave output listing is very useful for finding bad data. It lists together all reflections that are equivalent according to the specified symmetry. Any that fall out of line are flagged with warnings indicating they should be checked. They are identified by scan number in the listing. This is also a useful way of testing the symmetry of the data, if this is unknown from context, since the program can be run with different preassigned choices of assumed symmetry.

The *s(et)* menu has submenus for defining *c(olumns)*, *s(ymmetry)* from the list of all 17 possible plane groups from p1 to p6mm, and *a(veraging)* parameters and modes. Under *a(veraging)* is the *c(utoff)* parameter, which is the number of times that a structure factor must be larger than its statistical error for it to be used in the estimation of ε. The default value (number of standard deviations) is 2σ, and this should be increased to 10σ if the data are not strongly affected by counting statistics.

Fitting Procedure ROD

This is a large refinement program for fitting a structural model to crystallographic data. It is specific to surface diffraction in that it calculates CTR rather than bulk diffraction structure factors. It differentiates the "bulk" and "surface" parts of the structure and includes a simple description of surface roughness that is usually necessary to get a good fit. The bulk cell is periodic and invariant; it only contributes to the CTRs. The atoms in the surface cell can be selectively refined in position, occupancy, and Debye-Waller factor. It is important that the z coordinates in the surface be a smooth continuation from those of the bulk.

The input files needed for ROD are as follows. All are expected to have a header line that is ignored on input. The second line of the first two files must contain the six unit cell parameters in real space, a, b, and c (in angstroms) and α, β, and γ (in degrees). The program checks that both are the same:

1. bulk unit cell coordinates, *.bul;
2. surface unit cell coordinates, *.sur (fixed), or *.fit (parameterized);
3. structure factor data, *.dat (renamed from *.wgt above); and
4. model parameters, *.par (optional).

In order that small data sets can be compatible with large models, ROD allows a very flexible parameterization scheme. The same displacement, occupancy, and DW parameters can be assigned to more than one atom. This is very convenient for structures that contain repeated motifs or different symmetries of components within the whole unit cell. In fact, the program does not handle symmetry at all, since this can always be built into the model; all structures are assumed to have triclinic, 1×1 unit cells.

APPENDIX C: SUPER COMMAND FORMAT

Commands have a variable number of arguments, depending upon the specific application. Arguments follow on the same line as the command and are separated by spaces or commas. Most commands will prompt the user for the required arguments if no arguments are given. Multiple commands may be placed on one line provided they are separated by a semi-colon.

Values of important control parameters are stored as "variables." To obtain the current value of a variable,

type its name followed by =. To change the value stored in a variable, type the new value after the =, e.g., as = 1.65. Rudimentary algebra is also allowed, e.g., as = bs * 3. Space characters are required as delimiters around the mathematical operation.

A manual of allowed commands will be available at the beamline. A list of all valid "super" commands can be obtained by typing "help" at the "super" prompt or using the F8 key. Pressing F5 lists the orientation matrix and the values of string variables. The values of all other variables are listed using F6.

Basic Super Commands

mv Move motors (four arguments: TTH, TH, PHI, CHI).
ct Count (one optional argument = preset).
br Go to position defined by H, K, and L (three arguments).
wh Print current H, K, and L and values of the motor angles (no arguments).
ca Calculate motor angles from H, K, and L (three arguments).
ci Calculate H, K, and L from motor angles (five arguments).
end Exit the program.

Basic Super Variables

as Crystal reciprocal lattice parameter a^*
bs Crystal reciprocal lattice parameter b^*
cs Crystal reciprocal lattice parameter c^*
al Crystal reciprocal lattice angle α^*
be Crystal reciprocal lattice angle β^*
ga Crystal reciprocal lattice angle γ^*
wv Incident wave vector $(2\pi/\lambda)$
nobs Number of alignment reflections used to calculate OM
h1, k1, l1, h2, k2, ... HKLs of orientation reflections
t1, u1, p1, c1, a1, u2, p2, ... 2-Theta, theta, phi, chi, and alpha of orientation reflections
B Target value of incidence angle, α_i
B2 Target value of exit angle, α_f
alm Mode of calculating ALP for five-circle geometry (alm = 1 means ALP is frozen at its current value; alm = 2 means ALP is calculated)
bem Mode of constraining incidence angle for five-circle geometry (bem = 1 means *incidence* angle α_i is fixed at target value B; (bem = 2 means *exit* angle α_f is fixed at target value B2; bem = 3 means incidence and exit angles are made equal)
fchi CHI angle of optical surface orientation (laser alignment)
fphi PHI angle of optical surface orientation (laser alignment)

Orientation Matrix

There are two ways of achieving the sample orientation, depending on whether the reciprocal lattice parameters are known or unknown. In the more commonly used "lattice parameters known" mode, the six reciprocal lattice parameters are entered manually using syntax name = value. The names of the six reciprocal lattice constants are as, bs, cs, al, be, and ga. Two orientation reflections are required with h, k, l, TTH, TH, PHI, CHI, and ALP specified for each reflection. The meaning of the angles is slightly different in different modes of calculating the orientation matrix. We refer here only to the five-circle mode, which is set by the command *frz 5*.

Individual variables in the orientation matrix can be changed using syntax name = value. The variable names in this case consist of a single letter and a number (e.g., t1, t2, t3, ...). The number refers to the number of the orientation reflection. The letters are assigned as follows for the five-circle mode: t = 2-theta, u = theta, p = phi, c = chi, a = alpha, h,k,l = Miller indices, g = gamma (fixed out-of-plane detector position), and z = zeta (fixed sample offset angle).

The orientation reflections may also be entered using the *or* (orient) command. To enter the parameters for the current diffractometer position, type *or* followed by the number of the reflection (in this case 1 or 2) and the hkl. One may also manually enter all variables on one line by typing

or # h k l TTH TH PHI CHI ALP

The orientation matrix is a general 3×3 matrix with 9 degrees of freedom described by Busing and Levy (1967). Six degrees of freedom are constrained by the reciprocal lattice parameters, 2 more come from the first orientation reflection (the two polar angles), and the final degree of freedom comes from the second orientation reflection (the second reflection specifies the azimuthal rotation about the first vector). The direction of the first orientation reflection, called the primary reflection, will always be reproduced exactly. The angles of the second reflection may not necessarily be consistent with the lattice parameters, and consequently the program may not reproduce the angles of the secondary reflection exactly. The *sor n m* command allows reflections n and m to be swapped.

The "lattice parameters unknown" mode is set by toggling the *om* command. Then the six reciprocal lattice parameters are determined directly from the five angles specified for each of the orientation reflections (TTH, TH, PHI, CHI, and ALP). To define the nine matrix elements of the orientation matrix, at least three non-coplanar reflections must be used, so nobs = 3. If nobs3, the best matrix is constructed from the full list of nobs reflections using a least-squares fit and will usually improve as more reflections are added. Reciprocal lattice parameters that are consistent with the three observed reflections will then be calculated. In general, the reciprocal lattice parameters will correspond to a triclinic cell with interfacial angles nearly equal to those expected from the symmetry. As guide to selecting suitable reflections, an error is calculated for each, which is the distance (in reciprocal angstroms) of the reflection from the reciprocal lattice point derived from the orientation matrix. The orientation parameters may be entered manually or by using the *or* command.

Important Scan Types

ascan Angle scan—Scan four diffractometer motors TTH, TH, PHI, CHI.

mscan Motor scan—Scan up to three arbitrary motors.

iscan Index scan—Input three starting HKLs and three delta HKLs for the increments between points.

jscan Centered index scan—Like iscan except scan is *centered* on the given starting hkl value.

kscan Centered two-dimensional index scan—Like jscan it is *centered* on the given starting hkl value. Two sets of delta HKL step sizes are given.

rscan Rocking scan—Input HKL value, motor number (or hkl) and one delta value for the increment. The scan moves either one motor or one hkl component and is centered at the given HKL value.

lup Line-up scan—Input motor number and delta angle. Similar to a rocking scan, but the scan is centered at the current position of the motors. At the end of the scan, the motors move to the peak position, calculated as the center of gravity of the counts measured.

fpk Find peak scan—Like lup, except fewer para meters are required and the data are not saved on disk. After the scan, the motors moved to the peak calculated from the first moment.

vscan Variable scan—Input HKL, variable name, and starting value. The scan increments the value of the variable and moves the motors to the given HKL.

vlup Centered vscan—Input HKL, variable name, delta, and npts. The scan increments the value of the variable and moves the motors to the given HKL. The scan is centered on the current value of the variable beforehand, which is reset to the center of gravity afterward.

pkup Peakup reflection—The scan does a series of lup scans centered at each orientation reflection. At the end of the series, the angles of the orientation reflection are updated to reflect the new peak position.

The scan types *rscan* and *jscan* embody automatic "stretching" of the background, whereby the step size at the beginning and end of the scan is increased by a factor of 3. The *lup* and *vlup* scans are very useful during the alignment stages of the experiment because of their autocentering feature; they can also be used to make automatic realignments during an unattended batch procedure. All scans generate a video display of the data as they accumulate, which can be viewed at any time using the function keys: F1 (F2) to see the spectrum of the multichannel analyzer as it accumulates and F3 (F4) to see a histogram of the recorded counts on a linear (log) scale.

A scan may be executed by typing the command name followed by a list of arguments. To obtain help on the command format, type the name of the scan without arguments and the program will prompt you. Any scan can be interrupted during execution by ^C and continued by typing *co*. A series of commands (including scans) can be executed by using either the *sc* command or the *ex* command.

To use the *sc* command, one first has to enter a list of commands into the scan table. The list can contain almost any command executable from the command line. A scan list can be created in three ways. The most direct method is to use the *ds* (define scans) command. Alternatively, one can use the *es* (edit scans) command to edit the scan table directly. The command *gs* filename (get scans) will read a new scan table file. Once the scan table has been entered, a list of commands or scans can be executed with the *sc* command, which accepts a series of numbers in arbitrary order corresponding to the position of the command in the scan table. Before execution begins, the program does a "dry run" to check for setting calculation errors and limit problems. If the execution is halted, the *co* command will resume execution.

A series of commands or scans can also be executed with the *ex* command. This command diverts program input to a script file and executes the commands in order. There is no limit to the number of commands in the file. In contrast to *sc*, the *ex* command does not do a dry run before execution begins. If a problem is encountered during execution, the program skips to the next command in the list.

I. K. ROBINSON
University of Illinois
Urbana, Illinois

X-RAY DIFFRACTION TECHNIQUES FOR LIQUID SURFACES AND MONOMOLECULAR LAYERS

INTRODUCTION

X-ray and neutron scattering techniques are probably the most effective tools employed to determine the structure of liquid interfaces on molecular-length scales. These are not different in principle from conventional x-ray diffraction techniques that are usually applied to three-dimensional crystals, liquids, solid surfaces etc. However, special diffractometers that enable scattering from fixed horizontal surfaces are required to carry out the experiments. Indeed, systematic studies of liquid surfaces had not begun until the introduction of the first liquid surface reflectometer (Als-Nielsen and Pershan, 1983).

A basic property of a liquid-gas interface is the length scale over which the molecular density changes from the bulk value to that of the homogeneous gaseous medium. Molecular size and capillary waves, which depend on surface tension and gravity, are among the most important factors that shape the density profile across the interface and the planar correlations (Evans, 1979; Braslau et al., 1988; Sinha et al., 1988). In some instances the topmost layers of liquids are packed differently than in the bulk, giving rise to layering phenomena at the interface. Monolayers of compounds different than the liquid can be spread at the gas-liquid interface, and are termed Langmuir monolayers (Gaines, 1966; Swalen et al., 1987). The spread compound might "wet" the liquid surface to form a film of homogeneous thickness or cluster to form an inhomogeneous rough surface. The x-ray reflectivity (XR)

technique allows one to determine the electron density across such interfaces, from which the molecular density and the total thickness can be extracted. The grazing angle diffraction (GID) technique is commonly used to determine lateral arrangements and correlations of the topmost layers at interfaces. GID is especially efficient in cases where surface crystallization of the liquid or spread monolayers occurs. Both techniques (XR and GID) provide structural information that is averaged over macroscopic areas, in contrast to scanning probe microscopies (SPMs), where local arrangements are probed (see, e.g., SCANNING TUNNELING MICROSCOPY). For an inhomogeneous interface, the reflectivity is an incoherent sum of reflectivities, accompanied by strong diffuse scattering, which, in general, is difficult to interpret definitively and often requires complementary techniques to support the x-ray analysis. Therefore, preparation of well-defined homogeneous interfaces is a key to a more definitive and straightforward interpretation.

Competitive and Related Techniques

Although modern scanning probe microscopies (SPMs) such as scanning tunneling microscopy (STM; SCANNING TUNNELING MICROSCOPY; Binning and Rohrer, 1983) and atomic force microscopy (AFM; Binning, 1992) rival x-ray scattering techniques in probing atomic arrangements of solid surfaces, they have not yet become suitable techniques for free liquid surfaces (but see X-RAY DIFFRACTION TECHNIQUES FOR LIQUID SURFACES AND MONOMOLECULAR LAYERS). The large fluctuations due to the two-dimensional nature of the liquid interface, high molecular mobility, and the lack of electrical conductivity (which is necessary for STM) are among the main obstacles that make it difficult to apply these techniques to gas-liquid interfaces. In dealing with volatile liquids, inadvertent deposition or wetting of the probe by the liquid can occur, which may obscure the measurements. In addition, the relatively strong interaction of the probe with the surface might alter its pristine properties. For similar and other reasons, electron microscopy (see Section 11a) and electron diffraction techniques (LOW-ENERGY ELECTRON DIFFRACTION), which are among the best choices for probing solid surfaces, are not suitable to most liquid surfaces, and in particular aqueous interfaces. On the other hand, visible light microscopy techniques such as Brewster angle microscopy (BAM; Azzam and Bashara, 1977; Henon and Meunier, 1991; Hönig and Möbius, 1991) or fluorescence microscopy (Lösche and Möhwald, 1984) have been used very successfully in revealing the morphology of the interface on the micrometer-length scale. This information, in general, is complementary to that extracted from x-ray scattering. These techniques are very useful for characterizing inhomogeneous surfaces with two or more distinct domains, for which XR and GID results are usually difficult to interpret. However, it is impossible to determine the position of the domains with respect to the liquid interface, their thicknesses, or their chemical nature. Ellipsometry (ELLIPSOMETRY) is another technique that exploits visible light to allow determination of film thickness on a molecular length scale; it assumes that one knows the refractive index of the substrate and of the film (Azzam and Bashara, 1977; Ducharme et al., 1990). Either of these values might be different from the corresponding bulk value, and therefore difficult to determine.

In the following sections, theoretical background to the x-ray techniques is presented together with experimental procedures and data analysis concerning liquid surfaces. This unit is intended to provide a basic formulation which can be developed for further specific applications. Several examples of these techniques applied to a variety of problems are presented briefly to demonstrate the strengths and limitations of the techniques. It should be borne in mind that the derivations and procedures described below are mostly general and can be applied to solid surfaces, and, vice versa, many results applicable to solid surfaces can be used for liquid surfaces. X-ray reflectivity from surfaces and GID have been treated in recent reviews (Als-Nielsen and Kjaer, 1989; Russell, 1990; Zhou and Chen, 1995).

PRINCIPLES OF THE METHOD

We assume that a plane harmonic wave of frequency ω and wave-vector $\mathbf{k}_0$ (with electric field, $\mathbf{E} = \mathbf{E}_0 e^{i\omega t - i\mathbf{k}_0 \cdot r}$) is scattered from a distribution of free electrons, with a number density $N_e(\mathbf{r})$. Due to the interaction with the electric field of the x-ray wave, each free electron experiences a displacement proportional to the electric field, $\mathbf{X} = (-e/m\,\omega^2)\mathbf{E}$. This displacement gives rise to a polarization $\mathbf{P}(\mathbf{r})$ distribution vector

$$\mathbf{P}(\mathbf{r}) = N_e(\mathbf{r}) e \mathbf{X} \tag{1}$$

in the medium. For the sake of convenience we define the scattering length density (SLD), $\rho(\mathbf{r})$, in terms of the classical radius of the electron $r_0 = e^2/4\pi\varepsilon_0 mc^2 = 2.82 \times 10^{-13}$ cm as follows

$$\rho(\mathbf{r}) = N_e(\mathbf{r}) r_0 \tag{2}$$

The polarization then can be written as

$$\mathbf{P}(\mathbf{r}) = \frac{-N_e(\mathbf{r}) e^2}{\omega^2 m_e}\mathbf{E} = -\frac{4\pi\varepsilon_0}{k_0^2}\rho(\mathbf{r})\mathbf{E} \tag{3}$$

The scattering length density (or the electron density) is what we wish to extract from reflectivity and GID experiments and relate it to atomic or molecular positions at liquid interfaces. The displacement vector $\mathbf{D}$ can now be constructed as follows

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}(\mathbf{r}) = \varepsilon(\mathbf{r})\mathbf{E} \tag{4}$$

where $\varepsilon(\mathbf{r})$ is the permittivity of the medium, usually associated with the refractive index $n(\mathbf{r}) = \sqrt{\epsilon(\mathbf{r})}$. To account for absorption by the medium we introduce a phenomenological factor β that we calculate from the linear absorption coefficient μ (given in tables in Wilson, 1992) as follows: $\beta = \mu/(2k_0)$. Then the most general permittivity for x-rays becomes

$$\varepsilon(\mathbf{r}) = \varepsilon_0\left[1 - \frac{4\pi}{k_0^2}\rho(\mathbf{r})\right] + 2i\beta \tag{5}$$

Table 1. Electron Number Density, SDL, Critical Angles and Momentum Transfers, and Absorption Term for Water and Liquid Mercury

	$N_e(e/\mathring{A}^3)$	$\rho_s(\mathring{A}^{-2}\times_{10^{-5}})$	$Q_c(\mathring{A}^{-1})$	$\alpha_c(deg.)for\ \lambda=_{1.5404}\mathring{A}$	$\beta(\times_{10^{-8}})$
H_2O	0.334	0.942	0.02176	0.153	1.2
Hg	3.265	9.208	0.06803	0.478	360.9

Typical values of the SLD (ρ) and the absorption term (β) for water and liquid mercury are listed in Table 1. In the absence of true charges in the scattering medium (i.e., a neutral medium) and under the assumption that the medium is nonmagnetic (magnetic permeability $\mu = 1$) the wave equations that need to be solved to predict the scattering from a known SLD can be derived from the following Maxwell equations (Panofsky and Phillips, 1962)

$$\nabla \cdot \mathbf{D} = 0 \qquad \nabla \cdot \mathbf{H} = 0$$
$$\nabla \times \mathbf{E} = -\partial \mathbf{H}/\partial t \quad \nabla \times \mathbf{H} = \partial \mathbf{D}/\partial t \tag{6}$$

Under the assumption of harmonic plane waves, $\mathbf{E} = \mathbf{E}_0 e^{i\omega t - i\mathbf{k}\cdot\mathbf{r}}$, the following general equations are obtained from Equation 6

$$\nabla^2\mathbf{E} + [k_0^2 - 4\pi\rho(\mathbf{r})]\mathbf{E} = -\nabla(\nabla\ \ln\ \varepsilon(\mathbf{r}) \cdot \mathbf{E}) \tag{7}$$
$$\nabla^2\mathbf{H} + [k_0^2 - 4\pi\rho(\mathbf{r})]\mathbf{H} = -\nabla\ \ln\ \varepsilon(\mathbf{r}) \times (\nabla \times \mathbf{H}) \tag{8}$$

In some particular cases the right hand side of Equations 7 and 8 is zero. We notice then, that the term $4\pi\rho(\mathbf{r})$ in the equation plays a role similar to that of a potential $V(\mathbf{r})$ in wave mechanics. In fact, for most practical cases the right hand side of Equation 7 and Equation 8 can be approximated to zero; thus the equation for each component of the fields resembles a stationary wave equation. In those cases, general mathematical tools, such as the Born approximation (BA) and the distorted wave Born approximation (DWBA; see DYNAMICAL DIFFRACTION and SURFACE X-RAY DIFFRACTION) can be used (Schiff, 1968).

Reflectivity

In reflectivity experiments a monochromatic x-ray beam of wavelength λ [wavevector $k_0 = 2\pi/\lambda$ and $\mathbf{k_i} = (0, k_y, k_z)$] is incident at an angle α_i on a liquid surface and is detected at an outgoing angle α_r such that $\alpha_i = \alpha_r$, as shown in Figure 1, with a final wave vector $\mathbf{k_f}$. The momentum transfer is defined in terms of the incident and reflected beam as follows

$$\mathbf{Q} = \mathbf{k_i} - \mathbf{k_f} \tag{9}$$

where in the reflectivity case $\mathbf{Q}$ is strictly along the surface normal, with $Q_z = 2k_0 \sin\alpha = 2k_z$.

Single, Ideally Sharp Interface: Fresnel Reflectivity. Sol-
Solving the scattering problem exactly for the ideally sharp interface, although simple, is very useful for the derivation of more complicated electron density profiles across interfaces. The wavefunctions employed are also essential for inclusion of dynamical effects when dealing with non-specular scattering, i.e., GID and diffuse scattering. The following relates to the case of an s-polarized x-ray beam (see the Appendix at the end of this unit for a similar derivation in the case of the p-polarized x-ray beam).

For a stratified medium with an electron density that varies along one direction, z, $\rho(\mathbf{r}) = \rho(\mathbf{z})$, assuming no absorption, i.e., $\beta = 0$, an s-type polarized x-ray beam with the electrical field parallel to the surface (along the x axis; see Fig. 1) obeys the stationary wave equation as derived from Equation 7 and is simplified as follows

$$\nabla^2 E_x + [k_0^2 - V(z)]E_x = 0 \tag{10}$$

with an effective potential $V(z) = 4\pi\rho(z)$. The general solution to Equation 2 is then given by

$$E_x = E(z)e^{ik_y y} \tag{11}$$

where the momentum transfer along y is conserved when the wave travels through the medium, leading to the well-known Snell's rule for refraction. Inserting Equation 11 in

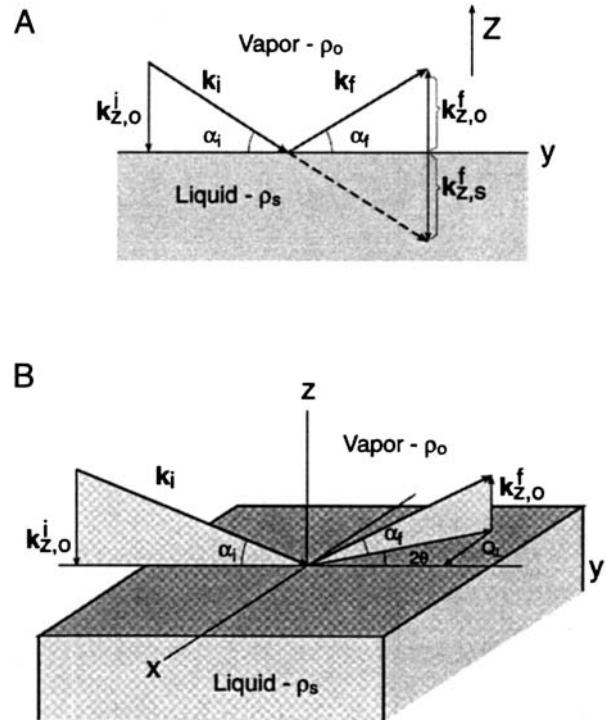


Figure 1. The geometry of incident and scattered beam in (**A**) specular reflectivity and (**B**) non-specular scattering experiments.

Equation 10 leads to a one-dimensional wave equation through a potential $V(z)$

$$\frac{d^2E}{dz^2} + [k_z^2 - V(z)]E = 0 \tag{12}$$

The simplest case of Equation 12 is that of an ideally flat interface, separating the vapor phase and the bulk scattering length density ρ_s, at $z = 0$. The solution of Equation 12 is then given by

$$E(z) = \begin{cases} e^{ik_{z,0}z} + r(k_{z,s})e^{-ik_{z,0}z} & z \geq 0 \text{ in gas} \\ t(k_{z,s})e^{ik_{z,s}z} & z \leq 0 \text{ in liquid} \end{cases} \tag{13}$$

where

$$k_{z,s} = \sqrt{k_{z,0}^2 - 4\pi\rho_s} = \sqrt{k_{z,0}^2 - k_c^2} \tag{14}$$

with $k_c \equiv 2\sqrt{\pi\rho_s}$. By applying continuity conditions to the wavefunctions and to their derivatives at $z = 0$, the Fresnel equations for reflectance, $r(k_{z,s})$, and transmission, $t(k_{z,s})$, are obtained

$$r(k_{z,s}) = \frac{k_{z,0} - k_{z,s}}{k_{z,0} + k_{z,s}}, \qquad t(k_{z,s}) = \frac{2k_{z,0}}{k_{z,0} + k_{z,s}} \tag{15}$$

The measured reflectivity from an ideally flat interface, R_F, is usually displayed as a function of the momentum transfer $Q_z = k_{z,0} + k_{z,s} \approx 2k_{z,0}$, and is given by

$$R_F(Q_z) = |r(k_{z,s})|^2 \tag{16}$$

Below a critical momentum transfer, $Q_c \equiv 2k_c \equiv 4\sqrt{\pi\rho_s}$, $k_{z,s}$ is an imaginary number and $R_F(Q_z) = 1$; thus, total external reflection occurs. Notice that whereas the critical momentum transfer does not depend on the x-ray wavelength, the critical angle for total reflection does, and it is given by $\alpha_c \approx \lambda\sqrt{\rho_s/\pi}$. Typical values of critical angles for x-rays of wavelength $\lambda_{CuKa} = 1.5404$ Å are listed in Table 1. For $Q_z \gg Q_c$, $R_F(Q_z)$ can be approximated to a form that is known as the Born approximation

$$R_F(Q_z) \sim \left(\frac{Q_c}{2Q_z}\right)^4 \tag{17}$$

This form of the reflectivity at large Q_z values is also valid for internal scattering, i.e., reflectivity from liquid into the vapor phase. However, total reflection does not occur for the internal reflectivity case. Calculated external and internal reflectivity curves from an ideally flat surface, R_F, displayed versus momentum transfer (in units of the critical momentum transfer) are shown in Figure 2A. Both reflectivities converge at large momentum transfer where they can be both approximated by Equation 17. The dashed line in the same figure shows the approximation $(Q_c/2Q_z)^4$, which fails in describing the reflectivity close to the critical momentum transfer.

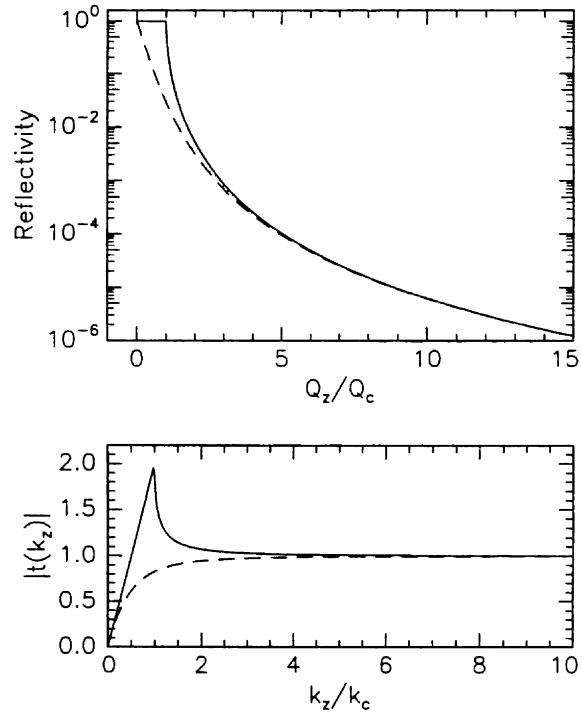


Figure 2. Calculated reflectivity curves for external (solid line) and internal (dashed line) scattering from an ideally flat interface versus momentum transfer given in units of the critical momentum transfer, $Q_c = 4(\pi\rho_s)^{1/2}$. The dotted line is kinematical approximation $(Q_c/2Q_z)^4$. The lower panel shows the amplitude of the wave in the medium for external (solid line) and external reflection (dashed line).

The photon transmission at a given $k_{z,s}$ is given by

$$T(k_{z,s}) = |t(k_{z,s})|^2 \frac{\mathrm{Re}(k_{z,s})}{\mathrm{Re}(k_{z,0})} \tag{18}$$

where the ratio on the right hand side accounts for the flux through the sample. In the case of external reflection, and for values of $k_{z,0}$ that are smaller than k_c, the real part of $k_{z,s}$ is zero, and there is no transmission, whereas above the critical angle, $k_{z,s}$ is real, and the transmission is given by

$$T(k_{z,s}) = \frac{4k_{z,0}k_{z,s}}{(k_{z,0} + k_{z,s})^2} \quad \text{for} \quad k_{z,0} > k_c \tag{19}$$

and the conservation of photons is fulfilled in the scattering process

$$T(k_{z,s}) + R(k_{z,s}) = 1 \tag{20}$$

In Figure 2B the transmission amplitude $|t(k_{z,s})|$ for external (solid line) and for internal (dashed line) reflections are shown. This amplitude modulates non-specular scattering processes at the interface as will be discussed later in this unit.

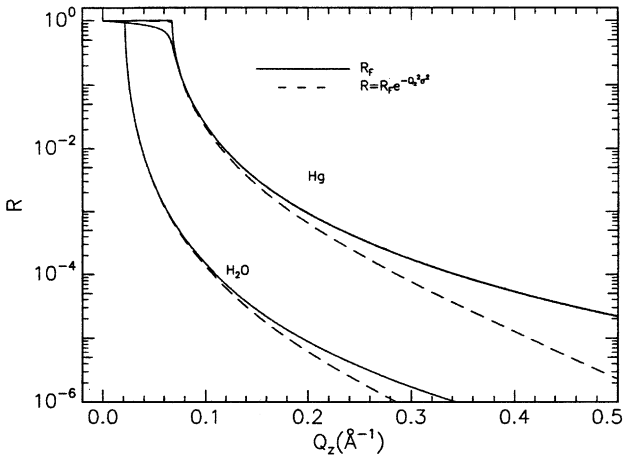


Figure 3. Calculated reflectivities from H_2O and liquid mercury (Hg) showing the effects of absorption and surface roughness. The absorption modifies the reflectivity near the critical momentum transfer for mercury with insignificant effect on the reflectivity from H_2O. The dashed line shows the calculated reflectivity from the same interfaces with root mean square surface roughness, $\sigma = 3$ Å.

The effect of absorption on the reflectivity can be incorporated by introducing β into the generalized potential in Equation 11, so that

$$k_{z,s} = \sqrt{k_{z,0}^2 - k_c^2 + 2i\beta} \tag{21}$$

is used in the Fresnel equations (Equation 15). Calculated reflectivities from water and liquid mercury, demonstrating that the effect of absorption is practically insignificant for the former, yet has the strongest influence near the critical angle for the latter, are shown in Figure 3.

Multiple Stepwise and Continuous Interfaces. On average, the electron density across a liquid interface is a continuously varying function, and is a constant far away on both sides of the interface, as shown in Figure 4.

The reflectivity for a general function $\rho(z)$ can be then calculated by one of several methods, classified into two major categories: dynamical and kinematical solutions. The dynamical solutions (see DYNAMICAL DIFFRACTION) are in general more exact, and include all the features of the scattering, in particular the low-angle regime, close to the critical angle where multiple scattering processes occur. For a finite number of discrete interfaces, exact solutions can be obtained by use of standard recursive (Parratt, 1954) or matrix (Born and Wolf, 1959) methods. These methods can be extended to compute, with very high accuracy, the scattering from any continuous potential by slicing it into a set of finite layers but with a sufficient number of interfaces. On the other hand, the kinematical approach (see KINEMATIC DIFFRACTION OF X-RAYS) neglects multiple scattering effects and fails in describing the scattering at small angles.

1. *The matrix method.* In this approach the scattering length density with variation over a characteristic length d_t is sliced into a histogram with N interfaces.

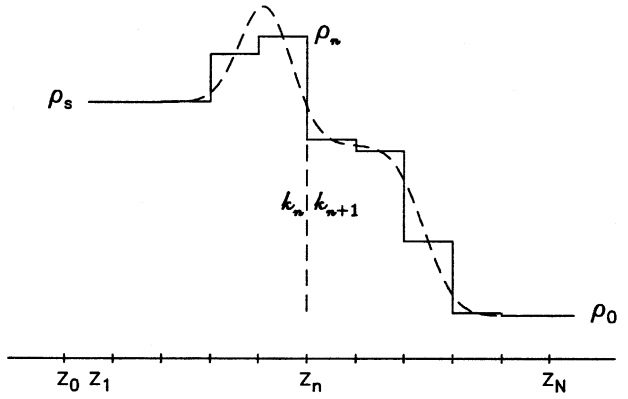


Figure 4. An illustration of a continuous scattering length density, sliced into a histogram.

The matrix method is practically equivalent to the Parratt formalism (Born and Wolf, 1959; Lekner, 1987). For each interface, the procedure described previously for the one interface is applied. Consider an arbitrary interface, n, separating two regions of a sliced SLD (as in Fig. 4), with ρ_{n-1}, and ρ_n at position $z = z_n$, with the following wavefunctions

$$\begin{array}{ccc|cc} \rho_{n-1} & & & & \rho_n \\ R_{n-1,n}e^{-ik_{n-1}z} & \leftarrow & | & \leftarrow & R_{n,n+1}e^{-ik_n z} \\ T_{n-1,n}e^{-ik_{n-1}z} & \rightarrow & | & \rightarrow & T_{n,n+1}e^{ik_n z} \\ & z = z_n & & & \end{array} \tag{22}$$

where

$$k_n \equiv \sqrt{k_{z,0}^2 - 4\pi\rho_n} \tag{23}$$

The effect of absorption can be taken into account as described earlier. For simplicity, the subscript z is omitted from the component of the wavevector so that $k_{z,n} = k_n$. The solution at each interface in terms of a transfer matrix, M_n, is given by

$$\begin{pmatrix} T_{n-1,n} \\ R_{n-1,n} \end{pmatrix} = \begin{pmatrix} e^{-i(k_{n-1}-k_n)z_n} & r_n e^{-i(k_{n-1}+k_n)z_n} \\ r_n e^{i(k_{n-1}+k_n)z_n} & e^{i(k_{n-1}-k_n)z_n} \end{pmatrix} \begin{pmatrix} T_{n,n+1} \\ R_{n,n+1} \end{pmatrix} \tag{24}$$

where

$$r_n = \frac{k_{n-1} - k_n}{k_{n-1} + k_n} \tag{25}$$

is the Fresnel reflection function through the z_n interface separating the ρ_{n-1} and ρ_n SLDs. The solution to the scattering problem is given by noting that beyond the last interface, (i.e., in the bulk), there is a transmitted wave for which only an arbitrary amplitude of the form $\binom{1}{0}$ can be assumed (i.e., the reflectivity is normalized to the incident beam anyway). The effect of all interfaces is calculated as follows

$$\begin{pmatrix} T_{0,1} \\ R_{0,1} \end{pmatrix} = (M_1)(M_2)\ \ldots(M_n)\ \ldots(M_{N+1}) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{26}$$

with

$$r_{N+1} = \frac{k_N - k_s}{k_N + k_s} \tag{27}$$

in the M_{N+1} matrix given in terms of the substrate k_s. The reflectivity is then given by the ratio

$$R(Q_z) = \left|\frac{R_{0,1}}{T_{0,1}}\right|^2 \tag{28}$$

Applying this procedure to the one-box model of thickness d with two interfaces yields

$$R(Q_z \equiv 2k_s) = \left|\frac{r_1 + r_2 e^{i2k_s d}}{1 + r_1 r_2 e^{i2k_s d}}\right|^2 \tag{29}$$

Figure 5 shows the calculated reflectivities from a flat liquid interface with two kinds of films (one box) of the same thickness d but with different scattering length densities, ρ_1 and ρ_2. The reflectivities are almost indistinguishable when the normalized

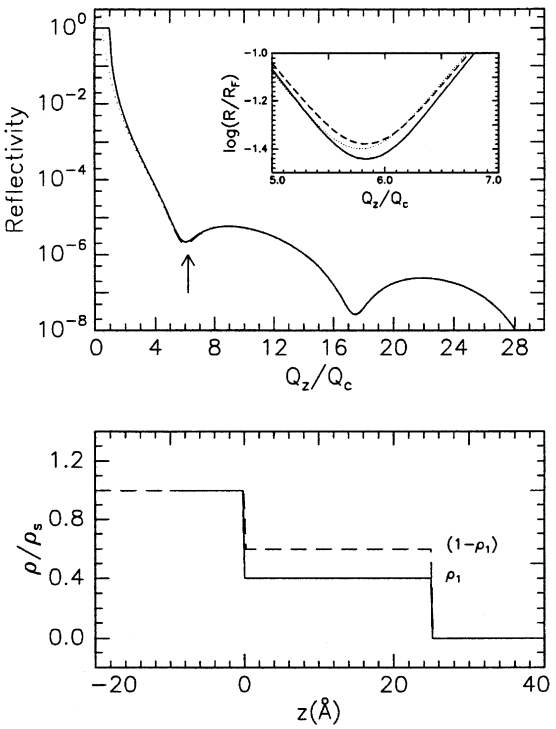


Figure 5. Calculated reflectivities for two films with identical thicknesses but with two distinct normalized electron densities, ρ_1 (solid line) and $1 - \rho_1$ (dashed line), and corresponding calculated reflectivities using the dynamical approach ($\sigma = 0$). The two reflectivities are almost identical except for a minute difference near the first minimum (see arrow in figure). The Born approximation (dotted line) for the two models yields identical reflectivities. The inset shows the normalized reflectivities near the first minimum. As Q_z is increased, the three curves converge. This is the simplest demonstration of the phase problem, i.e., the nonuniqueness of models where two different potentials give the same reflectivities.

SLDs (ρ_i/ρ_s) of the films are complementary to one ($\rho_1/\rho_s + \rho_2/\rho_s = 1$), except for a very minute difference near the first minimum. In the kinematical method described below, the two potentials shown in Figure 5 yield identical reflectivities.

The matrix method can be used to calculate the exact solution from a finite number of interfaces, and it is most powerful when used with computers by slicing any continuous scattering length density into a histogram. The criteria for determining the optimum number of slices to use is based on the convergence of the calculated reflectivity at a point where slicing the SLD into more boxes does not change the calculated reflectivity significantly.

2. *The kinematical approach.* The kinematical approach for calculating the reflectivity is only applicable under certain conditions where multiple scattering is not important. It usually fails in calculating the reflectivity at very small angles (or small momentum transfers) near the critical angle. The kinematical approach, also known as the Born approximation, gives physical insight in the formulation of $R(Q_z)$ by relating the Fresnel normalized reflectivity, R/R_F, to the Fourier transform of spatial changes in $\rho(z)$ across the interface (Als-Nielsen and Kjaer, 1989) as discussed below.

As in the dynamical approach, $\rho(z)$ is sliced so that

$$k(z) = \sqrt{k_{z,0}^2 - 4\pi\rho(z)} \quad \frac{dk}{dz} = -\frac{2\pi}{k(z)}\frac{d\rho}{dz} \tag{30}$$

and the reflectance across an arbitrary point z is given by

$$r(z) = \frac{k(z+\Delta z) - k(z)}{k(z+\Delta z) + k(z)} \approx -\frac{4\pi}{4k(z)^2}\frac{d\rho}{dz}dz \approx (Q_c/2Q_z)^2 \frac{1}{\rho_s}\frac{d\rho}{dz}dz \tag{31}$$

In the last step of the derivation, $r(z)$ was multiplied and divided by ρ_s, the SLD of the subphase, and the identity $Q_c^2 \equiv 16\pi\rho_s$ was used. Assuming no multiple scattering, the reflectivity is calculated by integrating over all reflectances at each point, z, with a phase factor $e^{iQ_z z}$ as follows

$$R(Q_z) = R_F(Q_z)\left|\frac{1}{\rho_s}\int\left(\frac{d\rho}{dz}\right)e^{iQ_z z}dz\right|^2 = R_F(Q_z)|\Phi(Q_z)|^2 \tag{32}$$

where $\Phi(Q_z)$ can be regarded as the generalized structure factor of the interface, analogous to the structure factor of a unit cell in 3-D crystals. This formula also can be derived by using the Born approximation, as is shown in the following section.

As an example of the use of Equation 32 we assume that the SLD at a liquid interface can be approximated by a sum of error functions as follows

$$\rho(z) = \rho_0 + \sum_{j=1}^{N}\frac{(\rho_j - \rho_{j-1})}{2}\left[1 + \mathrm{erf}\left(\frac{z - z_j}{\sqrt{2\sigma_j}}\right)\right] \tag{33}$$

where ρ_0 is the SLD of the vapor phase and $\rho_N = \rho_s$. Using Equation 32 the reflectivity is given by

$$R(Q_z) = R_{\mathrm{F}}(Q_z)\left|\sum_j \left(\frac{\rho_j - \rho_{j-1}}{\rho_s}\right) e^{-\frac{(Q_z\sigma_j)^2}{2}} e^{iQ_z z_j}\right|^2 \tag{34}$$

Assuming one interface at $z_1 = 0$ with surface roughness $\sigma_1 = \sigma$, the Fresnel reflectivity, $R_F(Q_z)$, is simply modified by a Debye-Waller-like factor

$$R(Q_z) = R_{\mathrm{F}}(Q_z)e^{-(Q_z\sigma)^2} \tag{35}$$

The effect of surface roughness on the reflectivities from water and from liquid mercury surfaces, assuming Gaussian smearing of the interfaces, is shown by the dashed lines in Figure 3. Braslau et al. (1988) have demonstrated that the Gaussian smearing of the interface due to capillary waves in simple liquids is sufficient in modeling the data, and that more complicated models cannot be supported by the x-ray data.

Applying Equation 34 to the one-box model discussed above (see Fig. 5), and assuming conformal roughness, $\sigma_j = \sigma$, the calculated reflectivity in terms of SLD normalized to ρ_s is

$$R(Q_z) = R_{\mathrm{F}}(Q_z)[(1-\rho_1)^2 + \rho_1^2 + 2\rho_1(1-\rho_1)\cos(Q_z d)]e^{-(Q_z\sigma)^2} \tag{36}$$

In this approximation, the roles of the normalized SLD of the one box, ρ_1, and of the complementary model, $\rho_2 = 1 - \rho_1$, are equivalent. This demonstrates that the reflectivities for both models are mathematically identical. This is the simplest of many examples where two or more distinct SLD models yield identical reflectivities in the Born approximation. When using the kinematical approximation to invert the reflectivity to SLD, there is always a problem of facing a nonunique result. For discussion of ways to distinguish between such models see Data Analysis and Initial Interpretation.

In some instances, the scattering length density can be generated by several step functions that are smeared with one Gaussian (conformal roughness $\sigma_j = \sigma$), representing different moieties of the molecules on the surface. The reflectivity can be calculated by using a combination of the dynamical and the kinematical approaches (Als-Nielsen and Kjaer, 1989). First, the exact reflectivity from the step-like functions ($\sigma = 0$) is calculated using the matrix method, $R_{\mathrm{dyn}}(Q_z)$, and the effect of surface roughness is incorporated by multiplying the calculated reflectivity with a Debye-Waller-like factor as follows (Als-Nielsen and Kjaer, 1989)

$$R(Q_z) = R_{\mathrm{dyn}}(Q_z)e^{-(Q_z\sigma)^2} \tag{37}$$

Non-Specular Scattering

The geometry for non-specular reflection is shown in Figure 1B. The scattering from a 2-D system is very weak, and enhancements due to multiple scattering processes at the interface are taken advantage of. As is shown in Figure 1B the momentum transfer $\mathbf{Q}$ has a finite component parallel to the liquid surface ($\mathbf{Q}_\perp \equiv \mathbf{k}_\perp^i - \mathbf{k}_\perp^{\mathbf{f}}$), enabling determination of lateral correlations in the 2-D plane. Exact calculation of scattering from surfaces is practically impossible except for special cases, and the Born approximation (BA; Schiff, 1968) is usually applied. When the incident beam or the scattered beam are at grazing angles (i.e., near the critical angle), multiple scattering effects modify the scattering, and these can be accounted for by a higher-order approximation known as the distorted wave Born approximation (DWBA). The features due to multiple scattering at grazing angles provide evidence that the scattering processes indeed occur at the interface.

The Born Approximation. In the BA for a general potential $V(\mathbf{r})$ the scattering length amplitude is calculated as follows (Schiff, 1968)

$$F(\mathbf{Q}) = \frac{1}{4\pi}\int V(\mathbf{r})e^{i\mathbf{Q}\cdot\mathbf{r}}d^3\mathbf{r} \tag{38}$$

where, in the present case, $V(\mathbf{r}) = 4\pi\rho(\mathbf{r})$. From the scattering length amplitude, the differential cross-section is calculated as follows (Schiff, 1968)

$$\frac{d\sigma}{d\Omega} = |F(\mathbf{Q})|^2 = \int [\rho(\mathbf{0})\rho(\mathbf{r})]e^{i\mathbf{Q}\cdot\mathbf{r}}d^3\mathbf{r} \tag{39}$$

where $[\rho(0)\ \rho(\mathbf{r})] \equiv \int[\rho(\mathbf{r}' - \mathbf{r})\rho(\mathbf{r}')]d^3\mathbf{r}'$ is the density-density correlation function. The measured reflectivity is a convolution of the differential cross-section with the instrumental resolution, as discussed below and in the literature (Schiff, 1968; Braslau et al., 1988; Sinha et al., 1988).

The scattering length density, ρ, for a liquid-gas interface can be described as a function of the actual height of the surface, $z(x,y)$, as follows

$$\rho(\boldsymbol{\mu}, z) = \begin{cases} \rho_s & \text{for} \quad z < z(\boldsymbol{\mu}) \\ 0 & \text{for} \quad z > z(\boldsymbol{\mu}) \end{cases} \tag{40}$$

where $\mu = (x, y, 0)$ is a 2-D in-plane vector. The height of the interface z is also time and temperature dependent due to capillary waves, and therefore thermal averages of z are used (Buff et al., 1965; Evans, 1979). Inserting the SLD (Equation 40) in Equation 39 and performing the integration over the z coordinate yields

$$F(\mathbf{Q}_\perp, Q_z) = \frac{\rho_s}{iQ_z}\int e^{i[\mathbf{Q}_\perp\cdot\boldsymbol{\mu} + Q_z z(\boldsymbol{\mu})]}d^2\boldsymbol{\mu} \tag{41}$$

where $\mathbf{Q}_\perp = (Q_x, Q_y, 0)$ is an in-plane scattering vector. This formula properly predicts the reflectivity from an ideally flat surface, $z(x,y) = 0$ within the kinematical approximation

$$F(\mathbf{Q}_\perp, Q_z) = \frac{4\pi^2\rho_s}{iQ_z}\delta^{(2)}(\mathbf{Q}_\perp) \tag{42}$$

with a 2-D delta-function ($\delta^{(2)}$) that guarantees specular reflectivity only. The differential cross-section is then given by

$$\frac{d\sigma}{d\Omega} = \pi^2 \left(\frac{Q_c^2}{4Q_z}\right)^2 \delta^{(2)}(\mathbf{Q}_\perp) \tag{43}$$

where $Q_c^2 \equiv 16\pi\rho_s$. This is the general form for the Fresnel reflectivity in terms of the differential cross-section $d\sigma/d\Omega$, which is defined in terms of the flux of the incident beam on the surface. In reflectivity measurements, however, the scattered intensity is normalized to the intensity of the incident beam, and therefore the flux on the sample is angle-dependent and is proportional to $\sin\alpha_i$. In addition, the scattered intensity is integrated over the polar angles α_f and 2θ with $k_0^2 \sin\alpha_f d\alpha_f d(2\theta) = dQ_x dQ_y$. Correcting for the flux and integrating

$$R_F(Q_z) \approx \sigma_{tot}(Q_z) = \int\int \left(\frac{d\sigma}{d\Omega}\right) \frac{dQ_x\, dQ_y}{4\pi^2 k_0^2 \sin\alpha_i \sin\alpha_f} = \left(\frac{Q_c}{2Q_z}\right)^4 \tag{44}$$

as approximated from the exact solution, given in Equation 17.

Taking advantage of the geometrical considerations above, the differential cross-section to the reflectivity measurement can be readily derived in the more general case of scattering length density that varies along z only, (i.e., $\rho(z)$). In this case, Equation 38 can be written as

$$\frac{d\sigma}{d\Omega} = 4\pi^2 \delta^{(2)}(\mathbf{Q}_\perp) \left| \int \rho(z) e^{iQ_z z} dz \right|^2 \tag{45}$$

If we normalize $\rho(z)$ to the scattering length density of the substrate, ρ_s, and use a standard identity between the Fourier transform of a function and its derivative, we obtain

$$\frac{d\sigma}{d\Omega} = \pi^2 \left(\frac{Q_c^2}{4Q_z}\right)^2 \left| \int \frac{1}{\rho_s} \frac{d\rho(z)}{dz} e^{iQ_z z} dz \right|^2 \delta^{(2)}(\mathbf{Q}_\perp) \tag{46}$$

which, with the geometrical corrections, yields Equation 32.

Thermal averages of the scattering length density under the influence of capillary waves and the assumption that the SLD of the gas phase is zero can be approximated as follows (Buff et al., 1965; Evans, 1979)

$$\rho(\mathbf{r}) \sim \frac{\rho_s}{2} \left(1 + \mathrm{erf}\left[\frac{z}{\sqrt{2}\sigma(\boldsymbol{\mu})}\right]\right) \tag{47}$$

where $\sigma(\mu)$ is the height-height correlation function.

Inserting Equation 47 into Equation 39, and integrating over z, results in the differential cross-section

$$\frac{d\sigma}{d\Omega} \sim \frac{Q_c^4}{Q_z^2} \int e^{i\mathbf{Q}_\perp \cdot \boldsymbol{\mu} - Q_z^2 \sigma^2(\boldsymbol{\mu})} d^2\boldsymbol{\mu} \tag{48}$$

and assuming isotropic correlation function in the plane yields (Sinha et al., 1988)

$$\frac{d\sigma}{d\Omega} \sim \frac{Q_c^4}{Q_z^2} \int \mu J_0(Q_\perp \mu) e^{-Q_z^2 \sigma^2(\mu)} d\mu \tag{49}$$

where J_0 is a Bessel function of the first kind. This expression was used by Sinha et al. (1988) to calculate the diffuse scattering from rough liquid surfaces with a height-height density correlation function that diverges logarithmically due to capillary waves (Sinha et al., 1988; Sanyal et al., 1991).

Distorted Wave Born-Approximation (DWBA). Due to the weak interaction of the electromagnetic field (x-rays) with matter (electrons), the BA is a sufficient approach to the scattering from most surfaces. However, as we have already encountered with the reflectivity, the BA fails (or is invalid) when either the incident beam or the scattered beam is near the critical angle where multiple scattering processes take place. The effect of the bulk on the scattering from the surface can be accounted for by defining the scattering length density as a superposition of two parts, as follows

$$\rho(\mathbf{r}) = \rho_1(z) + \rho_2(\boldsymbol{\mu}, z) \tag{50}$$

Here $\rho_1(z)$ is a step function that defines an ideally sharp interface separating the liquid and gas phases at $z = 0$, whereas the second term, $\rho_2(\mu, z)$, is a quasi–two-dimensional function in the sense that it has a characteristic average thickness, d_c, such that $\lim_{z \to \pm d_c/2} \rho_2(\mu, z) = 0$. It can be thought of as film-like and is a detailed function with features that relate to molecular or atomic distributions at the interface. Although the definition of ρ_2 may depend on the location of the interface, ($z = 0$) in ρ_1, the resulting calculated scattering must be invariant for equivalent descriptions of $\rho(\mathbf{r})$. In some cases ρ_2 can be defined as either a totally external or totally internal function with respect to the liquid bulk, (i.e., ρ_1). In other cases, especially when dealing with liquid surfaces, it is more convenient to locate the interface at some intermediate point coinciding with the center of mass of ρ_2 with respect to z. The effect of the substrate term $\rho_1(z)$ on the scattering from ρ_2 can be treated within the distorted wave Born approximation (DWBA) by using the exact solution from the ideally flat interface (see Principles of the Method) to generate the Green function for a higher-order Born approximation (Rodberg and Thaler, 1967; Schiff, 1968; Vineyard, 1982; Sinha et al., 1988). The Green function in the presence of an ideally flat interface, ρ_1, replaces the free particle Green function that is commonly used in the Born approximation. The scattering amplitude in this case is given by (Rodberg and Thaler, 1967; Schiff, 1968)

$$F_{DWBA}(\mathbf{Q}) = F_F(Q_z) + F_2(\mathbf{Q}) = -i\pi Q_z r_F(Q_z) + \int \tilde{\chi}^*_{k'}(\mathbf{r}) \rho_2(\mathbf{r}) \chi_k(\mathbf{r}) d\mathbf{r} \tag{51}$$

where the exact Fresnel amplitude $F_F(Q_z)$ is written in the form of a scattering amplitude so that Equation 51 reproduces the Fresnel reflectivity in the absence of ρ_2.

The exact solution of the step function ρ_1, $\chi_k(\mathbf{r})$, is given by

$$\chi_k(\mathbf{r}) = e^{ik^i_{\perp}\cdot\mu}\begin{cases} e^{ik^i_{z,0}z} + r^i(k^i_{z,s})e^{-ik^i_{z,0}z} & \text{for } z > 0 \\ t^i(k^i_{z,s})e^{ik^i_{z,s}z} & \text{for } z < 0 \end{cases} \tag{52}$$

and the $\tilde{\chi}^*_k(\mathbf{r})$ is the time-reversed and complex conjugate solution of an incident beam with $-\mathbf{k}_i$,

$$\tilde{\chi}^*_k(\mathbf{r}) = e^{-ik^f_{\perp}\cdot\mu}\begin{cases} e^{-ik^f_{z,0}z} + r^f(k^f_{z,0})e^{ik^f_{z,0}z} & \text{for } z > 0 \\ t^f(k^f_{z,0})e^{-ik^f_{z,0}z} & \text{for } z < 0 \end{cases} \tag{53}$$

In Equation 53, it is assumed that the scattered beam is detected only for $z > 0$, i.e., above the liquid surface. The notation for transmission and reflection functions indicate scattering of the wave from the air onto the subphase and vice-versa according to

$$k^i_{z,s} = \sqrt{(k^i_{z,0})^2 + k_c^2} \tag{54}$$

and

$$k^f_{z,0} = \sqrt{(k^f_{z,s})^2 + k_c^2} \tag{55}$$

respectively. In the latter case, total reflectivity does not occur except for the trivial case $k^f_{z,s} = 0$, and no enhancement due to the evanescent wave is expected. In this approximation, the final momentum transfer in the Q_z direction is a superposition of momentum transfers from ρ_2 (the film) and from the liquid interface, ρ_1. For instance, there could be a wave scattered with $Q_z = 0$ with respect to ρ_2 but reflected from the surface with a finite Q_z. This is due to multiple scattering processes between ρ_2 and ρ_1 and, therefore, in principle the amplitude of the detected beam contains a superposition of different components of the Fourier transform of $\rho_2(q_z)$ and interference terms between them. Detailed analysis of Equation 51 can be very cumbersome for the most general case, and is usually dealt with for specific cases of ρ_2. Assuming that the scattering from the film is as strong for $_Qz$ as for $-Q_z$ (as is the case for an ideal 2-D system with equal scattering along the rod, i.e., ρ_2 is symmetrical under the inversion of z), we can write $\tilde{\rho}_2(\mathbf{Q}_\perp, Q_z) \approx \tilde{\rho}_2(\mathbf{Q}_\perp, -Q_z)$. In this simple case, the cross-section can be approximated as follows (Vineyard, 1982; Sinha et al., 1988; Feidenhas'l, 1989)

$$\frac{d\sigma}{d\Omega} = |t^i(k^i_{z,s})\tilde{\rho}_2(\mathbf{Q}_\perp, Q'_z)t^f(k^f_{z,s})|^2 + |t^i(k^i_{z,s})\tilde{\rho}_2(\mathbf{Q}_\perp, Q''_z)t^f(k^f_{z,0})|^2 \tag{56}$$

where $\mathbf{Q}_\perp \equiv \mathbf{k}^{\mathbf{i}}_\perp - \mathbf{k}^{\mathbf{f}}_\perp$ and $Q'_z = k^i_{z,0} - k^f_{z,s}$ and $Q''_z = k^i_{z,s} - k^f_{z,0}$. Notice that the transmission functions modulate the scattering from the film (ρ_2), and in particular they give rise to enhancements as $k^i_{z,s}$ and $k^f_{z,s}$ are scanned around the critical angle as depicted in Figure 2B. Also,

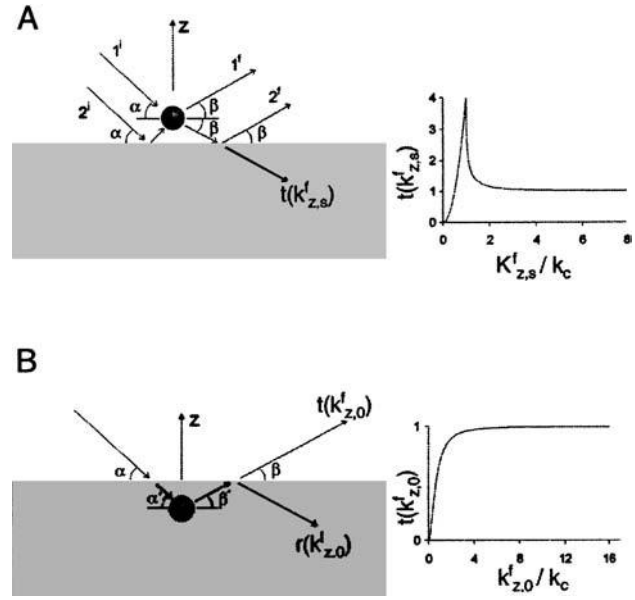


Figure 6. Illustration of wave paths for exterior (**A**) and interior (**B**) scatterer near a step-like interface. In both cases the scattering is enhanced by the transmission function when the angle of the incidence is varied around the critical angle. However, due to the asymmetry between external and internal reflectivity, the rod scan of the final beam modulates the scattering differently, as is shown on the right-hand side in each case.

it is only by virtue of the z symmetry of the scatterer that such enhancements occur for an exterior film. From this analysis we notice that there will be no enhancement due to the transmission function of the final wave for an interior film.

To examine the results from the DWBA method, we consider scattering from a single scatterer near the surface. The discussion is restricted to the case where the detection of the scattered beam is performed in the vapor phase only. The scatterer can be placed either in the vacuum ($z > 0$) or in the liquid (see Fig. 6). When the particle is placed in the vacuum there are two major relevant incident waves: a direct one from the source, labeled 1^i in Figure 6A, and a second one reflected from the surface before scattering from ρ_2, labeled 2^i. Assuming inversion symmetry along z, both waves scatter into a finite $\mathbf{Q}_\perp$ with similar strengths at Q_z and $-Q_z$, giving rise to an enhancement near the critical angle if the incident beam is near the critical angle. Another multiple scattering process that gives rise to enhancement at the critical angle is one in which beam 1^i does not undergo a change in the momentum transfer along z ($Q_z^{\text{film}} \approx 0$) before scattering from the liquid interface. The effect of these processes gives rise to enhancements if either the incident beam or the reflected beam are scanned along the z direction. Slight modifications of the momentum transfer along the z direction, such as

$$Q'_z = k^i_z + \sqrt{(k^f_z)^2 - k_c^2} \tag{57}$$

are neglected in the discussion above. The effective amplitude from the scatterer outside the medium is given by the following terms

$$|e^{iQ_z z} + e^{iQ'_z z} r(k^f_{z,s})| \approx |t(k^i_{z,s})| \tag{58}$$

where the approximation is valid since the phase factors can be neglected, and $Q_z \approx Q'_z$. At large angles the reflectivity is negligible and the transmission function approaches $t(k_{z,s}) \approx 1$. Similar arguments hold for the outgoing wave. Neglecting the small changes in the momentum transfer due to dynamical effects, the transmission function modulates the scattering as is shown in solid line in Figure 2B.

The scattered wave from a particle that is embedded in the medium is different due to the asymmetry between external and internal reflection from the liquid subphase. The wave scattered from the particle rescatters from the liquid-gas interface. Upon traversing the liquid interface, the index of refraction increases from $n = 1 - (2\pi/k_0^2)\rho$ to 1, and no total internal reflection occurs, as discussed earlier; thus there is no evanescent wave in the medium. The transmission function for this wave is given by $t(-k^i_{z,s})$, like that of a wave emanating from the liquid interface into the vapor phase. In this case, the transmission function is a real function for all $k^i_{z,s}$ and does not have the enhancements around the critical angle (as shown in Fig. 6B), with zero intensity at the horizon.

Grazing Incidence Diffraction (GID), and Rod Scans. In some instances, ordering of molecules at liquid interfaces occurs. Langmuir monolayers spread at the gas-water interface usually order homogeneously at high enough lateral pressures (Dutta et al., 1987; Kjaer et al., 1987, 1994). Surface crystallization of n-alkane molecules on molten alkane has been observed recently (Wu et al., 1993b; Ocko et al., 1997). In these cases, ρ_2 is a periodic function in x and y, and can be expanded as a Fourier series in terms of the 2-D reciprocal lattice vectors $\tau_\perp$ as follows

$$\rho_2(\boldsymbol{\mu}, z) = \sum_{\tau_\perp} F(\tau_\perp, z) e^{i\tau_\perp \cdot \boldsymbol{\mu}} \tag{59}$$

Inserting Equation 59 in Equation 56 and integrating yields the cross-section for quasi–2-D Bragg reflection at $\mathbf{Q}_\perp = \boldsymbol{\tau}_\perp$

$$\frac{d\sigma}{d\Omega} \sim P(\mathbf{Q}) |t(k^i_{z,s})|^2 \langle |F(\tau_\perp, Q_z)|^2 \rangle \mathrm{DW}(Q_\perp, Q_z) |t^f(k^f_{z,s})|^2 \delta(\mathbf{Q} - \boldsymbol{\tau}_\perp) \tag{60}$$

where P($\mathbf{Q}$) is a polarization correction and the 2-D unit cell structure factor is given as a sum over the atomic form factors $f_j(Q)$ with appropriate phase

$$F(\tau_\perp, Q_z) = \sum_j f_j(Q) e^{i\tau \cdot r_j + Q_z z_j} \tag{61}$$

The structure factor squared is averaged for multiplicity due to domains and weighted for orientation relative to the surface normal. The ordering of monolayers at the air-water interface is usually in the form of 2-D powder consisting of crystals with random orientation in the plane. From Equation 60 we notice that the conservation of momentum expressed with the delta-function allows for observation of the Bragg reflection at any Q_z. A rod scan can be performed by varying either the incident or reflected beam, or both. The variation of each will produce some modulation due to both the transmission functions and to the average molecular structure factor along the z-axis. The Debye-Waller factor (see KINEMATIC DIFFRACTION OF X-RAYS), DW($\mathbf{Q}_\perp$, Q_z), which is due to the vibration of molecules about their own equilibrium position with time-dependent molecular displacement $\mathbf{u}(t)$ is given by

$$\mathrm{DW}(\mathbf{Q}_\perp, Q_z) \sim e^{-(C_\perp Q_\perp^2 \langle \mathbf{u}_\perp^2 \rangle + Q_z^2 \sigma^2)} \tag{62}$$

The term due to capillary waves on the liquid surface is much more dominant than the contribution from the in-plane intrinsic fluctuations. The Debye-Waller factor in this case is an average over a crystalline size and might not reflect surface roughness extracted from reflectivity measurements, where it is averaged over the whole sample.

PRACTICAL ASPECTS OF THE METHOD

The minute sizes of interfacial samples on the sub-microgram level, combined with the weak interaction of x-rays with matter, result in very weak GID and reflectivity (at large Q_z) signals that require highly intense incident beams, which are available at x-ray synchrotron sources (see SURFACE X-RAY DIFFRACTION). A well prepared incident beam for reflectivity experiments at a synchrotron (for example, the X22B beam-line at the National Synchrotron Light Source at Brookhaven National Laboratory; Schwartz, 1992) has an intensity of 10^8 to 10^9 photons/sec, whereas, for a similar resolution, an 18-kW rotating anode generator produces 10^4 to 10^5 photons/sec. Although reflectivity measurements can be carried out with standard x-ray generators, the measurements are limited to almost half the angular range accessible at synchrotron sources, and they take hours to complete compared to minutes at the synchrotron. GID experiments are practically impossible with x-ray generators, since the expected signals (2-D Bragg reflections, for example) normalized to the incident beam are on the order of 10^{-8} to 10^{-10}.

Reflectivity

X-ray reflectivity and GID measurements of liquid surfaces are carried out on special reflectometers that enable manipulation of the incident as well as the outgoing beam. A prototype liquid surface reflectometer was introduced for the first time by Als-Nielsen and Pershan (1983). In order to bring the beam to an angle of incidence α_i with respect to the liquid surface, the monochromator is tilted by an angle χ either about the axis of the incident beam (indicated by χ_1 in Fig. 7) or about the axis normal to the reciprocal lattice wave vector of the monochromator, $\tau_0(\chi_2)$. Figure 7 shows the geometry that is used to deflect

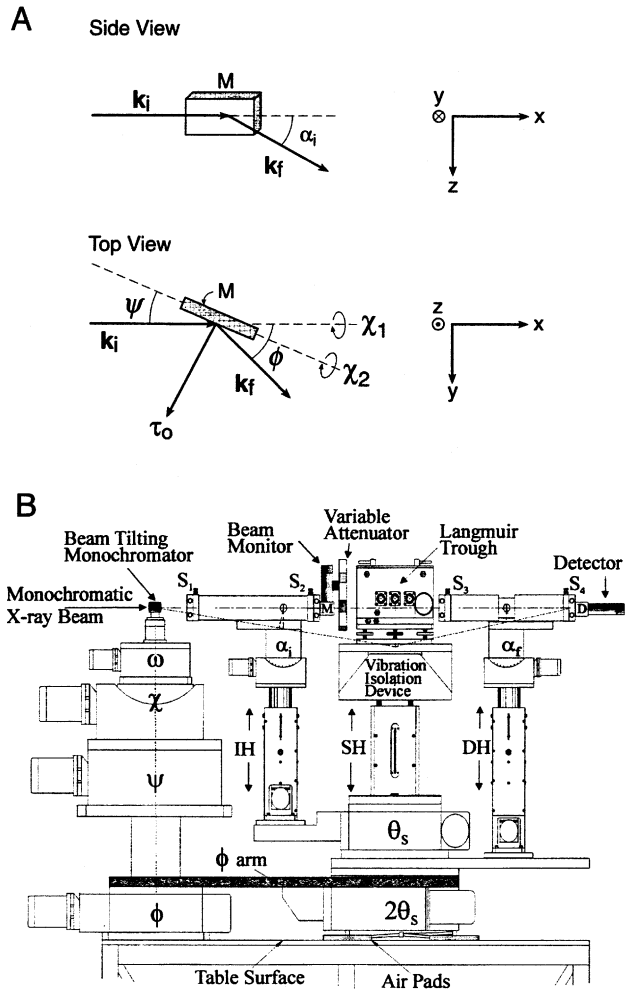


Figure 7. (**A**) Monochromator geometry to tilt a Bragg reflected beam from the horizon on a liquid surface. Two possible tilting configurations about the primary beam axis and about an axis along the surface of the reflecting planes are shown. (**B**) A side view diagram of the Ames Laboratory Liquid Surfaces Diffractometer at the 6-ID beam line at the Advanced Photon Source at Argonne National Laboratory.

the beam from the horizontal onto the liquid surface at an angle, α_i, by tilting the monochromator. At the Bragg condition, the surface of the monochromator crystal is at an angle ψ with respect to the incoming beam. Tilting over the incident beam axis is like tracing the Bragg reflection on the Debye-Scherer cone so that the ψ axis remains fixed, with a constant wavelength at different tilting angles. The rotated reciprocal lattice vector and the final wave vector in this frame are given by

$$\tau_0 = \tau_0(-\sin\psi, \cos\psi\cos\chi_1, \cos\psi\sin\chi_1)$$

$$\mathbf{k}_f = k_0(\cos\alpha_i\cos\phi, \cos\alpha_i\sin\phi, \sin\alpha_i) \qquad (63)$$

where ϕ is the horizontal scattering angle. The Bragg conditions for scattering are given by

$$\mathbf{k}_i + \tau_0 = \mathbf{k}_f; \qquad |\mathbf{k}_f| = k_0 \qquad (64)$$

Using Equations 63 and 64, the following relations for the monochromator axes are obtained

$$\begin{aligned} \sin\psi &= \frac{\tau_0}{2k_0} \\ \sin\chi_1 &= \frac{k_0}{\tau_0}\cos\psi\sin\alpha_i \\ \cos\phi &= \left(1 - \frac{\tau_0^2}{2k_0^2}\right)\Big/\cos\alpha_i \end{aligned} \qquad (65)$$

and we notice that the monochromator angle ψ is independent of α_i. However, the scattering angle ϕ has to be modified as α_i is varied. This means that the whole reflectometer arm has to be rotated. Similarly, for the configuration where the monochromator is tilted over the axis normal to τ_0 we get

$$\begin{aligned} \sin\psi &= \frac{\tau_0}{2k_0\cos\chi_2} \\ \sin\chi_2 &= \frac{k_0}{\tau_0}\sin\alpha_i \\ \cos\phi &= \left(1 - \frac{\tau_0^2}{2k_0^2}\right)\Big/\cos\alpha_i \end{aligned} \qquad (66)$$

From these relations, the conditions for a constant wavelength operation for any angle of incidence, α_i, can be calculated and applied to the reflectometer. Here, unlike the previous mode, deflection of the beam to different angles of incidence requires both the adjustment of ψ and of ϕ in order to maintain a constant wavelength. If ψ is not corrected in this mode of operation, the wavelength varies as χ is varied. This mode is sometimes desirable, especially when the incident beam hitting the monochromator consists of a continuous distribution of wavelengths around the wavelength at horizontal scattering, $\chi_2 = 0$. Such continuous wavelength distribution exists when operating with x-ray tubes or when the tilting monochromator is facing the white beam of a synchrotron. Although the variation in the wavelength is negligible as χ_2 is varied, without the correction of ψ, the exact wavelength and the momentum transfer can be computed using the relations in Equation 65 and Equation 66. In both modes of monochromator tilting, the surface height as well as the height of the slits are adjusted with vertical translations.

Figure 7, panel B, shows a side-view diagram of the Ames Laboratory Liquid Surfaces Diffractometer at the 6-ID beam line at the Advanced Photon Source at Argonne National Laboratory. A downstream Si double-crystal monochromator selects a very narrow-bandwidth energy ($\sim$1 eV) beam (in the range 4 to 40 keV) from an undulator-generated beam. The highly monochromatic beam is deflected onto the liquid surface to any desired angle of incidence, α_i, by the beam-tilting monochromator; typically Ge(111) or Ge(220) crystals are used. To simplify its description, the diffractometer can be divided into two main stages, with components that may vary slightly depending on the details of the experiment. In the first stage, the incident beam on the liquid surface is optimized. This part consists of the axes that adjust the beam tilting

monochromator (ψ, ϕ, χ), incident beam slits (S_i), beam monitor, and attenuator. The ω axis, below the monochromator crystal, is adjusted in the initial alignment process to ensure that the tilting axis χ is well defined. For each angle of incidence α_i, the ψ, ϕ, χ and the height of the incident beam arm (IH) carrying the S_1, S_2 slits, are adjusted. In addition, the liquid surface height (SH) is brought to the intersecting point between the incident beam and the detector arm axis of rotation $2\theta_s$. In the second stage of the diffractometer, the intensity of the scattered beam from the surface is mapped out. In this section, the angles α_f and $2\theta_s$ and the detector height (DH) are adjusted to select and detect the outgoing beam. The two stages are coupled through the ϕ-arm of the diffractometer. In general, θ_s is kept idle because of the polycrystalline nature of monolayers at liquid surfaces. Surface crystallization of alkanes proceeds with the formation of a few large single crystals, and the use of θ_s is essential to orienting one of these crystals with respect to the incoming beam. The scattering from liquid metals is complicated by the fact that their surfaces are not uniformly flat. For details on how to scatter from liquids with curved surfaces see Regan et al. (1997).

For aqueous surfaces, the trough (approximate dimensions $\sim 120 \times 270 \times 5$ mm^3) is placed in an airtight aluminum enclosure that allows for the exchange of the gaseous environment around the liquid surface. To get total reflectivity from water below the critical angle (e.g., $\alpha_c = 0.1538^\circ$ and 0.07688° at 8 keV and 16 keV respectively), the footprint of the beam has to be smaller than the specimen surface. A typical cross-section of the incident beam is 0.5×0.1 mm^2 with approximately 1010 to 1011 photons per second (6-ID beamline). At about $0.8\alpha_c$, the footprint of the beam with a vertical size (slit S_2) of about 0.1 mm is $\sim$47 and $\sim$94 mm at 8 keV and 16 keV, respectively, compared to 120 mm liquid-surface dimension in the direction of the primary beam. Whereas a 0.1 mm beam size is adequate for getting total reflectivity at 8 keV (exploiting about half of the liquid surface in the beam direction), a 0.05-mm beam size is more appropriate at 16 keV. This vertical beam size (slit S_2) can be increased at larger incident angles to maintain a relatively constant beam footprint on the surface.

The alignment of the diffractometer encompasses two interconnected iterated processes. First, the angles of the first stage (α_i, ψ, ϕ, χ, and ω) are optimized and set so that the x-ray flux at the monitor position is preserved upon deflection of the beam (tracking procedure). Second, the beam is steered so that it is parallel to the liquid surface. It should be emphasized that the beam, after the tracking procedure, is not necessarily parallel to the liquid surface. In this process, reflectivities from the liquid surface at various incident angles are employed to define the parallel beam, by adjustment of the angles and heights of the diffractometer. Then, repeatedly, the first and second processes are iterated until convergence is achieved (i.e., until corrections to the initial positions of motors are smaller than the mechanical accuracies).

The divergence of the monochromatic incident beam on the surface is determined by at least two horizontal slits located between the sample and the source. One of these slits is usually located as close as possible to the sample, and the other as close as possible to the source. These two slits determine the resolution of the incident beam. By deflecting the beam from the horizontal, the shape of the beam changes, and that may change incident beam intensity going through the slits; therefore the use of a monitor right after the slit in front of the sample is essential for the absolute determination of the reflectivity. The size of the two slits defining the incident beam is chosen in such a way that the footprint of the beam is much smaller than the width of the reflecting surface, so that total reflectivity occurs. Figure 8 shows the reflected beam and the direct beam from a flat surface of water, demonstrating total reflectivity at $Q_z = 0.85Q_c$. In this experiment the detector slit is wide open at $\sim$10 times the opening of the sample slit. As is demonstrated, the effect of absorption is negligible for water, and roughness is significantly reduced by damping surface waves. The damping can be achieved by reducing the height of the water film to $\sim$0.3 mm and placing a passive as well as an active antivibration unit underneath the liquid sample holder, suppressing mechanical vibrations (Kjaer et al., 1994).

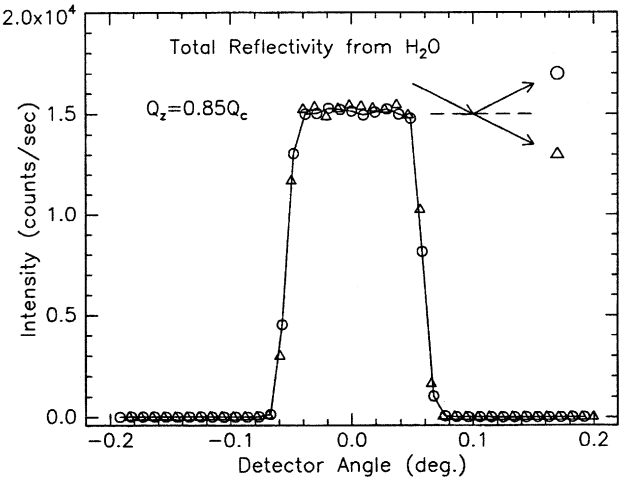


Figure 8. Superposition of the reflected beam (circles) below the critical angle and direct beam (triangles), demonstrating total reflectivity of x-rays from the surface of water. Severe surface roughness reduces the intensity and widens the reflected signal. A reduction from total reflectivity can also occur if the slits of the incident beam are too wide, so that the beam-footprint is larger than the surface sample.

Non-Specular Scattering: GID, Diffuse Scattering, and Rod Scans

X-ray GID measurements are performed at angles of incidence below the critical angle $\approx 0.9\alpha_c$. Operating with the incident beam below the critical angle enhances the signal from the surface with respect to that of the bulk by creating an evanescent wave in the medium that is exponentially decaying according to

$$E(z) = t(k_{z,s})e^{-z/\Lambda} \tag{67}$$

where

$$\frac{1}{\Lambda} = \sqrt{k_c^2 - k_{z,0}^2} \tag{68}$$

For water at $k_z \approx 0.9k_c$, $\Lambda \sim 100$ Å.

As illustrated in Figure 1, the components of the momentum transfer for GID are given by

$$\begin{aligned} Q_z &= k_0(\sin\alpha_i + \sin\alpha_f) \\ Q_x &= k_0(\cos\alpha_i - \cos\alpha_f)\cos 2\theta \\ Q_y &= k_0 \cos\alpha_f \sin 2\theta \end{aligned} \tag{69}$$

In most cases, the 2-D order on liquid surfaces is powder-like, and the lateral scans are displayed in terms of $Q_\perp$ which is given by

$$Q_\perp = k_0\sqrt{\cos^2\alpha_i + \cos^2\alpha_f - 2\cos\alpha_i\cos\alpha_f\cos 2\theta} \tag{70}$$

To determine the in-plane correlations, the horizontal resolution of the diffractometer can be adjusted with a Soller collimator consisting of vertical absorbing foils stacked together between the surface and the detector. The area that is probed at each scattering angle 2θ is proportional to $S_0/\sin 2\theta$, where S_0 is the area probed at $2\theta = \pi/2$. The probed area must be taken into account in the analysis of a GID scan that is performed over a wide range of angles.

Position-sensitive detectors (PSD) are commonly used to measure the intensity along the 2-D rods. It should be kept in mind that the intensity along the PSD is not a true rod scan of a Bragg reflection at a nominal $Q_\perp$ because of the variation in $Q_\perp$ as α_f is varied, as is seen in Equation 69.

DATA ANALYSIS AND INITIAL INTERPRETATION

The task of finding the SLD from a reflectivity curve is similar to that of finding an effective potential for Equation 7 from the modulus of the wave-function. Direct inversion of the scattering amplitude to SLD is not possible except for special cases when the BA is valid (Sacks, 1993, and references therein). If the modulus and the phase are known, they can be converted by the method of Gelfand-Levitan-Marchenko (Sacks, 1993, and references therein) to SLD (GLM method). However, in reflectivity experiments, the intensity of the scattered beam alone is measured, and phase information is lost.

Step-like potentials have been directly reconstructed recently by retrieving the phase from the modulus—i.e., reflectivity—and then using the GLM method (Clinton, 1993; also Sacks, 1993, and references therein). Model-independent methods which are based on optimization of a model to reflectivity, without requiring any knowledge of the chemical composition of the SLD at the interface, were also developed recently (Pedersen, 1992; Zhou and Chen, 1995). Such models incorporate a certain degree of objectivity. These methods are based on the kinematical and the dynamical approaches for calculating the reflectivity. One method (Pedersen, 1992) uses indirect Fourier transformation to calculate the correlation function of $d\rho/dz$, which is subsequently used in a square-root deconvolution model to construct the SLD model. Zhou and Chen (1995), on the other hand, developed a groove tracking method that is based on an optimization algorithm to reconstruct the SLD using the dynamical approach to calculate the reflectivity at each step.

The most common procedure to extract structural information from reflectivity is the use of standard nonlinear least squares refinement of an initial SLD model. The initial model is defined in terms of a P-dimensional set of independent parameters, $\mathbf{p}$, using all the information available in guessing $\rho(z, \mathbf{p})$. The parameters are then refined by calculating the reflectivity ($R[Q_z^i\mathbf{p}]$) with the tools described earlier, and by minimizing the $\chi^2(\mathbf{p})$ quantity

$$\chi^2(\mathbf{p}) = \frac{1}{N-P}\sum_{i=1}\left[\frac{R_{\exp}(Q_z^i) - R(Q_z^i, \mathbf{p})}{\varepsilon(Q_z^i)}\right]^2 \tag{71}$$

where $\varepsilon(Q_z^i)$ is the uncertainty of the measured reflectivity, $R_{\exp}(Q_z^i)$, and N is the number of measured points. The criteria for a good fit can be found in Bevington (1968). Uncertainties of a certain parameter can be obtained by fixing it at various values and for each value refining the rest of the parameters until χ^2 is increased by a factor of at least $1/(N-P)$.

The direct methods and model-independent procedures of reconstruction SLD do not guarantee uniqueness of the potential—i.e., there can be multiple SLD profiles that essentially yield the same reflectivity curve, as discussed with regard to Figure 5, for example. The uniqueness can be achieved by introducing physical constraints that are incorporated into the parameters of the model. Volume, in-plane density of electrons, etc., are among such constraints that can be used. Applying such constraints is discussed briefly below; see Examples (also see Vaknin et al., 1991a,b; Gregory et al., 1997). These constraints reduce the uncertainties and make the relationship of the SLD to the actual molecular arrangement apparent. In the dynamical approach, no two potentials yield exactly the same reflectivity, although the differences between two models might be too small to be detected in an experiment.

An experimental method for solving such a problem was suggested by Sanyal et al. (1992) using anomalous x-ray reflectivity methods. Two reflectivity curves from the same sample are measured with two different x-ray energies, one below and one above an absorption edge of the substrate atoms, thereby varying the scattering length density of the substrate. Subsequently the two reflectivity curves can be used to perform a direct Fourier reconstruction (Sanyal et al., 1992), or refinement methods can be used to remove ambiguities. This method is not efficient when dealing with liquids that consist of light atoms, because of the very low energy of the absorption edge with respect to standard x-ray energies. Another way to overcome the problem of uniqueness is by performing reflectivity experiments on similar samples with x-rays and with neutrons. In addition, the SLD, $\rho(z)$ across the

interface can be changed significantly, in neutron scattering experiments, by chemical exchange of isotopes that change $\rho(z)$, but maintain the same structure (Vaknin et al., 1991b; Penfold and Thomas, 1990). The reflectivities (x-ray as well as neutrons) can be fitted to one structural model that is defined in terms of geometrical parameters only, calculating the SLDs from scattering lengths of the constituents and the geometrical parameters (Vaknin et al., 1991b,c).

Examples

Since the pioneering work of Als-Nielsen and Pershan (1983), x-ray reflectivity and GID became standard tools for the characterization of liquid surfaces on the atomic length scales. The techniques have been exploited in studies of the physical properties of simple liquids (Braslau et al., 1988; Sanyal et al., 1991; Ocko et al., 1994), Langmuir monolayers (Dutta et al., 1987; Kjaer et al., 1987, 1989, 1994; Als-Nielsen and Kjaer, 1989; Vaknin et al., 1991b), liquid metals (Rice et al., 1986; Magnussen et al., 1995; Regan et al., 1995), surface crystallization (Wu et al., 1993a,b, 1995; Ocko et al., 1997), liquid crystals (Pershan et al., 1987), surface properties of quantum liquids (Lurio et al., 1992), protein recognition processes at liquid surfaces (Vaknin et al., 1991a, 1993; Lösche et al., 1993), and many other applications. Here, only a few examples are briefly described in order to demonstrate the strengths and the limitations of the techniques. In presenting the examples, there is no intention of giving a full theoretical background of the systems.

Simple Liquids. The term simple liquid is usually used for a monoatomic system governed by van der Waals–type interactions, such as liquid argon. Here, the term is extended to include all classical dielectric (nonmetallic) liquids such as water, organic solvents (methanol, ethanol, chloroform, etc.), and others. One of the main issues regarding dielectric liquids is the determination of the average density profile across the interface, $N(z)$. This density is the result of folding the intrinsic density $N_I(z)$ of the interface due to molecular size, viscosity, and compressibility of the fluid with density fluctuations due to capillary waves, $\Delta N_{\mathrm{CW}}(z)$. The continuous nature of the density across the interface due to capillary waves was worked out by Buff, Lovett, and Stillinger (BLS; Buff et al., 1965) assuming that $N_I(z)$ is an ideal step-like function. The probability for the displacement is taken to be proportional to the Boltzmann factor, $e^{-\beta U(z)}$, where U is the free energy necessary to disturb the surface from equilibrium state—i.e., $z(x,y) = 0$—and $\beta = 1/k_{\mathrm{B}}T$ where k_{B} is Boltzmann's constant. The free energy of an incompressible and nonviscous liquid surface consists of two terms; a surface tension (γ) term, which is proportional to the changes in area from the ideally flat surface and a gravitational term as follows

$$U = \int (\gamma[\sqrt{1 + |\nabla z|^2} - 1] + \frac{1}{2} m_s g z^2) d^2\boldsymbol{\mu} \approx \frac{1}{2} \int (\gamma |\nabla z|^2 + m_s g z^2) d^2\boldsymbol{\mu} \quad (72)$$

where m_{s} is the mass density of the liquid substrate. By using standard Gaussian approximation methods, Buff et al. (1965) find that

$$U(z) \sim \frac{z^2}{2\sigma_0^2} \quad (73)$$

Convolution of the probability with a step-like function, representing the intrinsic density of the liquid surface yields the following density function

$$N(z) = N_{\mathrm{s}} \mathrm{erfc}\left(\frac{z}{\sqrt{2}\sigma}\right) \quad (74)$$

with a form similar to the one given in Equation 47. The average surface roughness at temperature T, is then given by

$$\sigma_{\mathrm{CW}}^2 = \frac{k_{\mathrm{B}}T}{2\pi\gamma} \ln\left(\frac{L}{a_0}\right) \quad (75)$$

where a_0 is a molecular diameter and L is the size of the surface. Notice the logarithmic divergence of the fluctuations as the size of the surface increases, as expected of a 2-D system (Landau and Lifshitz, 1980). This model was further refined by assuming that the intrinsic profile has a finite width (Evans, 1979). In particular if the width due to the intrinsic profile is also expressed by a Gaussian then, the effective surface roughness is given by

$$\sigma_{\mathrm{eff}}^2 = \sigma_{\mathrm{I}}^2 + \sigma_{\mathrm{CW}}^2 \quad (76)$$

and the calculated reflectivity is similar to Equation 35 for an interface that is smeared like the error function

$$R_{\mathrm{CW}} = R_{\mathrm{F}}(Q_z) e^{-\alpha_{\mathrm{eff}}^2 Q_z^2} \quad (77)$$

Figure 9 shows the reflectivity from pure water measured at the synchrotron (D. Vaknin, unpub. observ.) where it is shown that, using Equation 77 for fitting the reflectivity data is satisfactory. This implies that the error function type of density profile (BLS model) for the liquid interface is sufficient. Only the surface roughness parameter, σ, is varied to refine the fit to the data ($\sigma = 2.54$ Å). This small roughness value depends on the attenuation of capillary waves by minimizing the depth of the water to $\sim$0.3 mm by placing a flat glass under the water (Kjaer et al., 1994). The validity of a Gaussian approximation of $N(z)$ (BLS model) was examined by various groups and for a variety of systems (Braslau et al., 1988; Sanyal et al., 1991; Ocko et al., 1994). Ocko et al. (1994) have measured the reflectivity of liquid alkanes over a wide range of temperatures, verifying that the surface roughness is of the form given in Equations 75 and 76.

Experimentally, the reflectivity signal at each Q_z from a rough interface is convoluted with the resolution of the spectrometer in different directions. The effect of the resolution in Q_z can be calculated analytically or convoluted numerically. For simplicity, we consider that the resolution functions can be approximated as a Gaussian with a

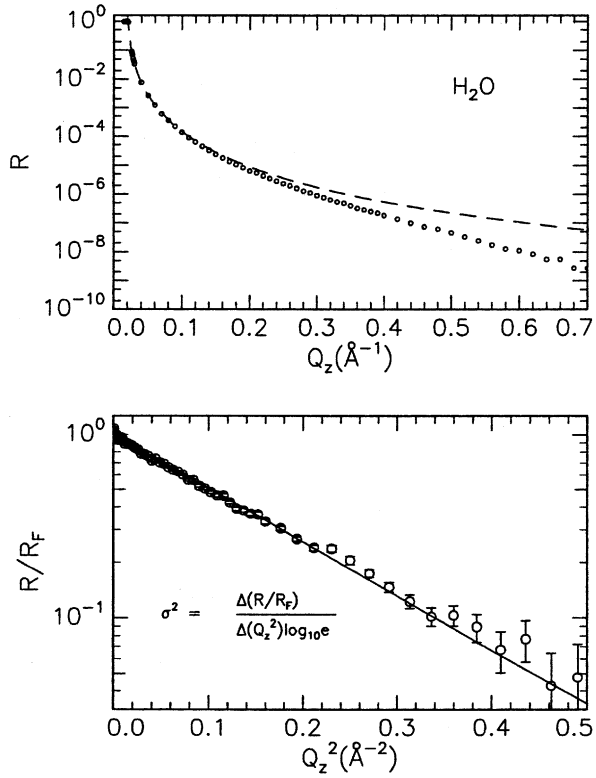


Figure 9. Experimental reflectivity from the surface of water. The dashed line is the calculated Fresnel reflectivity from an ideally flat water-interface, R_F. The normalized reflectivity versus Q_z^2 is fitted to a the form $R/R_F = e^{-(Q_z\sigma)^2}$, demonstrating the validity of the capillary-wave model (Buff et al., 1965).

width of ΔQ_z along Q_z can be taken as $e^{-Q_z^2/\Delta Q_z^2}$ with appropriate normalization factor (Bouwman and Pederesen, 1996). The resolution, ΔQ_z, is Q_z-dependent as the angles of incidence and scattering are varied (Ocko et al., 1994). However, if we assume that around a certain Q_z value the resolution is a constant and we measure σ_{exp}, the convolution of the true reflectivity with the resolution function yields the following relation

$$\frac{1}{\sigma_{exp}^2} \approx \frac{1}{\sigma_{eff}^2} + \Delta Q_z^2 \tag{78}$$

from which the effective roughness can be extracted as follows

$$\sigma_{eff} \approx \frac{\sigma_{exp}}{\sqrt{1 - \sigma_{exp}^2 \Delta Q_z^2}} \tag{79}$$

Thus, if the resolution is infinitely good, i.e., with a $\Delta Q_z = 0$, the measured and effective roughness are the same. However, as the resolution is relaxed, the measured roughness gets smaller than the effective roughness. The effect of the resolution on the determination of true surface roughness was discussed rigorously by Braslau et al. (1988).

Diffuse scattering from liquid surfaces is practically inevitable, due to the presence of capillary waves. Calculation of the scattering from disordered interfaces of various characteristics are treated in Sinha et al. (1988). In the Born approximation, true specular scattering from liquid surfaces exists only by virtue of the finite cutoff length of the mean-square height fluctuations (Equation 75). In other words, the fluctuations due to capillary waves diverge logarithmically, and true specular reflectivity is observed only by virtue of the finite instrumental resolution. The theory for the diffuse scattering from fractal surfaces and other rough surfaces was developed in Sinha et al. (1988).

Langmuir Monolayers. A Langmuir monolayer (LM) is a monomolecular amphiphilic film spread at the air-water interface. Each amphiphilic molecule consist of a polar head group (hydrophilic moiety) and a nonpolar tail, typically hydrocarbon (hydrophobic) chains (Gaines, 1966; Swalen et al., 1987; Möhwald, 1990). Typical examples are fatty acids, lipids, alcohols, and others. The length of the hydrocarbon chain can be chemically varied, affecting the hydrophobic character of the molecule. On the other hand, the head group can be ionic, dipolar, or it may have with a certain shape that might attract specific compounds present in the aqueous solution. One important motivation for studying LMs is their close relationship to biological systems. Membranes of all living cells and organelles within cells consist of lipid bilayers interpenetrated with specific proteins, alcohols, and other organic compounds that combine to give functional macromolecules which determine transport of matter and energy through them. It is well known that biological functions are structural, and structures can be determined by XR and GID. In addition, delicate surface chemistry can be carried out at the head-group interface with the molecules in the aqueous solution. From the physics point of view, the LM belongs to an important class of quasi–2-D system, by means of which statistical models that depend on the dimension of the system can be examined.

In this example, results from a simple lipid, dihexadecyl hydrogen phosphate (DHDP), consisting of a phosphate head group (PO_4^-) and its two attached hydrocarbon chains, are presented. Figure 10A displays the normalized reflectivity of a DHDP monolayer at the air-water interface at a lateral pressure of 40 mN/m. The corresponding electron density profile is shown in the inset as a solid line. The profile in the absence of surface roughness ($\sigma = 0$) is displayed as a dashed line. The bulk water subphase corresponds to $z < 0$, the phosphate headgroup region is at $0 \geq z \geq 3.4$ Å, and the hydrocarbon tails are at the 3.4 Å $\geq z \geq$ 23.1 Å region. As a first-stage analysis of the reflectivity, a model SLD with minimum number of boxes, $i = 1, 2, 3\ldots$, is constructed. Each box is characterized by a thickness d_i and an electron density $N_{e,i}$, and one surface roughness, σ for all interfaces. Refinement of the reflectivity with Equation 37 shows that the two-box model is sufficient. In order to improve the analysis, we can take advantage of information we know of the monolayer, i.e., the constituents used and the molecular area determined from the lateral-pressure versus molecular area isotherm. If the monolayer is homogeneous and not necessarily ordered, we can assume an average area per

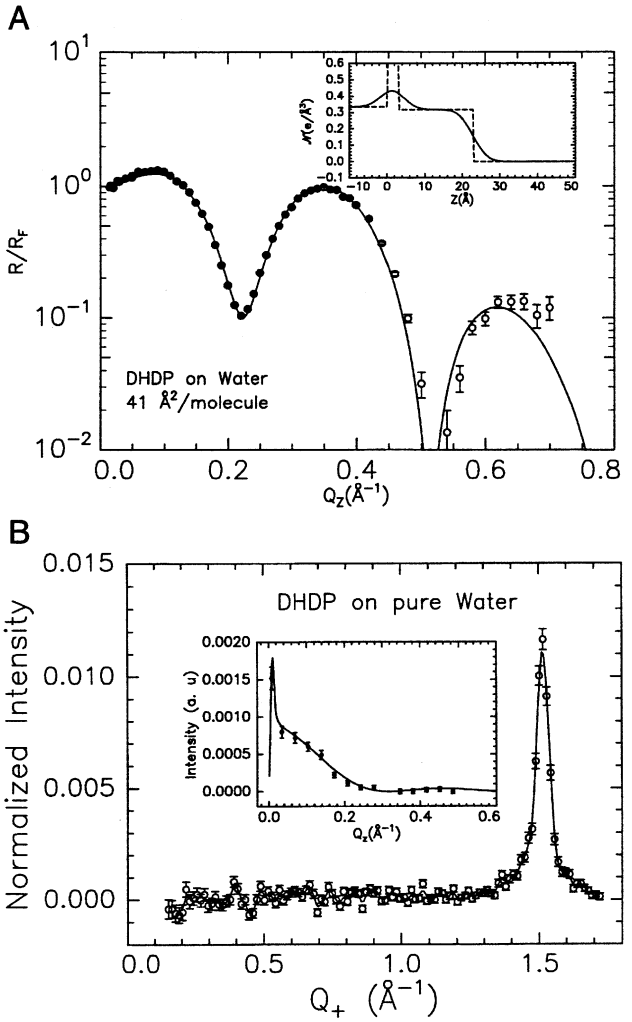


Figure 10. (**A**) Normalized x-ray reflectivity from dihexadecyl phosphate (DHDP) monolayer at the air-water interface with best-fit electron density, N_e, shown with solid line in the inset. The calculated reflectivity from the best model is shown with a solid line. The dashed line in the inset shows the box model with no roughness $\sigma = 0$. (**B**) A diffraction from the same monolayer showing a prominent 2-D Bragg reflection corresponding to the hexagonal ordering of individual hydrocarbon chains at $Q_{\perp}^{B} = 1.516$ Å^{-1}. The inset shows a rod scan from the quasi-2D Bragg reflection at $Q_{\perp}^{B}$, with a calculated model for tilted chains denoted by solid line (see text for more details).

molecule at the interface A, and calculate the electron density of the tail region as follows

$$\rho_{\text{tail}} = N_{e,\text{tail}}\, r_0/(A d_{\text{tail}}) \tag{80}$$

where $N_{e,\text{tail}}$ is the number of electrons in the hydrocarbon tail and d_{tail} is the length of the tail in the monolayer. The advantage of this description is two-fold: the number of independent parameters can be reduced, and constraints on the total number of electrons can be introduced. However, in this case, the simple relation $\rho_{\text{head}} = N_{e,\text{phosphate}}/(A d_{\text{head}})$ is not satisfactory, and in order to get a reasonable fit, additional electrons are necessary in the head-group region. These additional electrons can be associated with water molecules that interpenetrate the head-group region, which is not densely packed. The cross-section of the phosphate head group is smaller than the area occupied by the two hydrocarbon tails, allowing for water molecules to penetrate the head group region. We therefore introduce an extra parameter N_{H_2O}, the number of water molecules with ten electrons each. The electron density of the head group region is given by

$$\rho_{\text{head}} = (N_{e,\text{phosphate}} + 10N_{H_2O})/(A d_{\text{head}}) \tag{81}$$

This approach gives a physical insight into the chemical constituents at the interface. In modeling the reflectivity with the above assumptions, we can either apply volume constraints or, equivalently, examine the consistency of the model with the literature values of closely packed moieties. In this case the following volume constraint can be applied

$$V_{\text{headgroup}} = A\, d_{\text{head}} = N_{H_2O} V_{H_2O} + V_{\text{phosphate}} \tag{82}$$

where $V_{H_2O} \approx 30$ Å^2 is known from the density of water. The value of $V_{\text{phosphate}}$ determined from the refinement should be consistent within error with known values extracted from crystal structures of salt phosphate (Gregory et al., 1997).

Another parameter that can be deduced from the analysis is the average tilt angle, t, of the tails with respect to the surface. For this the following relation is used

$$d_{\text{tail}}/l_{\text{tail}} = \cos t \tag{83}$$

where l_{tail} is the full length of the extended alkyl chain evaluated from the crystal data for alkanes (Gregory et al., 1997). Such a relation is valid under the condition that the electron density of the tails when tilted is about the same as that of closely packed hydrocarbon chains in a crystal, $\rho_{\text{tail}} \approx 0.32 e/\text{Å}^3 r_0$, as observed by Kjaer et al. (1989). Such a tilt of the hydrocarbon tails would lead to an average increase in the molecular area compared to the cross-section of the hydrocarbon tails (A_0)

$$A_0/A = \cos t \tag{84}$$

Gregory et al. (1997) found that at lateral pressure $\pi = 40$ mN/m, the average tilt angle is very close to zero ($\approx 7 \pm 7^{\circ}$), and extracted an $A_0 \approx 40.7$ Å^2 compared with a value of 39.8 Å^2 for closely packed crystalline hydrocarbon chains. The small discrepancy was attributed to defects at domain boundaries.

The GID for the same monolayer is shown in Figure 10B, where a lowest-order Bragg reflection at 1.516 Å^{-1} is observed. This reflection corresponds to the hexagonal ordering of the individual hydrocarbon chains (Kjaer et al., 1987, 1989) with lattice constant $d = 4.1144$ Å, and molecular area per chain $A_{\text{chain}} = 19.83$ Å^2. Note that in DHDP the phosphate group is anchored to a pair of hydrocarbon chains with molecular area $A = 39.66$ Å^2, and it is surprising that ordering of the head group with a larger unit cell (twice that of the hydrocarbon unit cell) is not observed, as is evident in Figure 10B. Also shown in

the inset of Figure 10B is a rod scan of the Bragg reflection. To model the rod scan in terms of tilted chains, the procedure developed in Kjaer et al. (1989) is followed. The structure factor of the chain can be expressed as

$$F_{\text{chain}}(\mathbf{Q}'_{\perp}, Q'_z) = F(Q_{\perp}) \frac{\sin(lQ'_z/2)}{(lQ'_z/2)} \tag{85}$$

where $F(Q'_{\perp})$ is the in-plane Fourier transform of the cross-section of the electron density of chain, weighted with the atomic form factors of the constituents. The second term accounts for the length of the chain, and is basically a Fourier transform of a one-dimensional aperture of length l. If the chains are tilted with respect to the surface normal (in the $y-z$ plane) by an angle t, the $\mathbf{Q}'$ should be rotated as follows

$$\begin{aligned} Q'_x &= Q_x \cos t + Q_z \sin t \\ Q'_y &= Q_y \\ Q'_z &= -Q_x \sin t + Q_z \cos t \end{aligned} \tag{86}$$

Applying this transformation to the molecular structure factor, Equation 85, and averaging over all six domains (see more details in Kjaer et al., 1989) with the appropriate weights to each tilt direction, we find that at 40 mN/m the hydrocarbon chains are practically normal to the surface, consistent with the analysis of the reflectivity.

In recent Brewster Angle Microscopy (BAM) and x-ray studies of C_{60}-propylamine spread at the air-water interface (see more details on fullerene films; Vaknin, 1996), a broad in-plane GID signal was observed (Fukuto et al., 1997). The GID signal was analyzed in terms of a 2-D radial distribution function that implied short-range positional correlations extending to only few molecular distances. It was demonstrated that the local packing of molecules on water is hexagonal, forming a 2-D amorphous solid.

Surface Crystallization of Liquid Alkanes. Normal alkanes are linear hydrocarbon chains $(CH_2)_n$ terminating with CH_3 groups similar to fatty acids and lipids. The latter compounds, in contrast to alkanes, possess a hydrophilic head group at one end. Recent extensive x-ray studies of pure and mixed liquid alkanes (Wu et al., 1993a,b, 1995; Ocko et al., 1997) reveal rich and remarkable properties near their melting temperature, T_f. In particular, a single crystal monolayer is formed at the surface of an isotropic liquid bulk up to $\approx 3^{\circ}$C above T_f for a range of hydrocarbon number n. The surface freezing phenomenon exists for a wide range of chain lengths, $16 \geq n \geq 50$. The molecules in the ordered layer are hexagonally packed and show three distinct ordered phases: two rotator phases, one with the molecules oriented vertically ($16 \geq n \geq 30$), and the other tilted toward nearest neighbors ($30 \geq n \geq 44$). The third phase ($44 \geq n$) orders with the molecules tilted towards next-nearest neighbors. In addition to the 2-D Bragg reflections observed in the GID studies, reflectivity curves from the same monolayers were found to be consistent with a one-box model of densely packed hydrocarbon chains, and a thickness that corresponds to slightly tilted chains. This is an excellent demonstration of a case where no other technique but the x-ray experiments carried out at a synchrotron could be applied to get the detailed structure of the monolayers. Neutron scattering from this system would have yielded similar information; however, the intensities available today from reactors and spallation sources are smaller by at least a factor of 10^5 counts/sec for similar resolutions, and will not allow observation of any GID signals above background levels.

Liquid Metals. Liquid metals, unlike dielectric liquids, consist of the classical ionic liquid and quantum free-electron gas. Scattering of conduction electrons at a step-like potential (representing the metal-vacuum interface) gives rise to quantum interference effects and leads to oscillations of the electron density across the interface (Lang and Kohn, 1970). This effect is similar to the Friedel oscillations in the screened potential arising from the scattering of conduction electrons by an isolated charge in a metal. By virtue of their mobility, the ions in a liquid metal can in turn rearrange and conform to these oscillations to form layers at the interface, not necessarily commensurate with the conduction electron density (Rice et al., 1986). Such theoretical predictions of atomic layering at surfaces of liquid metals have been known for a long time, and have only recently been confirmed by x-ray reflectivity studies for liquid gallium and liquid mercury (Magnussen et al., 1995; Regan et al., 1995). X-ray reflectivities of these liquids were extended to $Q_z \sim 3\text{Å}^{-1}$, showing a single peak that indicates layering with spacing on the order of atomic diameters. The exponential decay for layer penetration into the bulk of Ga (6.5Å) was found to be larger than that of Hg (~3Å). Figure 11 shows a peak in the reflectivity of liquid Ga under in situ UHV oxygen-free surface cleaning (Regan et al., 1995, 1997). The normalized reflectivity was fitted to a model scattering length density shown in Figure 11B, of the following oscillating and exponentially decaying form (Regan et al., 1995)

$$\rho(z)/\rho_s = \text{erf}[(z - z_0)/\sigma] + \theta(z) A \sin(2\pi z/d) e^{-z/\xi} \tag{87}$$

where $\theta(z)$ is a step function, d is the inter-layer spacing, ξ is the exponential decay length, and A is an amplitude. Fits to this model are shown in Figure 11 with $d = 2.56$ Å, $\xi = 5.8$ Å. The layering phenomena in Ga showed a strong temperature dependence. Although liquid Hg exhibits layering with a different decay length, the reflectivity at small momentum transfers, Q_z, is significantly different than that of liquid Ga, indicating fundamental differences in the surface structures of the two metals. The layering phenomena suggest in-plane correlations that might be different than those of the bulk, but had not been observed yet with GID studies.

SPECIMEN MODIFICATION AND PROBLEMS

In conducting experiments from liquid surfaces, the experimenter faces problems that are commonly encountered in other x-ray techniques. A common nuisance in

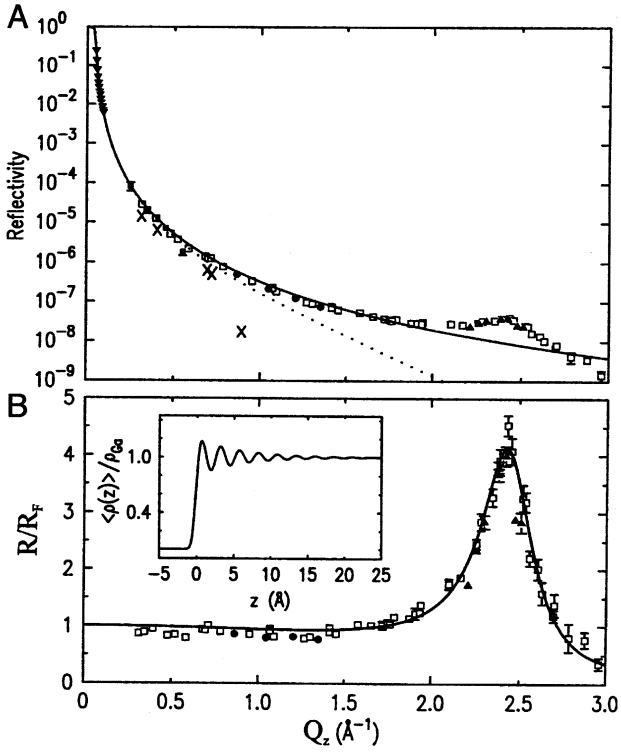


Figure 11. (**A**) Measured reflectivity for liquid gallium (Ga). Data marked with X were collected prior to sample cleaning whereas the other symbols correspond to clean surfaces (for details see Regan et al., 1995). Calculated Fresnel reflectivity from liquid Ga surface convoluted with a surface roughness due to capillary waves ($\sigma = 0.82$ Å), and the atomic form factor for Ga is denoted with a solid line. (**B**) The normalized reflectivity, with a solid line that was calculated with the best fit by an exponentially decaying sine model shown in the inset (courtesy of Regan et al., 1995).

an undulator beam line is the presence of high harmonic components that accompany the desired monochromatic beam. These high harmonics can skew the data as they affect the counting efficiency of the detector, can damage the specimen or the detector, and can increase undesirable background. The higher harmonic components can be reduced significantly (but not totally eliminated) by inserting a mirror below the critical angle for total reflection, or by detuning the second crystal of the double-crystal monochromator (XAFS SPECTROSCOPY). However, there are also problems that are more specific to reflectivity and GID in general, and to liquid surfaces in particular. Such problems can arise at any stage of the study, initially during the alignment of the diffractometer, or subsequently during data collection, and also in the final data analysis.

Poor initial alignment of the diffractometer will eventually result in poor data. Accurate tracking, namely small variation (5% to 15%) in the count rate of the monitor as the incident angle α_i is varied (workable range of α_i is 0° to 10°) is a key to getting reliable data. Although seemingly simple, one of the most important factors in getting a good alignment is the accurate determination of relevant distances in the diffractometer. For the specific design in Figure 7B, these include the distances from the center of the tilting monochromator to the center of the IH elevator and to the sample SH, or the distances from the center of the liquid surface to the detector (S_4) and to the elevator (DH). These distances are used to calculate the translations of the three elevators (IH, SH, and DH) for each angle. Although initial values measured directly with a ruler can be used, effective values based on the use of the x-ray beam yield the best tracking of the diffractometer.

Radiation damage to the specimen is a common nuisance when dealing with liquid surfaces. Many of the studies of liquid surfaces and monolayers involve investigations of organic or biomaterials that are susceptible to chemical transformations in general and, in particular, in the presence of the intense synchrotron beam. Radiation damage is of course not unique to monolayers on liquid surfaces; other x-ray techniques that involve organic materials (protein, polymer, liquid crystals, and others) face similar problems. Radiation damage to a specimen proceeds in two steps. First, the specimen or a molecule in its surroundings is ionized (by the photoelectric, Auger, or Compton effects) or excited to higher energy levels (creating radicals). Subsequently, the ionized/excited product can react with a nearby site of the same molecule or with a neighboring molecule to form a new species, altering the chemistry as well the structure at the surface. In principle, recovery without damage is also possible, but there will be a recovery time involved. The remedies that are proposed here are in part specific to liquid surfaces and cannot be always fulfilled in view of the specific requirements of an experiment. To minimize the primary effects (photoelectric, Auger, and Compton scattering) one can employ one or all of the following remedies. First, the exposure time to the x-ray beam can be minimized. For instance, while motors are still moving to their final positions, the beam can be blocked. Reduced exposure can be also achieved by attenuating the flux on the sample to roughly match it to the expected signal, so that the full intense beam is used only for signals with cross-sections for scattering that require it. That, of course, requires rough knowledge of signal intensity, which is usually the case in x-ray experiments. Monolayers on liquid surfaces are inevitably disposable and it takes a few of them to complete a study, so that in the advanced stage of a study many peak intensities are known. Another approach to reducing the effect of the primary stage is by operating at high x-ray energies. It is well known that the cross-section for all the primary effects is significantly reduced with the increase of x-ray energy. If the experiment does not require a specific energy, such as in resonance studies (anomalous scattering), it is advantageous to operate at high x-ray energies. However, operation at higher x-ray energies introduces technical difficulties. Higher mechanical angular resolutions, and smaller slits, are required in order to achieve reciprocal space resolutions comparable to those at lower energies. As discussed above, slit S_2 at 16 keV has to be set at about 50 μm width, which cannot be reliably achieved with variable slits. Finally, reducing the high harmonic component in the x-ray beam will also reduce the primary effect of the radiation.

Reducing the effects of secondary processes also depends on the requirements of a specific experiment. Air surrounding the sample has probably the worst effect on the integrity of an organic film at the liquid interface. The x-ray radiation readily creates potent radicals in air (such as monatomic oxygen) that are highly diffusive and penetrant and can interact with almost any site of an organic molecule. Working in an He environment or under vacuum can significantly reduce this source of radiation damage. Another approach to consider for reducing the secondary effect of radiation is scattering at low temperatures. It is well documented in protein crystallography that radiation damage is significantly reduced at liquid nitrogen temperatures. However, such low temperatures are not a choice when dealing with aqueous surfaces, and variations in temperature can also lead to dramatic structural transition in the films which may alter the objectives of the study.

The liquid substrate, even water, can create temporary radicals that can damage the monolayer: in particular, the head group region of lipids. Water under intense x-ray radiation can give many reactive products such as H_2O_2 or monatomic oxygen that can readily interact with the monolayer. Thus, some radiation damage, with an extent that may vary from sample to sample, is inevitable, and fresh samples are required to complete a study. Moving the sample underneath the footprint is a quick fix in that regard, assuming that the radiation damage is mostly localized around the illuminated area. To accomplish that, one can introduce a translation of the trough (perpendicular to the incident beam direction at $\alpha_i = 0$) to probe different parts of the surface. It should be noted that for lipid monolayers, a common observation suggests that radiation damage is much more severe at lower in-plane densities than for closely packed monolayers.

Another serious problem concerning scattering from surfaces is background radiation that can give count rates comparable to those expected from GID or rod scan signals. Background can be classified into two groups, one due to room background and the other due to the sample and its immediate surroundings. Although it is very hard to locate the sources of room background, it is important to trace and block them, as they reduce the capability of the diffractometer. The specimen can give unavoidable background signal due to diffuse or incoherent scattering from the liquid substrate, and this needs to be accounted for in the interpretation of the data. An important source of background is the very immediate environment of the liquid surface that does not include the sample but is included in the scattering volume of the beam: worst of all is air. Working under vacuum is not an option with monolayers, and therefore such samples are kept under an He environment. Air scattering in the trough can give rise to background levels that are at least two or three orders of magnitude higher than the expected signal from a typical 2-D Bragg reflection in the GID.

As discussed previously (see Data Analysis and Initial Interpretation), reflectivity data can give ambiguous SLD values, although that rarely happens. More often, however, the reflectivity is overinterpreted, with details in the SLD that cannot be supported by the data. The reflectivity from aqueous surfaces can be at best measured up to a momentum transfer Q_z of ~ 1 Å^{-1}, which, roughly speaking, in an objective reconstruction of the SLD, should give uncertainties on the order of $2\pi/Q_z \sim 6$ Å. It is only by virtue of complementary knowledge on the constituents of the monolayer that these uncertainties can be lowered to about 1 to 2 Å. Another potential pitfall for overinterpretation lies in the fact that in the GID experiments only a few peaks are observed, and without the knowledge on the 3-D packing of the constituents it is difficult to interpret the data unequivocally.

ACKNOWLEDGMENTS

The author would like to thank Prof. P. S. Pershan for providing a copy of Fig. 11 for this publication. Ames Laboratory is operated by Iowa State University for the U.S. Department of Energy under Contract No. W-7405-Eng-82. The work at Ames was supported by the Director for Energy Research, Office of Basic Energy Sciences.

LITERATURE CITED

Als-Nielsen, J. and Kjaer, K. 1989. X-ray reflectivity and diffraction studies of liquid surfaces and surfactant monolayers. *In* Phase Transitions in Soft Condensed Matter (T. Riste and D. Sherrington, eds.).. pp. 113–138. Plenum Press, New York.

Als-Nielsen, J. and Pershan, P. S. 1983. Synchrotron X-ray diffraction study of liquid surfaces. *Nucl. Instrum. Methods* 208:545–548.

Azzam, R. M. A. and Bashara, N. M. 1977. Ellipsometry and Polarized Light. North-Holland Publishing, New York.

Bevington, P. R. 1968. Data Reduction and Error Analysis. McGraw-Hill, New York.

Binnig, G. 1992. Force microscopy. *Ultramicroscopy* 42–44:7–15.

Binnig, G. and Rohrer, H., 1983. Scanning tunneling microscopy. *Surf. Sci.* 126:236–244.

Born, M. and Wolf, E., 1959. Principles of Optics. MacMillan, New York.

Bouwman, W. G. and Pedersen, J. S., 1996. Resolution function for two-axis specular neutron reflectivity. *J. Appl. Crystallogr.* 28:152–158.

Braslau, A., Pershan, P. S., Swislow, G., Ocko, B. M., and Als-Nielsen, J. 1988. Capillary waves on the surface of simple liquids measured by X-ray reflectivity. *Phys. Rev. A* 38:2457–2469.

Buff, F. P., Lovett, R. A., and Stillinger, Jr., F. H. 1965. Interfacial density profile for fluids at the critical region. *Phys. Rev. Lett.* 15:621–623.

Clinton, W. L. 1993. Phase determination in X-ray and neutron reflectivity using logarithmic dispersion relations. *Phys. Rev. B* 48:1–5.

Ducharme, D., Max, J.-J., Salesse, C., and Leblanc, R. M. 1990. Ellipsometric study of the physical states of phosphotadylcholines at the air-water interface. *J. Phys. Chem.* 94:1925–1932.

Dutta, P., Peng, J. B., Lin, B., Ketterson, J. B., Prakash, M. P., Georgopoulos, P., and Ehrlich, S. 1987. X-ray diffraction studies of organic monolayers on the surface of water. *Phys. Rev. Lett.* 58:2228–2231.

Evans, R. 1979. The nature of the liquid-vapor interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Adv. Phys.* 28:143–200.

Feidenhas'l, R. 1989. Surface structure determination by X-ray diffraction. *Surf. Sci. Rep.* 10:105–188.

Fukuto, M., Penanen, K., Heilman, R. K., Pershan, P. S., and Vaknin, D. 1997. C_{60}-propylamine adduct monolayers at the gas/water interface: Brewster angle microscopy and x-ray scattering study. *J. Chem. Phys.* 107:5531.

Gaines, G. 1966. Insoluble Monolayers at Liquid Gas Interface. John Wiley & Sons, New York.

Gregory, B. W., Vaknin, D., Gray, J. D., Ocko, B. M., Stroeve, P., Cotton, T. M., and Struve, W. S. 1997. Two-dimensional pigment monolayer assemblies for light harvesting applications: Structural characterization at the air/water interface with X-ray specular reflectivity and on solid substrates by optical absorption spectroscopy. *J. Phys. Chem. B* 101:2006–2019.

Henon, S. and Meunier, J. 1991. Microscope at the Brewster angle: Direct observation of first-order phase transitions in monolayers. *Rev. Sci. Instrum.* 62:936–939.

Hönig, D. and Möbius, D. 1991. Direct visualization of monolayers at the air-water interface by Brewster angle microscopy. *J. Phys. Chem.* 95:4590–4592.

Kjaer, K., Als-Nielsen, J., Helm, C. A., Laxhuber, L. A., and Möhwald, H. 1987. Ordering in lipid monolayers studied by synchrotron x-ray diffraction and fluorescence microscopy. *Phys. Rev. Lett.* 58:2224–2227.

Kjaer, K., Als-Nielsen, J., Helm, C. A., Tippman-Krayer, P., and Möhwald, H. 1989. Synchrotron X-ray diffraction and reflection studies of arachidic acid monolayers at the air-water interface. *J. Phys. Chem.* 93:3202–3206.

Kjaer, K., Als-Nielsen, J., Lahav, M., and Leiserowitz, L. 1994. Two-dimensional crystallography of amphiphilic molecules at the air-water interface. *In* Neutron and Synchrotron Radiation for Condensed Matter Studies, Vol. III (J. Baruchel, J.-L. Hodeau, M. S. Lehmann, J.-R. Regnard, and C. Schlenker, eds.). pp. 47–69. Springer-Verlag, Heidelberg, Germany.

Landau, L. D. and Lifshitz, E. M. 1980. Statistical Physics. p. 435. Pergamon Press, Elmsford, New York.

Lang, N. D. and Kohn, W. 1970. Theory of metal surfaces: Charge density and surface energy. *Phys. Rev. B1* 4555–4568.

Lekner, J. 1987. Theory of Reflection of Electromagnetic Waves and Particles. Martinus Nijhoff, Zoetermeer, The Netherlands.

Lösche, M. and Möhwald, H. 1984. Fluorescence microscope to observe dynamical processes in monomolecular layers at the air/water interface. *Rev. Sci. Instrum.* 55:1968–1972.

Lösche, M., Piepenstock, M., Diederich, A., Grünewald, T., Kjaer, K., and Vaknin, D. 1993. Influence of surface chemistry on the structural organization of monomolecular protein layers adsorbed to functionalized aqueous interfaces. *Biophys. J.* 65:2160–2177.

Lurio, L. B., Rabedeau, T. A., Pershan, P. S., Silvera, I. S., Deutsch, M., Kosowsky, S. D., and Ocko, B. M., 1992. Liquid-vapor density profile of helium: An X-ray study. *Phys. Rev. Lett.* 68:2628–2631.

Magnussen, O. M., Ocko, B. M., Regan, M. J., Penanen, K., Pershan, P. S., and Deutsch, M. 1995. X-ray reflectivity measurements of surface layering in mercury. *Phys. Rev. Lett.* 74 4444–4447.

Möhwald, H. 1990. Phospholipid and phospholipid-protein monolayers at the air/water interface. *Annu. Rev. Phys. Chem.* 41:441–476.

Ocko, B. M., Wu, X. Z., Sirota, E. B., Sinha, S. K., and Deutsch, M., 1994. X-ray reflectivity study of thermal capillary waves on liquid surfaces. *Phys. Rev. Lett.* 72:242–245.

Ocko, B. M., Wu, X. Z., Sirota, E. B., Sinha, S. K., Gang, O., and Deutsch, M., 1997. Surface freezing in chain molecules: I. Normal alkanes. *Phys. Rev. E* 55:3166–3181.

Panofsky, W. K. H. and Phillips, M. 1962. Classical Electricity and Magnetism. Addisson-Wesley, Reading, Mass.

Parratt, L. G. 1954. Surface studies of solids by total reflection of X-rays. *Phys. Rev.* 95:359–369.

Pedersen, J. S. 1992. Model-independent determination of the surface scattering-length-density profile from specular reflectivity data. *J. Appl. Crystallogr.* 25:129–145.

Penfold, J. and Thomas, R. K., 1990. The application of the specular reflection of neutrons to the study of surfaces and interfaces. *J. Phys. Condens. Matter* 2:1369–1412.

Pershan, P. S., Braslau, A., Weiss, A. H., and Als-Nielsen, 1987. Free surface of liquid crystals in the nematic phase. *Phys. Rev. A* 35:4800–4813.

Regan, M. J., Kawamoto, E. H., Lee, S., Pershan, P. S., Maskil, N., Deutsch, M., Magnussen, O. M., Ocko, B. M., and Berman, L. E. 1995. Surface layering in liquid gallium: An X-ray reflectivity study. *Phys. Rev. Lett.* 75:2498–2501.

Regan, M. J., Pershan, P. S., Magnussen, O. M., Ocko, B. M., Deutch, M., and Berman, L. E. 1997. X-ray reflectivity studies of liquid metal and alloy surfaces. *Phys. Rev. B* 55:15874–15884.

Rice, S. A., Gryko, J., and Mohanty, U. 1986. Structure and properties of the liquid-vapor interface of a simple metal. *In* Fluid Interfacial Phenomena (C. A. Croxton, ed.) pp. 255–342. John Wiley & Sons, New York.

Rodberg, L. S. and Thaler R. M. 1967. Introduction to the Quantum Theory of Scattering. Academic Press, New York.

Russell, T. P. 1990. X-ray and neutron reflectivity for the investigation of polymers. *Mater. Sci. Rep.* 5:171–271.

Sacks, P. 1993. Reconstruction of step like potentials. *Wave Motion* 18:21–30.

Sanyal, M. K., Sinha, S. K., Huang, K. G., and Ocko, B. M. 1991. X-ray scattering study of capillary-wave fluctuations at a liquid surface. *Phys. Rev. Lett.* 66:628–631.

Sanyal, M. K., Sinha, S. K., Gibaud, A., Huang, K. G., Carvalho, B. L., Rafailovich, M., Sokolov, J., Zhao, X., and Zhao, W. 1992. Fourier reconstruction of the density profiles of thin films using anomalous X-ray reflectivity. *Europhys. Lett.* 21:691–695.

Schiff, L. I. 1968. Quantum Mechanics, McGraw-Hill, New York.

Schwartz, D. K., Schlossman, M. L., and Pershan, P. S. 1992. Reentrant appearance of the phases in a relaxed Langmuir monolayer of the tetracosanoic acid as determined by X-ray scattering. *J. Chem. Phys.* 96:2356–2370.

Sinha, S. K., Sirota, E. B., Garof, S., and Stanely, H. B. 1988. X-ray and neutron scattering from rough surfaces. *Phys. Rev. B* 38:2297–2311.

Swalen, J. D., Allra, D. L., Andrade, J. D., Chandross, E. A., Garrof, S., Israelachvilli, J., Murray, R., Pease, R. F., Wynne, K. J., and Yu, H. 1987. Molecular monolayers and films. *Langmuir* 3:932–950.

Vaknin, D., Als-Nielsen, J., Piepenstock, M., and Lösche, M. 1991a. Recognition processes at a functionalized lipid surface observed with molecular resolution. *Biophys. J.* 60:1545–1552.

Vaknin, D., Kjaer, K., Als-Nielsen J., and Lösche, M. 1991b. A new liquid surface neutron reflectometer and its application to the

study of DPPC in a monolayer at the air/water interface. *Makromol. Chem. Macromol. Symp.* 46:383–388.

Vaknin, D., Kjaer, K., Als-Nielsen, J., and Lösche, M. 1991c. Structural properties of phosphotidylcholine in a monolayer at the air/water interface: Neutron reflectivity study and reexamination of X-ray reflection experiments. *Biophys. J.* 59:1325–1332.

Vaknin, D., Kjaer, K., Ringsdorf, H., Blankenburg, R., Piepenstock, M. Diederich, A., and Lösche, M. 1993. X-ray and neutron reflectivity studies of a protein monolayer adsorbed to a functionalized aqueous surface. *Langmuir* 59:1171–1174.

Vaknin, D. 1996. C_{60}-amine adducts at the air-water interface: A new class of Langmuir monolayers. *Phys. B* 221:152–158.

Vineyard, G. 1982. Grazing-incidence diffraction and the distorted-wave approximation for the study of surfaces. *Phys. Rev. B* 26:4146–4159.

Wilson, A. J. C. (eds.). 1992. International Tables For Crystallography Volume C. Kluwer Academic Publishers, Boston.

Wu, X. Z., Ocko, B. M., Sirota, E. B., Sinha, S. K., Deutsch, M., Cao, B. H., and Kim, M. W. 1993a. Surface tension measurements of surface freezing in liquid normal-alkanes. *Science* 261:1018–1021.

Wu, X. Z., Sirota, E. B., Sinha, S. K., Ocko, B. M., and Deutsch, M. 1993b. Surface crystallization of liquid normal-alkanes. *Phys. Rev. Lett.* 70:958–961.

Wu, X. Z., Ocko, B. M., Tang, H., Sirota, E. B., Sinha, S. K., and Deutsch, M. 1995. Surface freezing in binary mixtures of alkanes: New phases and phase transitions. *Phys. Rev. Lett.* 75:1332–1335.

Zhou, X. L. and Chen, S. H. 1995. Theoretical foundation of X-ray and neutron reflectometry. *Phys. Rep.* 257:223–348.

KEY REFERENCES

Als-Neilsen and Kjaer, 1989. See above.

An excellent, self-contained, and intuitive review of x-ray reflectivity and GID from liquid surfaces and Langmuir monolayers by pioneers in the field of liquid surfaces.

Braslau et al., 1988. See above.

The first comprehensive review examining the properties of the liquid-vapor interface of simple liquids (water, carbon tetrachloride, and methanol) employing the reflectivity technique. The paper provides many rigorous derivations such as the Born approximation, the height-height correlation function of the surface, and surface roughness due to capillary waves. Discussion of practical aspects regarding resolution function of the diffractometer and convolution of the resolution with the reflectivity signal is also provided.

Sinha et al., 1988. See above.

This seminal paper deals with the diffuse scattering from a variety of rough surfaces. It also provides a detailed account of the diffuse scattering in terms of the distored-wave Born approximation (DWBA). It is relevant to liquid as well as to solid surfaces.

APPENDIX: P-POLARIZED X-RAY BEAM

A p-polarized x-ray beam has a magnetic field component that is parallel to the stratified medium (along the x axis, see Fig. 1), and straightforward derivation of the wave equation (Equation 7) yields

$$\frac{d}{dz}\left(\frac{dB}{\varepsilon dz}\right) + [k_z^2 - V(z)]B = 0 \tag{88}$$

By introducing a dilation variable Z such that $dZ = \varepsilon dz$, Equation 88 for B can be transformed to a form similar to Equation 12

$$\frac{d^2B}{dZ^2} = \left[\frac{k_z^2 - V(z)}{\varepsilon}\right]B = 0 \tag{89}$$

The solution of Equation 89 for an ideally flat interface in terms of $r_p(k_{z,s})$ and $t_p(k_{z,s})$ is then simply given by

$$r_p(k_{z,s}) = \frac{k_{z,0} - k_{z,s}/\varepsilon}{k_{z,0} + k_{z,s}/\varepsilon}, \quad t_p(k_z, s) = \frac{2k_{z,0}}{k_{z,0} + k_{z,s}/\varepsilon} \tag{90}$$

The critical momentum transfer for total external reflectivity of the p-type x-ray beam is $Q_c = 2k_c = \sqrt{4\pi\rho_s}$, identical to the one derived for the s-type wave. Also, for $2k_z^B \gg Q_z \gg Q_c$, (k_z^B is defined below), $R_F(Q_z)$ can be approximated as

$$R_F(Q_z) \sim \left(\frac{Q_c}{2Q_z}\right)^4 \left(\frac{2}{1+\varepsilon}\right)^2 \tag{91}$$

The factor on the right hand side is equal to 1 for all practical liquids, and thus the Born approximation is basically the same as for the s-polarized x-ray beam (Equation 17). The main difference between the s-type and p-type waves occurs at larger angles near a Brewster angle that is given by $\theta_B = \sin^{-1}(k_z^B/k_0)$. At this angle, total transmission of the p-type wave occurs ($r_p(k_z, s) = 0$). Using Equations 14 and 90, k_z^B can be derived

$$\frac{k_z^B}{k_0} = \frac{1}{\sqrt{2 - 4\pi\rho_s/k_0^2}} \tag{92}$$

The Brewster angle for x-rays is then given by $\theta_B = \sin^{-1}(k_z^B/k_0) \approx \pi/4$. This derivation is valid for solid surfaces, including crystals, where the total transmission effect of the p-polarized wave at a Bragg reflection is used to produce polarized and monochromatic x-ray beams.

DAVID VAKNIN
Iowa State University
Ames, Iowa

This page intentionally left blank

ELECTRON TECHNIQUES

INTRODUCTION

This part describes how electrons are used to probe the microstructures of materials. These methods are arguably the most powerful and flexible set of tools available for materials characterization. For the characterization of structures internal to materials, electron beam methods provide capabilities for determining crystal structure, crystal shapes and orientations, defects within the crystals, and the distribution of atoms within these individual crystals. For characterizing surfaces, electron methods can determine structure and chemistry at the level of the atomic monolayer.

The methods in this part use electrons with kinetic energies spanning the range from 1 to 1,000,000 eV. The low energy range is the domain of scanning tunneling microscopy (STM). Since STM originates with quantum mechanical tunneling of electrons through potential barriers, this method differs from the others in this part, which employ electron beams incident on the material. Progressively higher energies are used for the electron beams of low energy electron diffraction (LEED), Auger spectrometry, reflection high energy electron diffraction (RHEED), scanning electron microscopy (SEM), and transmission electron microscopy (TEM). Electron penetration through the solid increases strongly over this broad energy range. STM, Auger, and LEED techniques are used for probing the monolayer of surface atoms (although variants of STM can probe sub-surface structures). On the other hand, the penetration capability of high energy electrons makes TEM primarily a bulk technique. Nevertheless, even with TEM it is often unrealistic to study samples having thicknesses much greater than a fraction of a micron.

Since electron beams can be focused with high accuracy, the incident beam can take various forms. A tightly focused beam, rastered across the sample, is one useful form. The acquisition of various signals in synchronization with this raster pattern provides a spatial map of the signal. Signals can be locations of characteristic x-ray emission, secondary-electron emission, or electron energy losses, to name but three. A plane wave is another useful form of the incident electron beam. Plane wave illumination is typically used for image formation with conventional microscopy and diffraction. It turns out that for complementary optical designs, the same diffraction effects in the specimen that produce visible features in images will occur with either a point- or plane-illumination mode. There are, however, advantages to one mode of operation versus another, and instruments of the scanning type and imaging type are typically used for different types of measurements. Atomic resolution imaging, for example, can be performed with a plane wave illumination method in high resolution electron microscopy (HREM), or with a probe mode in Z-contrast imaging. A fundamental difference between these techniques is that the diffracted waves interfere coherently in the case of HREM imaging, and the phase information of the scattering is used to make interference patterns in the image. On the other hand, the operation of the scanning transmission electron microscope for Z-contrast imaging makes use of incoherent imaging, where the intensities of the scatterings from the individual atoms add independently to the image.

Electron scattering from a material can be either elastic or inelastic. In general, elastic scattering is used for imaging measurements, whereas inelastic scattering provides for spectroscopy. (Electron beam techniques are enormously versatile, however, so there are many exceptions to this generality.) Chemical mapping, for example, makes images by use of inelastic scattering. For chemical mapping, the energy loss of the electron itself may be used to obtain the chemical information, as in electron energy loss spectrometry (EELS). Alternatively, the subsequent radiations from atoms ionized by the inelastic scattering may be as signals containing the chemical information. A component of the characteristic atomic x-ray spectrum, measured with an energy dispersive x-ray spectrometer (EDS), is a particularly useful signal for making chemical maps with a scanning electron microscope. The intensity of selected Auger electrons emitted from the excited atoms provides the chemical mapping capability of scanning Auger microscopy. Auger maps are highly surface-sensitive, since the Auger electrons lose energy as they traverse only a short distance through the solid.

The chapters in this part enumerate some requirements for sample preparation, but it is remarkable that any one of these techniques permits studies on a large variety of samples. All classes of materials (metals, ceramics, polymers, and composites) have been studied by all of these electron techniques. Except for some problems with the lightest (or heaviest) elements in the periodic table, the electron beam techniques have few compositional limitations. There are sometimes problems with how the sample may contaminate the vacuum system that is integral to all electron beam methods. This is especially true for the surface techniques that employ ultrahigh vacuums.

When approaching a new problem in materials characterization, electron beam methods (especially SEM), provide some of the best reward/effort ratios of any materials characterization technique. Many, if not all, of the techniques in this part should therefore be familiar to everyone interested in materials characterization. Some of the methods are available through commercial laboratories on a routine basis, such as SEM, EDS and Auger spectroscopy. TEM services can sometimes be found at companies offering asbestos analysis. Manufacturers of electron beam instruments may also provide analysis services at reasonable rates. All the electron methods of this part are available at materials research laboratories at universities and national laboratories. Preliminary assessments of the applicability of an electron method to a materials problem can often be made by contacting someone at a local institution. These institutions usually have established policies for fees and service for outside users. For value in capital investment, perhaps a scanning electron

microscope with an energy dispersive spectrometer is a best buy. The surface techniques are more specialized, and the expense and maintenance of their ultrahigh vacuum systems can be formidable. TEM is also a technique that cannot be approached casually, and some contact with an expert in the field is usually the best way to begin.

BRENT FULTZ

SCANNING ELECTRON MICROSCOPY

INTRODUCTION

The scanning electron microscope (SEM) is one of the most widely used instruments in materials research laboratories and is common in various forms in fabrication plants. Scanning electron microscopy is central to microstructural analysis and therefore important to any investigation relating to the processing, properties, and behavior of materials that involves their microstructure. The SEM provides information relating to topographical features, morphology, phase distribution, compositional differences, crystal structure, crystal orientation, and the presence and location of electrical defects. The SEM is also capable of determining elemental composition of micro-volumes with the addition of an x-ray or electron spectrometer (see ENERGY-DISPERSIVE SPECTROMETRY and AUGER ELECTRON SPECTROSCOPY) and phase identification through analysis of electron diffraction patterns (see LOW-ENERGY ELECTRON DIFFRACTION). The strength of the SEM lies in its inherent versatility due to the multiple signals generated, simple image formation process, wide magnification range, and excellent depth of field.

Lenses in the SEM are not a part of the image formation system but are used to demagnify and focus the electron beam onto the sample surface. This gives rise to two of the major benefits of the SEM: range of magnification and depth of field in the image. Depth of field is that property of SEM images where surfaces at different distances from the lens appear in focus, giving the image three-dimensional information. The SEM has more than 300 times the depth of field of the light microscope. Another important advantage of the SEM over the optical microscope is its high resolution. Resolution of 1 nm is now achievable from an SEM with a field emission (FE) electron gun. Magnification is a function of the scanning system rather than the lenses, and therefore a surface in focus can be imaged at a wide range of magnifications from 3× up to 150,000×. The higher magnifications of the SEM are rivaled only by the transmission electron microscope (TEM) (see TRANSMISSION ELECTRON MICROSCOPY and SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING), which requires the electrons to penetrate through the entire thickness of the sample. As a consequence, TEM sample preparation of bulk materials is tedious and time consuming, compared to the ease of SEM sample preparation, and may damage the microstructure. The information content of the SEM and TEM images is different, with the TEM image showing the internal structure of the material.

Due to these unique features, SEM images frequently appear not only in the scientific literature but also in the daily newspapers and popular magazines. The SEM is relatively easy to operate and affordable and allows for multiple operation modes, corresponding to the collection of different signals. The following sections review the SEM instrumentation and principles, its capabilities and applications, and recent trends and developments.

PRINCIPLES OF THE METHOD

Signal Generation

The SEM electron beam is a focused probe of electrons accelerated to moderately high energy and positioned onto the sample by electromagnetic fields. The SEM optical column is utilized to ensure that the incoming electrons are of similar energy and trajectory. These beam electrons interact with atoms in the specimen by a variety of mechanisms when they impinge on a point on the surface of the specimen. For inelastic interactions, energy is transferred to the sample from the beam, while elastic interactions are defined by a change in trajectory of the beam electrons without loss of energy. Since electrons normally undergo multiple interactions, the inelastic and elastic interactions result in the beam electrons spreading out into the material (changing trajectory from the original focused probe) and losing energy. This simultaneous energy loss and change in trajectory produces an interaction volume within the bulk (Fig. 1). The size of this interaction volume can be estimated by Monte Carlo simulations, which incorporate probabilities of the multiple possible elastic and inelastic interactions into a calculation of electron trajectories within the specimen.

The signals resulting from these interactions (e.g., electrons and photons) will each have different depths within the sample from which they can escape due to their unique

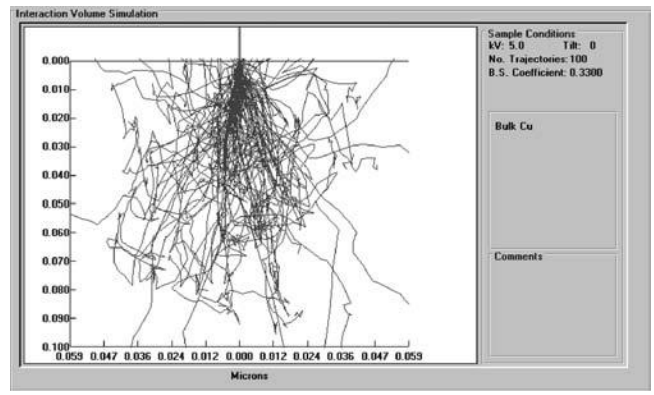


Figure 1. A Monte Carlo simulation of electron beam interaction with a bulk copper target at 5 keV shows the interaction volume within the specimen. The electron trajectories are shown, and the volume is symmetric about the beam due to the normal incidence angle (derived from Goldstein et al., 1992; Monte Carlo simulation using Electron Flight Simulator, Small World, D. Chernoff).

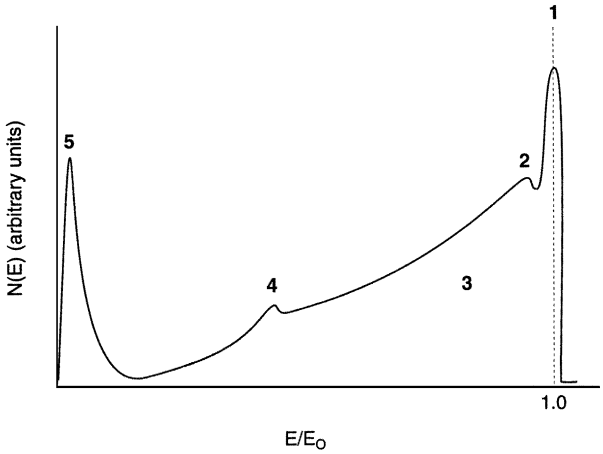


Figure 2. Distribution of relative energies (E/E_0; E_0 = incident beam energy) of electrons ejected from a surface by an incident electron beam (not to scale). The peak (1) is the elastic peak, BSEs that have lost no energy in the specimen. Slightly lower in energy is the plasmon loss region (2). The BSEs (3) are spread over the entire energy range from 0 eV to the energy of the incident beam. The characteristic Auger electron peak (4) is usually a small peak superimposed on the backscattered curve. Secondary electrons emitted from specimen atoms are responsible for the large peak at low energy (5).

physical properties and energies. For example, a secondary electron (SE) is a low-energy (2- to 5-eV) electron ejected from the outer shell of a sample atom after an inelastic interaction. These low-energy electrons can escape the surface only if generated near the surface. Thus we have an "interaction volume" in which beam electrons interact with the specimen and a "sampling volume" from which a given signal escapes the solid, which is some fraction of the interaction volume. It is this sampling volume and the signal distribution within it that determine the spatial resolution of the technique (Joy, 1984). We can thus expect different spatial resolutions for the various types of signals generated in the SEM.

Backscattered electrons (BSEs) are electrons from the incident probe that undergo elastic interactions with the sample, change trajectory, and escape the sample. These make up the majority of electrons emitted from the specimen at high beam voltage, and their average energy is much higher than that of the SEs (Fig. 2). The depth from which BSEs escape the specimen is dependent upon the beam energy and the specimen composition, but >90% generally escape from less than one-fifth the beam penetration depth. The intensity of the BSE signal is a function of the average atomic number (Z) of the specimen, with heavier elements (higher Z samples) producing more BSEs. It is thus a useful signal for generating compositional images, in which higher Z phases appear brighter than lower Z phases. The BSE intensity and trajectory are also dependent upon the angle of incidence between the beam and the specimen surface. The topography, or physical features of the surface, is then imaged by using these properties of the BSE signal to generate BSE topographic images. Due to the relatively high energy of the BSE signal, the sampling volume (sample depth) is greater than that of SEs.

Secondary electrons are due to inelastic interactions, are low energy (typically 2 to 5 eV), and are influenced more by surface properties than by atomic number. The SE is emitted from an outer shell of a specimen atom upon impact of the incident electron beam. The term "secondary" thus refers to the fact that this signal is not a scattered portion of the probe, but a signal generated within the specimen due to the transfer of energy from the beam to the specimen. In practice, SEs are arbitrarily defined as those electrons with <50 eV energy (Fig. 2). The depth from which SEs escape the specimen is generally between 5 and 50 nm due to their low energy. Secondary electrons are generated by both the beam entering the specimen and BSEs as they escape the specimen; however, SE generation is concentrated around the initial probe diameter. The SE sampling volume is therefore smaller than the BSE and provides for a high-resolution image of the specimen surface. Secondary electron intensity is a function of the surface orientation with respect to the beam and the SE detector and thus produces an image of the specimen morphology. The SE intensity is also influenced by chemical bonding, charging effects, and BSE intensity, since BSE-generated SEs are a significant part of the SE signal.

Characteristic x rays and Auger electrons are generated by inelastic interactions of the probe electrons in which an inner shell electron is emitted from a specimen atom (ionization). Following the creation of a hole in an inner shell, the atom relaxes by a transition in which an outer shell electron fills the inner shell (relaxation). Since the outer shell electron was at a higher energy level, this relaxation results in a release of energy from the atom (emission). This emission may be in the form of a photon (x ray) or may go into the emission of an electron (the Auger electron, AUGER ELECTRON SPECTROSCOPY) from the atom. These two processes therefore compete, since a single ionization can result in either an x ray or an Auger electron. The x-ray energy would then equal the difference in binding energy of the two electron levels within the atom and is thus characteristic of that element. The Auger electron generated from this ionization is emitted from an outer shell. It would thus have a kinetic energy equal to that of the x ray minus the binding energy of the outer shell from which it is emitted. The Auger sample depth (1 to 3 nm) is much less than the x-ray sample depth (0.1 to 100 μm) since electrons have a higher probability of interaction with the specimen than photons (x rays) of similar energy, and energy loss interactions will render them non-characteristic. X rays may interact with the specimen through absorption, in which the entire energy is absorbed by the specimen, but do not undergo partial energy loss interactions. These signals can be used for imaging as well as the identification of elements within the sample volume (see ENERGY-DISPERSIVE SPECTROMETRY and AUGER ELECTRON SPECTROSCOPY). It is common for the SEM in a materials laboratory to include an x-ray spectrometer as one of the detectors, and scanning Auger microanalysis combines the SEM with an electron spectrometer. Electron energy loss spectroscopy is a transmission electron technique,

since multiple energy loss interactions occur for a single beam electron in the bulk samples characteristic of SEM and deconvolution of these energy losses is not possible.

Other signals originating from electron beam interactions with the sample include plasmons, phonons, bremsstrahlung radiation (noncharacteristic x rays), and cathodoluminescence. Cathodoluminescence is the emission of light from a material under the electron beam. The wavelength of the light is a function of the material's band gap energy, since the generation involves the production and recombination of electron-hole pairs. Cathodoluminescence is therefore useful in the characterization of semiconductors, especially in cases where a contactless, high-spatial-resolution method is needed (Yacobi and Holt, 1986).

The sample may also absorb a significant portion of the electron beam current, and the specimen current image can thus provide the same information as the combined BSE and SE signals (Newbury, 1976). Charge collection scanning electron microscopy with the specimen as the detector can also be used to study the electron-beam-induced current/ conductivity (EBIC), which has been used to determine carrier lifetime, diffusion length, defect energy levels, and surface recombination velocities and to locate *p-n* junctions and recombination sites (Leamy, 1982).

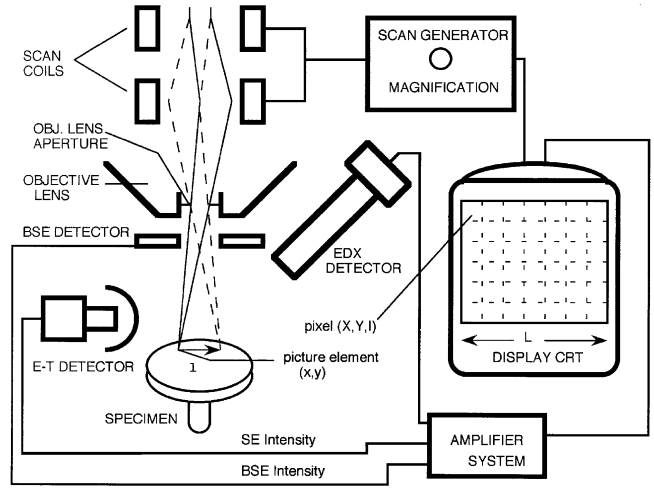


Figure 3. Schematic of the image formation system of the SEM. The electron beam is shown at two positions in time, scanned in the X direction to define the distance l on the specimen surface. Reproduction of this scan on the CRT, with a width L, gives a magnification $M = L/l$. The picture element is an area on the specimen surface, its location given by (x, y). Signal from this area (SE, BSE, x ray, or other) is transferred to the CRT and converted to an intensity value for pixel (X, Y, I). This simple image formation process is highly versatile in that the magnification is only a function of the scan length on the sample, and any signal may be used to provide the intensity variation, thus producing an image. (E-T, Everhart-Thornley.)

Image Formation

Image construction in the SEM is accomplished by mapping intensity of one of these signals (usually SE and/or BSE) from the specimen onto a viewing screen or film (Fig. 3). There is a point-to-point transfer of this intensity information, since the SEM scan generator simultaneously drives an electron beam across the surface of the specimen and an electron beam on the viewing cathode-ray tube (CRT) or recording device. The intensity information from the detector is translated to a gray scale, with higher signal intensity being brighter on the display. The image is therefore an array of pixels, with each pixel defined by x and y spatial coordinates and a gray value proportional to the signal intensity. Magnification of the image is defined as the length of the scan line on the monitor or recording device divided by the length of the scan line on the specimen. Magnification is therefore independent of the lenses. Magnification is adjusted by changing the size of the area scanned on the specimen while the monitor or film size is held constant. Thus a smaller area scanned on the sample will produce a higher magnification. This is a major strength of SEM imaging, and with optimized beam conditions and focus, the image magnification can be changed through its entire range without loss of image quality. A typical magnification range for the SEM is 10× to 100,000×; however, the higher magnifications are dependent upon acquiring sufficient resolution.

The region on the specimen from which information is transferred to a single pixel of the image is called a picture element. The size of the picture element is determined by the length of the scan on the specimen divided by the number of pixels in a line of the image. Picture element therefore refers to a position on the specimen, and pixel refers to the corresponding portion of the recorded image. The intensity of the signal is recorded for a single pixel of the image while the beam is addressed to the center of a single picture element. Consider a digital image that is 10 cm on a side and 100× magnification. The image represents an area on the sample that is 1 mm (1000 μm) wide, and the scale for the image is 1 cm = 100 μm. The picture element width is 1000 μm (the length of the scan on the specimen) divided by the number of pixels in the digital image. A common digital resolution for the SEM is 1024×1024, and thus the picture element width for the 100× image would be 0.98 μm. The picture element width would be 0.098 μm at 1000× and 0.0098 μm (9.8 nm) at 10,000×. The SEM image will appear in focus if the sampling volume is smaller in diameter than this picture element size. If the probe diameter or resultant sampling volume is larger than the picture element, blurring of the image occurs due to overlap of information from neighboring picture elements. This results in loss of resolution. When magnification is increased, the picture element size is reduced, and this overlap of information can result. The sampling volume is therefore a limiting factor on resolution for the SEM.

The SEM image conveys three-dimensional information due to the depth of field in the image. The depth of field depends upon the electron beam divergence angle, which is defined by the diameter of the objective lens aperture and the distance between the specimen and the aperture (Fig. 4). This angle is exceptionally small compared to that in a light microscope and results in a small change

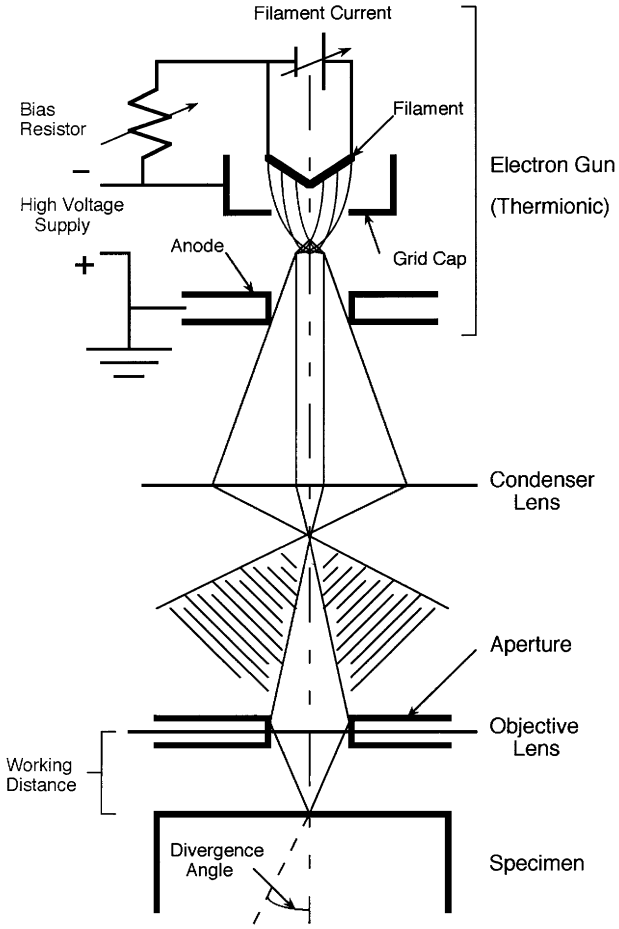


Figure 4. Schematic of the probe-forming system of the SEM. The thermionic electron gun is a triode system, with the initial crossover of electrons from the filament due to the field created by the filament, grid cap, and anode. This initial beam diameter is reduced by the condenser and objective lenses. A portion of the beam (shaded in the schematic) will not pass through the final aperture. Increasing the strength of the condenser lens will decrease the probe diameter but will also decrease the probe current. The objective lens will produce a smaller probe diameter when the working distance is reduced (the lens is stronger) and provide higher resolution. A longer working distance will reduce the divergence angle and therefore increase depth of field in the image.

in diameter of the beam with change in distance (depth) from the lens. This means that the sampling volume can likewise remain small over this range of depth. All points will appear in focus where the sampling volume diameter is smaller than the picture element. Depth of field is increased by reducing the beam divergence angle, which can be done by increasing the working distance or decreasing the aperture diameter. The excellent depth of field in SEM images is as important to the usefulness of scanning electron microscopy as the resolution and provides the topographic/morphological information well known to anyone who has admired any SEM image. Depth of field is typically between 10% and 60% of the field width.

Resolution

The resolving power of the SEM is dependent upon the sampling volume, since this sampling volume is the portion of the specimen from which the signal originates. It was stated under Image Formation above that the image appears in focus if the sampling volume is smaller than the picture element. This is also a statement of the resolution; if the sampling volume is smaller than the feature of interest, then the feature can be resolved, or distinguished from its surroundings. This simple model of resolution works well for most situations, but for higher magnifications the actual signal distribution within the sampling volume becomes critical, since the signal density is not uniform across the sampling volume. The sampling volume is a function of the probe (energy and diameter) and the specimen (composition and orientation). Resolution is therefore not an instrument constant, and it varies significantly with application. Microscope manufacturers quote a resolution value that is determined under ideal conditions on a specimen produced specifically for the purpose of testing high resolution. These values are used to characterize the instrument rather than to predict results expected from a given specimen. It is the job of the microscopist to prepare the sample and control the instrument in such a way as to provide the optimum or maximum resolution.

Modern SEMs with field emission electron guns (FESEMs) are capable of resolutions near 1 nm on appropriate samples using the SE signal. Tungsten filament instruments can routinely obtain resolutions of 3 to 4 nm on appropriate specimens. Samples that require special conditions, such as low voltage or low dosage, and samples that are inherently poor signal generators due to their composition may fall far short of this resolution. The signal from these specimens can often be enhanced by depositing a thin film of metal such as chromium or gold onto the surface. This increases contrast in the SE signal and allows for higher resolution imaging. Operating conditions that favor one signal, for instance, the high beam currents and energies used for x-ray analysis, can also reduce image resolution. High-resolution BSE images can be obtained by selecting only high-energy BSEs (low-energy-loss BSEs), since these have undergone few interactions with the sample and are thus from a smaller sampling volume.

PRACTICAL ASPECTS OF THE METHOD

SEM Instrumentation

The production of an electron probe requires an electron gun to emit and accelerate electrons and a series of lenses to de-magnify and focus the beam (Fig. 4). "Scanning" involves positioning of the probe on the specimen surface by electromagnetic (or electrostatic) deflection and simultaneously positioning a beam that carries signal intensity information onto a viewing device (camera or monitor). When the two beams are scanned in a regular pattern covering a rectangular area (raster), the signal output onto the viewing monitor becomes a map of signal intensity coming from the sample, and this "map" is the SEM image.

This can be done in an analog or digital manner. Scanning a small area on the sample and mapping the resultant signal intensity information onto a larger viewing surface result in magnification.

Three types of electron guns are in common use today: tungsten and lanthanum hexaboride thermionic electron guns, and the field emission gun. Tungsten and lanthanum hexaboride are used as filaments in thermionic electron guns, in which the filament (cathode) is heated by applying a current. Emission of electrons occurs at high temperatures and requires that the gun be operated in a high vacuum to avoid oxidation. The emitted electrons are accelerated to high energy by the application of a negative voltage to the cathode and a bias voltage to the grid cap with the anode grounded. This triode system accelerates the electrons and provides an initial focusing into a probe. Tungsten is popular for its low cost, ease of replacement, and high current yield, while lanthanum hexaboride provides a brighter source (higher current density) useful for higher resolution. In the field emission gun, a fine tip is placed near a grid and electrons escape the surface due to the high extracting field strength created by the first grid. A second grid provides the accelerating voltage. Field emission requires ultrahigh vacuum and is more expensive, but it provides much higher brightness with a resultant increase in resolution.

Scanning, positioning, and focusing of the electron beam after it leaves the electron gun is accomplished by the use of electromagnetic or electrostatic fields. A simple electromagnetic lens requires only a coil of wire wrapped concentrically around the optic axis. A current through this coil produces a field that deflects electrons toward the center of the lens. This field focuses the electrons into a smaller area, thus reducing the diameter of the electron beam produced by the electron gun. Lenses in modern instruments are far from simple, utilizing shroud and pole piece designs that maximize lens strength while minimizing lens aberrations (spherical, chromatic, and diffraction; see Goldstein et al., 1992).

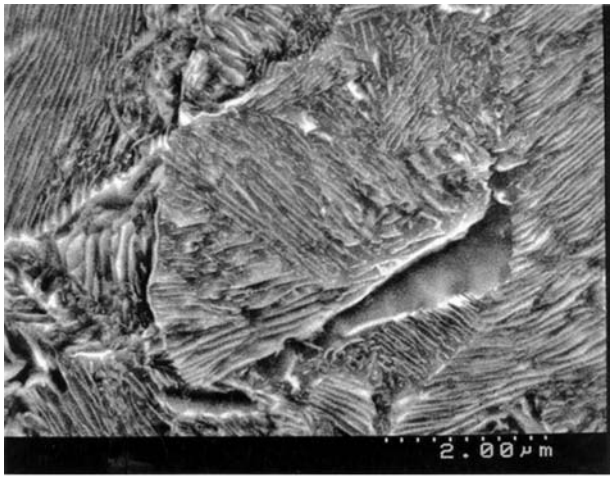
The "lenses" in the SEM all combine to produce a fine probe diameter. The condenser lens (often a system of several lenses adjusted by a single control) is used as the primary means of reducing the beam diameter produced by the electron gun. The condenser lens is usually directly below the electron gun and above a series of apertures necessary to define a small probe diameter (Fig. 4). Increasing the strength of the condenser lens will decrease the probe diameter but will also decrease the probe current due to interception of part of the beam by these apertures. This decrease in probe current with reduction of the probe diameter is a serious limitation in the SEM, since signal intensity is a fraction of probe intensity. High resolution requires a small probe diameter as well as a signal of high enough intensity that the differences in intensity are within the detector's range. It is this limitation that makes the brightness of the electron gun such an important consideration. The control knob for the condenser lens system may be labeled "spot size," "beam current," "probe diameter," or various other similar descriptive terms referring to its dual but interconnected effect on the beam.

The SEM objective lens is in basic respects no different than the condenser lens, in that it reduces and focuses the probe. Its position in the instrument, however, is such that its primary purpose is to reduce the probe to the smallest diameter at the surface of the sample (Fig. 4), which results in the clearest, or "focused," image. In the SEM there is no image formed behind the lenses as in the optical microscope. The lenses act on the electron beam rather than on the image. The objective lens aperture, often called the probe-forming aperture, defines the convergence (divergence) angle of the probe, which is important for the depth of field. Adjustment of the objective lens does not have a large effect on beam current as adjustment of the condenser lens does.

Electromagnetic or electrostatic fields that are not concentric with the optic axis will deflect the electron beam. This is utilized to center the beam onto the optic axis (alignment) and to scan or position the beam on the sample surface. Alignment coils are commonly utilized to position the electron beam coming from the electron gun along the optic axis of the instrument. These coils may also be used to "blank" the beam by directing it away from the sample toward a grounded aperture. The scan coils are usually located near or within the objective lens assembly and are used for many of the operator-controlled functions of the SEM. The scan coils control the position of the beam on the sample and are used for scanning of the beam to produce an image, determining magnification of the image, electronic shifting of the imaged area, and positioning a probe for x-ray analysis, among other functions. A final component of the optical system, the stigmator, utilizes paired coils arranged around the optic axis to eliminate astigmatism. Astigmatism arises due to irregularities in the electron-lens fields, which can be caused by aperture irregularities or contamination. The operator corrects for this irregularity by providing an equal but opposite irregularity with the coils of the stigmator. This operation should be done in conjunction with focusing, since it is necessary for producing a small, symmetric probe diameter on the sample surface.

Contrast and Detectors

Microstructure for most solids can be imaged in the SEM with minimum specimen preparation, although interpretation of the microstructure can be aided by careful attention to specimen preparation, such as the polishing and etching of metals (Fig. 5). The SEM offers a choice of detectors and operating parameters that can be taken advantage of to produce a variety of images showing microstructure. Contrast in the SEM image is the difference in intensity, or brightness, of the pixels that make up the image. The difference in intensity represents a difference in signal from corresponding picture elements on the specimen. This signal intensity is dependent upon the signal generated within the specimen and the type of detector used. It is therefore necessary to recognize both the beam-specimen interaction that produces the signal and the detection and amplification process that converts the signal intensity to image brightness. Contrast components of the signal electrons—their number, trajectory,



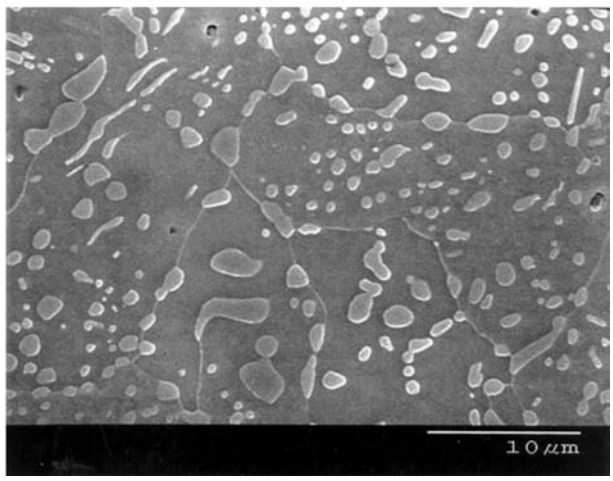


Figure 5. Microstructures of two steel samples of identical chemical composition (0.97% carbon by weight) but with different cooling and annealing histories. Scanning electron microscopy reveals the characteristic structure that makes the steel in the top image extremely hard and brittle, while the steel in the bottom image can be used for springs and is comparatively soft (Killick, 1996).

and energy—are a useful means to describe and understand how this information from the specimen is translated to an intensity value on the image. The number component is expressed by higher numbers of electrons from the specimen resulting in higher brightness on the image. Trajectory is equally important, since detectors have limited size and shape and therefore detect only a portion of the signal emitted. A high signal generated in the specimen that is directed away from the detector will result in a low detected signal. The energy component is also critical, since the ability to collect and detect the signal is a function of the signal energy. The combination of these three components (number, trajectory, and energy) creates a unique situation for each probe-specimen-detector combination. The most widely used images produced by SEM imaging are the SE image, the BSE compositional image, and the BSE topographic image. It is important to note that many images will have a

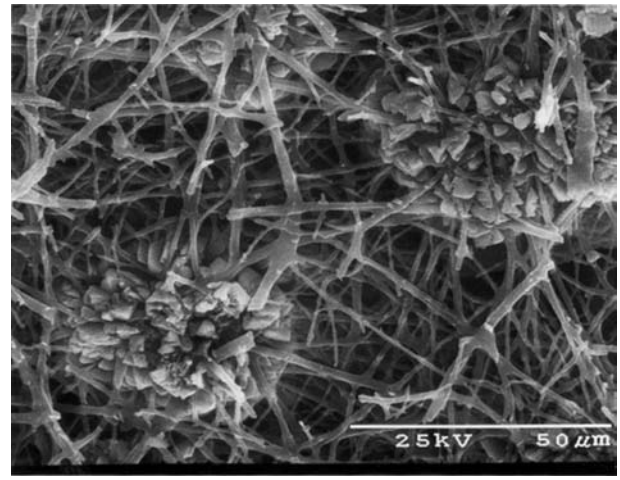


Figure 6. Secondary electron image of the internal structure of a dove eggshell. Note the bright edges of the fibers (collagen) and the differing brightness from crystal faces of the calcium carbonate.

combination of contrast components that contribute to the final image and that the description here has summarized only the most common images.

The most popular and useful detector for observing surfaces and morphology is the Everhart-Thornley (ET) detector, composed of a scintillator, collector, light pipe, and photomultiplier tube. Electrons striking the positively biased scintillator are converted to a burst of photons that travel through the light pipe to the photomultiplier tube, where the signal is amplified. The collector is a grid or collar around the scintillator that is positively biased to draw the low-energy SEs toward the detector. The image produced from the ET detector with a positively biased collector is dominated by SE contrast, and this detector is often referred to as the SE detector. The SE image is dominated by surface features due to the shallow escape depth of SEs and the dependence of the SE signal on the incidence angle of the probe with the specimen (Fig. 6). Surfaces tilted away from the normal to the beam allow for more SE escape. A noticeable feature of the SE image is that edges appear bright. This is due to the escape of SEs from the top and side surfaces when the interaction volume intercepts these two surfaces. A surface facing away from the detector can still appear illuminated, since some portion of the SEs can be detected due to the bias on the collector. The traditional SEM instrument has an ET detector within the specimen chamber. The ET detector can also be placed within the optical column of the SEM, and with an alteration to the lens design, the SE signal can be trapped within the field of the objective lens and collected by the biased detector. The SE image from the ET detector, often coupled with enhancement of the signal by deposition of metal thin films on the specimen, results in high-resolution SEM images.

Common BSE detectors are the solid-state detector and the scintillator detector. These detectors are not responsive to low-energy electrons and therefore exclude the SE signal. Both detectors are designed such that they can be placed above the specimen, where the geometric collection

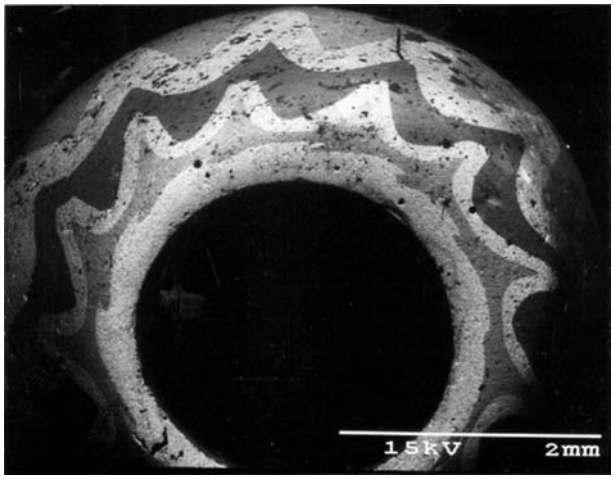


Figure 7. Backscattered electron image of a glass trade bead. The bands are different colored glasses. In this image, contrast is due to the different intensities of BSEs, which are due to differences in the average atomic number of the glasses. X-ray analysis coupled with this image provides a complete characterization of the specimen.

efficiency can be optimized to collect a high proportion of the BSE signal. The solid-state detector is based on the principle that high-energy electrons striking the silicon produce electron-hole pairs in the material. A bias on the detector separates the electrons and holes, and the resultant charge can be collected. The intensity of the charge is proportional to the energy of the incoming electrons (in silicon it takes $\sim$3.6 eV to produce one electron-hole pair) and the number of electrons striking the detector. The solid-state detector is usually an array of detectors that are symmetrically placed around the optic axis, and the signals can be mixed in additive or subtractive modes to enhance different contrast components. The scintillator detector relies on the interaction of BSEs with the detector to produce photons, similar to the ET detector. Unlike the ET detector, however, SEs are not detected due to their lower energy. The geometry or position of the detector itself can be altered to select for contrast components.

Both of these BSE detectors can be operated in two modes. In the first mode, placing the BSE detector above the specimen and symmetrically around the optic axis will eliminate the trajectory component of contrast by collecting all trajectories equally. This produces the BSE compositional image, in which the "number" component of contrast predominates and the signal intensity increases with increasing average atomic number of the sample (Fig. 7). This imaging strategy works best with flat, polished surfaces where information on the distribution of phases is easily interpreted. In the second mode, the trajectory of the signal can be accentuated by nonsymmetric detector geometry. This is accomplished in the solid-state detector by subtracting the signal from one part of the array from the other or by differential weighting of the signals from different sides of the specimen chamber. The scintillator detector can be partially removed from its position directly above the sample to create this same enhancement of trajectory differences. The resulting BSE topographic image shows in high contrast minor irregularities on the surface. Figure 7 exhibits this topographic contrast, with surface pitting shown in sharp relief, superimposed upon a background of compositional contrast. Since the BSE trajectories are essentially straight due to their high energy, this image also shows "shadowing" where material between the detector and a surface region will block the signal reaching the detector. The ET detector can also be used to produce this BSE topographic image, but its small size and position away from the specimen leads to poor collection efficiency and therefore a noisy image.

Special Techniques and Recent Innovations

The SEM is an exceptionally adaptable and versatile instrument, and many special application areas have been developed. Its unequaled depth of field can be enhanced by stereographic techniques, in which two images of the same area are recorded at different specimen tilt angles (Gabriel, 1985). The resulting stereopair can be viewed with the aid of filters or a stereo viewer to produce a three-dimensional image. Measurements of distances in three dimensions are possible if the initial geometry is recorded, and software programs are available to simplify the task. Metrology, particularly the measurement of critical dimensions in microelectronic devices, is possible with refined SEM scanning systems and appropriate standards and control. The SEM, by nature, is a bulk analysis technique, and many *in situ* experiments can be devised. Electronic testing in the SEM has the benefit of precise placement of the probe (beam), control of the beam voltage and/or current, and avoidance of mechanical damage and impedance loading associated with mechanical probes (Menzel and Buchanan, 1987). Physical testing of materials within the SEM chamber is also possible, with heating, cooling, strain, and fracture stages custom built for the microscope.

The diffraction of electrons by a crystalline sample can lead to electron channeling patterns in the SEM image (Joy et al., 1982; Newbury et al., 1986) or Kikuchi patterns in the BSE distribution (Adams et al., 1993; Michael and Goehner, 1993; Thaveeprungsriporn et al., 1994; Goehner and Michael, 1996). Electron diffraction techniques in the SEM can be used to identify phases, determine their orientation, and study crystal damage and defects (Dingley, 1981).

Variable-pressure SEM is a technique in which the specimen chamber pressure is controlled independently of the SEM column (Mathieu, 1996). The introduction of air (or other gases) into the specimen chamber neutralizes charging on the surface of insulating materials, allowing them to be imaged and analyzed (with decreased resolution) without having to deposit a conducting thin film (Fig. 8). The higher chamber pressures also allow for the observation of materials with high vapor pressures, including biological material, without extensive sample preparation to remove volatile components. A special form of variable-pressure scanning electron microscopy is referred to as environmental scanning electron microscopy, in which

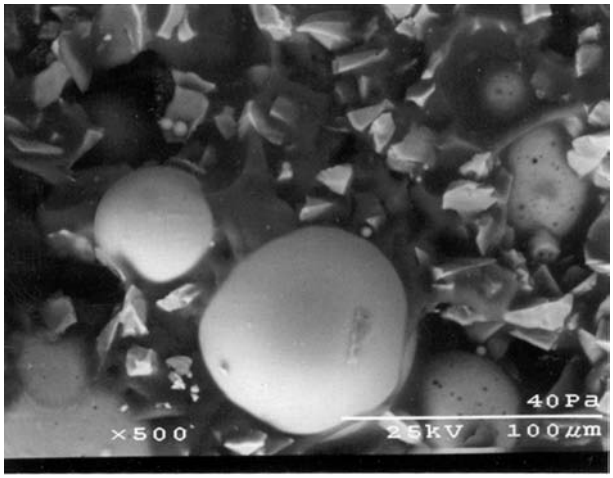


Figure 8. Variable-pressure SEM image of an abrasive. The abrasive is composed of an epoxy resin, silicon nitride particles, and ceramic balloons. The ability to control the pressure in the chamber of the SEM (40 Pa for this image) allows for operation in an atmosphere capable of eliminating charge buildup on the specimen. This specimen is also a problem for conventional SEM due to outgassing of the epoxy and ceramic balloons. (Image by David Johnson, University of Arizona.)

chamber atmosphere can also be altered (Danilatos, 1988, 1993).

The FESEM deserves special mention due to its enormous impact on the field. The field emission (FE) electron gun allows for the creation of an exceptionally bright (small with high-current-density) electron beam. This in itself has made scanning electron microscopy competitive with transmission electron microscopy for materials characterization in the nanometer range (Fig. 9). The FE source is often coupled with a special lens design such

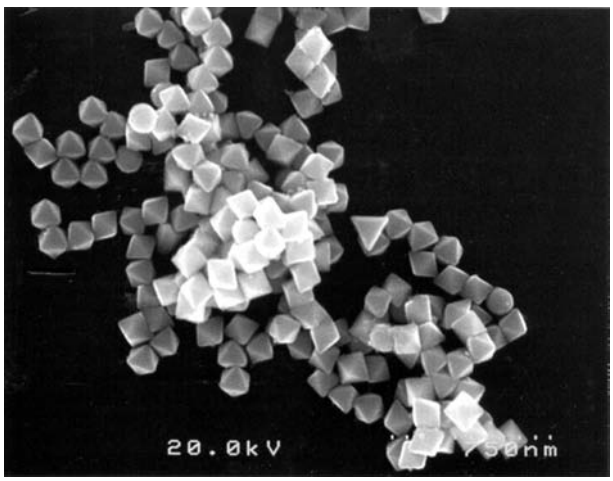


Figure 9. Field emission SEM image of gold particles in dichromic glass taken at 400× magnification. The image is at the upper range of magnification for an SEM with a thermionic electron gun but a normal range for the FESEM. The particle geometry and size are apparent, and also some irregularities on the surfaces and where two particles join. (Image by Toshinori Kokubu and G. Chandler, University of Arizona.)

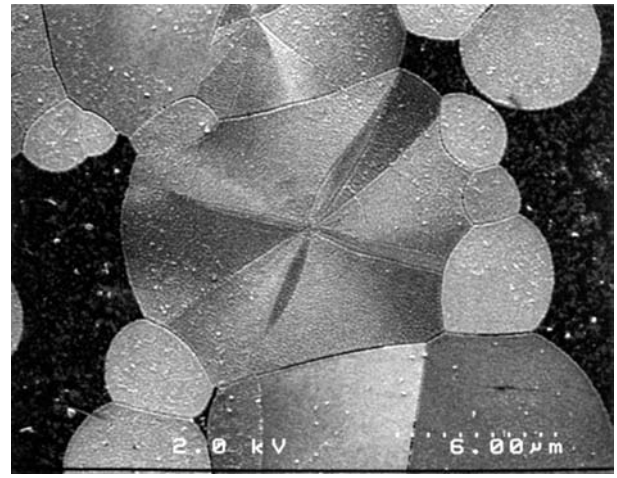


Figure 10. Low-voltage image of PZT on platinum. The low accelerating voltage used here (2 kV) gives excellent surface contrast of this thin material. (Image by Mike Orr and G. Chandler, University of Arizona.)

that the specimen is in the field of the lens and the detector is within the column rather than the sample chamber. The FESEM also has the ability to produce a small probe diameter at low voltage, opening the way to many applications that were difficult with thermionic electron guns (Greenhut and Friel, 1997). Low-voltage scanning electron microscopy has the advantage of revealing more details of the surface since there is less penetration of the beam into the specimen (Fig. 10) and is less damaging to the material. The SE yield from the specimen is a function of voltage, and at lower beam voltages the specimen reaches a point at which emission of secondary and backscattered electrons equals the beam current. This is a unique voltage for each material at which there is no specimen charging, and the surface can be imaged without a conductive coating. This phenomenon has also been used as a voltage contrast mechanism in materials such as copolymers and ceramic composites that were traditionally difficult to observe with the SEM.

METHOD AUTOMATION

The SEM used as a research tool typically requires the operator to make many value judgments during a viewing session. There are many situations, however, where automated routines are useful. Instruments with automatic electron gun start-up, automatic focus and astigmatism correction, and automatic contrast and brightness may be the choice for situations in which repetitive sampling is necessary, such as in quality control. Systematic sampling, in which a motorized stage positions the specimen below the beam or the beam is moved to positions on the specimen, is available and useful for precise, repetitive analysis. Automatic search routines are available that can identify points of interest with the SEM image and return to those areas for compositional analysis with a spectrometer. The SEM is commonly fitted with automated sample handling in industrial settings such as

microelectronic fabrication plants. Instrument manufacturers offer extensive options for automated specimen preparation, handling, manipulation, sampling, instrument operation, and data collection, presentation, storage, processing, and analysis.

DATA ANALYSIS AND INITIAL INTERPRETATION

Interpretation of SEM images is dependent upon the geometry of the probe, specimen, and detector. The specimen appears to the viewer as if looking down the optic axis, with the illumination coming from the detector. This is most clearly seen in the BSE topographic image but is also important for the SE image. The most common imaging mode utilizes SEs for information on the surface morphology (Fig. 6). This SE image is often an extremely complex convolution of contrast-producing factors but is amazing for its ease of interpretation. Surfaces at different angles to the probe produce different brightness levels and edges appear brighter. "Holes" are darker while raised areas are brighter. The image generally looks like a black-and-white optical view of the surface.

The BSE signal can also be used to show surface topography but is perhaps more often used to show atomic number contrast, in which areas of higher average atomic number produce more BSEs, and thus appear brighter, than lower atomic number regions. Figure 7 shows this contrast clearly on a glass trade bead. The bead was produced by wrapping layers of different colored glass around a core, forming the pattern seen in the image. The contrast in this image is not due to color differences but is produced by different intensities of electrons backscattering from the layers with different average atomic number.

The SE image is sensitive to changes in potential on (or near) the surface of the specimen. This can create problems when viewing insulators, but it can also be used to enhance features of the specimen. In Figure 11, the probe energy was set to create a buildup of electrons within an insulating layer below the surface of the specimen. This charged area then emits more SEs than surrounding, grounded material and appears brighter. The same example can be given to demonstrate that features below the surface can contribute to the image contrast. The operator can control the depth that the beam penetrates by changing the energy of the electron probe through changing voltage.

Specimens with multiple phases or grains, twinning, and other microstructure not associated with composition may exhibit contrast due to causes such as differences in crystal orientation or magnetic domain. These contrast mechanisms are superimposed upon the SE or BSE contrast, and careful attention to sample preparation and instrument set-up is necessary for interpretation (Newbury et al., 1986).

Imaging with the SEM has benefited greatly from the incorporation of digital techniques. Multiple low-dosage frames can be averaged to produce a low-noise image on specimens that are sensitive to electron dosage. Digital filtering techniques bring out details in the signals, which can vary over many orders of magnitude, and are not visible when converted directly to a visible gray scale as in conventional photography. The independent collection and storage of multiple signals (SE, BSE, and x ray) from a digital array can greatly enhance the productivity and utility of the SEM and allows for postsampling manipulation and study of the results. Digital images from the SEM can be analyzed with numerous commercially available and free-ware software products.

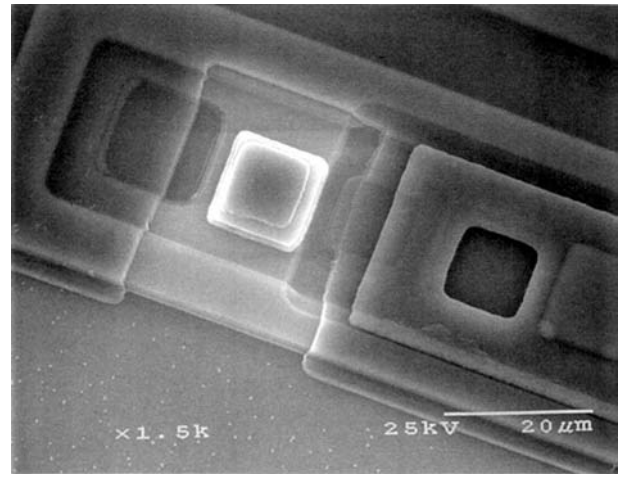


Figure 11. Secondary electron image of microelectronic device. The accelerating voltage of the electron beam was set such that a large portion of the beam electrons were absorbed within an insulating material below the surface. The resulting buildup of charge creates a potential that causes an increased emission of SEs, thus providing a way to highlight the structure. (Image by Ward Lyman, University of Arizona.)

SAMPLE PREPARATION

Surface topology, such as in fractography, may require no sample preparation if the specimen can withstand the low chamber pressure and electron beam bombardment of the SEM. The microstructure may be greatly enhanced, however, by such methods as polishing and selective etching of the surface, as with the metals in Figure 5. Most SEMs have sample chambers with limited dimensions, and the specimen must be affixed to a stage holder for orientation and manipulation within the chamber. Conductive adhesives with low vapor pressure or mechanical devices are used to mount the specimen. Special chambers and stages are available to suit most needs.

The most common form of sample preparation for the SEM is the deposition of a metal thin film onto the specimen surface. Vacuum evaporation and ion sputtering of metals are common methods of depositing these thin films (Goldstein et al., 1992). The metal thin film provides electrical conductivity, enhances the signal (if a higher-Z metal is used), and may add strength to the specimen. If x-ray spectroscopy is coupled with scanning electron microscopy, the metal film may be replaced with a carbon thin film or deposited to a thickness that is not detectable by the spectrometer. The deposition of very thin (1-nm)

films of metal may be necessary to obtain the highest resolution from an FESEM.

Materials with volatile or high-vapor-pressure components, such as biological specimens, must be prepared for viewing in the SEM. Sample preparation is often tedious and extensive, and special protocols are required to preserve the material's structure while eliminating the unwanted components. Elimination of water from biological specimens, for example, may employ freeze drying, critical point drying, or chemical means. Cryo-microscopy, in which a liquid nitrogen- or liquid helium-cooled stage is employed, is an option for especially challenging materials. Polymers often require some protection from the beam, and staining techniques similar to those employed for biological materials can increase contrast in the image.

Specimens such as living organisms, museum pieces, or very large objects can be surface replicated using polymer films or dental casting resins if they cannot be placed within the SEM chamber. These materials will closely replicate the surface features of the specimen and, after hardening, can be coated with a metal thin film for SEM viewing.

Extensive expertise on sample preparation is available through the Microscopy Society of America Listserver (see Internet Resources).

SPECIMEN MODIFICATION

Scanning electron microscopes vary widely in the specimen chamber pressure but are traditionally in the high-vacuum range due to requirements of the electron optics. This limits materials that can be analyzed using the SEM; however, extensive sample preparation protocols have been developed for this purpose. The variable-pressure and environmental SEMs are designed to allow operation at higher pressures (still vacuum) and can therefore eliminate some sample preparation.

The electron beam can deposit significant energy within the specimen, and damage may occur due to heating, buildup of electrical charge, or breaking of bonds. Sample preparation, in particular coating with thin metal films, can eliminate many of these problems. Electron beam lithography in integrated circuit production is an example of where this damage is used to produce a pattern in a thin polymer film. Controlled scanning of the beam over the surface produces a pattern within the film. It is then exposed to a mild etch, so that the regions "damaged" by the electron beam etch at a different rate than regions not exposed, and the pattern becomes a mask for subsequent deposition procedures.

Specimens observed in the SEM often become contaminated with a thin film of carbon deposited on the area being scanned with the beam. This is often explained as being due to oil vapor from the vacuum system entering the specimen chamber. This will occur even in oil-free systems if the specimen itself has carbon compounds on its surface. Under electron beam bombardment, these molecules are broken down into constituent species, and while oxygen and nitrogen are removed by the vacuum system, the carbon remains on the surface. Migration of carbon compounds toward the area under electron beam bombardment creates a buildup of carbon.

PROBLEMS

The sense of perspective described above (see Data Analysis and Interpretation) can be critical for specimens with simple surface geometry, since topography can appear inverted to some observers if the image is rotated 180°. Three-dimensional images produced by viewing stereopairs also require that the operator control this geometry of probe, specimen, and detector and in addition record sample tilt and tilt axis alignment.

Image quality in the SEM usually refers to the amount of noise, or graininess, of the image. There is a trade-off in SEM imaging between resolution and image quality in which higher resolution is only obtained by accepting noisier images. This situation is due to the SEM optics described earlier (see Practical Aspects of the Method), in which a stronger condenser lens results in a smaller probe diameter and a decreased probe current. The decrease in probe current will result in a decrease in signal being generated. Noise is inherent in any electronic system, and decreasing the signal will decrease the signal-to-noise ratio. Thus, when one decreases the probe diameter to obtain higher resolution, the signal must be amplified more, and the noise is apparent in the image. This results in grainy images. The use of higher-brightness electron guns, especially the FE sources, makes dramatic improvements in the image quality at higher magnifications by increasing the current density (brightness) of the probe.

The scan length on the specimen can be calculated from two measurements and used to provide an estimate of the magnification. Working distance is the distance between the specimen and the objective lens. Assuming the image is in focus, the current supplied to the objective lens can be sampled to estimate the working distance. A similar reading of the scanning coil current will give an estimate of the scan deflection angle. The length of the scan line on the sample can then be calculated from this angle and the working distance. The magnification readout on most instruments is a calculated value derived from these two measurements, and instruments vary significantly in their accuracy. This problem can easily be overcome by the use of standards to calibrate magnification. Images are also subject to distortions due to irregular scanning, variations in the working distance due to sample topography or tilt, and the projection of the scan (angular) onto a flat surface (film). The projection distortion lessens as magnification increases (scan angle decreases), and distortions are normally apparent only when viewing highly symmetrical samples such as grids. Use of the SEM for critical dimension measurement therefore requires special instruments and techniques.

Traditionally, SEM images were produced at high voltages, where probe electrons could build up a significant charge on insulating materials. This surface charge can manifest itself in many ways in the image. A negatively charged surface usually appears brighter due to increased

emission of SEs from the region. The positioning of the beam on the surface may be altered by a nonuniform charge, creating a distorted image. Beam energy is altered by specimen bias from charge buildup and will affect imaging and spectroscopy. Particles on the specimen surface that collect charge can produce a local field that deflects the SE signal away from the detector, resulting in a "shadowed" region. Sporadic discharge of electrons from a charged surface may create bright dots and lines on the image that are not associated with the sample topography. Due to these many detrimental effects of charging, specimens are often coated with thin films of heavy metals (or carbon if x-ray analysis is desired) to provide a conductive path to ground. Newer instruments, particularly the FESEM, have the ability to produce images at low accelerating voltages. Specimen charging can be reduced or eliminated with control of the accelerating voltage, and positive charging is possible when the intensity of secondary and backscattered electron emission exceeds the probe current. Constructive use of sample charging is used to create contrast in materials such as ceramics and microelectronics.

Contamination on the surface of the specimen plays a critical role in SE images, since the signal is influenced most by the surface characteristics. The most troublesome contamination is the buildup of carbon on the surface. Carbon compounds from oil vapor in the vacuum pumps, exposure to atmosphere, specimen handling, and solvents used to clean the specimen can migrate toward the area that the beam is striking. The high-energy electrons from the beam can break bonds within these materials, producing a carbon buildup on the surface and gases that are easily removed by the vacuum pumps. It is common to see this carbon buildup on the surface of the specimen while viewing the live image. This contaminant layer obscures and blurs the surface detail, and the image becomes darker, since much of the signal is within this low-Z material on the surface. Contamination becomes worse at lower voltages. Careful attention to sample preparation and handling, rapid signal collection, and the use of clean vacuum techniques (see GENERAL VACCUM TECHNIQUES) will combat contamination problems.

LITERATURE CITED

Adams, B. L., Wright, S. I., and Kunze, K. 1993. Orientation imaging: The emergence of a new microscopy. *Met. Trans.* 24A:819–831.

Danilatos, G. D. 1988. Foundation of environmental scanning electron microscopy. *Adv. Electron. Electron Phys.* 71:109–250.

Danilatos, G. D. 1993. Introduction to the ESEM instrument. *Microsc. Res. Tech.* 25:354–361.

Dingley, D. J. 1981. A comparison of diffraction techniques for the SEM. *SEM/1981* 4:273–286.

Gabriel, B. L. 1985. SEM: A Users Manual for Materials Science. American Society for Metals, Menlo Park, Ohio.

Goehner, R. P. and Michael, J. R. 1996. Phase identification in a scanning electron microscope using backscattered electron Kikuchi patterns. *J. Res. NIST* 101:301-308.

Goldstein, J. I., Newbury, D. E., Echlin, P., Joy, D. C., Romig, A. D., Lyman, C. E., Fiori, C., and Lifshin, E. 1992. Scanning Electron Microscopy and X-Ray Microanalysis. Plenum, New York.

Greenhut, V. A. and Friel, J. J. 1997. Application of advanced field emission scanning electron microscopy to ceramic materials. *USA Microsc. Anal.* March:15–17.

Joy, D. C. 1984. Beam interactions, contrast, and resolution in the SEM. *J. Microsc.* 136:241–258.

Joy, D. C., Newbury, D. E., and Davidson, D. L. 1982. Electron channeling patterns in the scanning electron microscope. *J. Appl. Phys.* 53(8):R81–R119.

Killick, D. 1996. Optical and electron microscopy in material culture studies. *In* Learning from Things (W. D. Kingery, ed.). pp. 204–230. Smithsonian Institution Press, Washington.

Leamy, H. J. 1982. Charge collection scanning electron microscopy. *J. Appl. Phys.* 53(6):R51–R80.

Mathieu, C. 1996. Principles and applications of the variable pressure scanning electron microscope. *USA Microsc. Anal.* Sept.:15–16.

Menzel, E. and Buchanan, R. 1987. Noncontact testing of integrated circuits using an electron beam probe. *J. SPIE* 795:188–200.

Michael, J. R. and Goehner, R. P. 1993. Crystallographic phase identification in the scanning electron microscope: Backscattered electron Kikuchi patterns imaged with a CCD-based detector. *MSA Bull.* 23:168–175.

Newbury, D. 1976. The utility of specimen current imaging in the SEM. *SEM/1976* 1:111–120.

Newbury, D. E., Joy, D. C., Echlin, P., Fiori, C. E., and Goldstein, J. I. 1986. Advanced Scanning Electron Microscopy and X-Ray Microanalysis. Plenum, New York.

Thaveeprungsriporn, V., Mansfield, J. F., and Was, G. S. 1994. Development of an economical backscattering diffraction system for an environmental scanning electron microscope. *J. Mater. Res.* 9:1887–1894.

Yacobi, B. G. and Holt, D. B. 1986. Cathodoluminescence scanning electron microscopy of semiconductors. *J. Appl. Phys.* 59(4): R1–R24.

KEY REFERENCES

Goldstein et al., 1992. See above.

Provides a very complete coverage of basics and written for a wide audience. It includes chapters on instrumentation, basic imaging principles, electron optics, and sample preparation. The authors have extensive experience in teaching electron microscopy to thousands of students and have addressed the novice as well as providing more advanced sections.

Newbury et al., 1986. See above.

A companion volume to the previous textbook and provides in-depth coverage of specific topics. Topics include modeling of electron beam-sample interactions, microcharacterization of semiconductors, electron channeling contrast, magnetic contrast, computer-aided imaging, specimen coating, biological specimen preparation, cryomicroscopy, and alternative microanalytical techniques.

Lyman, C. E., Newbury, D. E., Goldstein, J. I., Williams, D. B., Romig, A. D., Armstrong, J. T., Echlin, P., Fiori, C. E., Joy, D. C., Lifschin, E., and Peters, K.-R. 1990. Scanning Electron Microscopy, X-Ray Microanalysis, and Analytical Electron Microscopy. A Laboratory Workbook. Plenum, New York.

A detailed laboratory workbook with questions, problems, and solutions to the problems given for each of the exercises.

Although written as a guide for a course, this is a helpful reference for those who will be operating the instrument, as it provides detailed, step-by-step instructions for obtaining specific results from the instrument. Exercises include basic imaging, measurement of beam parameters, image contrast, stereomicroscopy, BSE imaging, transmitted electron imaging, low-voltage, high-resolution, SE signal components, electron channeling, magnetic contrast, voltage contrast, electron beam-induced current, environmental scanning electron microscopy, and computer-aided imaging.

INTERNET RESOURCES

The Microscopy and Microanalysis WWW server site: http://www.amc.anl.gov/

This site is hosted by Argonne National Laboratory and is a complete guide to the Internet for microscopy. The site is also home of the Microscopy Society of America listserver, providing a forum for novice and expert microscopists to ask questions of their colleagues. Other links at this site include the MSA software library, ftp sites, a list of upcoming meetings, conferences and workshops, and lists of national and international www sites. This provides access to educational, government, commercial, and society resources.

APPENDIX: GUIDELINES FOR SELECTING AN SEM

The selection of a SEM can be quite involved, with ten companies offering new instruments, and some of these having as many as 15 models from which to choose. The market is highly competitive, and a thorough search by the purchaser will be very beneficial. Many options may be available from companies specializing in custom equipment rather than from the instrument manufacturer, and a potential purchaser should also contact these vendors for information. A quick look at the Microscopy Society of America web pages and a few days on their list server should give one a feeling for the variety of options available. Table 1 lists some features that a buyer should check, although it is not possible to include all options, adaptations, add-on systems, and interfaces available. The list of criteria allows for a comparison between microscopes, and the buyer must use some ranking system for the individual items, such as imperative, useful, a nice feature, or unnecessary. A cost analysis should also be conducted, and the many laboratories that offer SEM services are an alternative to purchase. The determination of employee needs, training, maintenance and repair costs, and instrument reliability/lifetime should be a part of consideration to purchase and are not covered here. A list such as the one below must be customized for each individual purchase, and performance of the microscope on real samples should always be the final test. Most manufacturers will arrange for a demonstration of their product, and the buyer should prepare well for this demonstration, including bringing in your most problematic specimens. Some general notes on the major systems of the SEM are included here for background information on the criteria.

Vacuum System and Specimen Handling

The vacuum system for the SEM is designed to meet the requirements of the electron gun and probe-forming system, while the buyer must decide if the system is compatible with requirements of the specimen. Traditional SEM pressures are on the order of 1×10^{-6} torr, with the electron gun and specimen chamber kept at the same pressure. Modern instruments may offer differential pumping systems such that ultrahigh vacuum can be obtained for a field emission source, or low-vacuum "environmental" chambers can be maintained for the specimen chamber, while other portions of the system are held at traditional high-vacuum levels. The cleanliness of the vacuum system and component gases may be equally important to a particular application as the pressure. Oil-free pumping systems, special sample introduction systems, inert gas back-filling, or environmental chambers with control of gas composition are examples of the options available.

The specimen-handling system is closely related to the vacuum system and should therefore be designed/chosen with the specimen and the vacuum in mind. Most instruments are equipped with specimen stages that are adequate for manipulating a specimen within the chamber but will not provide the fine control necessary for many special techniques. Stage options are available for eucentric geometry, large samples, *in situ* experiments, fine movement, cryo-microscopy, nonstandard detection systems, and other special requirements. Introduction of a specimen into the instrument can also be accomplished in a variety of ways, from manually opening the entire chamber to completely automated handling and transfer from one instrument/device to another under, e.g., vacuum or at low temperature.

Electron Gun

Three electron gun types are commonly available today. The tungsten filament thermionic source is inexpensive, requires little in the way of vacuum systems, produces a stable beam current, and is appropriate for most routine SEM imaging. The lanthanum hexaboride thermionic source offers higher brightness than the tungsten source, with resulting higher resolution. It requires lower operating pressures, and although it has a longer lifetime, replacement cost is high. The field emission electron gun is much more expensive but produces the brightest and smallest probe diameter and thus the highest resolution. The FE electron gun is also the choice for low-voltage applications.

Imaging System

The manufacturer will usually quote a resolution for the microscope that is a "best case" situation and is normally obtained using the SE signal on a high-contrast, standard specimen. The resolution of an image from the SEM is dependent on the beam parameters (accelerating voltage, beam diameter), the signal and detector used (SE, BSE, x ray), and the characteristics of the specimen (composition, structure, preparation). It is therefore imperative

Table 1. Some Criteria for Evaluation of a Scanning Electron Microscope

- Vacuum system
  - Types of vacuum pumps
  - Chamber operating pressure(s)
  - Pump-down time/sample exchange time
  - Contamination rate (on your sample)
  - Vacuum meter/status indicator light
  - Safety interlock to high-voltage supply
  - Valve closure on power failure
  - Column isolation valve
  - Gun isolation valve
  - Differential pumping
  - Inert gas back-filling of chamber
  - Pump filters or venting system
  - Sample introduction system/airlock
- Chamber/stage
  - Maximum specimen size
  - Range of movement (X,Y,Z)
  - Range of rotation and tilt
  - Eucentric point or points
  - *In situ* capabilities
- Vacuum feed-through for external controls
  - Location and number of ports
  - Energy dispersive spectrometry (EDS) angle/working distance
  - Simultaneous BSE/SE/EDS
  - Automation
  - Vibration isolation
- Electron gun and probe-forming system
  - Type of gun
  - Brightness
  - Accelerating voltage range/steps
  - Minimum probe diameter (specify working distance, probe current, accelerating voltage)
  - Probe current range
  - Special requirements/maintenance needs
  - Adjustable grid cap
  - Adjustable anode
  - Gun alignment (mechanical/electromagnetic)
  - Column liner
  - Column aperture
  - Aperture alignment (mechanical/electromagnetic)
  - Current-measuring device
  - Beam stability (keV and current)
- Imaging system
  - Image quality (your sample)
  - Magnification range/steps
  - Scan speeds (view/photo)
  - Digital resolution (view/photo/store)
  - Video out/peripheral devices
  - Waveform monitor for photo settings
  - Filament image
  - Channeling patterns
  - Viewing screen(s)
  - Scanning mode options
  - Electromagnetic image shift
  - Spot mode
  - Selected area
  - Line scan
  - Dual image/split screen
  - Dual magnification
  - Scan rotation
  - Image data display
  - Micrometer marker/ranges and format
  - Data entry (text on image)
  - Automatic features (manual override available)
- Focus
  - Stigmation
  - Start-up
  - Gun bias
  - Contrast/brightness
- Image modes and detectors
  - SE detector(s) and position
  - SE resolution (kV/working distance)
  - BSE detector
  - BSE resolution (kV/working distance)
  - Tested resolution (on your samples)
    - SE (kV/working distance)
    - BSE (kV/working distance)
  - X ray (kV/element line)
  - BSE Z resolution (atomic number)
  - Specimen current amplifier
- Additional options and customizing
- Live image processing
  - Gamma correction
  - Y-modulated image
  - Differential filter
  - Dynamic focus
  - Tilt correction
- Camera system and film type(s)
- Digital image storage system
- Computer control/interface
- Other
  - Training in operation and maintenance
  - Installation
- Warranty
  - Service contract
    - Response time
    - Parts covered
    - Phone consultation
  - Parts availability
  - Guarantee of resolution/performance
  - Other instruments in area
  - Documentation, instructions, schematics
  - Tool kit
  - Internal diagnostics
  - Ease of routine maintenance/repair
- Ease and comfort of use
- Environmental needs/stability
  - Cooling water
  - Gases
  - Power
  - Vibration
  - Fields
  - Temperature/humidity
- EDS compatibility
  - Apertures
  - Working distance/detector geometry
- Compatibility with other systems/instruments
- Availability of other options

that the instrument be tested with "real" specimens if resolution is a critical factor in the selection.

All modern instruments will offer a wide array of capabilities to the user and will meet basic needs. Criteria such as magnification range and scanning speeds are important to consider, but the operator's preference and attention to most common usage should guide choices. Several types of detector are available and may be offered as options by the manufacturer or as an add-on system from a third party. Probably the most rapidly changing portion of the SEM is the system for recording, storing, and presenting images. Many laboratories are currently using a dual-system approach, with traditional photography and digital imaging available on the instrument. There are no standards for software control, digital image format, image data format, or printing. The SEM control and recording systems are rapidly evolving, and one should pay close attention to possible interfacing and compatibility problems with existing or desired equipment, computers, and software.

GARY W. CHANDLER
SUPAPAN SERAPHIN
University of Arizona
Tucson, Arizona

TRANSMISSION ELECTRON MICROSCOPY

INTRODUCTION

Transmission electron microscopy (TEM) is the premier tool for understanding the internal microstructure of materials at the nanometer level. Although x-ray diffraction techniques generally provide better quantitative information than electron diffraction techniques, electrons have an important advantage over x rays in that they can be focused using electromagnetic lenses. This allows one to obtain real-space images of materials with resolutions on the order of a few tenths to a few nanometers, depending on the imaging conditions, and simultaneously obtain diffraction information from specific regions in the images (e.g., small precipitates) as small as 1 nm. Variations in the intensity of electron scattering across a thin specimen can be used to image strain fields, defects such as dislocations and second-phase particles, and even atomic columns in materials under certain imaging conditions. Transmission electron microscopy is such a powerful tool for the characterization of materials that some microstructural features are defined in terms of their visibility in TEM images.

In addition to diffraction and imaging, the high-energy electrons (usually in the range of 100 to 400 keV of kinetic energy) in TEM cause electronic excitations of the atoms in the specimen. Two important spectroscopic techniques make use of these excitations by incorporating suitable detectors into the transmission electron microscope.

1. In energy-dispersive x-ray spectroscopy (EDS), an x-ray spectrum is collected from small regions of the specimen illuminated with a focused electron probe using a solid-state detector. Characteristic x rays of each element are used to determine the concentrations of the different elements present in the specimen (Williams and Carter, 1996). The principles behind this technique are discussed in details (ENERGY-DISPERSIVE SPECTROMETRY).
2. In electron energy loss spectroscopy (EELS), a magnetic prism is used to separate the electrons according to their energy losses after having passed through the specimen (Egerton, 1996). Energy loss mechanisms such as plasmon excitations and core-electron excitations cause distinct features in EELS. These can be used to quantify the elements present as well as provide information about atomic bonding and a variety of other useful phenomena.

In scanning transmission electron microscopy (STEM), a focused beam of electrons (typically <1 nm in diameter) is scanned in a television-style raster pattern across the specimen, as in a scanning electron microscope (SCANNING ELECTRON MICROSCOPY). In synchronization with the raster scan, emissions resulting from the interaction of the electron beam with the specimen are collected, such as x rays or secondary or backscattered electrons, to form images. Electrons that pass through the specimen can also be detected to form images that are similar to conventional TEM images. An annular detector can be used to collect the scattered transmitted electrons, which leads to Z-contrast imaging (discussed in SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). The STEM mode of operation is particularly useful for spectroscopic analysis, since it permits the acquisition of a chemical map of the sample typically with a resolution of a few nanometers. For example, one can make an image of the distribution of Fe in a sample by recording, in synchronization with the raster pattern, either the emission from the sample of Fe $K\alpha$ x rays (with the EDS spectrometer) or transmitted electrons with energy losses greater than that of the Fe L edge (with the EELS spectrometer). The STEM mode of operation is different from the conventional TEM mode in that the objective lens is operated in tandem with the illumination lens system to assist in the formation of a focused electron probe on the specimen (Keyse et al., 1998).

A fully equipped transmission electron microscope has the capability to record the variations in image intensity across the specimen using mass thickness or diffraction contrast techniques, to reveal the atomic structure of materials using high-resolution (phase-contrast) imaging or Z-contrast (incoherent) imaging, to obtain electron diffraction patterns from small areas of the specimen using a selected-area aperture or a focused electron probe, and to perform EELS and EDS measurements with a small probe. Additional lenses can be installed in conjunction with an EELS spectrometer to create an energy filter, enabling one to form energy-filtered TEM images (EFTEM—Krivanek et al., 1987; Reimer, 1995). These images enable mapping of the chemical composition of a specimen with nanometer spatial resolution. A block diagram of such a transmission electron microscope is shown in Figure 1.

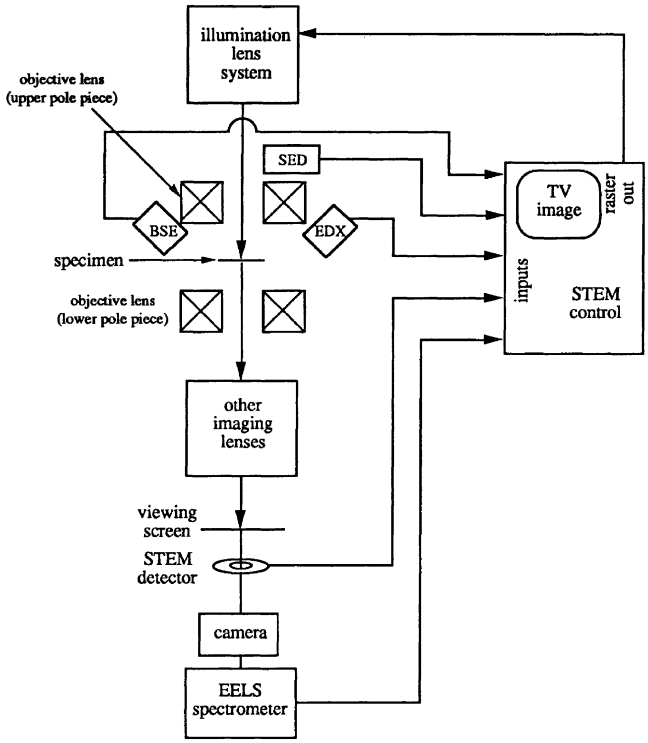


Figure 1. Typical transmission electron microscope with STEM capability. It is also possible to perform scanning electron microscopy (SEM) in a STEM using backscattered electron (BSE) and secondary electron detectors (SEDs) located above the specimen.

In addition to the main techniques of (1) conventional imaging, (2) phase-contrast imaging, (3) *Z*-contrast imaging, (4) electron diffraction, (5) EDS, and (6) EELS, in TEM many other analyses are possible. For example, when electrons pass through a magnetic specimen, they are deflected slightly by Lorentz forces, which change direction across a magnetic domain wall. In a method known as Lorentz microscopy, special adjustments of lens currents permit imaging of these domain walls (Thomas and Goringe, 1979). Phase transformations and microstructural changes in a specimen can be observed directly as the specimen is heated, cooled, or deformed in the microscope using various specimen stages (Butler and Hale, 1981). Differential pumping can be used to allow the introduction of gases into the microscope column surrounding a thin foil, making it possible to follow chemical reactions in TEM *in situ*. Many of these techniques can be performed at spatial resolutions of a few tenths of a nanometer. The possibilities are almost endless, and that is why TEM continues to be an indispensible tool in materials research. Since it is not possible to cover all of these techniques, this unit focuses on the theory and practice of conventional TEM electron diffraction and imaging techniques, including some examples of high-resolution TEM (HRTEM). A few imaging and diffraction techniques offer the resolution and versatility of TEM and are competitive or complementary in some respects.

Transmission electron microscopy is not able to readily image single point defects in materials. In contrast, the field ion microscope (FIM) can be used to study vacancies and surface atoms and therefore extends the available resolution to atomic dimensions (Miller and Smith, 1989). When combined with a mass spectrometer, the FIM becomes an atom probe (APFIM), easily capable of compositional analysis of regions ~1 nm wide and now approaching atom-by-atom analysis of local areas of samples. In the FIM, a high potential is applied to a very fine pointed specimen. A low pressure of inert gas is maintained between the specimen and imaging screen/spectrometer. Positively charged gas ions generated by the process of field ionization are used to produce images of the atoms on the surface of the specimen. Successive atom layers of material may be ionized and removed from the specimen surface by the process of field evaporation, enabling the three-dimensional (3D) structure of the material to be imaged in atomic detail and also to provide the source of ions for mass spectrometry.

Another instrument capable of imaging the internal structure of materials is the scanning acoustic microscope (SAM). In this microscope, a lens is used to focus the acoustic emission from an ultrasonic transducer onto a small region of a specimen through a coupling medium such as water. The same lens can be used to collect the acoustic echo from defects below the surface of the specimen. Either the lens or the specimen is scanned mechanically to form a two-dimensional (2D) image, the resolution of which depends on the wavelength of the acoustic wave in the specimen and thus on the frequency of the ultrasound and the velocity of sound in the specimen. Since the depth to which sound waves can penetrate decreases as their frequency increases, the choice of frequency depends on a compromise between resolution and penetration. Typical values might be a resolution of 40 to 100 μm at a depth of 5 mm below the surface for 50-MHz frequency, or a resolution of 1 to 3 μm at a similar depth for a frequency of 2 GHz. The resolution is clearly not as good as in TEM although the depth penetration is greater.

Lastly, x-ray diffraction and microscopy are alternative techniques to TEM and offer the advantage of greater penetration through materials. The main obstacle to high spatial resolution in both techniques is the difficulty of focusing x-ray beams, since x rays are uncharged and cannot be focused by electromagnetic or electrostatic lenses. While x-ray diffraction is commonly used for characterization of bulk samples in materials science and can be used to determine the atomic structures of materials (McKie and McKie, 1986), x-ray microscopy is not common, largely due to the better resolution and versatility of TEM.

PRINCIPLES OF THE METHOD

Here we develop the theoretical basis necessary to understand and quantify the formation of diffraction patterns and images in TEM. The theory developed is known as the "kinematical theory" of electron diffraction, and the contrast that arises in TEM images due to electron diffraction is "diffraction contrast." These concepts are then utilized in examples of the method and data analyses.

Structure Factor and Shape Factor

In both real space and in reciprocal space, it is useful to divide a crystal into parts according to the formula (Fultz and Howe, 2000)

$$\mathbf{r} = \mathbf{r}_g + \mathbf{r}_k + \delta_{\mathbf{r}_{g,k}} \quad (1)$$

For a defect-free crystal the atom positions $\mathbf{R}$ are given by

$$\mathbf{R} = \mathbf{r}_g + \mathbf{r}_k \quad (2)$$

where the lattice is one of the 14 Bravais lattice types (see SYMMETRY IN CRYSTALLOGRAPHY) and the basis is the atom group associated with each lattice site (Borchardt-Ott, 1995). Here we calculate the scattered wave ψ ($\Delta\mathbf{k}$) for the case of an infinitely large, defect-free lattice with a basis:

$$\psi(\Delta\mathbf{k}) = \sum_{\mathbf{R}} f_{at}(\mathbf{R}) \exp(-i2\pi\ \Delta\mathbf{k} \cdot \mathbf{R}) \quad (3)$$

where the scattered wave vector $\Delta\mathbf{k}$ is defined as the difference between the diffracted wave vector $\mathbf{k}$ and the incident wave vector $\mathbf{k}_0$ (refer to Fig. 2), or

$$\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0 \quad (4)$$

and $f_{at}(\mathbf{R})$ is the atomic scattering factor for electrons from an atom. We decompose our diffracted wave into a lattice component and a basis component with

$$\psi(\Delta\mathbf{k}) = \sum_{(\mathbf{r}_g + \mathbf{r}_k)} f_{at}(\mathbf{r}_g + \mathbf{r}_k) \exp[-i2\pi\ \Delta\mathbf{k} \cdot (\mathbf{r}_g + \mathbf{r}_k)] \quad (5)$$

Since the atom positions in all unit cells are identical, $f_{at}(\mathbf{r}_g + \mathbf{r}_k)$ cannot depend on $\mathbf{r}_g$, so $f_{at}(\mathbf{r}_g + \mathbf{r}_k) = f_{at}(\mathbf{r}_k)$, and thus

$$\psi(\Delta\mathbf{k}) = \sum_{\mathbf{r}_g} \exp(-i2\pi\ \Delta\mathbf{k} \cdot \mathbf{r}_g) \sum_{\mathbf{r}_k} f_{at}(\mathbf{r}_k) \exp(-i2\pi\ \Delta\mathbf{k} \cdot \mathbf{r}_k) \quad (6)$$

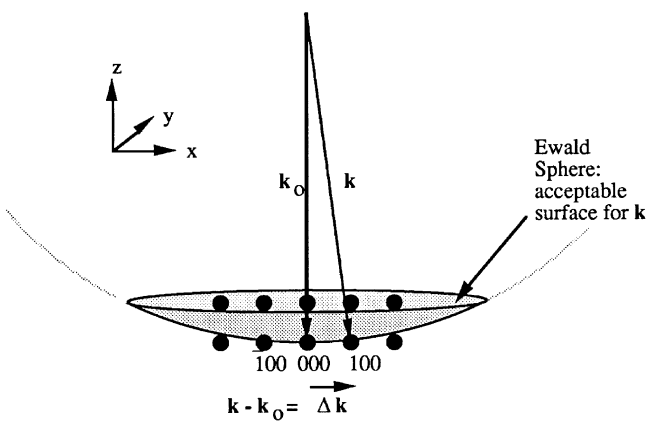


Figure 2. Ewald sphere construction showing incident $\mathbf{k}_0$ and scattered $\mathbf{k}$ wave vectors joined tail to tail and definition of $\Delta\mathbf{k}$.

or, simplifying,

$$\Psi(\Delta\mathbf{k}) = \mathcal{S}(\Delta\mathbf{k})\mathcal{F}(\Delta\mathbf{k}) \quad (7)$$

In writing Equation 7, we have given formal definitions to the two summations in Equation 6. The first sum, which is over all the lattice sites of the crystal (all unit cells), is known as the shape factor $\mathcal{S}$. The second sum, which is over the atoms in the basis (all atoms in the unit cell), is known as the structure factor $\mathcal{F}$. The notation ($\Delta\mathbf{k}$) is used to indicate the dependence of these terms on $\Delta\mathbf{k}$. The decomposition of the diffracted wave into the shape factor and the structure factor parallels the decomposition of the crystal into a lattice plus a basis.

Calculation of the structure factor $\mathcal{F}(\Delta\mathbf{k})$ for a unit cell is discussed in detail in KINEMATIC DIFFRACTION OF X RAYS and in standard books on diffraction (Schwartz and Cohen, 1987) and is not developed further here. Because in TEM we examine thin foils often containing small particles with different shapes, it is useful to examine the shape factor $\mathcal{S}(\Delta\mathbf{k})$ in Equations 6 and 7 in further detail. The shape factor $\mathcal{S}(\Delta\mathbf{k})$ is not very interesting for an infinitely large crystal where it becomes a set of delta functions centered at the various values of $\Delta\mathbf{k}$, where $\Delta\mathbf{k} = \mathbf{g}$ ($\mathbf{g}$ is a reciprocal lattice vector), but it is interesting for small crystals, which give rise to various spatial distributions of the diffracted electron intensity. The full 3D expression for the kinematical diffracted intensity due to the shape factor is given as

$$\mathcal{S}^*\mathcal{S}(\Delta\mathbf{k}) = \frac{\sin^2(\pi\ \Delta k_x a_x N_x)}{\sin^2(\pi\ \Delta k_x a_x)} \frac{\sin^2(\pi\ \Delta k_y a_y N_y)}{\sin^2(\pi\ \Delta k_y a_y)} \frac{\sin^2(\pi\ \Delta k_z a_z N_z)}{\sin^2(\pi\ \Delta k_z a_z)} \quad (8)$$

where a_x, a_y, and a_z are the magnitudes of the primitive translation vectors of the unit cell expressed along orthonormal x, y, and z axes; N_x, N_y, and N_z are the number of unit cells along the same axes, i.e., $a_x N_x = t_x$, the crystal thickness along the x direction and similarly for y and z; and Δk_x, Δk_y, and Δk_z are the components of $\Delta\mathbf{k}$ expressed along the x, y, and z axes. Equation 8 is valid for a crystal shaped as a rectangular prism.

The function in Equation 8 becomes large when the denominator goes to zero. This occurs when the argument of the sine function is equal to π or to some integral multiple of it, expressed (in the x direction only) as

$$\Delta k_x a_x = \text{integer} \quad (9)$$

Since similar conditions are expected for y and z, this condition requires that $\Delta\mathbf{k}$ is a reciprocal lattice vector $\mathbf{g}$. In other words, the kinematical intensity $\mathcal{S}^*\mathcal{S}$ is large when the Bragg condition is exactly satisfied. Since the denominator varies slowly with respect to the numerator, we can make the following approximation, which is valid near the center of the main peaks (expressed in only one direction):

$$\mathcal{S}^*\mathcal{S}(\Delta\mathbf{k}) \cong \frac{\sin^2(\pi\ \Delta k\ aN)}{(\pi\ \Delta k\ a)^2} \quad (10)$$

This function describes an envelope of satellite peaks situated near the main Bragg diffraction peaks. By examining

the numerator, we see that the positions of the satellite peaks get closer to the main peak in proportion to $1/Na$, and the position of the first minimum in the intensity is located on either side of the main peak at the position $\Delta k = 1/Na$. Similarly, the widths of the main peaks and satellite peaks also decrease as $1/Na$. Thus, large crystal dimensions in real space lead to sharp diffracted intensities in reciprocal space, and vice versa (Hirsch et al., 1977; Thomas and Goringe, 1979).

Ewald Sphere Construction

The Bragg condition for diffraction, $\Delta\mathbf{k} = \mathbf{g}$, where $\mathbf{g}$ is a reciprocal lattice vector and $\Delta\mathbf{k}$ is any possible scattered wave vector, can be implemented in a geometrical construction due to Ewald (McKie and McKie, 1986). The Ewald sphere depicts the incident wave vector $\mathbf{k}_0$ and all possible $\mathbf{k}$ for the diffracted waves. The tip of the wave vector $\mathbf{k}_0$ is always placed at a point of the reciprocal lattice that serves as the origin. To obtain $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, we would normally reverse the direction of $\mathbf{k}_0$ and place it tail to head with the vector $\mathbf{k}$, but in the Ewald sphere construction in Figure 2, we draw $\mathbf{k}$ and $\mathbf{k}_0$ tail to tail. The vector $\Delta\mathbf{k}$ is the vector from the head of $\mathbf{k}_0$ to the head of $\mathbf{k}$. If the head of $\mathbf{k}$ touches any reciprocal lattice point, the Bragg condition ($\Delta\mathbf{k}$ = $\mathbf{g}$) is satisfied and diffraction occurs. In elastic scattering, the length of $\mathbf{k}$ equals the length of $\mathbf{k}_0$, since there is no change in wavelength (magnitude of the wave vector). The tips of all possible $\mathbf{k}$ vectors lie on a sphere whose center is at the tails of $\mathbf{k}$ and $\mathbf{k}_0$. By this construction, one point on this sphere always touches the origin of the reciprocal lattice. Whenever another point on the Ewald sphere touches a reciprocal lattice point, the Bragg condition is satisfied and diffraction occurs. As illustrated in Figure 2, the Bragg condition is approximately satisfied for the (100) and ($\bar{1}00$) diffraction spots.

The Ewald sphere is strongly curved for x-ray diffraction because $|\mathbf{g}|$ (typically 5 to 10 nm^{-1}) is comparable to $|\mathbf{k}_0|$. Electron wave vectors, on the other hand, are much longer than the spacings in the reciprocal lattice (100-keV electrons have a wavelength of 0.0037 nm). So, for high-energy electrons, the Ewald sphere is approximately a plane, and consequently, $\Delta\mathbf{k}$ is nearly perpendicular to $\mathbf{k}_0$. In practice, the diffracted intensity distribution, such as that in Equation 8, is located in a finite volume around the reciprocal lattice points, so $\Delta\mathbf{k}$ need not equal $\mathbf{g}$ exactly in order for diffraction to occur. The shape factor intensity $\mathcal{S}^*\mathcal{S}(\cdot\mathbf{k})$ serves in effect to broaden the reciprocal lattice points. With a crystal oriented near a zone axis, i.e., along a direction parallel to the line of intersection of a set of crystal planes, the Ewald sphere passes through many of these small volumes around the reciprocal lattice points, and many diffracting spots are observed. Figure 3 provides an example of Ewald sphere constructions for two orientations of the specimen with respect to $\mathbf{k}_0$, together with their corresponding selected-area diffraction (SAD) patterns. Figure 3 is drawn as a 2D slice (the x-z plane) of Figure 2. The crystal on the right is oriented precisely along a zone axis, but the crystal on the left is not.

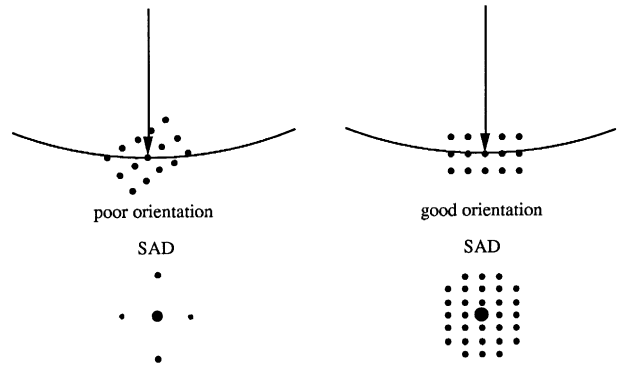


Figure 3. Two orientations of reciprocal lattice with respect to Ewald sphere and corresponding SAD patterns.

Deviation Vector and Deviation Parameter

From the previous discussion, it is apparent that the diffracted intensity observed in an electron diffraction pattern depends on exactly how the Ewald sphere cuts through the diffracted intensity around the Bragg position. To locate the exact position of intersection, we introduce a new parameter, the deviation vector $\mathbf{s}$, defined as

$$\mathbf{g} = \Delta\mathbf{k} + \mathbf{s} \tag{11}$$

where $\mathbf{g}$ is a reciprocal lattice vector and $\Delta\mathbf{k}$ is the diffraction vector whose end lies on the Ewald sphere ($\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$). For high-energy electrons, the shortest distance between the Ewald sphere and a reciprocal lattice point $\mathbf{g}$ is parallel to the $\mathbf{z}$ direction, so we often work with only the magnitude of $\mathbf{s}$, i.e., $|\mathbf{s}| = s$, referred to as the deviation parameter. We choose a sign convention for s that is convenient when we determine s by measuring the positions of Kikuchi lines (Kikuchi, 1928). Positive s means that $\mathbf{s}$ points along positive $\mathbf{z}$. (By convention, $\mathbf{z}$ points upward toward the electron gun.) Figure 4 shows that s is positive when the reciprocal lattice point lies inside the Ewald sphere and negative when the reciprocal lattice point lies outside the sphere.

The parameter s is useful because it is all that we need to know about the diffraction conditions to calculate the kinematical shape factor. Using Equation 11 for $\Delta\mathbf{k}$ in Equation 7 yields

$$\mathcal{S}(\Delta\mathbf{k}) = \sum_{\mathbf{r}_g} \exp(-i2\pi\mathbf{g}\cdot\mathbf{r}_g)\exp(+i2\pi\mathbf{s}\cdot\mathbf{r}_g) \tag{12}$$

$$\mathcal{F}(\Delta\mathbf{k}) = \sum_{\mathbf{r}_k} f_{at}(\mathbf{r}_k, \Delta\mathbf{k})\exp(-i2\pi\mathbf{g}\cdot\mathbf{r}_k)\exp(+i2\pi\mathbf{s}\cdot\mathbf{r}_k) \tag{13}$$

$$\mathcal{F}(\Delta\mathbf{k}) \cong \sum_{\mathbf{r}_k} f_{at}(\mathbf{r}_k, \Delta\mathbf{k})\exp(-i2\pi\mathbf{g}\cdot\mathbf{r}_k) = \mathcal{F}(\mathbf{g}) \text{ or } \mathcal{F}_g \tag{14}$$

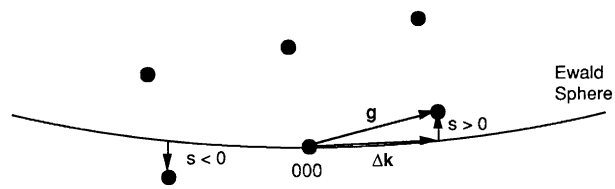


Figure 4. Convention for defining deviation vector $\mathbf{s}$ and deviation parameter s, from Ewald sphere to reciprocal lattice spot $\mathbf{g}$.

where we made the last approximation for the structure factor $\mathcal{F}(\mathbf{g})$ because $\mathbf{s} \cdot \mathbf{r}_k$ is small when the unit cell is small. Since $\mathbf{g}$ is a reciprocal lattice vector, $\exp(-i2\pi\mathbf{g} \cdot \mathbf{r}_g) = 1$ for all $\mathbf{r}_g$, which simplifies the shape factor to

$$\mathcal{S}(\Delta\mathbf{k}) = \mathcal{S}(\mathbf{s}) \sum_{\mathbf{r}_g} \exp(+i2\pi\mathbf{s} \cdot \mathbf{r}_g) \tag{15}$$

and the diffracted wave $\psi(\mathbf{g}, \mathbf{s})$, which is a function of the reciprocal lattice vector and deviation parameter, becomes

$$\psi(\mathbf{g}, \mathbf{s}) = \mathcal{F}_g \sum_{\mathbf{r}_g} \exp(+i2\pi\mathbf{s} \cdot \mathbf{r}_g) \tag{16}$$

The shape factor depends only on the deviation vector $\mathbf{s}$ and not on the particular reciprocal lattice vector $\mathbf{g}$. When we substitute for the components of $\mathbf{s}$ and $\mathbf{r}_g$ along the x, y, and z axes, we obtain an equation for the shape factor intensity that is similar to Equation 8:

$$\mathcal{S} * \mathcal{S}(\mathbf{s}) = \frac{\sin^2(\pi s_x a_x N_x)}{\sin^2(\pi s_x a_x)} \frac{\sin^2(\pi s_y a_y N_y)}{\sin^2(\pi s_y a_y)} \frac{\sin^2(\pi s_z a_z N_z)}{\sin^2(\pi s_z a_z)} \tag{17}$$

For high-energy electron diffraction, we can make the following simplifications: (1) the deviation vector is very nearly parallel to the z axis, so s_z is simply equal to s; (2) the denominator is given as $(\pi s_z)^2$; and (3) the quantity $a_z N_z$ is the crystal thickness t. Ignoring the widths along x and y of the diffracting columns, a useful expression for the shape factor intensity is

$$\mathcal{S} * \mathcal{S}(\mathbf{s}) = \frac{\sin^2(\pi s t)}{(\pi s)^2} \tag{18}$$

Combining Equations 14 and 18 then gives the resulting diffracted intensity for the vector $\mathbf{g}$:

$$\mathbf{I}_g = |\psi(\mathbf{g}, \mathbf{s})|^2 = |\mathcal{F}(\mathbf{g})|^2 \frac{\sin^2(\pi s t)}{(\pi s)^2} \tag{19}$$

This intensity depends only on the diffracting vector $\mathbf{g}$ and the deviation parameter s. Below we present many examples of how $\mathbf{I}_g$ depends on the deviation parameter s and the sample thickness t. The dependence of $\mathbf{I}_g$ on the position (x,y) of the diffracting column provides diffraction contrast in bright-field (BF) or dark-field (DF) images. In a two-beam condition, where only the incident beam and one diffracting beam have appreciable intensity, the intensities of the transmitted beam $\mathbf{I}_0$ and diffracted beam $\mathbf{I}_g$ are complementary, and with the incident intensity normalized to 1, we have the relationship

$$\mathbf{I}_0 = 1 - \mathbf{I}_g \tag{20}$$

The kinematical diffraction theory developed above is valid when the intensity of the diffracted beam is much less than the transmitted beam, or

$$\mathbf{I}_g \ll \mathbf{I}_0 \tag{21}$$

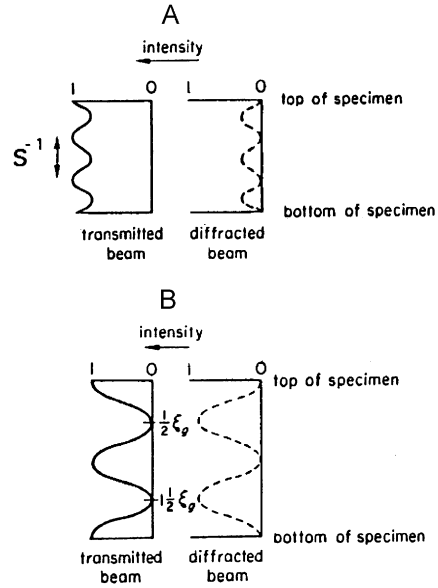


Figure 5. Extinction distance and amplitudes of transmitted and diffracted beams for two-beam condition in relatively thick crystal (**A**) under kinematical conditions with $s \gg 0$ and (**B**) under dynamical conditions with $s = 0$. (After Edington, 1974.)

When this condition is not satisfied, one must resort to the dynamical theory of electron diffraction (DYNAMICAL DIFFRACTION; Hirsch et al., 1977; Williams and Carter, 1996). Equations 19 through 21 tell us that, for a foil of constant thickness t, we expect the intensities of the transmitted and diffracted beams to vary with depth in the foil and the period of depth variation is approximately s^{-1}. This behavior is illustrated schematically in Figure 5A. The larger is the deviation parameter, the smaller is the period of oscillation and vice versa.

Extinction Distance

A more general way of writing Equation 19 is

$$\mathbf{I}_g = \left(\frac{\pi}{\xi_g}\right)^2 \frac{\sin^2(\pi s t)}{(\pi s)^2} \tag{22}$$

where the extinction distance ξ_g is given by the expression

$$\xi_g = \frac{\pi V \cos\theta}{\lambda \mathcal{F}_g} \tag{23}$$

and V is the volume of the unit cell, θ is the Bragg angle for the diffraction vector $\mathbf{g}$, λ is the electron wavelength, and $\mathcal{F}_g$ is the structure factor for the diffracting vector $\mathbf{g}$. Note that ξ_g increases with increasing order of diffracting vectors because $\mathcal{F}_g$ decreases as θ increases.

The quantity ξ_g defined in Equation 23 is an important length scale in the kinematical and dynamical theories of electron diffraction. The extinction distance is defined as twice the distance in the cystal over which 100% of the incident beam is diffracted when $s = 0$ and the Bragg condition is satisfied exactly. The magnitude of ξ_g depends on

Table 1. Extinction Distances ξ_g (nm) for fcc Metals with 100-kV Electrons (from Hirsch et al., 1977)

Diffracting Plane	Al	Ni	Cu	Ag	Pt	Au	Pb
111	55.6	23.6	24.2	22.4	14.7	15.9	24.0
200	67.3	27.5	28.1	25.5	16.6	17.9	26.6
220	105.7	40.9	41.6	36.3	23.2	24.8	35.9
311	130.0	49.9	50.5	43.3	27.4	29.2	41.8
222	137.7	52.9	53.5	45.5	28.8	30.7	43.6
400	167.2	65.2	65.4	54.4	34.3	36.3	50.5
331	187.7	74.5	74.5	61.1	38.5	40.6	55.5
420	194.3	77.6	77.6	63.4	39.8	42.0	57.2
422	219.0	89.6	89.7	72.4	45.3	47.7	63.8
511	236.3	98.3	98.5	79.2	49.4	51.9	68.8
333	236.3	112.0	112.6	90.1	55.8	58.7	77.2
531	279.8	119.6	120.6	96.4	59.4	62.6	82.2
600	285.1	122.1	123.2	98.4	60.6	63.8	83.8
442	285.1	122.1	123.2	98.4	60.6	63.8	83.8

the atomic form factors; the stronger the scattering, the shorter is ξ_g. Table 1 shows some values of ξ_g for different diffracting planes in pure metals with a face-centered cubic (fcc) crystal structure (Hirsch et al., 1977). Notice how ξ_g increases with the indices of the diffracting vectors *hkl* and decreases with increasing atomic number. The values of ξ_g generally range from a few tens to a few hundred nanometers.

The extinction distance defines the depth variation of the transmitted and diffracted intensities in a crystal when the diffraction is strong. The extinction distance ξ_g applies to the dynamical situation where $s = 0$, as illustrated in Figure 5B, not just the kinematical situation where $s \gg 0$. In the development of the dynamical theory of electron diffraction, the dependence of the diffracted intensity on the deviation parameter and sample thickness is much the same as in Equation 22 with the following modification: The extinction distance in Equation 23 is made dependent on s by transforming ξ_g into an effective extinction distance, ξ_{geff},

$$\xi_{geff} = \frac{\xi_g}{\sqrt{1 + s^2\xi_g^2}} \tag{24}$$

and another quantity known as the effective deviation parameter is defined as

$$s_{eff} = \sqrt{s^2 + \xi_g^{-2}} \tag{25}$$

Equation 24 shows that the effective extinction distance $\xi_{geff} = \xi_g$ when $s = 0$, but ξ_{geff} decreases with increasing deviation from the exact Bragg position. With large deviations, $s \gg 1/\xi_g$, and $\xi_{geff} = s^{-1}$, so the kinematical result is recovered for large s.

Diffraction Contrast from Lattice Defects

The following are important variables in the diffraction contrast from defects in crystals:

$\mathcal{F}(\mathbf{g}) \equiv$ structure factor of unit cell
$t \equiv$ specimen thickness
$\Delta\mathbf{k} \equiv$ diffraction vector
$\mathbf{g} \equiv$ reciprocal lattice vector
$\mathbf{s} \equiv$ deviation vector ($\mathbf{g} = \Delta\mathbf{k} + \mathbf{s}$)
$\mathbf{r} \equiv$ actual atom centers
$\mathbf{R} \equiv$ atom centers in a perfect crystal ($\mathbf{R} = \mathbf{r}_g + \mathbf{r}_k$, where $\mathbf{r}_g$ refers to the lattice and $\mathbf{r}_k$ to the basis)
$\delta\mathbf{r} \equiv$ displacements of atoms off the ideal atom centers ($\mathbf{r} = \mathbf{R} + \delta\mathbf{r}$).

Note that the actual atom centers $\mathbf{r}$ as defined above are the locations of the atoms off of the atom positions in a perfect crystal $\mathbf{R}$ that are given by the lattice (plus basis) points, i.e., the ideal mathematical positions. Spatial variations in these variables (e.g., an x dependence) can produce diffraction contrast in an image. Examples include the following:

$\mathcal{F}$: $d\mathcal{F}/dx$ causes chemical (compositional) contrast,
t: dt/dx causes thickness contours,
$\mathbf{g}$: $d\mathbf{g}/dx$ causes bend contours,
$\mathbf{s}$: $d\mathbf{s}/dx$ causes bend contours, and
$\delta\mathbf{r}$: $d\delta\mathbf{r}/dx$ causes strain contrast.

Up to this point, we have only considered diffraction occurring from perfect crystals. We now consider the displacements in atom positions $\delta\mathbf{r}$ caused by strain fields around defects in a crystal. These displacements represent the time-averaged position of the atoms during electron scattering. We decompose $\mathbf{r}$ into components from the lattice vectors, basis vectors, and distortion vectors as

$$\mathbf{r} = \mathbf{r}_g + \mathbf{r}_k + \delta\mathbf{r} \tag{26}$$

and we use the familiar expression for our diffracting vector as the difference of a reciprocal lattice vector and a deviation vector:

$$\Delta\mathbf{k} = \mathbf{g} - \mathbf{s} \tag{27}$$

By using Equations 26 and 27 and noting that $\mathbf{r} = \mathbf{R} + \delta\mathbf{r}$, we can rewrite Equation 3 as

$$\psi(\Delta\mathbf{k}) = \sum_{\mathbf{r}} f_{\text{at}}(\mathbf{r}) \exp[-i2\pi(\mathbf{g} - \mathbf{s}) \cdot (\mathbf{r}_{\text{g}} + \mathbf{r}_{\text{k}} + \delta\mathbf{r})] \quad (28)$$

We then arrange Equaton 28 into a product of a structure factor $\mathcal{F}(\Delta\mathbf{k})$ and something like a shape factor, as we did previously. In doing so, we assume that the distortion $\delta\mathbf{r}$ is the same for all atoms in a particular unit cell,

$$\psi(\Delta\mathbf{k}) = \sum_{\mathbf{r}_{\text{g}}} \exp[-i2\pi(\mathbf{g} - \mathbf{s}) \cdot (\mathbf{r}_{\text{g}} + \delta\mathbf{r})] \times \sum_{\mathbf{r}_{\text{k}}} f_{\text{at}}(\mathbf{r}_{\text{k}}) \exp[-i2\pi(\mathbf{g} - \mathbf{s}) \cdot \mathbf{r}_{\text{k}}] \quad (29)$$

When there are only a few atoms in the unit cell, we can neglect the factor of $\exp(i2\pi\mathbf{s} \cdot \mathbf{r}_{\text{k}})$. We then identify the second sum in Equation 29 as the familiar structure factor of the unit cell $\mathcal{F}(\Delta\mathbf{k})$, and rewriting Equation 29 gives

$$\psi(\Delta\mathbf{k}) = \sum_{\mathbf{r}_{\text{g}}} \exp[-i2\pi(\mathbf{g} \cdot \mathbf{r}_{\text{g}} + \mathbf{g} \cdot \delta\mathbf{r} - \mathbf{s} \cdot \mathbf{r}_{\text{g}} - \mathbf{s} \cdot \delta\mathbf{r})] \mathcal{F}(\Delta\mathbf{k}) \quad (30)$$

First, note that the product $\mathbf{g} \cdot \mathbf{r}_{\text{g}}$ is an integer, so it yields a factor of unity in the exponential. Second, because both $\mathbf{s}$ and $\delta\mathbf{r}$ are small, their product is negligible compared to the other two terms in the exponential in Equation 30. Thus, one obtains

$$\psi_{\text{g}} - F(\mathbf{g}) \sum_{\mathbf{r}_{\text{g}}} \exp[-i2\pi(\mathbf{g} \cdot \delta\mathbf{r} - \mathbf{s} \cdot \mathbf{r}_{\text{g}})] \quad (31)$$

It is convenient and appropriate to substitute $\mathcal{F}(\mathbf{g})$ for $\mathcal{F}(\Delta\mathbf{k})$ in Equation 31, since we use all of the diffracted intensity around a specific reciprocal lattice point in forming a DF image, as defined in the next section. Equation 31 is useful for determining the diffraction contrast from defects such as dislocations, stacking faults, and precipitates. Such defects possess strain fields, and it is these strain fields that produce diffraction contrast in TEM images. Equation 31 shows that, in addition to the local atomic displacements $\delta\mathbf{r}$, the defect image is also controlled by the deviation vector $\mathbf{s}$.

The diffraction condition $\mathbf{g} \cdot \delta\mathbf{r} = 0$ is particularly important for the study of defects with diffraction contrast. The condition $\mathbf{g} \cdot \delta\mathbf{r} = 0$ is a null-contrast condition. For this condition, $\delta\mathbf{r}$ lies in the diffracting plane, and the interplanar spacing d (and thus $|\mathbf{g}|$) is therefore unaltered. The condition $\mathbf{g} \cdot \delta\mathbf{r}$ is the condition for zero diffraction contrast originating from the displacement $\delta\mathbf{r}$. Even when $\mathbf{g} \cdot \delta\mathbf{r}$ is not exactly zero, the magnitude of $\mathbf{g} \cdot \delta\mathbf{r}$ must be sufficient to change the local intensity from the background level so that contrast is visible in the image. This typically requires an intensity change of $\sim$10%. A rule of thumb is that if $\mathbf{g} \cdot \delta\mathbf{r} \leq \frac{1}{3}$, there is no visible contrast associated with $\delta\mathbf{r}$. This criterion is adequate for the analysis of diffracton contrast from many defects of interest in materials, such as dislocations, dislocation loops, precipitates, stacking faults, domain boundaries, and grain and interphase boundaries, and its use is illustrated below (see Data Analysis and Initial Interpretation).

PRACTICAL ASPECTS OF THE METHOD

Bright-Field/Dark-Field Imaging

A ray diagram for making an image with a conventional TEM (CTEM) containing two lenses is shown in Figure 6B. In this diagram, the intermediate lens is focused on the image plane of the objective lens. We assume the illumination system provides rays (electrons) that travel straight down the microscope (parallel to the optic axis) before hitting the specimen. In Figure 6A, all transmitted and diffracted rays leaving the specimen are combined to form an image at the viewing screen, much as in a conventional optical microscope. In this simple mode of imaging, the specimen generally shows little contrast because all of the diffracted intensity reaches the viewing screen. By tracing the individual rays in Figure 6A, one can see how each point in the back focal plane of the objective lens contains rays from all parts of the specimen. Not all of the rays in the back focal plane are therefore required to form an image. An image can be formed with only those rays passing through one point in the back focal plane. What distinguishes the points located in the back focal plane is that all rays entering a given point have been scattered by the specimen into the same angle. By positioning an objective aperture at a specific location in the back focal plane, an image will be made with only those electrons that have been diffracted by a specific angle. (A similar principle applies in EFTEM, except that an aperture is used to allow electrons of only a certain energy loss to form the image.) This defines two basic imaging modes, which are also illustrated in Figure 6:

1. When the objective aperture is positioned to pass only the transmitted (undiffracted) electrons, a BF image is formed, as illustrated in Figure 6B.
2. When the objective aperture is positioned to pass only some diffracted electrons, a DF image is formed, as illustrated in Figure 6C.

The best way to form a DF image is to tilt the incident illumination on the specimen by an angle equal to the angle of the particular diffraction used for making the DF image, as illustrated in Figure 6C. On the back focal plane, the position of the transmitted beam was tilted into the position of the diffracted spot on the left, and the diffracted spot on the right is now used to form the DF image. Notice how these diffracted rays remain near the optic axis. This minimizes the detrimental effect of spherical aberration on the image resolution (see below). The following section illustrates how one can visualize tilting either the specimen or the electron beam in the TEM mode using the Ewald sphere construction discussed previously.

In most CTEM studies of crystalline materials, features in the image originate primarily from diffraction contrast.

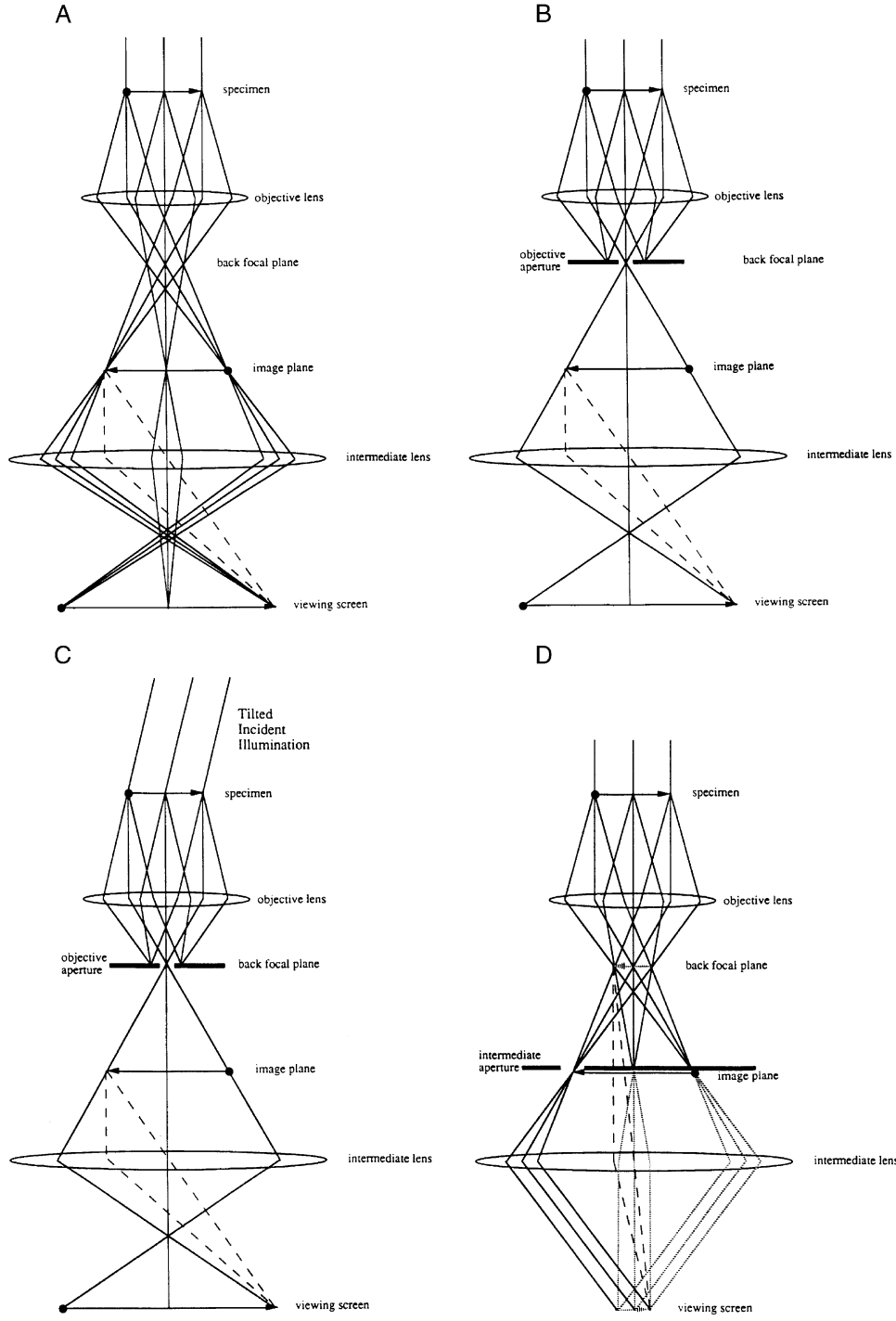


Figure 6. (**A**) Image mode for TEM containing objective and intermediate lens (**B**) BF imaging mode, (**C**) axial DF imaging mode, and (**D**) SAD mode in TEM.

Diffraction contrast is the variation in the diffracted intensity of particular electrons across the specimen, observed by inserting an objective aperture into the path of the beam. Upon doing so, features in the image become far more visible; without the objective aperture, the image tends to be gray and featureless. The physical reason that the diffraction contrast of a BF (or DF) image (Fig. 6B or C) is so much better than that of an apertureless image (Fig. 6A) is as follows: When there is a large intensity in the diffracted beams, there is a large complementary loss of intensity in the transmitted beam, so either the BF or DF image alone will show strong diffraction contrast. Without the objective aperture, however, the diffracted intensity recombines with the transmitted intensity at the viewing screen. This recombination eliminates the observed diffraction contrast.

For thick specimens, an apertureless image mainly shows contrast caused by incoherent scattering in the

sample. This generic type of contrast is called mass-thickness contrast because it increases with the atomic mass (squared) and the thickness of the material (Reimer, 1993). Mass-thickness contrast is particularly useful in biology, where techniques have been developed for selectively staining the different cell organelles with heavy elements to increase their mass-thickness contrast. In the case of very thin specimens (say $<\sim$10 nm), if a large objective aperture is used (or no aperture at all, as in Fig. 6A) and the transmitted and diffracted electron beams (waves) are allowed to recombine (interfere), the resulting image contrast depends on the relative phases of the beams involved; this mode of imaging is often called phase-contrast imaging. If the microscope resolution is less than the atomic spacings in the sample and both the sample and microscope conditions are optimized, phase-contrast imaging can be used to resolve the atomic structures of materials. This forms the basis of HRTEM.

In the microscope, typical objective apertures range from $\sim$0.5 to 20 μm in diameter. The apertures are moveable with high mechanical precision and can be positioned around selected diffraction spots in the back focal plane of the objective lens. The practice of positioning an objective aperture requires changing the operating mode of the microscope to the diffraction mode (described in the next section). In the diffraction mode, the images of both the diffraction pattern and the aperture are visible on the viewing screen, and the objective aperture can then be moved until it is in the desired position. Once the objective aperture is positioned properly, the microscope is switched back into the image mode, and either a DF or a BF image is formed.

Tilting Specimens and Tilting Electron Beams

Many problems in the geometry of diffraction can be solved by employing an Ewald sphere and a reciprocal lattice. When working problems, it is important to remember the following:

1. The Ewald sphere and the reciprocal lattice are connected at the origin of the reciprocal lattice. Tilts of either the Ewald sphere or the reciprocal lattice are performed about this pivot point.
2. The reciprocal lattice is affixed to the crystal (for cubic crystals the reciprocal lattice directions are along the real-space directions). Tilting the specimen is represented by tilting the reciprocal lattice by the same angle and in the same direction.
3. The Ewald sphere surrounds the incident beam and is affixed to it. Tilting the direction of the incident beam is represented by tilting the Ewald sphere by the same amount.

These three facts are useful during practical TEM work. It is handy to think of the viewing screen as a section of the Ewald sphere, which shows a disk-shaped slice of the reciprocal lattice of the specimen. When we tilt the sample, the Ewald sphere and the viewing screen stay fixed, but the different points on the reciprocal lattice of the sample move into the viewing screen. For small tilts of the specimen, the diffraction pattern does not move on the viewing screen, but the diffraction spots change in intensity. Alternatively, when we tilt the incident beam, we rotate the transmitted beam and the Ewald sphere. We can think of this operation as moving the disk-shaped viewing screen around the surface of the Ewald sphere. The diffraction spots therefore move in a direction opposite to the tilt of the incident beam, and the diffraction pattern moves as though the central beam were pivoting about the center of the sample.

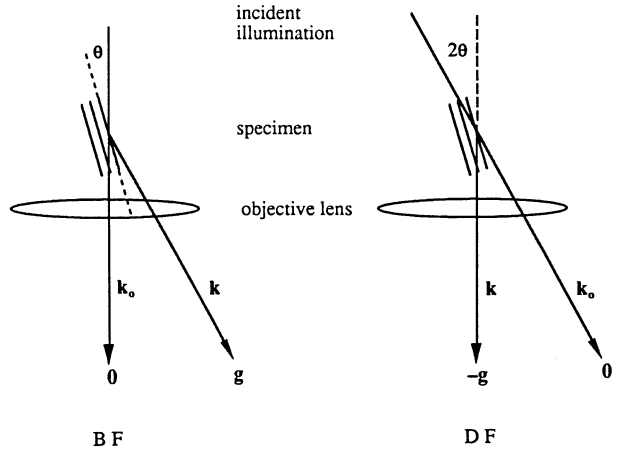


Figure 7. Procedures for axial BF and DF imaging. Specimen remains stationary as beam is tilted 2θ to center $-\mathbf{g}$ diffracted spot on optic axis.

Tilted Illumination and Diffraction: How to Do Axial DF Imaging

To form DF images with good resolution, it is necessary to ensure that the image-forming diffracted rays travel straight down the optic axis. This requires that the direction of the incident electrons ($\mathbf{k}_0$) be tilted away from the optic axis by an angle 2θ, as shown in Figure 7. After tilting the illumination, there is a difference in the positions and intensities of the diffraction spots for the BF and DF modes. By tilting the illumination, we in fact tilt the Ewald sphere about the origin of the reciprocal lattice, as shown in Figure 8. Seen on the viewing screen, we tilt the transmitted beam into the former position of the diffraction spot **g**. In doing so, the diffraction spot **g** moves far from the optic axis, but the diffraction spot $-\mathbf{g}$ becomes active, and its rays travel along the optic axis. This procedure seems counterintuitive, so why did we use it? The answer is that if the active diffraction spot **g** is tilted into the center of the viewing screen, the diffraction spot **g** becomes weak, and the diffraction spot 3**g** becomes strong. Since the diffraction spot **g** is weak, it would be difficult to use for making a DF image. This latter procedure can be useful to improve resolution in DF analysis of defects and is referred to as weak-beam dark-field (WBDF) imaging (Cockayne et al., 1969).

Selected-Area Diffraction

Figure 6D is a ray diagram for making a diffraction pattern with the simplified two-lens transmission electron

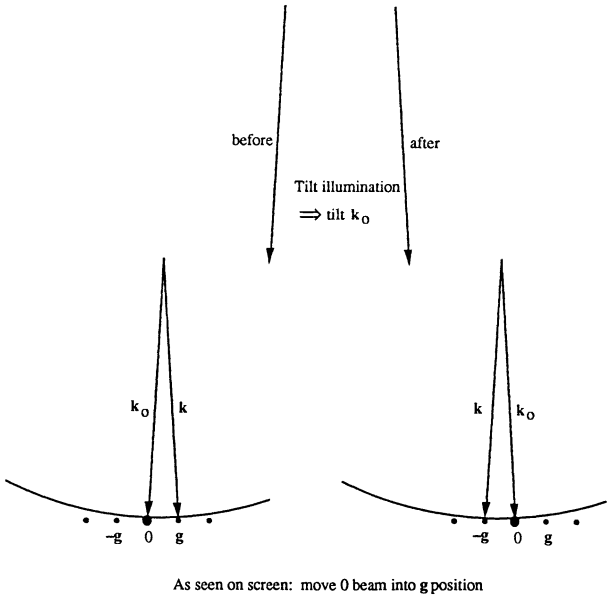


Figure 8. Procedure to set up axial DF image as in Figure 7. As seen on viewing screen, transmitted beam is moved into former position of diffracted spot **g** and $-\mathbf{g}$ diffracted spot becomes intense.

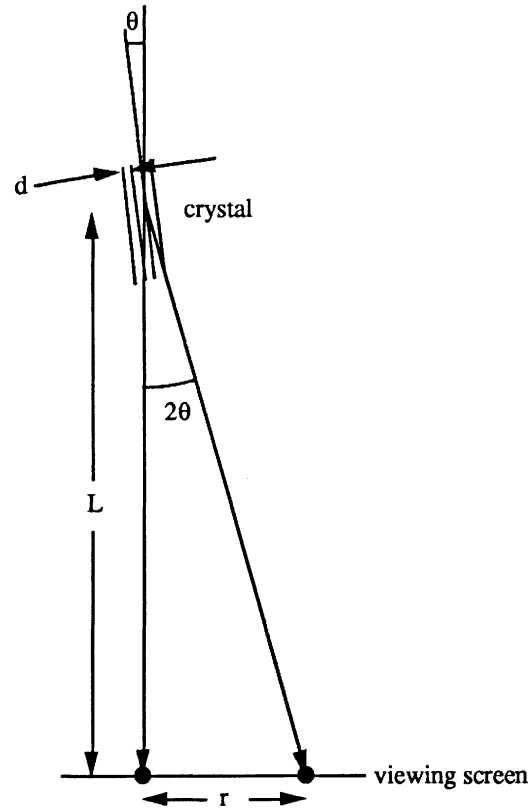


Figure 9. Geometry for electron diffraction and definition of cameral length L.

microscope. The intermediate lens is now focused on the back focal plane of the objective lens. The transmitted beam and all of the diffracted beams are imaged on the viewing screen. In this configuration, a second aperture, called an intermediate (or selected-area) aperture, can be positioned in the image plane of the objective lens, and this provides a means of confining the diffraction pattern to a selected area of the specimen. This technique of SAD is usually performed in the following way. The specimen is first examined in image mode until a region of interest is found (the tip of the solid arrow in the image plane in Fig. 6D). The intermediate aperture is then inserted and positioned around this feature. Owing to spherical aberration, it may be necessary to slightly underfocus the objective lens to ensure that the SAD pattern comes from the region of interest (Hirsch et al., 1977; Thomas and Goringe, 1979). The microscope is then switched into the diffraction mode. The SAD pattern that appears on the viewing screen originates from the area selected in the image mode (the tip of the solid arrow in the image plane). Selected-area diffraction can be performed on regions a micrometer or so in diameter, but spherical aberration of the objective lens limits the technique to regions not much smaller than this. For diffraction work from smaller regions, it is necessary to use small-probe techniques such as convergent-beam electron diffraction (CBED).

We can use the separation of the diffraction spots on the viewing screen to determine interplanar spacings in crystals, and this can be used to help determine the lattice parameter of a crystal or to identify an unknown phase. To do this, we need to know the camera length of the microscope. Consider the geometry of the SAD pattern in Figure 9, which shows the camera length L that is characteristic of the optics of the microscope. Bragg's law is given as

$$2d \sin \theta = n\lambda \tag{32}$$

where d is the interplanar spacing, λ is the electron wavelength, and n is an integer. The value of θ is on the order of a degree for low-order diffraction spots from most materials using 100-keV electrons ($\lambda = 0.0037$ nm). For such small angles,

$$\sin \theta \sim \tan \theta \sim \frac{1}{2}\tan(2\theta) \tag{33}$$

and from the geometry in Figure 9,

$$\tan 2\theta = \frac{r}{L} \tag{34}$$

where r is the separation of the diffraction spots on the viewing screen. If we substitute Equation 34 into Bragg's law (Equation 32) and arrange terms, we obtain

$$rd = \lambda L \tag{35}$$

Equation 35 is referred to as the camera equation. It allows one to determine an interplanar spacing d by measuring the separation of diffraction spots r in the

transmission electron microscope (or on a photographic negative). To do this, we need to know the product λL, known as the camera constant (in units of nanometers-centimeters), and its approximate value ($\pm 5\%$) can be found on a console display of a modern transmission electron microscope. For higher precision, one should perform a calibration of the camera constant using a standard such as an evaporated Au film and insert the unknown specimen into the microscope, keeping the lens conditions fixed while adjusting the specimen focus using the height (Z) control on the goniometer. Better still is to evaporate the film directly on part of the unknown specimen, or if the unknown is a precipitate in a matrix, for example, the matrix material can be used for calibration, as illustrated below (see Data Analysis and Initial Interpretation).

Indexing Diffraction Patterns

The reciprocal lattices of materials are three dimensional, so their diffraction patterns are three dimensional. Diffraction data from TEM are obtained as near-planar sections through diffraction space. The magnitude of the diffraction vector, $|\Delta\mathbf{k}|$, is obtained from the angle between the transmitted and diffracted beams. The interpretation of these data is simplified because the large electron wave vector provides an Ewald sphere that is nearly flat. This allows the convenient approximation that the diffraction pattern is sampling the diffaction intensities from a planar section of the reciprocal lattice. Two degrees of orientational freedom are required for the sample in TEM. These degrees of freedom are typically obtained with a double-tilt specimen holder, which provides two tilt axes oriented perpendicular to each other. Modern TEMs provide two methods for obtaining diffraction patterns from individual crystallites. One method is SAD, which is useful for obtaining diffraction patterns from regions as small as 0.5 μm in diameter. The second method is CBED (or microdiffraction/nanodiffraction), in which a focused electron beam is used to obtain diffraction patterns from regions as small as 1 nm. Both techniques provide a 2D pattern of diffraction spots, which can be highly symmetrical when a single crystal is oriented precisely along a zone axis orientation, as on the right in Figure 3.

We now describe how to index the planar sections of single-crystal diffraction patterns; i.e., label the individual diffraction spots with their appropriate values of h, k, and l. It is helpful to imagine that the single-crystal diffraction pattern is a picture of a plane in the reciprocal space of the material. (See SYMMETRY IN CRYSTALLOGRAPHY for more details on point groups.) Indexing begins with the identification of the transmitted beam, or the (000) diffraction spot. This is usually the brightest spot in the center of the diffraction pattern. Next, we need to index two independent (i.e., noncolinear) diffraction spots nearest to the (000) spot. Once these two vectors are determined, we can make linear combinations of them to obtain the positions and indices of all the other diffraction spots. To complete the indexing of a diffraction pattern, we also specify the normal to the plane of the spot pattern; this normal is the zone axis. By convention, the zone axis points toward the electron gun (i.e., upward in most transmission electron microsopes). The indexing of a diffraction pattern is not unique. If a crystal has high symmetry, so does its reciprocal lattice (although these symmetries may be different). A high symmetry leads to a multiplicity of different, but correct, ways to index a given diffraction pattern. For example, a vector cube axis can be a [100], [010], or [001] vector. Once the orientation of the diffraction pattern is specified, it is important to index the pattern consistently.

Most of the work in indexing diffraction patterns involves measuring angles and distances between diffraction spots and then comparing these measurements to geometrical calculations of angles and distances. When indexing a diffraction pattern, one must remember that structure factor rules eliminate certain diffraction spots. For consistency, one must also satisfy the "right-hand rule," which is given by the vector cross-product relation: $\mathbf{x} \times \mathbf{y} \parallel \mathbf{z}$. The procedures are straightforward for low-index zone axes of simple crystal structures but become increasingly difficult for crystal structures with low symmetry and for high-index zone axes, where many different combinations of interplanar spacings and angles provide diffraction patterns that look similar. In these cases, a computer program to calculate such diffraction patterns is extremely helpful (see Method Automation; also see Data Analysis and Initial Interpretation).

The procedure to index a diffraction pattern is illustrated by example below. Suppose we need to index the diffraction pattern in Figure 10A and we know that it is from an fcc crystal. The first step is to consult a reference and find an indexed diffraction pattern that looks like the one in Figure 10A. In this method, we guess the zone axis and its diffraction pattern. This method is most useful when the diffraction pattern shows an obvious symmetry, such as a square or hexagonal array of spots for a cubic crystal. With experience, one eventually remembers the symmetries for common fcc and body-centered cubic (bcc) diffraction patterns shown in Table 2. Without turning immediately to published SAD patterns, however, we first note that the pattern in Figure 10A is less symmetrical than those in Table 2. Nevertheless, we note that the

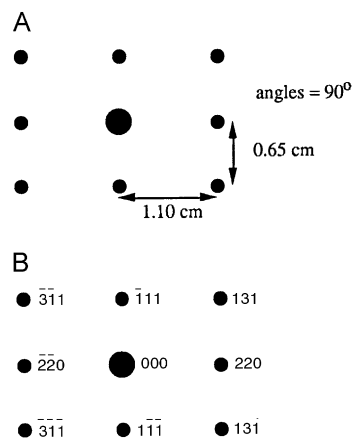


Figure 10. (**A**) An fcc diffraction pattern ready for indexing and (**B**) successful indexing of diffraction pattern in (**A**).

Table 2. Some Symmetrical Diffraction Patterns

Zone Axis	[100]	[110]	[111]
Symmetry	square	rectangular 1:$\sqrt{2}$ for bcc (almosthexagonal for fcc)	hexagonal

density of spots seems reasonably high, so we expect that we have a fairly low-order zone axis. The lowest order zone axes are

[100] [110] [111] [210] [211] [310]
[200] [220] [222]
[300]

Since the [200], [220], and [222] vectors point along the directions of [100], [110], and [111], we need only consider the lower index [100], [110], and [111] directions as candidate zone axes. We eliminate the first three zone axes because the pattern does not have the required symmetry as listed in Table 2. The diffraction pattern from the [310] zone axis is not rectangular, so the [210] and [211] patterns seem most promising.

If we knew the camera constant λL, it would be appropriate to work with absolute distances for the spot spacings. In this case, we could rearrange the camera equation $rd = \lambda L$ (Equation 35) to obtain the measured distance r of a diffraction spot from the transmitted beam by

$$r = \frac{\lambda L}{a}\sqrt{h^2 + k^2 + l^2} \tag{36}$$

Here we work with relative spacings instead. Equation 36 shows that the ratio of the spot distances must equal the ratio of $\sqrt{h^2 + k^2 + l^2}$. We first measure the spacings to the vertical and horizontal spots as reference distances (0.65 and 1.10 cm, respectively, from Fig. 10A originally). We then seek the ratios of $\sqrt{h^2 + k^2 + l^2}$ from the allowed (hkl) of an fcc crystal by making a list of possible distances, as shown in Table 3, and looking for two diffraction spots, preferably low-order ones, whose spacings are in the ratio of $0.65/1.10 = 0.591$. One can find by trial and error that $\sqrt{3}/\sqrt{8} = 0.61$. This ratio corresponds to the (111) and (220) diffraction spots, which seem promising candidates for further work. Note also that the diffraction pair (200) and (311), the pair (200) and (222), and the pair (220) and (422) also have good ratios of their spot spacing, but they contain higher order reflections.

Table 3. List of Allowed fcc (hkl) Reflections and Their Relative Interplanar Spacings

Allowed fcc (hkl)	$\sqrt{h + k^2 + l^2}$	Relative Spacing
(111)	$\sqrt{3}$	= 1.732
(200)	$\sqrt{4}$	= 2.000
(220)	$\sqrt{8}$	= 2.828
(311)	$\sqrt{11}$	= 3.317
(222)	$\sqrt{12}$	= 3.464
(400)	$\sqrt{16}$	= 4.000
(331)	$\sqrt{19}$	= 4.359
(420)	$\sqrt{20}$	= 4.472
(422)	$\sqrt{24}$	= 4.899

We need to choose specific vectors in the $\langle 111\rangle$ and $\langle 220\rangle$ families that provide the correct angles in the diffraction pattern. There are a number of ways to do this; here we use $[\bar{1}11]$ and [220]. Note that $[\bar{1}11] \cdot [220] = 0$, so these spots are consistent with the observed 90° angles. It turns out that we can eliminate two of our other three candidate pairs of diffraction spots—the pair (200) and (311) and the pair (200) and (222)—because no vectors in their families are at 90° angles. (For nonperpendicular normalized vectors, we find the angle between them by taking the arccosine of their dot product.) Now we complete the diffraction pattern by labeling the other diffraction spots by vector addition, as illustrated in Figure 10B. Finally, we get the zone axis of the crystal from the vector cross-product:

$$[220] \times [\bar{1}11] = (2-0)\hat{x} + (0-2)\hat{y} + (2--2)\hat{z} = [2\bar{2}4] = [1\bar{1}2] \tag{37}$$

The astute reader may wonder what happened to the candidate pair of (220) and (422) reflections, which also have good ratios of their spot spacings, and a 90° angle is formed by [220] and $[2\bar{2}4]$. We could have gone ahead and constructed a candidate diffraction pattern with these diffraction vectors. The zone axis is

$$[220] \times [2\bar{2}4] = (8-0)\hat{x} + (0-8)\hat{y} + (-4-4)\hat{z} = [8\bar{8}\bar{8}] = [1\bar{1}\bar{1}] \tag{38}$$

This should seem suspicious, however, because a $\langle 111\rangle$ zone axis provides a diffraction pattern with hexagonal symmetry (Table 2), quite unlike the rectangular symmetry of Figure 10. These {220} diffracted spots make a hexagonal pattern around the transmitted beam. Once the zone axis is identified, it is important to check again all expected diffraction spots and ensure that the diffraction pattern accounts for all of them. Once we have done so, it is clear that the (220) and (422) spots are inappropriate for indexing the pattern in Figure 10.

Having gone through the exercise of indexing the diffraction pattern in Figure 10, one can appreciate how tedious the practice could be for low-symmetry patterns with nonorthogonal vectors. As mentioned above, several excellent computer programs are available to help simplify the task (but do not trust them blindly). It is also important to note that the eye is able to judge distances to ~0.1 mm, particularly with the aid of a 10× calibrated magnifier, so the measurement accuracy of spot spacings in a diffraction pattern is typically about that good. A diffraction spot 10 mm from the center of the pattern yields a measurement error of 1%. For the highest accuracy in determining spot spacings, it is often preferable to measure the distance between sharper, higher order spots and then divide by the number of spots separating them (plus 1). Photographic printing can distort spot spacings, so measurements should be performed directly on negatives or the photographic enlarger should be checked for possible distortions.

Kikuchi Lines and Specimen Orientation

Electron Diffuse Scattering. In thick specimens, features are seen in electron diffraction patterns in addition to the usual Bragg diffraction spots and their fine structure. Inelastic scattering contributes a diffuse background to the diffraction pattern, even at moderate specimen thicknesses. More interestingly, intersecting sets of straight lines appear on top of the diffuse background. These are called Kikuchi lines (Kikuchi, 1928). They may be either bright or dark and are straight and regularly arranged. We can explain the existence of Kikuchi lines by a combination of two electron-scattering processes, the first one inelastic, followed by elastic Bragg diffraction. Although the first electron-scattering process is inelastic, a 100-keV electron does not lose too much energy, typically < 30 eV, owing to plasmon excitation, so the magnitude of its wave vector is nearly unaffected. Even for an energy loss of 1 keV, a 100-keV electron will undergo a change in wave vector of only $\sim 0.5\%$. The direction of the inelastically scattered electron deviates from the direction of the incident beam, but usually only by a small amount. There is a distribution of directions for inelastically scattered electrons, but it is important to note that this distribution is peaked along the incident direction, as illustrated at the top of Figure 11A. A diffuse background in the diffraction pattern originates from the inelastically scattered electrons that are not Bragg diffracted but are forward peaked, as shown at the bottom in Figure 11A. This diffuse background has a higher intensity for thicker specimens, at least until electron transparency is lost.

Origin of Kikuchi Lines. Kikuchi lines are features in the inelastic background that show the crystallographic structure of the sample. To create Kikuchi lines, a few of the inelastically scattered electrons must undergo a second scattering, which is an elastic Bragg diffraction event. It is important to recognize that most of the inelastically scattered electrons have lost relatively little energy and undergo a minimal change in their wavelength λ. Of these electrons, some may be diverted by the first inelastic scattering event so they become incident at a Bragg angle to some crystal planes. These electrons can then be Bragg diffracted.

The two rays drawn in Figure 11B (labeled "two inelastically scattered rays") are special ones, being rays in the plane of the paper that are oriented properly for Bragg diffraction by the crystal planes (hkl). Notice the two directions for the electrons coming out of the crystal, labeled K_{hkl} and $\mathbf{K}_{\bar{h}\bar{k}\bar{l}}$ in Figure 11B. The beam $\mathbf{K}_{hkl}$ consists of electrons that were Bragg diffracted out of the forward-scattered inelastic beam plus those electrons that were scattered inelastically at a modest angle but not diffracted. The other beam $\mathbf{K}_{\bar{h}\bar{k}\bar{l}}$ consists of the forward-scattered electrons plus those electrons that were Bragg diffracted into the forward direction (after first having been inelastically scattered by a modest angle). The important point is that the forward beam is stronger, so there are more electrons lost by secondary Bragg diffraction from the forward beam, $\mathbf{K}_{\bar{h}\bar{k}\bar{l}}$, than vice versa. Electron intensity in the diffuse background is therefore taken from the beam $\mathbf{K}_{\bar{h}\bar{k}\bar{l}}$ and added to the beam $\mathbf{K}_{hkl}$. As shown at the bottom of Figure 11B, the diffuse background has a deficit of intensity on the right and an excess of intensity on the left. The excess and deficit beams are separated by the angle 2θ.

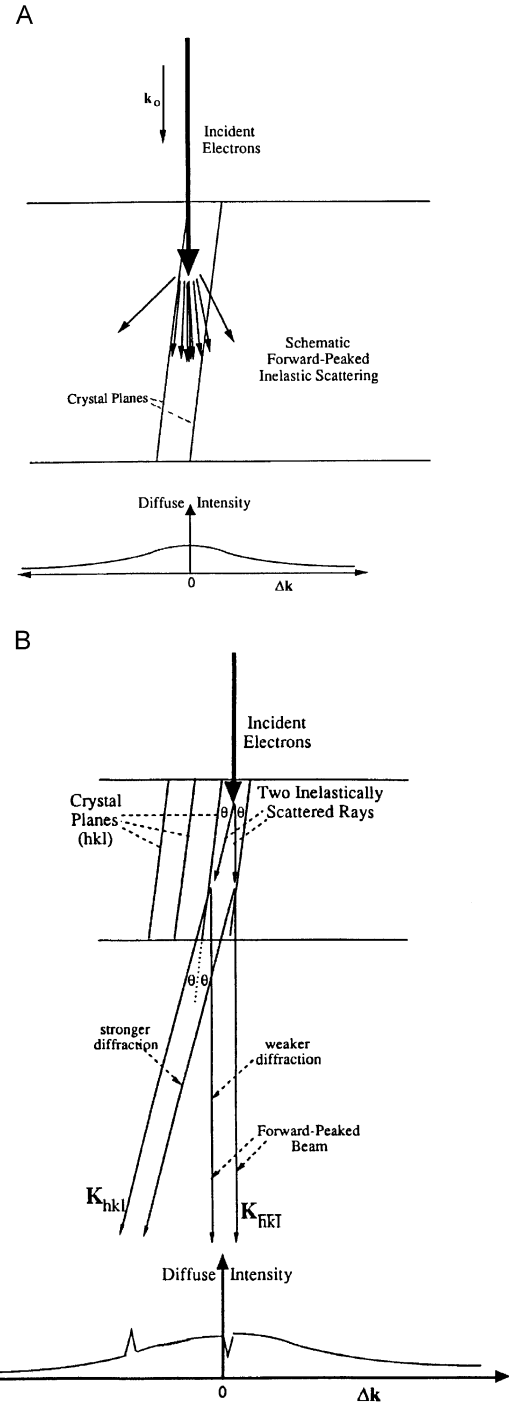


Figure 11. (**A**) Top: Electron paths through specimen. Bottom: Diffuse intensity recorded in diffraction pattern from forward-peaked inelastic scattering. (**B**) Origin of Kikuchi lines. Top: Electron paths through specimen for electrons subject to inelastic scattering followed by secondary Bragg diffraction. Bottom: Pattern of scattered intensity showing sharp modulations in diffuse intensity caused by Bragg diffraction.

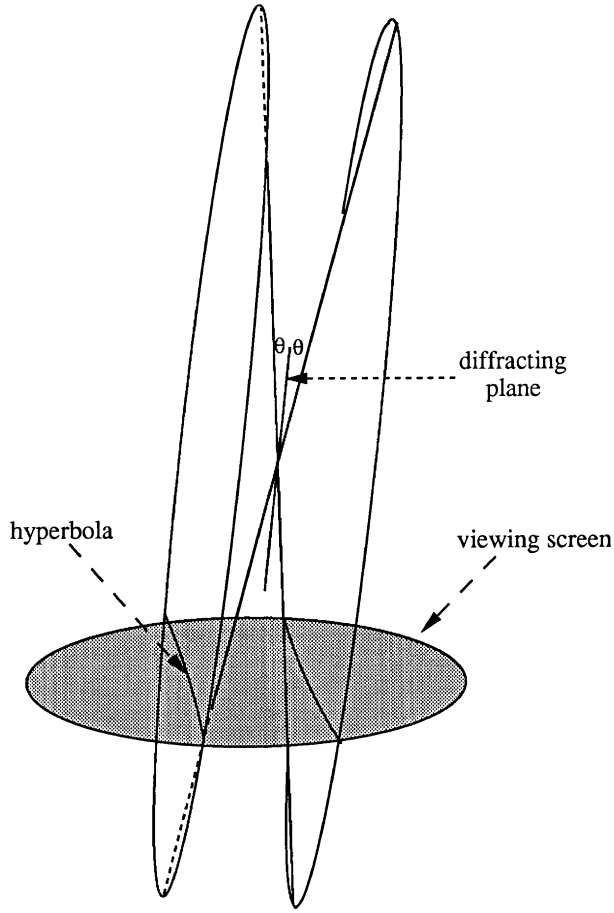


Figure 12. Intersection of Kossel cones with viewing screen of TEM.

Excesses and deficits in the inelastic background intensity can be produced in any direction that is at the Bragg angle to the crystal planes. In three dimensions, the set of these rays form pairs of Kossel cones, which are oriented symmetrically about the diffracting planes. All lines from the vertices of these cones make the Bragg angle θ with respect to the diffracting plane. In Figure 12, note how the minimum angle between the surfaces of the two cones is 2θ. The intersections of these two cones with the viewing screen of the microscope are nearly hyperbolas. Since θ is so small and the viewing screen is so far away, these hyperbolas appear as straight lines. The line closer to the transmitted beam is darker than the background (the deficit line); the other line is brighter (the excess line).

Because Kikuchi lines originate from inelastically scattered electrons, their intensity increases with the diffuse background in the diffraction pattern. Kikuchi lines are found to be weak or nonexistent in very thin regions of a specimen where the diffraction spot pattern is without much diffuse background. As one translates a specimen in TEM to observe thicker regions, the Kikuchi lines become more prominent, often becoming more observable than the diffraction spots themselves. The Kikuchi lines decrease in intensity when the sample becomes so thick that it is no longer electron transparent. Lastly, we note that by the explanation of Figure 11, there should be no excess or deficit bands in the symmetric case where the Kikuchi lines straddle the transmitted beam. Kikuchi lines are in fact observed, however, showing that our simple explanation of their origin is not adequate. A much more complicated explanation based on dynamical diffraction theory with absorption is necessary to fully account for the intensities of Kikuchi lines (Hirsch et al., 1977; Williams and Carter, 1996). Regardless of this, we can use the crystallographic symmetry of the Kikuchi lines in the TEM analysis of crystals.

Indexing Kikuchi Lines. For a specific set of crystal planes (hkl), the complement of the two vertex angles of the Kossel cones is 2θ (see Fig. 12). This is the same angle as between the transmitted beam and the (hkl) diffracted beam. On the viewing screen, the separation between the two Kikuchi lines will therefore be the same as the separation of the (hkl) diffraction spot from the (000) spot. We can index the Kikuchi lines by measuring their separations in much the same way as we index diffraction spots. For two different pairs of Kikuchi lines, from the planes $(h_1k_1l_1)$ and $(h_2k_2l_2)$, their separations between excess and deficit lines p_1 and p_2 are in the ratio

$$\frac{p_1}{p_2} = \frac{\sqrt{h_1^2 + k_1^2 + l_1^2}}{\sqrt{h_2^2 + k_2^2 + l_2^2}} \tag{39}$$

Figure 13 shows ratios of $\sqrt{6}$, 2, and $\sqrt{2}$ for indexed (211), (200), and (110) Kikuchi line pairs. The angles between the Kikuchi line pairs can be measured with precision. High-energy electrons have very obtuse vertex angles for their Kossel cones (small 2θ); it is useful to think of the Kossel cones as extensions of the crystal planes onto the viewing screen. At the same time it is useful to think of

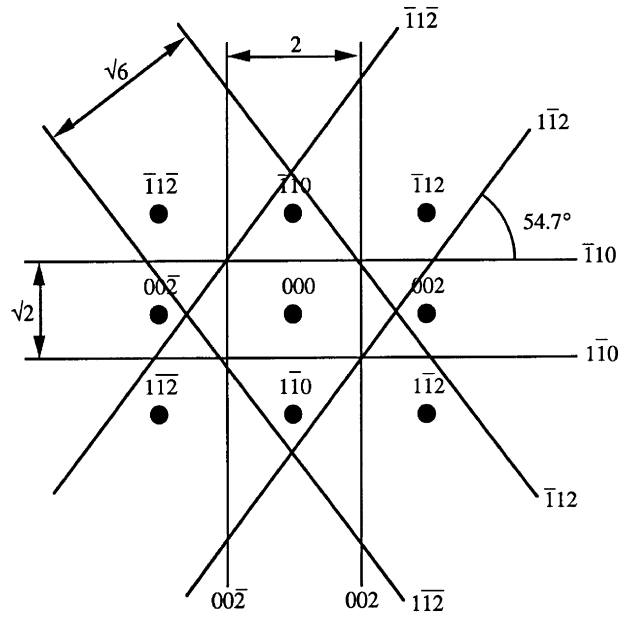


Figure 13. Some indexed Kikuchi lines for bcc diffraction pattern on a[110] zone axis.

the diffraction spots as representing the normals to the diffracting planes. It is then clear that the angles between intersecting Kikuchi line pairs will be the same as the angles between their corresponding diffraction spots, at least so long as the Kikuchi lines are not too far from the center of the viewing screen. These angles are helpful for indexing Kikuchi lines in the same way that the angles between pairs of diffraction spots were useful for indexing diffraction patterns above. For example, the angle ϕ between the $(1\bar{1}2)$ and $(1\bar{1}0)$ Kikuchi lines is given as

$$\phi = \arccos\left(\frac{1}{\sqrt{6}}(1\bar{1}2)\cdot\frac{1}{\sqrt{2}}(1\bar{1}0)\right) = 54.7^\circ \qquad (40)$$

Specimen Orientation and Deviation Parameter. Kikuchi lines are useful for precise determination of the specimen orientation in TEM. Recall that when we tilt the specimen, we tilt its reciprocal lattice. Tilts of the reciprocal lattice with respect to a stationary Ewald sphere do not cause any substantial changes in the positions of the diffraction spots, but the individual spots increase or decrease in intensity. Conversely, the positions of the Kikuchi lines vary sensitively with the specimen tilt and make an excellent tool for accurately determining specimen orientation. We see from Figure 12 that as the diffracting plane is tilted, the Kossel cones move by exactly the same angle. The Kikuchi lines behave as though they were affixed to the bottom of the crystal. Usually, we use a rather long camera length for diffraction work, so there is significant movement of the Kikuchi lines on the viewing screen when the specimen is tilted.

Figure 14 shows how the Kikuchi lines can be used to determine the sign and magnitude of the deviation parameter s, which quantifies how accurately the Bragg condition is satisfied for a particular reflection $\mathbf{g}$ (discussed above). This in turn determines the diffraction contrast and appearance of BF and DF TEM images. As illustrated in Figure 14, the distance between the Kikuchi lines of order $\mathbf{g}$ is r, where

$$r = 2\theta L \qquad (41)$$

This is the same distance r between the 000 and $\mathbf{g}$ (hkl) diffraction spots given in Equation 36, and L is the camera length shown in Figure 9.

Consider first the special situation where the Kikuchi lines intersect the transmitted beam and the diffraction spot $\mathbf{g}$ exactly, as shown in Figure 14A. This special situation corresponds to the exact Bragg condition because the transmitted beam is oriented at the angle θ with respect to the diffracting planes. In this special case, $\mathbf{s} = 0$. Now tilt the crystal counterclockwise into the arrangement on the right. The angle of tilt is

$$\phi = \frac{x}{L} \qquad (42)$$

where x is the distance between the diffracted spot and its corresponding bright Kikuchi line. Since the Kikuchi lines move as though they are affixed to the crystal, the bright

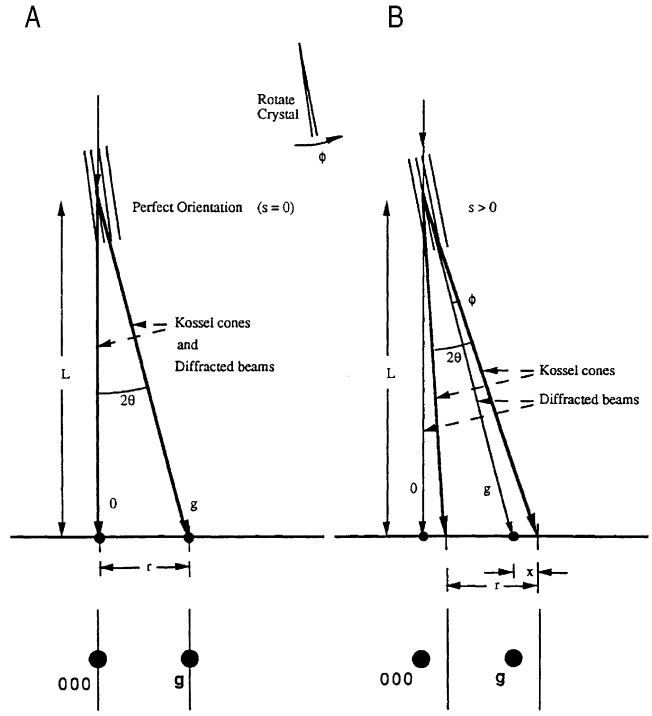


Figure 14. Geometry of crystal rotation and position of Kikuchi lines with respect to diffraction spots. Bottom portion of each figure illustrates positions of Kikuchi lines and diffraction spots as seen on viewing screen. (**A**) Intersection of Kikuchi lines with $\mathbf{g}$ vector occurs when specimen is at exact Bragg orientation for that $\mathbf{g}$ vector. (**B**) When crystal in (**A**) is rotated counterclockwise by angle ϕ, Kikuchi lines move right, in the same sense as crystal rotation.

Kikuchi line moves to the right of the spot, as illustrated in Figure 14B. When we rotate the crystal by the angle ϕ, we also rotate the reciprocal lattice with respect to the Ewald sphere by the same amount, as shown in Figure 15. By using Figure 15, we can find the relation between the magnitudes of $\mathbf{s}$ and $\mathbf{g}$:

$$\phi = \frac{s}{g} \qquad (43)$$

where g is the magnitude of $\mathbf{g}$, i.e., $|\mathbf{g}|$. Combining Equations 42 and 43 yields

$$\frac{s}{g} = \frac{x}{L} \qquad \text{or simply} \qquad s = \frac{gx}{L} \qquad (44)$$

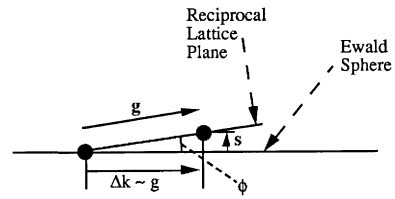


Figure 15. Relationship between deviation vector $\mathbf{s}$ and rotation of crystal by angle ϕ (as in Fig. 14).

where x is the distance from the bright Kikuchi line to the corresponding spot **g**. Note that it is possible to eliminate the camera length from Equation 44 by inserting Equation 41 for L, so the magnitude of **s** is given by

$$s = g2\theta\frac{x}{r} \tag{45}$$

Equation 45 shows how one can obtain a value for the deviation parameter s from the position of the Kikuchi lines with respect to the diffraction spots. Because θ (in radians) is small, we can determine very small values of s with accuracy. In electron diffraction, typical units for both s and g are reciprocal nanometers.

The Sign of *s*. We say that $s > 0$ if the excess Kikuchi line lies outside its corresponding diffraction spot **g**. In this case, the reciprocal lattice point lies inside the Ewald sphere, as was the case in Figures 14B and 15. Alternatively, we say that $s < 0$ if the excess Kikuchi line lies inside its corresponding diffraction spot, and we say that $s = 0$ when the Kikuchi line runs exactly through its corresponding diffraction spot, e.g., Figure 14A. The convention is that s points from the Ewald sphere to the reciprocal lattice point, and s is positive if **s** points upward along the positive z axis. This is consistent with the relationship $\mathbf{g} = \Delta\mathbf{k} + \mathbf{s}$ given previously in Equation 11, and shown in the Ewald sphere construction in Figure 4.

Lens Defects and Resolution

Important performance figures in TEM are the smallest spatial features that can be resolved in a specimen, or the smallest probes that can be formed on a sample. Electromagnetic lenses have aberrations that limit their performance, and to understand resolution, we must first understand lens aberrations. We can then consider the cumulative effect of all lens defects on the performance of a TEM.

Spherical Aberration. Spherical aberration distorts the focus of off-axis rays. The further the ray deviates from the optic axis, the greater is its error in focal length. All magnetic lenses have a spherical aberration coefficient that is positive so that those rays furthest from the optic axis are focused most strongly. The angle of illumination into the lens is defined as the aperture angle α; in paraxial imaging conditions, α is small. For reference, we define the true (or Gaussian) image plane as the image plane for paraxial imaging conditions. Spherical aberration causes an enlargement of the image of a point P to a distance QQ', as illustrated in Figure 16. The minimum enlargement of point P occurs just in front of QQ' and is termed the disk of least confusion. For a magnetic lens, the diameter d_s of the disk of least confusion caused by spherical aberration corresponds to a distance d_s on the specimen (Hirsch et al., 1977; Reimer, 1993; Williams and Carter, 1996):

$$d_s = 0.5C_s\alpha^3 \tag{46}$$

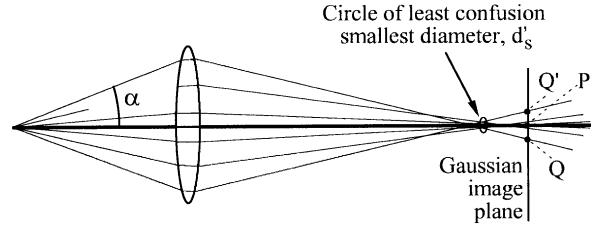


Figure 16. Lens with positive spherical aberration, i.e., positive C_s, forming disk of confusion with diameter d_s.

where C_s is the spherical aberration coefficient (usually 1 to 2 mm) and α is the semiangle of convergence of the electron beam.

Chromatic Aberration. Magnetic lenses are not achromatic. Electrons entering the lens along the same path but having different energies have different focal lengths. The spread in focal lengths is proportional to the spread in energy of the electrons. There are two main sources of this energy distribution. First, the electron gun does not produce monochromatic electrons. Typically, $<\pm 1$ eV of energy spread can be attributed to irregularities of the high-voltage supply. There is also a thermal distribution of electron energies as the electrons are thermionically emitted from a hot filament, which contributes on the order of ± 0.1 eV of energy spread. With high beam currents, there are also Coulomb repulsions of the electrons at the condenser cross-over. This contributes an energy spread of ± 1 eV through a phenomenon known as the Boersch effect. The second main cause of nonmonochromaticity of electrons is the specimen itself. Inelastic scatterings of the high-energy electrons by plasmon excitations are a common way for electrons to lose 10 to 20 eV. Thin specimens will help minimize the blurring of TEM images caused by chromatic aberration. The disk of least confusion for chromatic aberration has a diameter d_c at the specimen,

$$d_c = \frac{\Delta E}{E}C_c\alpha \tag{47}$$

where $\Delta E/E$ is the fractional variation in electron beam voltage and C_c is the chromatic aberration coefficient ($\sim$1 mm).

Diffraction from Apertures. Diffraction of the electron wave from the edge of an aperture contributes a disk of confusion corresponding to a distance at the specimen of d_d:

$$d_d = \frac{0.61\lambda}{\alpha_{OA}} \tag{48}$$

where λ is the electron wavelength and α_{OA} is the aperture angle of the objective lens. Equation 48 is the classic Rayleigh criterion that sets the resolution in light optics (Smith and Thomson, 1988). In essence, Equation 48 states that when the intensity between two point (Gaussian) sources of light reaches 0.81 of the maximum intensity of the sources, they can no longer be resolved.

Astigmatism. Astigmatism occurs when the focusing strength of a lens varies with angle θ around the optic axis, again leading to a spread of focus and a disk of least confusion. Two lenses of the transmission electron microscope require routine correction for astigmatism. To obtain a circular probe on the specimen, we must correct for astigmatism in the second condenser lens. Similarly, objective lens astigmatism blurs the image and degrades resolution, and it is important to correct astigmatism of the objective lens. Unlike spherical aberration, it is possible to use stigmators to correct for astigmatism. This correction can in fact be performed so well that astigmatism has a negligible effect on image resolution.

A stigmator in a modern transmission electron microscope is a pair of magnetic quadrupole lenses arranged one above the other and rotated with respect to each other. Correction of objective lens astigmatism is one of the more difficult skills to learn in electron microscopy. This correction is particularly critical in HRTEM, where the image detail depends on the phases of the beams and hence on the symmetry of the magnetic field of the objective lens about the optic axis. The astigmatism correction is tricky because three interdependent adjustments are needed: (1) main focus (adjustment of objective lens), (2) adjustment of x stigmator, and (3) adjustment of y stigmator. These adjustments must be performed in an iterative manner using features in the image as a guide. The procedure to accomplish this is a bit of an art and a matter of personal preference. A holey carbon film is an ideal specimen for practicing this correction. When the objective lens is overfocused (strong current) or underfocused (weak current) with respect to the Gaussian image plane, dark and bright Fresnel fringes, respectively, appear around the inside of a hole.

Resolution. Here we collect the results of lens defects (besides astigmatism) described above to obtain a general expression for the resolution of the transmission electron microscope for its two important modes of operation. In one case, we are concerned with the smallest probe that we can put on a specimen. In high-resolution imaging, we are concerned with the smallest feature in the specimen that can be resolved. We begin by obtaining an expression for the minimum probe diameter on a specimen, d_0.

If no aberrations were present in the optics of an electron microscope, the minimum probe size at the specimen could be readily calculated. The probe diameter d_0 is related to the total probe current I_p by the expression (Reimer, 1993)

$$I_\mathrm{p} = \frac{1}{4}\pi d_0^2 j_0 \tag{49}$$

where j_0 is the current density (in amperes per square centimeter) in the probe and I_p is j_0 times the area of the probe. Conservation of gun brightness on the optic axis implies that

$$j_0 = \pi\beta\alpha_\mathrm{p}^2 \tag{50}$$

where α_p is the semiangle of beam convergence of the electron probe on the specimen and β is the brightness, defined as current density per solid angle. Substituting the second expression into the first and solving for d_0 yield

$$d_0 = \frac{\sqrt{4I_\mathrm{p}/\beta}}{\pi\alpha_\mathrm{p}} = \frac{C_0}{\alpha_\mathrm{p}} \tag{51}$$

So for a given probe current I_p, small values of the probe diameter d_0 are obtained by increasing the brightness β and semiangle of convergence α_p. Because of lens aberrations, however, α_p has a maximum value, and β is limited by the design of the electron gun.

A general expression for the minimum probe size and image resolution can be obtained by summing in quadrature all diameters of the disks of least confusion, d_s, d_c, d_d, and d_0, as

$$d_\mathrm{p}^2 = d_\mathrm{s}^2 + d_\mathrm{c}^2 + d_\mathrm{d}^2 + d_0^2 \tag{52}$$

Substituting the diameters of these disks of least confusion from Equations 46, 47, 48, and 51 yields

$$d_\mathrm{p}^2 = \frac{C_0^2 + (0.61\lambda)^2}{\alpha_\mathrm{p}^2} + 0.25C_\mathrm{s}^2\alpha_\mathrm{p}^6 + \left(\alpha_\mathrm{p}C_\mathrm{c}\frac{\Delta E}{E}\right)^2 \tag{53}$$

For a thermionic gun with a large probe current, $C_0 \gg \lambda$ and the contributions of d_d and d_c can be neglected. In this case, the diameters d_0 and d_s superpose to produce a minimum probe diameter d_min at an optimum aperture angle α_opt for a constant I_p. The optimum aperture angle is found by setting $(d/d\alpha_\mathrm{p})d_\mathrm{p} = 0$, giving

$$\alpha_\mathrm{opt} = \left(\frac{4}{3}\right)^{1/8}\left(\frac{C_0}{C_\mathrm{s}}\right)^{1/4} \tag{54}$$

and substitution in Equation 53 yields

$$d_\mathrm{min} = \left(\frac{4}{3}\right)^{3/8} C_0^{3/4}C_\mathrm{s}^{1/4} \cong 1.1C_0^{3/4}C_\mathrm{s}^{1/4} \tag{55}$$

In the case of image resolution in TEM (or for a field emission gun where the total probe current is lower so that $C_0 \ll \lambda$), the contributions of d_0 and d_c can be neglected. Superposition of the remaining terms again yields a minimum that is less than with d_0. In this case, α_opt and d_min are given by

$$\alpha_\mathrm{opt} = 0.9\left(\frac{\lambda}{C_\mathrm{s}}\right)^{1/4} \tag{56}$$

$$d_\mathrm{min} = 0.8\lambda^{3/4}C_\mathrm{s}^{1/4} \tag{57}$$

These expressions can be used to calculate the optimum aperture angle and the resolution limit (or minimum probe size) of a transmission electron microscope. (Note that the same equations with slightly different values of the leading constants, which are all near 1, are found in the literature depending on the exact expressions used for the

quantities in Equation 52.) In a modern field emission gun TEM, it is common to achieve spatial resolutions <0.2 nm and probe sizes of 0.5 nm for a 200-keV microscope with $C_s = 1.0$ mm. It is important to note that these quantities represent the instrumental resolution limits and are well-defined, calculated quantities. The actual resolution that one can achieve for any given sample depends on a variety of specimen parameters, such as the sample thickness, the diffracting conditions, and how localized scattering is in the sample, so that the image resolution actually achieved in a typical BF or DF image is usually one or two orders of magnitude larger than the instrument resolution. For example, the width of the dark contrast associated with an edge dislocation in a BF image is $\sim \xi_g/3$ (Hirsch et al., 1977), or 10 nm wide for low-index reflections in metals (refer to Table 2).

Detailed alignment procedures for the transmission electron microscope are provided by the instrument manufacturers. In addition, the book by Chescoe and Goodhew (1984) and Chapters 5 to 9 in Williams and Carter (1996) explain the basic TEM components and their alignment procedures.

METHOD AUTOMATION

Transmission electron microscopy traditionally has been operator intensive with little computer automation except during postmicroscope analysis of data or when the microscope is operated in the scanning (STEM) mode. This trend is changing rapidly since the evolution of digital charge-coupled diode (CCD) cameras and faster computers now makes it possible to acquire and process images in real time on the microscope (Egerton, 1996; Williams and Carter, 1996). Companies such as Gatan offer several different types of CCD cameras and software packages such as Autotuning and DigitalMicrograph, which allow the user to correct objective lens astigmatism and process images on the microscope. Some TEM manufacturers have hardware/software that can be installed to step a probe across the sample to perform x-ray analyses across an interface or to automatically measure the spot spacings and angles in a diffraction pattern on the screen, for example. The specimen stages in TEM can now be computer controlled, allowing the user to input or recall particular specimen positions and/or tilts in the microscope. In addition, considerable effort is being made to establish remote microscopy centers in national electron microscope laboratories (Voelkl et al., 1998). Microscopes are becoming automated in many other ways, and it is likely that this trend will continue, although the operator will still need to plan the desired investigation.

DATA ANALYSIS AND INITIAL INTERPRETATION

The theories underlying electron diffraction and imaging were outlined (see Principles of the Method) and practical considerations associated with implementation of the theory in TEM were described (see Practical Aspects of the Method). The present section uses several examples to illustrate how we can analyze and interpret typical data from TEM. It should be emphasized that a variety of computer programs are available to calculate, process, and analyze diffraction patterns and images.

Effect of Shape Factor

One of the most important aspects in our discussion of the shape factor in Equations 8 to 10 was that the width of diffracted intensity in reciprocal space, i.e., the SAD pattern, varies inversely with the dimensions of a crystal in real space. This shape factor effect is often observed in diffraction patterns from multiphase materials containing small plate-shaped or rod-shaped precipitates, as illustrated below.

Figure 17 shows an example of the shape factor effect in an Al-Cu alloy, where very thin Cu-rich precipitates form in the Al-rich matrix. These precipitates, called Guinier-

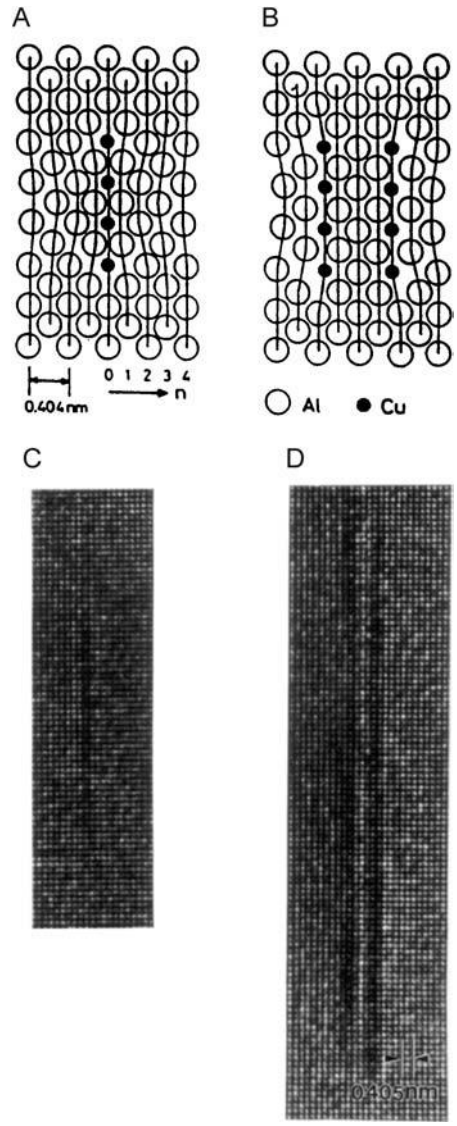


Figure 17. (**A**) Illustration of GP(1) zone, (**B**) illustration of GP(2) zone, (**C**) HRTEM image of GP(1) zone in Al-Cu, and (**D**) HRTEM image of GP(2) zone in Al-Cu. (From Chang, 1992.)

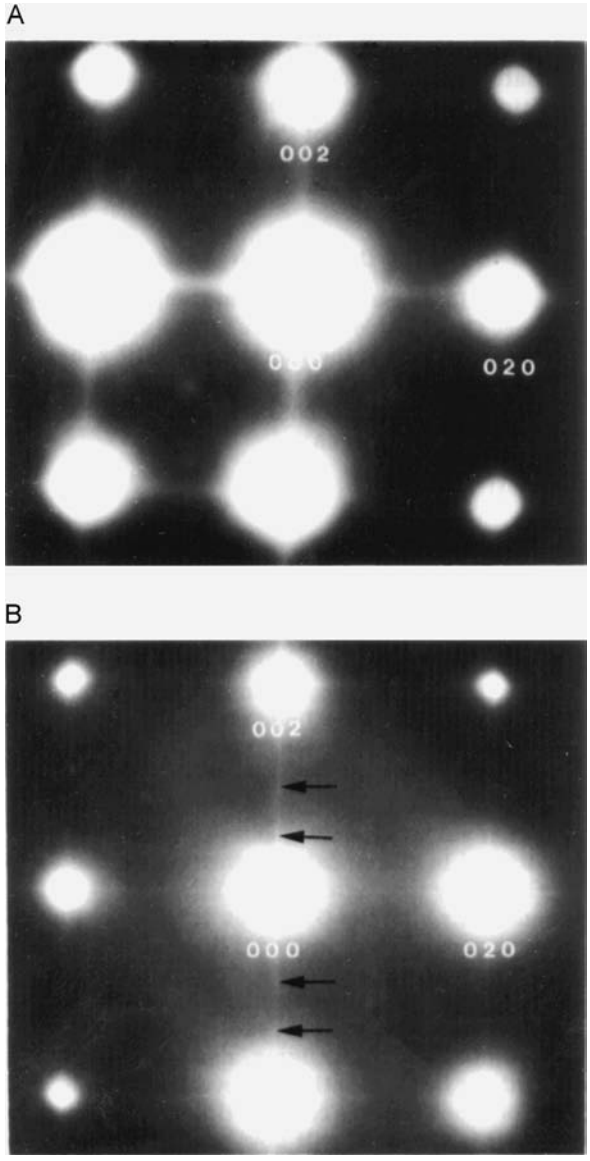


Figure 18. SAD pattern from sample containing (**A**) GP(1) zones and (**B**) GP(2) zones in Al-Cu alloy. (From Rioja and Laughlin, 1977.)

Preston (GP) zones, lie on the {100} planes and have a thin disk shape. When the disks have a thickness of only an atom or two, the broadened diffraction spots appear as streaks rather than as discrete spots, and two types of GP zones are found. As illustrated in Figures 17A and B, a GP(1) zone consists of a single layer of Cu atoms that have substituted for Al on a {100} plane, while a GP(2) zone contains two such Cu layers separated by three {100} planes of Al atoms. Experimental HRTEM images of these two types of GP zones, taken along a ⟨100⟩ matrix direction in an Al–4 wt% Cu alloy, are shown in Figures 17C and D (Chang, 1992). The atomic columns in the high-resolution images are white and the Cu-rich layers appear as darker planes in the images.

Figure 18 shows two diffraction patterns from samples that were aged to contain GP(1) and GP(2) zones. In Figure 18A, the streaks from the single layer of Cu atoms are practically continuous along the two ⟨100⟩ directions in the plane of the figure because the precipitate is essentially a monolayer of Cu atoms, as in Figures 17A and C. In Figure 18B, the streaks along ⟨100 are no longer continuous but have maxima at 1/4⟨100⟩ positions in the diffraction pattern. This periodicity arises because the Cu planes in the GP(2) zone are spaced four {100} planes apart, as illustrated in Figures 17B and D. (Note that the precipitate reflections still appear as streaks because these plates are very thin.) This illustrates the important point that for every real-space periodicity in a specimen, there is a corresponding reciprocal space intensity, and diffraction patterns and corresponding images should be carefully compared to interpret microstructures.

Variations in Specimen Thickness and Deviation Parameter

As described below (see Sample Preparation), many TEM specimens are thinned to perforation using techniques such as electropolishing or ion milling, so that the specimens are thin foils that vary in thickness from the edge of the perforation (hole) inward. In such specimens, it is common to observe thickness fringes running parallel to the edge of the hole, particularly when the specimen is oriented in a strong two-beam diffracting condition with $s = 0$ for a particular $\mathbf{g}$. These fringes are a result of the variation in diffracted intensity $\mathbf{Ig}$ (and $\mathbf{I}_0$) in Equation 22 as a function of t for constant s, i.e., for a fixed specimen orientation. Figures 19A and B show complementary BF and DF images of thickness fringes in Al, taken by placing a small objective aperture around the optic axis and tilting the beam to obtain the DF image, as illustrated in Figures 6 through 8. Note that the intensity in the BF and DF images is complementary, as expressed by Equation 20. For a constant value of s, the intensity varies periodically with t, becoming zero each time the product st in Equation 22 is an integer, as illustrated schematically in Figure 5. Thus, the fringes in a DF image are dark for a thickness $t = n/s$, where n is an integer and bright for $t = (n + \frac{1}{2})/s$; the reverse is true for a BF image. We can determine the magnitude of $\mathbf{s}$ from the position of the Kikuchi line in a diffraction pattern (refer to Equation 45 and Fig. 14). If we also know the value of ξ_g (Table 1), we can estimate the sample thickness anywhere in the foil based on the position of the thickness fringes. If $s = 0$ in a BF image, the foil is $\frac{1}{2}\xi_g$ thick in the middle of the first dark fringe, $\frac{3}{2}\xi_g$ in the middle of the second dark fringe, etc. (Fig. 19C); the reverse is true for the DF image. It is possible to calculate the periodicity of the fringes for quantitative comparison with experimental images such as those in Figure 19 using Equations 22 to 25. If the sample is tilted in the microscope, the thickness fringes remain parallel to the edge of the foil, but their spacing changes because s is changing.

Transmission electron microscopy foils of ductile materials such as metal alloys often bend during specimen preparation or handling, so that bend contours are observed in the TEM foil. These intensity fringes are a result of the variation in diffracted intensity $\mathbf{Ig}$ (and $\mathbf{I}_0$) in Equation 22 as a function of s for constant t, i.e., changing specimen

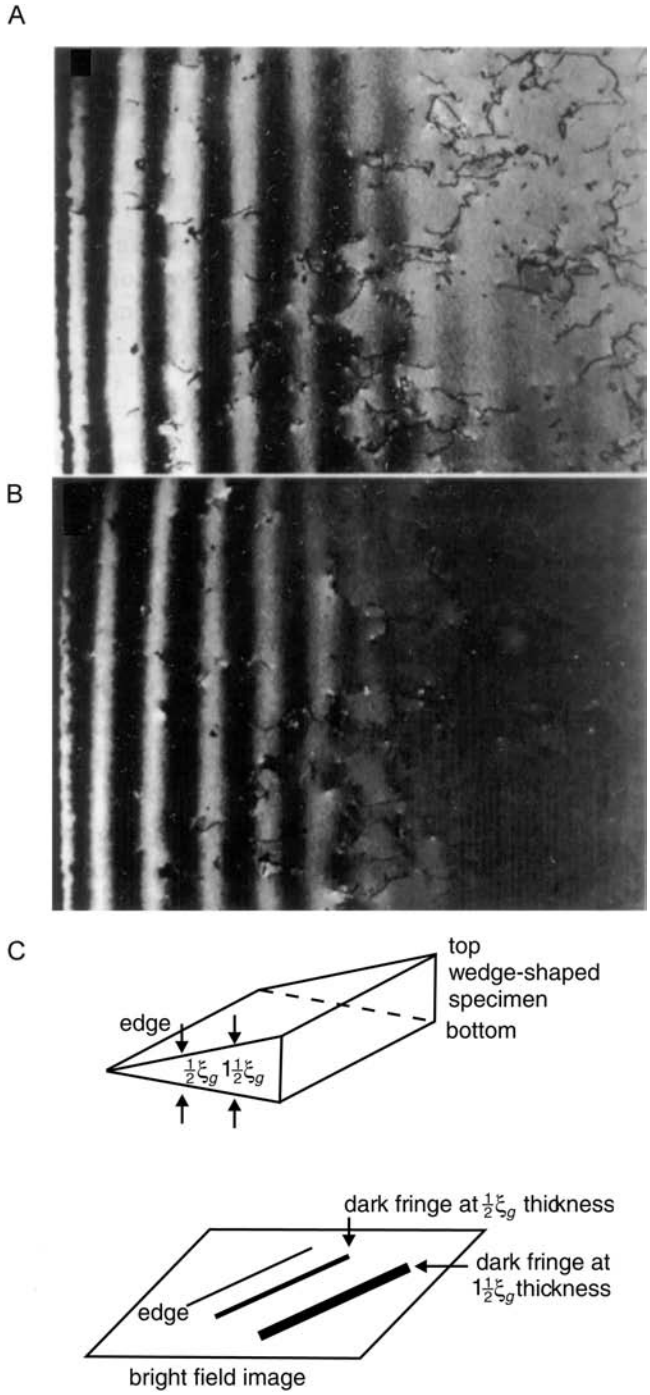


Figure 19. Complementary (**A**) BF and (**B**) DF images of thickness fringes in Al taken under dynamical conditions with $s = 0$, and (**C**) schematic showing position of fringes in BF image relative to wedge-shaped specimen. Foil edge is on left.(From Edington, 1974.)

orientation for a fixed thickness. Figure 20 shows complementary BF and DF images of bend contours in a metal foil taken under conditions similar to those in Figure 19. The specimen thickness is approximately constant horizontally across the field of view, and the specimen is slightly curved (bent) about a vertical axis in the center of the images. Thus, for a constant value of t, the intensity varies periodically with s horizontally across the images. However, unlike thickness fringes where t continually increases, s varies from negative to zero to positive as the specimen bends through orientation space. The DF image in Figure 20B shows only the left contour as a bright line for the diffracting planes $-\mathbf{g}$. Diffracted intensity $\mathbf{I}_{-g}$ has a symmetric profile about the maximum at $s = 0$, indicated by a line in the figure. In contrast to the thickness fringes in Figure 19, the corresponding BF image in Figure 20A is not symmetric across $-g$ (or $\mathbf{g}$) or complementary to the DF image, for reasons that can be accounted for by the dynamical theory of electron diffraction. Examining the BF image in Figure 20A shows that the crystal is generally darker between $\mathbf{g}$ and $-\mathbf{g}$ when $s < 0$ than when $s > 0$, with the darkest line of intensity lying approximately at $s = 0$, again indicated by lines in the figure. Unlike thickness fringes, bend contours can lie along any direction with respect to the edge of the foil and generally move across the sample when it is tilted.

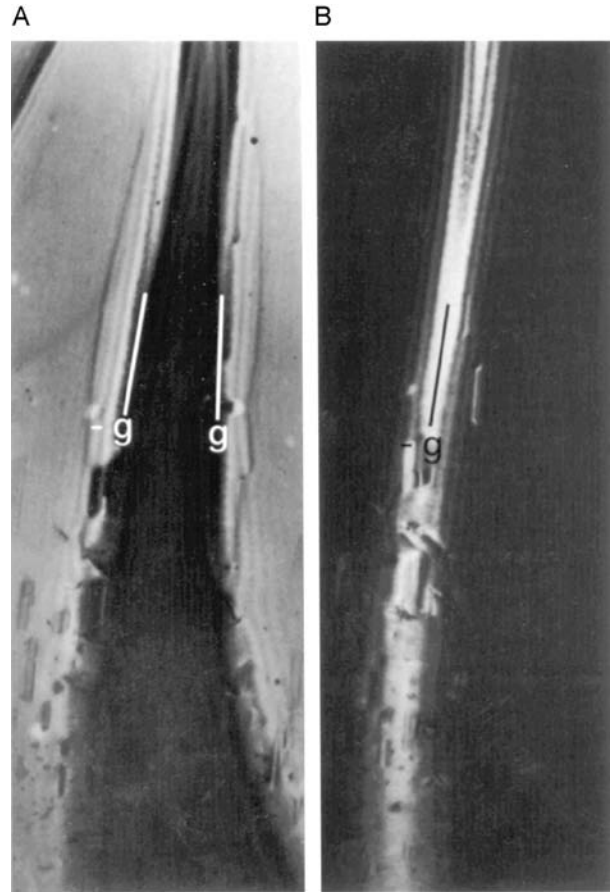


Figure 20. Complementary (**A**) BF and (**B**) DF images of bend contour, with thickness constant horizontally across images. Left side of bend contour in (**A**) is bright in (**B**). (From Thomas and Goringe, 1979.)

Use of Bright-Field, Dark-Field, and Selected-Area Diffraction as Complementary Techniques

Using BF/DF imaging and SAD as complementary techniques to understand the microstructures of materials is one

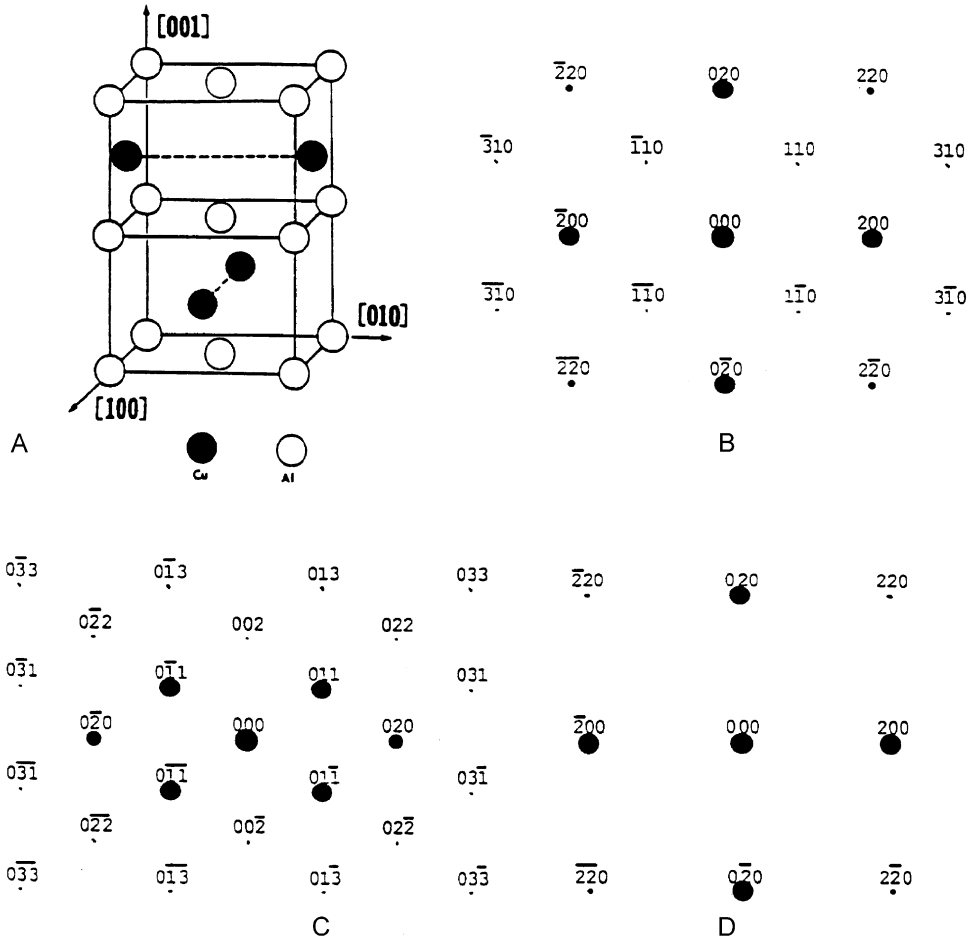


Figure 21. (**A**) Crystal structure of the θé phase, (**B**) calculated diffraction pattern for θé phase along [001] and (**C**) along [100], and (**D**) calculated diffraction pattern for [100] Al.

of the most important aspects of CTEM. Here we illustrate the use of these techniques to understand the microstructure of an Al-Cu alloy (the same alloy described in the discussion of shape factors above) heat treated to form θ′ precipitate plates on the {100} planes in the fcc Al-rich matrix. The θ′ plates have the tetragonal crystal structure illustrated in Figure 21A with a composition of Al_2Cu. The [001] axis of the θ′ plates contains the fourfold axis while the [100] and [010] axes of the θ′ phase have twofold rotation axes and are crystallographically indistinguishable. Diffraction patterns calculated for the θ′ phase along the [001] and [100] zone axes are shown in Figures 21B and C, and a calculated [100] diffraction pattern for Al is shown in Figure 21D.

The θ′ phase forms in the Al-rich matrix such that the [001] axis of θ′ aligns along each of the three possible ⟨100⟩ directions in the matrix. When the sample is viewed along a ⟨100⟩ matrix direction, one variant of the θ′ plates is seen face-on, while the other two variants are seen edge-on. Figure 22A shows a BF TEM image of all three variants of the θ′ precipitates, and Figure 22B shows a corresponding SAD pattern, taken with a large aperture containing all three variants of precipitates. The square pattern of bright spots is from the Al-rich matrix, which is the majority phase, while the weaker spots, particularly noticeable in the middle of the square pattern of Al spots, arise from the three variants of θ′ plates. A schematic of the diffraction pattern that distinguishes the three variants of spots from the θ′ plates is shown in Figure 22C.

It is possible to determine which precipitate spots in the SAD pattern correspond to each variant by forming DF images with the precipitate spots. This procedure is illustrated in Figure 22D, where a small objective aperture was centered on the optic axis and the electron beam was tilted (see Practical Aspects of the Method) to move the precipitate spot indicated by an arrow in Figures 22B and C into the aperture. As a result, only the vertical plates in Figure 22D appear bright, verifying that the reflection came from that variant. It is also possible to determine which precipitate spots correspond to a particular variant of θ′ by placing a selected-area aperture around one variant and recording its corresponding SAD pattern. This procedure is illustrated in Figures 22E and F, where an aperture placed around the vertical plate reveals that its spots form the rectangular diffraction pattern outlined in Figure 22F. This pattern is identical to the calculated pattern for θ′ in Figure 21C. This procedure can be repeated on the other variants to account for all of the precipitate spots in Figure 22A.

It is also possible to determine the lattice parameters of the θ′ phase using the Al-rich matrix spots for calibration and Equation 35. For example, Al has a lattice parameter

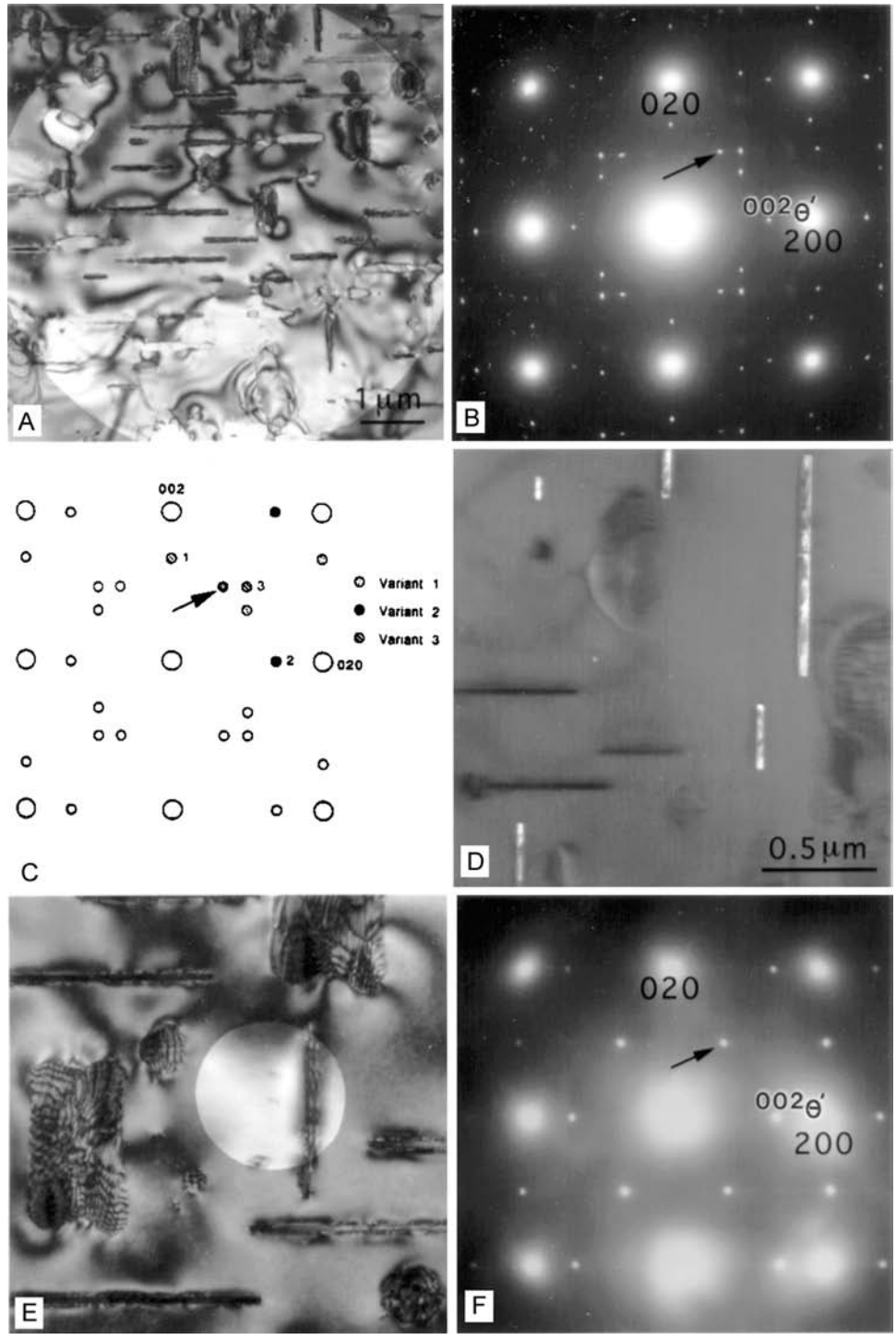


Figure 22. (**A**) BF TEM image of θ′ plates viewed along $a\langle 100\rangle$ Al direction; (**B**) corresponding SAD pattern containing weak spots from all three variants of θ′ plates and stronger spots from Al matrix (labeled); (**C**) schematic of diffraction pattern in (**B**) with arrow indicating which precipitate spot was used for axial DF imaging; (**D**) corresponding DF image showing bright, vertical θ′ plates; (**E**) SAD placed around vertical θ′ plate (revealed by double exposure); and (**F**) corresponding SAD pattern, similar to that in Figure 21C.

of 0.40496 nm so that the {200} interplanar spacing $d_{\{200\}} = 0.2025$ nm. The {200} spots in Figure 22B have a spacing $r = 21.0$ mm on the original image. Substituting these values for r and $d_{\{200\}}$ into Equation 35 gives $\lambda L = 0.425\ \text{nm}\cdot\text{cm}$. The [002] spots of the θ′ phase (labeled in Fig. 22F) have a spacing $r = 14.8$ mm, which, when substituted into Equation 35 with λL then gives $d_{(002)} = 0.575$ nm. This is 0.8% different than the reported value for the c lattice parameter of the θ′ phase of 0.58 nm (Weatherly and Nicholson, 1968). Since the [200] spots of θ′ phase coincide with the {200} spots of the Al matrix, the a lattice parameter of θ′ must be approximately the same as that of Al, or 0.40496 nm. This good matching of lattice parameters between the two phases is why the phase forms as plates on the {100} planes of Al.

Use of $\mathbf{g}\cdot\delta\mathbf{r} = 0$ in Defect Analysis

The same θ′ plates discussed above can be used to illustrate the application of the $\mathbf{g}\cdot\delta\mathbf{r}$ criterion for defect

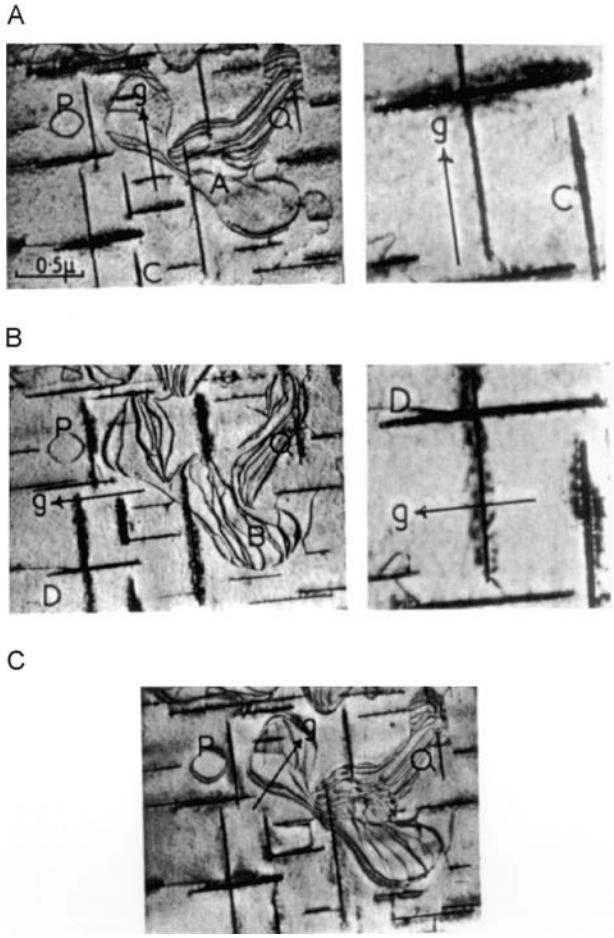


Figure 23. Burgers vector determination for misfit dislocations A, B, C, and D on flat faces of θ′ plates in Al-Cu alloy aged 1 day at 275°C. P represents dislocations around periphery of plates and Q complex dislocations arising from agglomeration of two plates: (**A**) $\mathbf{g} = [200]$, dislocations A visible with enlargement showing dislocations C; (**B**) $\mathbf{g} = [020]$, dislocations B visible with enlargement showing dislocations D; (**C**) $\mathbf{g} = [220]$, all dislocations visible. (From Weatherly and Nicholson, 1968.)

analysis in TEM (see Diffraction Contrast from Lattice Defects, above). Figure 23 shows the three variants of θ′ plates viewed along a ⟨100⟩ matrix direction at fairly high magnification (Weatherly and Nicholson, 1968). Four types of dislocation arrays are visible in the BF TEM images.

1. Complex arrays with no simple Burgers vector (for a definition of the Burgers vector, refer to Hull and Bacon, 1984) that seem to be associated with the agglomeration of two or more θ′ plates (Q in Fig. 23).
2. Dislocations associated with the periphery of the θ′ plate (P in Fig. 23). These frequently show a characteristic double or triple image and a line of no contrast that is typical of pure edge dislocations having a Burgers vector parallel to the electron beam (Hirsch et al., 1977). The Burgers vector of these dislocations is $a[001]$ for a precipitate plate lying parallel to (001); these dislocations accommodate the large misfit (∼4.3%) associated with the peripheral interface of the plate and the c direction of the θ′ lattice.
3. Dislocations associated with the flat face of the θ′ plate (A and B in Fig. 23). These are affected by the contours of the θ′ plates but lie predominantly in the [100] and [010] directions for a plate on (001). It is shown below that the Burgers vectors of these dislocations are $a[100]$ and $a[010]$.
4. Dislocations associated with plastic deformation of the foil (either accidental or deliberate). They tend to lie in the [110] or [110] directions for a θ′ plate on (001) and have Burgers vectors of the type $a/2[110]$.

Here we concentrate on the genuine misfit dislocations associated with the flat interface of the θ′ plate, type 3. The Burgers vector analysis that follows is thus confined to dislocations lying away from the periphery of the plates.

The four sets of dislocations A, B, C, and D are marked in Figure 23. If the dislocations are discrete loops wrapped around the θ′ plates, an analysis in a [001] foil gives a complete determination. The results in Figure 23 show that when $\mathbf{g} = [\bar{2}00]$, dislocations A and C are visible (Fig. 23A); when $g = [0\bar{2}0]$, dislocations B and D are visible (Fig. 23B); and when $\mathbf{g} = [\bar{2}20]$, all four loops are visible (Fig. 23C). Applying the $\mathbf{g} \cdot \delta\mathbf{r} = 0$ criterion (which becomes the $\mathbf{g} \cdot \mathbf{b} = 0$ criterion since the displacement $\delta\mathbf{r} = \mathbf{b}$, where $\mathbf{b}$ is the Burgers vector) for dislocation invisibility to the network AB, it must consist of dislocations with Burgers vectors in directions $\pm[h0l]$ and $\pm[h0l]$, where h, k, and l are undetermined. The θ′ plate with the dislocations marked D lies on (100) and, by analogy with the θ′ plate on (001), must be covered by a network of dislocations with Burgers vectors in directions $\pm[hk0]$ and $\pm[h0l]$. The dislocations marked D have Burgers vectors in directions $\pm[hk0]$ because they are visible when $\mathbf{g} = [0\bar{2}0]$. Because dislocations D are invisible when $\mathbf{g} = [\bar{2}00]$, $l = 0$. Hence the network AB for a θ′ plate on (001) has Burgers vectors in directions $\pm[100]$ and $\pm[010]$. There are no contrast effects associated with the interface to suggest that the Burgers vectors of the dislocations are not lattice translation vectors, and hence the most likely Burgers vectors are $\pm a[100]$ and $\pm a[010]$ for a (001) plate. The dislocations are mainly edge in character since their line direction is perpendicular to the Burgers vectors and are mainly accommodating a small misfit (<1%) in the plane of the flat interface.

Kikuchi Lines and the Effect of Deviation Parameter on Defect Contrast

One of the most important uses of the deviation parameter s is for obtaining and quantifying high-quality TEM images of defects in materials. Figure 24 shows three diffraction conditions with (a) $s = 0$, (b) $s < 0$, and (c) $s > 0$ and the corresponding BF images of dislocations in Al, respectively. Note that the best images are formed when $s > 0$, because the dislocations display sharp, strong contrast above a bright background. The magnitude of $\mathbf{s}$ must be determined to calculate images of such dislocations (Head et al., 1973).

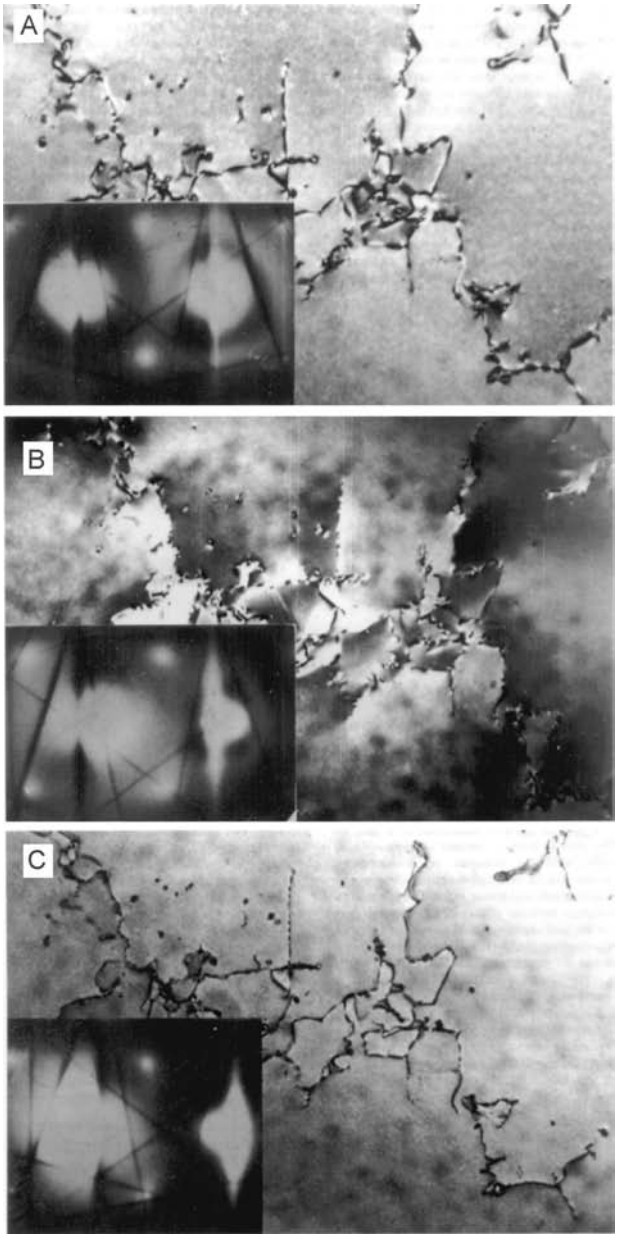


Figure 24. Changes in diffraction patterns and corresponding images of dislocations with changes in s: (**A**) $s = 0$, (**B**) $s60; 0$, and (**C**) $s > 0$. (From Edington, 1974.)

SAMPLE PREPARATION

The ideal TEM specimen is usually a thin, parallel-sided foil that is stable in the electron microscope as well as during storage. A gradually tapered wedge is often obtained in practice, and this too is quite useful. The TEM foil should be representative of the bulk material being studied and thus unaltered by sample preparation. The exact thickness of the specimen depends on the type of analysis desired and may range from a few nanometers for HRTEM to tens of nanometers for CTEM imaging. The surface condition is important, particularly for very thin specimens. It is difficult to meet all of these requirements, but the quality of the results depends on the specimen quality, so specimen preparation is an important aspect of TEM. The range of possible specimen preparation techniques is enormous, and the discussion below is intended to provide an overview of some of the more common techniques available to materials scientists. More comprehensive discussions of specimen preparation procedures are provided by Edington (1974), Goodhew (1984), and Williams and Carter (1996).

There are usually two stages in the preparation of thin specimens from bulk material. The first is to prepare, often by cutting and grinding, a thin slice of the material that is a few millimeters in diameter and a tenth of a millimeter or so thick. The main consideration in the first stage is to avoid using a coarse cutting technique so that damage is introduced throughout the specimen. A diamond wafer saw is often used to cut thin specimens of metals and ceramics, and SiC paper with successively finer grit (e.g., down to 600 grit) is used to grind these materials to the required thickness. The second stage involves thinning this specimen to electron transparency for which several techniques are available. Four of the most common—electropolishing, chemical thinning, ion milling, and Ga^+ ion beam thinning—are briefly described next.

Electropolishing and Chemical Thinning

The most common technique for thinning electrically conductive materials such as metal alloys is electropolishing. The principle behind the method is that the specimen is made the anode in an electrolytic cell so that when current is passed through the cell, the solution dissolves the alloy and deposits it on the cathode. The main considerations in electropolishing are the composition, temperature and flow of the electrolyte across the sample, and the potential difference applied between the sample (anode) and cathode. Under the right conditions, the foil becomes smoother and thinner until perforation is achieved near the center of the specimen. The technique is fairly simple, but sophisticated electropolishing units that allow the operator to carefully control the polishing variables and terminate polishing immediately upon perforation are commercially available. The commercial units are designed to accommodate 3-mm-diameter disks, which is the size required for most TEM specimen holders.

A major limitation of electropolishing is that it cannot be used on nonconducting materials such as ceramics and semiconductors. Chemical thinning, using mixtures of acids without an applied potential, is frequently used for these materials. Both electropolishing and chemical thinning share the advantage that the thinning rate is high, so they are quick and generally nondamaging, although it is possible to preferentially leach out elements or introduce hydrogen into some materials, so as with any technique, one must exercise some caution when utilizing these procedures.

Ion Milling and Focused Ga^+ Ion Beam Thinning

It is common to thin specimens to electron transparency by bombarding them with energetic ions. The ions knock atoms off the surface in a process called sputtering,

gradually thinning the material. The most common ions used are Ar^+, and these are accelerated toward the sample with an applied potential ranging from 4 to 6 keV at angles ranging from 5° to 20° from the sample surface, i.e., at a glancing angle. Usually two ion guns are used so that the sample can be thinned from both sides simultaneously, and the sample is often cooled with liquid nitrogen to temperatures of ~ -180°C to avoid excessive heating by the ion beams. The samples are usually rotated during thinning to prevent preferential sputtering along one direction. The milling rate is usually only a few micrometers per hour, so compared to electropolishing, sample preparation is much more time consuming. Therefore, a dimpler is often used to grind the center of the sample to a few tens of micrometers in thickness to reduce the ion milling time. Ion milling is very useful for thinning nonconducting samples, for thinning materials that consist of phases that electropolish at much different rates, for thinning cross-sectional samples of thin films on substrates such as semiconductor and multilayer materials, or for cleaning the surfaces of samples prepared by other methods. Ion milling produces an amorphous layer on the surfaces of most materials, which can be detrimental to some analyses.

A focused ion beam (FIB) microscope is becoming popular for preparing thin membranes for TEM examination (Basile et al., 1992). A FIB microscope is similar to a scanning electrom microscope (SCANNING ELECTRON MICROSCOPY) except that it uses electrostatic lenses to focus a Ga^+ ion beam on the specimen instead of electromagnetic lenses and electrons. The Ga^+ ions are typically accelerated toward the specimen at 30 keV so that rapid sputtering of the sample occurs. The Ga^+ ion probe can be focused to a size as small as 10 nm, and a computer can be used to raster the beam across the sample in a variety of patterns, so that it is possible to fabricate very specific specimen geometries. This technique can be used on both conducting and nonconducting materials, and it is particularly useful for preparing cross-sectional samples of semiconductors and thin films. There is evidence that FIB milling can produce a damaged layer on the surfaces of materials (Susnitzky and Johnson, 1998), but the extent of this behavior and its relation to specimen and microscope parameters remain to be fully determined.

Another instrument that is particularly useful for thinning semiconductor materials to electron transparency is a tripod polisher (Klepeis et al., 1988). The tripod polisher, so called because it has three micrometer feet, is used to hold the specimen while it is mechanically thinned to electron transparency on a polishing wheel using diamond lapping films. This "wedge technique" can be used to examine particular depths in a sample, such as a microelectronic device. Even if not used for final sample preparation, the use of a tripod polisher can dramatically reduce the time for the final thinning step.

Ultramicrotomy

An ultramicrotome utilizes a fixed diamond knife to slice a thin section from a specimen, which is typically $\sim 0.5 \times 0.5$ mm in cross-section and can be mounted in epoxy for support. The resulting slices are collected in a liquid-filled trough and mounted on copper grids before being inserted into the transmission electron microscope. Ultramicrotomy is widely used to prepare samples of polymers for TEM and can also be used for other materials such as powders, metals, and composites. Slicing generally damages the material (e.g., introducing dislocations into metal alloys), but chemical information is retained, so ultramicrotomy is well suited for preparing thin sections for analytical techniques such as EDS in TEM. Although it only takes a few minutes to obtain a slice from a specimen, considerable expertise is required to mount the specimen and obtain a thin section, so ultramicrotomy can be a fairly time consuming preparation method.

Replication

A thin film of carbon can be deposited on the surface of a specimen, typically by striking an arc between two carbon rods in vacuum. The carbon layer conforms to the shape of the surface and is allowed to build up to a few tens of nanometers in thickness. The carbon film is then scored into squares a few millimeters wide, removed from the surface by etching, and collected on a copper grid for examination in the transmission electron microscope. If the sample contains highly dispersed phases that can be left in relief on the surface by etching, the carbon film can be deposited on the surface, and the phases are then embedded in the carbon film when it is removed from the surface by further etching. This specimen is called an extraction replica. Since the particles are present in the replica, they can be analyzed by diffraction or analytical techniques without interference from the surrounding matrix. A common alternative to carbon media is cellulose acetate, which can be deposited on the sample as a sheet with solvent.

Dispersion

One final technique that is simple and works well for small particles such as catalysts or brittle ceramics that can be ground to a fine powder is to disperse a small quantity of the material in ethanol by sonication and then place a few drops of the dispersion on a carbon support film mounted on a grid. When the solution evaporates, the particles cling to the support film and can be examined in the transmission electron microscope. Depending on the material, it may be preferable to use a solvent other than ethanol or a grid material other than copper, but the procedure is generally the same.

SPECIMEN MODIFICATION

Inelastic collisions of electrons with a specimen often produce the undesirable effect of beam damage. The damage that affects the structure and/or chemistry of the material depends mainly on the beam energy. Damage usually occurs through one of two main mechanisms: (1) radiolysis, in which inelastic scattering breaks chemical bonds in the material, or (2) knock-on damage, in which atoms are displaced off their sites, creating point defects.

The first mechanism generally decreases with increasing beam energy whereas the second increases with increasing beam energy. It is important to remember that, under certain conditions, in TEM it is possible to damage any material. Sometimes this is useful as it allows one to simulate various irradiation conditions or perform *in situ* studies, but more often it produces unwanted side effects. Some of the more important aspects of beam damage are summarized below. This is followed by a brief discussion of intentional specimen modification in TEM during *in situ* experiments.

Specimen heating can occur in TEM. Temperature is difficult to measure experimentally because the temperature is affected by many variables, such as specimen thickness, thermal conductivity, surface condition, contact with the holder, beam energy, and beam current. Hobbs (1979) has performed calculations on the effects of beam current and thermal conductivity on specimen temperature. The results of these calculations indicate that beam heating is generally negligible for good thermal conductors such as metals but substantial for insulating materials. To minimize beam heating, one can (1) operate at the highest accelerating potential to reduce the cross-section for inelastic scattering, (2) cool the specimen to liquid nitrogen or liquid helium temperatures in a cooling holder, (3) coat the specimen with a conducting film, (4) use low-dose imaging techniques, or (5) all of these techniques (Sawyer and Grubb, 1987).

Polymers are highly sensitive to the process of radiolysis in TEM. Electrons usually cause the main polymer chain to break up or side groups to break off, leading to breakdown or crosslinking of the polymer, respectively. In the first case, the polymer continually loses mass under irradiation, whereas in the second case, the polymer ultimately ends up as a mass of carbon. If the polymer is crystalline, radiation damage causes a loss of crystallinity, which can be measured experimentally by the loss of diffraction contrast in the image or the reduction in intensity of diffraction spots in the diffraction pattern and the appearance of an amorphous pattern. Radiolysis can also occur in ionic and covalent materials such as ceramics and minerals. This often results in the formation of new compounds or the amorphization of the original material, processes that can be observed in TEM. Radiolysis is not affected by heat transfer considerations, so the main ways to minimize this process are to increase the accelerating voltage and minimize the specimen thickness and exposure to the electron beam.

Metals are mainly affected by knock-on damage in the transmission electron microscope. Knock-on damage is directly related to the beam energy and the atomic weight of the atoms (Hobbs, 1979). Accelerating potentials $>\sim$200 kV can damage light metals such as aluminum, and the effect on heavier elements continually increases as the beam energy increases above this value. Knock-on damage usually manifests itself by the formation of small vacancy or interstitial clusters and dislocation loops in the specimen, which can be observed by diffraction contrast. The main way to avoid displacment damage is to use lower accelerating voltages. Knock-on damage can also occur in polymers and minerals where there is generally a trade-off between knock-on damage and radiolysis. Since knock-on damage creates vacancies and interstitials, it can be used to study the effects of irradiation *in situ*.

Another important aspect of specimen modification in TEM is the ability to perform *in situ* studies to directly observe the behavior of materials under various conditions (Butler and Hale, 1981). This potentially useful effect for electron irradiation studies was mentioned above. In addition, specimen holders designed to heat, cool, and strain (and combinations of these) a TEM sample in the microscope are available commercially from several manufacturers. It is also possible to introduce gases into the transmission electron microscope to observe chemical reactions *in situ* by using apertures and differential pumping in the microscope. Once they have equilibrated, most commercial TEM holders are sufficiently stable that it is possible to perform EDS and EELS on foils to obtain compositional information *in situ* (Howe et al., 1998). To reproduce the behavior of bulk material in TEM, it is often desirable to perform *in situ* experiments in high-voltage TEM (e.g., 1 MeV) so that foils on the order of 1 μm thick can be used. Using thicker foils can be particularly important for studies such as straining experiments. As with most techniques, one needs to be careful when interpreting *in situ* data, and it is often advisable to compare the results of *in situ* experiments with parallel experiments performed in bulk material. For example, if we want to examine the growth behavior of precipitates during *in situ* heating in TEM, we might compare the size of the precipitates as a function of time and temperature obtained *in situ* in the transmission electron microscope with that obtained by bulk aging experiments. The temperature calibration of most commercial holders seems to be reliable to $\pm 20^{\circ}$C, and calibration specimens need to be used to obtain greater temperature accuracy. One problem that can arise during heating experiments, for example, is that solute may preferentially segregate to the surface. Oxidation of some samples during heating can also be a problem. In spite of all these potential difficulties, *in situ* TEM is a powerful technique for observing the mechanisms and kinetics of reactions in solids and the effects of electron irradiation on these phenomena.

PROBLEMS

A number of limitations should be considered in the TEM analysis of materials, including but not limited to (1) the sampling volume, (2) image interpretation, (3) radiation damage, (4) specimen preparation, and (5) microscope calibration. This section briefly discusses some of these factors. More thorough discussion of these topics can be found in Edington (1974), Williams and Carter (1996), and other textbooks on TEM (see Literature Cited).

It is important to remember that in TEM only a small volume of material is observed at high magnification. If possible, it is important to examine the same material at lower levels of resolution to ensure that the microstructure observed in TEM is representative of the overall specimen. It is often useful to prepare the TEM specimen by more than one method to ensure that the microstructure was

not altered during sample preparation. As discussed above (see Specimen Modification), many materials damage under a high-energy electron beam, and it is important to look for signs of radiation damage in the image and diffraction pattern. In addition, as shown above (see Data Analysis and Initial Interpretation), image contrast in TEM varies sensitively with diffracting conditions, i.e., the exact value of the deviation parameter s. Therefore, one must carefully control and record the diffracting conditions during imaging in order to quantify defect contrast.

It is important to remember that permanent calibration of TEM features such as image magnification and camera length, which are typically performed during installation, are only accurate to within $\pm 5\%$. If one desires higher accuracy, then it is advisable to perform *in situ* calibrations using standards. A variety of specimens are available for magnification calibration. The most commonly used are diffraction grating replicas for low and intermediate magnifications ($100\times$ to $200,000\times$) and direct lattice images of crystals for high magnifications ($>200{,}000\times$). Similarly, knowledge of the camera constant in Equation 35 greatly simplifies indexing of diffraction patterns and is essential for the identification of unknown phases. The accuracy of calibration depends on the experiment, but an error of $\pm 1\%$ can be achieved relatively easily, but this can be improved to $\pm 0.1\%$ if care is used in operation of the microscope and measurement of the diffraction patterns. An example of this calibration was given above (see Data Analysis and Initial Interpretation), and it is common to evaporate a metal such as Au directly onto a sample or to use a thin film as a standard. Permanent calibrations are normally performed at standard lens currents that must be subsequently reproduced during operation of the microscope for the calibration to be valid. This is accomplished by setting the lens currents and using the height adjustment (z control) on the goniometer to focus the specimen. Other geometric factors that can introduce errors into the magnification of diffraction patterns are discussed by Edington (1974). There are also a number of other calibrations that are useful in TEM, including calibration of (1) the accelerating voltage, (2) specimen drift rate, (3) specimen contamination rate, (4) sense of specimen tilt, (5) focal increments of the objective lens, and (6) beam current, but these are not detailed here.

ACKNOWLEDGMENTS

The authors are grateful for support of this work by the National Science Foundation, JMH under Grant DMR-9630092 and BTF under Grant DMR-9415331.

LITERATURE CITED

Basile, D. P., Boylan, R., Hayes, K., and Soza, D. 1992. FIBX-TEM—Focussed ion beam milling for TEM sample preparation. *In* Materials Research Society Symposium Proceedings, Vol. 254 (R. Anderson, B. Tracy, and J. Bravman, eds.). pp. 23–41. Materials Research Society, Pittsburgh.

Borchardt-Ott, W. 1995. Crystallography, 2nd ed. Springer-Verlag, New York.

Butler, E. P. and Hale, K. F. 1981. Dynamic Experiments in the Electron Microscope, Vol. 9. *In* Practical Methods in Electron Microscopy (A. M. Glauert, ed.). North-Holland, New York.

Chang, Y.-C. 1992. Crystal structure and nucleation behavior of (111) precipitates in an AP-3.9Cu-0.5Mg-0.5Ag alloy. Ph.D. thesis, Carnegie Mellon University, Pittsburgh.

Chescoe, D. and Goodhew, P. J. 1984. The Operation of the Transmission Electron Microscope. Oxford University Press, Oxford.

Cockayne, D. J. H., Ray, I. L. F., and Whelan, M. J. 1969. Investigation of dislocation strain fields using weak beams. *Philos. Mag.* 20:1265–1270.

Edington, J. W. 1974. Practical Electron Microscopy in Materials Science, Vols. 1-4. Macmillan Philips Technical Library, Eindhoven.

Egerton, R. F. 1996. Electron Energy-Loss Spectroscopy in the Electron Microscope, 2nd ed. Plenum Press, New York.

Fultz, B. and Howe, J. M. 2000. Transmission Electron Microscopy and Diffractometry of Materials. Springer-Verlag, Berlin.

Goodhew, P. J. 1984. Specimen Preparation for Transmission Electron Microscopy of Materials. Oxford University Press, Oxford.

Head, A. K., Humble, P., Clarebrough, L. M., Morton, A. J., and Forwood, C. T. 1973. Computed Electron Micrographs and Defect Identification. North-Holland, Amsterdam, The Netherlands.

Hirsch, P. B., Howie, A., Nicholson, R. B., Pashley D. W., and Whelan, M. J. 1977. Electron Microscopy of Thin Crystals, 2nd ed. Krieger, Malabar.

Hobbs, L. W. 1979. Radiation effects in analysis of inorganic specimens by TEM. *In* Introduction to Analytical Electron Microcopy (J. J. Hren, J. I. Goldstein, and D. C. Joy, eds.) pp. 437–480. Plenum Press, New York.

Howe, J. M., Murray, T. M., Csontos, A. A., Tsai, M. M., Garg, A., and Benson, W. E. 1998. Understanding interphase boundary dynamics by *in situ* high-resolution and energy-filtering transmision electron microscopy and real-time image simulation. *Microsc. Microanal.* 4:235–247.

Hull, D. and Bacon, D. J. 1984. Introduction to Dislocations, 3rd ed. (see pp. 17–21). Pergamon Press, Oxford.

Keyse, R. J., Garratt-Reeed, A. J., Goodhew, P. J., and Lorimer, G. W. 1998. Introduction to Scanning Transmission Electron Microscopy. Springer-Verlag, New York.

Kikuchi, S. 1928. Diffraction of cathode rays by mica. *Jpn. J. Phys.* 5:83–96.

Klepeis, S. J., Benedict, J. P., and Anderson, R. M. 1988. A grinding/ polishing tool for TEM sample preparation. *In* Specimen Preparation for Transmission Electron Microscopy of Materials (J. C. Bravman, R. M. Anderson, and M. L. McDonald, eds.). pp. 179–184. Materials Research Society, Pittsburgh.

Krivanek, O. L., Ahn, C. C., and Keeney, R. B. 1987. Parallel detection electron spectrometer using quadrupole lenses. *Ultramicroscopy* 22:103–116.

McKie, D. and McKie, C. 1986. Essentials of Crystallography, p 208. Blackwell Scientific Publications, Oxford.

Miller, M. K. and Smith, G. D. W. 1989. Atom Probe Microanalysis: Principles and Applications to Materials Problems. Materials Research Society, Pittsburgh.

Reimer, L. 1993. Transmission Electron Microscopy: Physics of Image Formation and Microanalysis, 3rd ed. Springer-Verlag, New York.

Reimer, L. (ed.). 1995. Energy-Filtering Transmission Electron Microscopy. Springer-Verlag, Berlin.

Rioja, R. J. and Laughlin, D. E. 1977. The early stages of GP zone formation in naturally aged Al-4 wt pct Cu alloys. *Metall. Trans.* 8A:1257–1261.

Sawyer, L. C. and Grubb, D. T. 1987. Polymer Microscopy. Chapman and Hall, London.

Schwartz, L. H. and Cohen, J. B. 1987. Diffraction from Materials, 2nd ed. Springer-Verlag, New York.

Smith, F. G. and Thomson, J. H. 1988. Optics, 2nd ed. John Wiley & Sons, Chichester.

Susnitzky, D. W. and Johnson, K. D. 1998. Focused ion beam (FIB) milling damage formed during TEM sample preparation of silicon. *In* Microscopy and Microanalysis 1998 (G. W. Bailey, K. B. Alexander, W. G. Jerome, M. G. Bond, and J. J. McCarthy, eds.) pp. 656–667. Springer-Verlag, New York.

Thomas, G. and Goringe, M. J. 1979. Transmission Electron Microscopy of Metals. John Wiley & Sons, New York.

Voelkl, E., Alexander, K. B., Mabon, J. C., O'Keefe, M. A., Postek, M. J., Wright, M. C., and Zaluzec, N. J. 1998. The DOE2000 Materials MicroCharacterization Collaboratory. *In* Electron Microscopy 1998, Proceedings of the 14th International Congress on Electron Microscopy (H. A. Calderon Benavides and M. Jose Yacaman, eds.) pp. 289–299. Institute of Physics Publishing, Bristol, U.K.

Weatherly G. C. and Nicholson, R. B. 1968. An electron microscope investigation of the interfacial structure of semi-coherent precipitates. *Philos. Mag.* 17:801–831.

Williams, D. B. and Carter, C. B. 1996. Transmission Electron Microscopy: A Textbook for Materials Science. Plenum Press, New York.

KEY REFERENCES

Edington, 1974. See above.

Reprinted edition available from Techbooks, Fairfax, Va. Filled with examples of diffraction and imaging analyses.

Fultz and Howe, 2000. See above.

An integrated treatment of microscopy and diffraction, with emphasis on principles.

Hirsch et al., 1977. See above.

For many years, the essential text on CTEM.

Reimer, 1993. See above.

Excellent reference with emphasis on physics of electron scattering and TEM.

Shindo, D. and Hiraga, K. 1998. High Resolution Electron Microscopy for Materials Science. Springer-Verlag, Tokyo.

Provides numerous high-resolution TEM images of materials.

Williams and Carter, 1996. See above.

A current and most comprehensive text on modern TEM techniques.

INTERNET RESOURCES

http://www.amc.anl.gov

An excellent source for TEM information on the Web in the United States. Provides access to the Microscopy ListServer and a Software Library as well as a connection to the Microscopy & Microanalysis FTP Site and Libraries plus connections to many other useful sites.

http://cimewww.epfl.ch/welcometext.html

A similar site based at the Ecole Polytechnique Federale de Lausanne in Switzerland that contains software and a variety of electron microscopy information.

http://www.msa.microscopy.com

Provides access to up-to-date information about the Microscopy Society of America, affiliated societies, and microscopy resources that are sponsored by the society.

http://rsb.info.nih.gov/nih-image

Public domain software developed by National Institutes of Health (U.S.) for general image processing and manipulation. Available from the Internet by anonymous ftp from zippy.nimh.nih.gov or on floppy disk from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161, part number PB93-504868.

JAMES M. HOWE
University of Virginia
Charlottesville, Virginia

BRENT FULTZ
California Institute of Technology
Pasadena, California

SCANNING TRANSMISSION ELECTRON MICROSCOPY: *Z*-CONTRAST IMAGING

INTRODUCTION

As its name suggests, the scanning transmission electron microscope is a combination of the scanning electron microscope and the transmission electron microscope. Thin specimens are viewed in transmission, while images are formed serially by the scanning of an electron probe. In recent years, electron probes have become available with atomic dimensions, and, as a result, atomic resolution images may now be achieved in this instrument. The nature of the images obtained in scanning transmission electron microscopy (STEM) can differ in significant ways from those formed by the more widespread conventional transmission electron microscopy (CTEM). The key difference lies in their modes of image formation; the STEM instrument can be configured for almost perfect incoherent imaging whereas CTEM provides almost perfect coherent imaging. The latter technique is generally referred to as high-resolution electron microscopy (HREM), though both methods now provide atomic resolution.

The difference between coherent and incoherent imaging was first discussed over one hundred years ago in the context of light microscopy by Lord Rayleigh (1896). The difference depends on whether or not permanent phase relationships exist between rays emerging from different parts of the object. A self-luminous object results in perfect incoherent imaging, as every atom emits independently, whereas perfect coherent imaging occurs if the entire object is illuminated by a plane wave, e.g., a point source at infinity. Lord Rayleigh noted the factor of 2 improvement in resolution available with incoherent

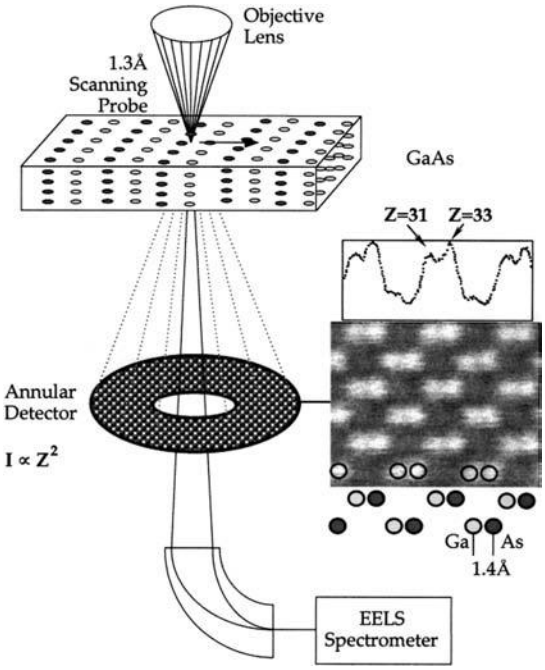


Figure 1. Schematic showing *Z*-contrast imaging and atomic resolution electron energy loss spectroscopy with STEM. The image is of GaAs taken on the 300-kV STEM instrument at Oak Ridge National Laboratory, which directly resolves and distinguishes the sublattice, as shown in the line trace.

imaging and also the lack of artifacts caused by interference phenomena that could be mistaken for real detail in the object. Of particular importance in the present context, he appreciated the role of the condenser lens (1896, p. 175): "It seems fair to conclude that the function of the condenser in microscopic practice is to cause the object to behave, at any rate in some degree, as if it were self-luminous, and thus to obviate the sharply-marked interference bands which arise when permanent and definite phase relationships are permitted to exist between the radiations which issue from various points of the object." A large condenser lens provides a close approximation to perfect incoherent imaging by ensuring a range of optical paths to neighboring points in the specimen.

A hundred years later, STEM can provide these same advantages for the imaging of materials with electrons. A probe of atomic dimensions illuminates the sample (see Fig. 1), and a large annular detector is used to detect electrons scattered by the atomic nuclei. The large angular range of this detector performs the same function as Lord Rayleigh's condenser lens in averaging over many optical paths from each point inside the sample. This renders the sample effectively self-luminous; i.e., each atom in the specimen scatters the incident probe in proportion to its atomic scattering cross-section. With a large central hole in the annular detector, only high-angle Rutherford scattering is detected, for which the cross-section depends on the square of the atomic number (Z); hence this kind of microscopy is referred to as *Z*-contrast imaging.

The concept of the annular detector was introduced by Crewe et al. (1970), and spectacular images of single heavy

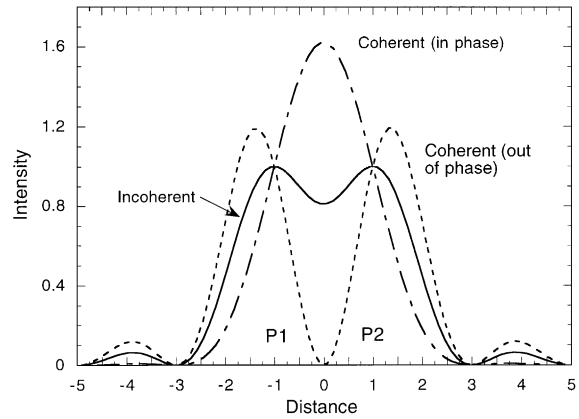


Figure 2. Image intensity for two point objects, P1 and P2, illuminated coherently in phase, 180° out of phase, and with incoherent illumination. (After Lord Rayleigh, 1896.)

atoms were obtained (see, e.g., Isaacson et al., 1979). In the field of materials, despite annular detector images showing improved resolution (Cowley, 1986a) and theoretical predictions of the lack of contrast reversals (Engel et al., 1974), it was generally thought impossible to achieve an incoherent image at atomic resolution (Cowley, 1976; Ade, 1977). Incoherent images of thick crystalline materials were first reported by Pennycook and Boatner (1988), and the reason for the preservation of incoherent characteristics despite the strong dynamical diffraction of the crystal was soon explained (Pennycook et al., 1990), as described below.

An incoherent image provides the most direct representation of a material's structure and, at the same time, improved resolution. Figure 2 shows Lord Rayleigh's classic result comparing the observation of two point objects with coherent and incoherent illumination. Each object gives rise to an Airy disc intensity distribution in the image plane, with a spatial extent that depends on the aperture of the imaging system. The two point objects are separated so that the first zero in the Airy disc of one coincides with the central maximum of the other, a condition that has become known as the Rayleigh resolution criterion. With incoherent illumination, there are clearly two peaks in the intensity distribution and a distinct dip in between; the two objects are just resolved, and the peaks in the image intensity correspond closely with the positions of the two objects. With coherent illumination by a plane-wave source (identical phases at the two objects), there is no dip, and the objects are unresolved. Interestingly, however, if the two objects are illuminated 180° out of phase, then they are always resolved, with the intensity dropping to zero half-way between the two. Unfortunately, this desirable result can only be achieved in practice for one particular image spacing (e.g., by illuminating from a particular angle), and other spatial frequencies will have different phase relationships and therefore show different contrast, a characteristic that is generic to coherent imaging. Note also that the two peaks are significantly displaced from their true positions.

In an incoherent image, there are no fixed phase relationships, and the intensity is given by a straightforward

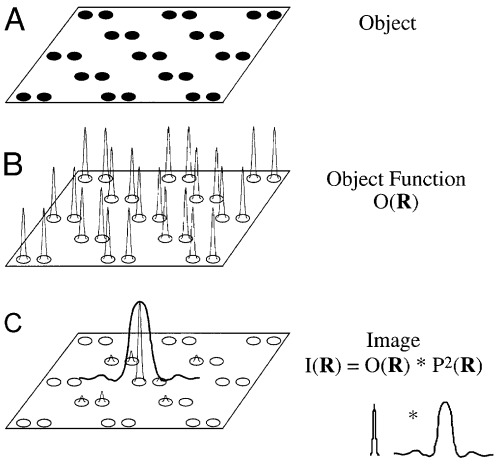


Figure 3. Schematic showing incoherent imaging of a thin specimen with STEM: (**A**) monolayer raft of Si⟨110⟩; (**B**) the *Z*-contrast object function represents the high-angle scattering power localized at the atomic sites; (**C**) illumination of the sites for a 1.26-Å probe located over one atom in the central dumbbell. As the probe scans, it maps out the object function, producing an incoherent image.

convolution of the electron probe intensity profile with a real and positive specimen object function, as shown schematically in Figure 3. With Rutherford scattering from the nuclei dominating the high-angle scattering, the object function is sharply peaked at the atomic positions and proportional to the square of the atomic number. In the figure, a monolayer raft of atoms is scanned by the probe, and each atom scatters according to the intensity in the vicinity of the nucleus and its high-angle cross-section. This gives a direct image with a resolution determined by the probe intensity profile. Later it will be shown how crystalline samples in a zone axis orientation can also be imaged incoherently. These atomic resolution images show similar characteristics to the incoherent images familiar from optical instruments such as the camera; in a *Z*-contrast image, atomic columns do not reverse contrast with focus or sample thickness. In Figure 1, columns of Ga can be distinguished directly from columns of As simply by inspecting the image intensity. Detailed image simulations are therefore not necessary.

This unit focuses on *Z*-contrast imaging of materials at atomic resolution. Many reviews are available covering other aspects of STEM, e.g., microanalysis and microdiffraction (Brown, 1981; Pennycook, 1982; Colliex and Mory, 1984; Cowley, 1997). Unless otherwise stated, all images were obtained with a 300-kV scanning transmission electron microscope, a VG Microscopes HB 603U with a 1.26-Å probe size, and all spectroscopy was performed on a 100-kV VG Microscopes HB 501UX STEM.

Competitive and Related Techniques

The absence of phase information in an incoherent image is an important advance; it allows the direct inversion from the image to the object. In coherent imaging, the structural relationships between different parts of the object are encoded in the phase relationships between the various diffracted beams. These are lost when the intensity is recorded, giving rise to the well-known phase problem. Despite this, much effort has been expended in attempts to measure the phase relationships between different diffracted beams in order to reconstruct the object (e.g., Coene et al., 1992; Orchowski et al., 1995; Möbus, 1996; Möbus and Dehm, 1996).

However, electrons interact with the specimen potential, which is a real quantity. There is no necessity to involve complex quantities; in an incoherent image, there are no phases and so none can be lost. Information about the object is encoded in the image intensities, and images may be directly inverted to recover the object. As seen in Figure 2, intensity maxima in an incoherent image are strongly correlated with atomic positions, so that often this inversion can be done simply by eye, exactly as we interpret what we see around us in everyday life. With the additional benefit of strong *Z* contrast, structures of defects and interfaces in complex materials may often be determined directly from the image.

In a phase-contrast image, the contrast changes dramatically as the phases of the various diffracted beams change with specimen thickness or objective lens focus, which means that it is much more difficult to determine the object uniquely. It is often necessary to simulate many trial objects to find the best fit. The situation is especially difficult in regions such as interfaces or grain boundaries, where atomic spacings deviate from those in the perfect crystal, which can also cause atomic columns to reverse contrast. If one does not think of the correct structure to simulate, obviously the fit will be spurious. Unexpected phenomena such as the formation of new interfacial phases or unexpected atomic bonding are easily missed. Such phenomena are much more obvious in an incoherent image, which is essentially a direct image of atomic structure. As an example, Figure 4 shows the detection of van der Waals bonding between an epitaxial film and its substrate from the measurement of film-substrate atomic separation (Wallis et al., 1997). At 3.2 Å, this is significantly larger than bond lengths in covalently bonded semiconductors, which are typically in the range of 2.3 to 2.6 Å.

Closely related to STEM, scanning electron microscopy (SEM) gives an image of the surface (or near-surface)

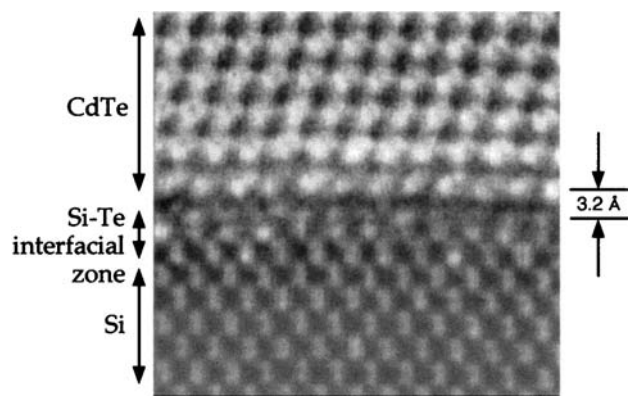


Figure 4. *Z*-contrast image of an incommensurate CdTe(111)/Si(100) interface showing a 3.2-Å spacing between film and substrate indicating van der Waals bonding.

region of a bulk sample, whereas STEM gives a transmission image through a thin region of the bulk, which by careful preparation we hope is representative of the original bulk sample. The information from the two microscopes is therefore very different, as is the effort required for sample preparation. Whereas in STEM we detect the bright field and dark field transmitted electrons to reveal the atomic structure, in SEM we detect secondary electrons, i.e., low-energy electrons ejected from the specimen surface. The resolution is limited by the interaction volume of the beam inside the sample, and atomic resolution has not been achieved. Probe sizes are therefore much larger, ~1 to 10 nm. Secondary electron images show topographic contrast with excellent depth of field. Backscattered primary electrons can also be detected to give a *Z*-contrast image analogous to that obtained with the STEM annular detector, except at lower resolution and contrast. Other signals such as cathodoluminescence or electron beam-induced conductivity can be used to map optical and electronic properties, and x-ray emission is often used for microanalysis. These signals can also be detected in STEM (Pennycook et al., 1980; Pennycook and Howie, 1980; Fathy and Pennycook, 1981), except that signals tend to be weaker because the specimen is thin.

Atomic resolution images of surfaces are produced by scanning tunneling microscopy (STM) and atomic force microscopy (AFM). Here, the interaction is with the surface electronic structure, although this may be influenced by defects in underlying atomic layers to give some subsurface sensitivity. In scanning probe microscopy, it is the valence electrons that give rise to an image; it is not a direct image of the surface atoms, and interpretation must be in terms of the surface electronic structure. The STEM *Z*-contrast image is formed by scattering from the atomic nuclei and is therefore a direct structure image. Valence electrons are studied in STEM through electron energy loss spectroscopy (EELS) and can be investigated at atomic resolution using the *Z*-contrast image as a structural reference, as discussed below. Scanning transmission electron microscopy can study a wider range of samples than STM or AFM as it has no problem with rough substrates and can tolerate substrates that are quite highly insulating.

Data from EELS provide information similar to that from x-ray absorption spectroscopy. The position and fine structure on absorption edges give valuable information on the local electronic structure. The EELS data give such information from highly localized regions, indeed, from single atomic columns at an interface (Duscher et al., 1998a). Instead of providing details of bulk electronic structure, it provides information on how that structure is modified at defects, interfaces, and grain boundaries and insight into the changes in electronic, optical, and mechanical properties that often determine the overall bulk properties of a material or a device.

The detailed atomic level characterization provided by STEM, with accurate atomic positions determined directly from the *Z*-contrast image and EELS data on impurities, their valence, and local band structure, represents an ideal starting point for theoretical studies. This is particularly valuable for complex materials; it would be impractical to explore all the possible configurations of a complex extended defect with first-principles calculations, even with recent advances in computational capabilities. As the complexity increases, so the number of required trial structures grows enormously, and experiment is crucial to cut the possibilities down to a manageable number. Combining experiment and theory leads to a detailed and comprehensive picture of complex atomistic mechanisms, including equilibrium structures, impurity or stress-induced structural transformations, and dynamical processes such as diffusion, segregation, and precipitation. Recent examples include the observation that As segregates at specific sites in a Si grain boundary in the form of dimers (Chisholm et al., 1998) and that Ca segregating to an MgO grain boundary induces a structural transformation (Yan et al., 1998).

PRINCIPLES OF THE METHOD

Comparison to Coherent Phase-Contrast Imaging

The conventional means of forming atomic resolution images of materials is through coherent phase-contrast imaging using plane-wave illumination with an (approximately) parallel incident beam (Fig. 5A). The objective aperture is behind the specimen and collects diffracted beams that are brought to a focus on the microscope screen where they interfere to produce the image contrast. The electrons travel from top to bottom in the figure. Not shown are additional projector lenses to provide higher magnification. In Figure 5B, the optical path of the STEM is shown, with the electrons traveling from bottom to top. A point source is focused into a small probe by the objective lens, which is placed before the specimen. Not shown are the condenser lenses (equivalent to projector lenses in the CTEM) between the source and the objective lens to provide additional demagnification of the source. Transmitted electrons are then detected through a defined angular range. For the small axial collector aperture shown, the two microscopes have identical optics, apart from the fact that the direction of electron propagation is reversed. Since image contrast in the electron microscope is dominated by elastic scattering, no energy loss is involved and time reversal symmetry applies. With equivalent apertures, the image contrast is independent of the direction of electron propagation, and the two microscopes are optically equivalent: the STEM bright-field image will be the same image, and be described by the same imaging theory, as that of a conventional TEM with axial illumination. This is the principle of reciprocity, which historically was used to predict the formation of high-resolution lattice images in STEM (Cowley, 1969; Zeitler and Thomson, 1970).

For phase-contrast imaging, the axial aperture (illumination aperture in TEM, collection aperture in STEM) must be much smaller than a typical Bragg diffraction angle. If β is the semiangle subtended by that aperture at the specimen, then the transverse coherence length in the plane of the specimen will be of order λ/β, where λ is the electron wavelength. Coherent illumination of

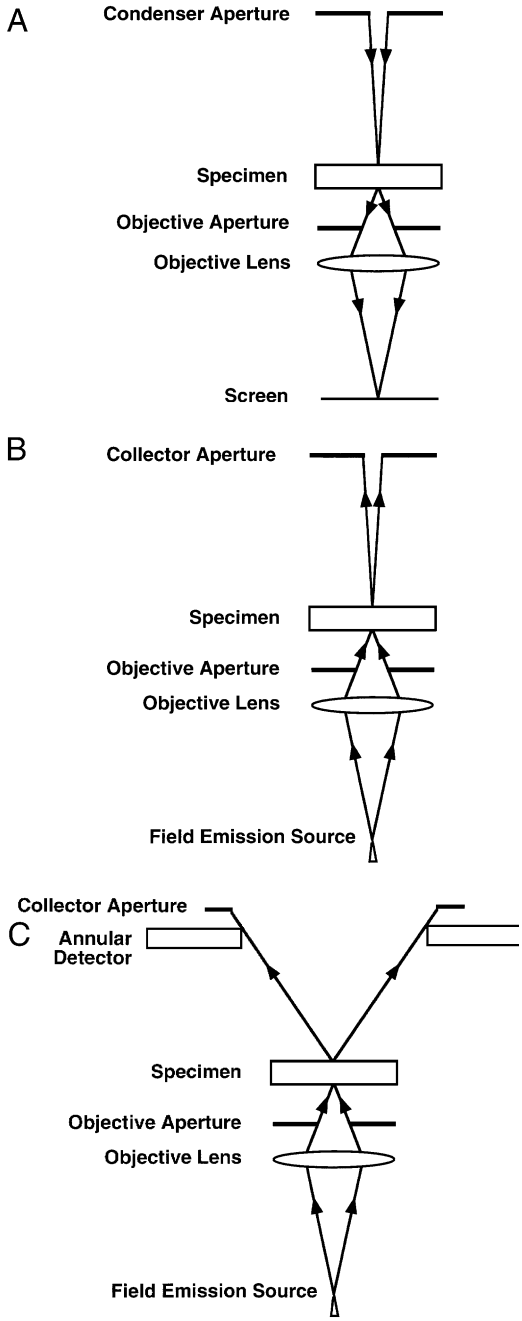


Figure 5. Schematics of the electron-optical arrangement in (**A**) CTEM and (**B**) STEM with a small axial detector. Note the direction of the electron propagation is reversed in the two microscopes. (**C**) A large-angle bright-field detector or annular dark-field detector in STEM gives incoherent imaging.

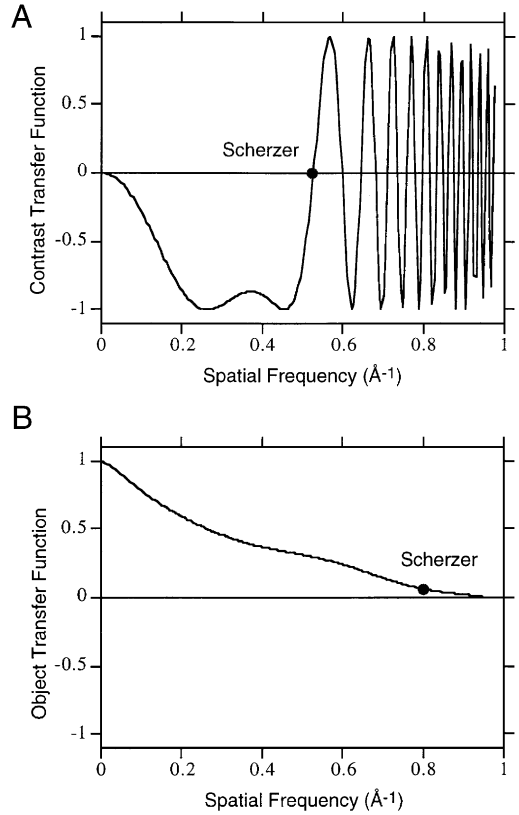


Figure 6. Contrast transfer functions for a 300-kV microscope with an objective lens with $C_s = 1$ mm; (**A**) coherent imaging conditions; (**B**) incoherent imaging conditions. Curves assume the Scherzer (1949) optimum conditions shown in Table 1: (**A**) defocus −505 Å; (**B**) defocus −438 Å, aperture cutoff 0.935 $Å^{-1}$.

neighboring atoms spaced a distance d apart requires $d \ll \lambda/\beta$, or $\beta \ll \lambda/d \approx 2\theta_B$, where θ_B is the Bragg angle corresponding to the spacing d. Similarly, if the collection aperture is opened much wider than a typical Bragg angle (Fig. 5C), then the transverse coherence length at the specimen becomes much less than the atomic separation d, and the image becomes incoherent in nature, in precisely the same way as first described by Lord Rayleigh. Now the annular detector can be seen as the complementary dark-field version of the Rayleigh case; conservation of energy requires the dark-field image $I = 1 - I_{BF}$, where I_{BF} is the bright-field image and the incident beam intensity is normalized to unity. In practice, it is more useful to work in the dark-field mode for weakly scattering objects. If $I \ll 1$, the incoherent bright-field image shows weak contrast on a large background due to the unscattered incident beam. The dark-field image avoids this background, giving greater contrast and better signal-to-noise ratio.

The difference between coherent and incoherent characteristics is summarized in Figure 6, where image contrast is plotted as a function of spatial frequency in a microscope operating at an accelerating voltage of 300 kV. In both modes some objective lens defocus is used to compensate for the very high aberrations of electron lenses, and the conditions shown are the optimum conditions for each mode as defined originally by Scherzer (1949). In the coherent imaging mode (Fig. 6A), the contrast transfer function is seen to oscillate rapidly with increasing spatial frequency (smaller spacings). Spatial frequencies where the transfer function is negative are imaged in reverse contrast; those where the transfer function crosses zero will be absent. For this reason, atomic positions in distorted regions such as grain boundaries may reverse contrast or be absent from a phase-contrast image. In the incoherent mode (Fig. 6B), a smoothly

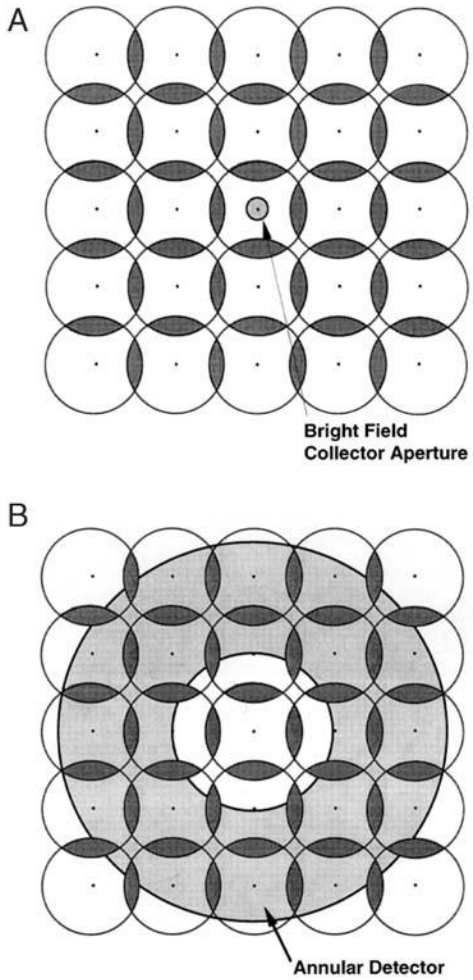


Figure 7. Diffraction pattern in the detector plane from a simple cubic crystal of spacing such that the angle between diffracted beams is greater than the objective aperture radius. (**A**) An axial bright-field detector shows no contrast. (**B**) Regions of overlapping discs on the annular detector produce atomic resolution in the incoherent image.

decaying positive transfer is obtained (referred to as an object transfer function to distinguish from the coherent case), which highlights the lack of contrast reversals associated with incoherent imaging. It is also apparent from the two transfer functions that incoherent conditions give significantly improved resolution.

To demonstrate the origin of the improved resolution available with incoherent imaging, Figure 7 shows the diffraction pattern formed on the detector plane in the STEM from a simple cubic crystal with spacing d. Because the illuminating beam in the STEM is a converging cone, each Bragg reflection appears as a disc with diameter determined by the objective aperture. Electrons, unlike light, are not absorbed by a thin specimen but are only diffracted or scattered. The terms are equivalent, diffraction being reserved for scattering from periodic assemblies of atoms when sharp peaks or discs are seen in the scattered intensity. The periodic object provides a good demonstration of Abbé theory (see, e.g., Lipson and Lipson, 1982) that all image contrast arises through interference; as the probe is scanned across the atomic planes, only the regions of overlap between the discs change intensity as the phases of the various reflections change (Spence and Cowley, 1978).

For the crystal spacing corresponding to Figure 7A, coherent imaging in CTEM or STEM using a small axial aperture will not detect any overlaps and will show a uniform intensity with no contrast. The limiting resolution for this mode of imaging requires diffraction discs to overlap on axis, which means they must be separated by less than the objective aperture radius. Any detector that is much larger than the objective aperture will detect many overlapping regions, giving an incoherent image. Clearly, the limiting resolution for incoherent imaging corresponds to the discs just overlapping, i.e., to a crystal spacing such that the diffracted beams are separated by the objective aperture diameter (as opposed to the radius), double the resolution of the coherent image formed by an axial aperture. The approximately triangular shape of the contrast transfer function in Figure 6B results from the simple fact that a larger crystal lattice produces closer spaced diffraction discs with more overlapping area, thus giving more image contrast.

In practice, of course, one would tend to use larger objective apertures in the phase-contrast case to improve resolution. The objective aperture performs a rather different role in the two imaging modes; in the coherent case the objective aperture is used to cut off the oscillating portion of the contrast at high spatial frequencies, and the resolution is directly determined by the radius of the objective aperture, as mentioned before. In the incoherent case, again it is clear that the available resolution will be limited to the aperture diameter, but there is another consideration. To take most advantage of the incoherent mode requires that the image bear a direct relationship to the object, so that atomic positions correlate closely with peaks in the image intensity. This is a rather more stringent condition and necessitates some loss in resolution. The optimum conditions worked out long ago by Scherzer (1949) for both coherent and incoherent imaging conditions are shown in Figure 5 and summarized in Table 1. The Scherzer resolution limit for coherent imaging is now only 50% poorer than for the incoherent case.

The role of dynamical diffraction is also very different in the two modes of imaging. As will be seen later, dynamical diffraction manifests itself in a *Z*-contrast image as a columnar channeling effect, not by altering the incoherent nature of the *Z*-contrast image, but simply by scaling the columnar scattering intensities. In coherent imaging, dynamical diffraction manifests itself through contrast

Table 1. Optimum Conditions for Coherent and Incoherent Imaging

Parameter	Coherent Imaging	Incoherent Imaging
Resolution limit	$0.66 C_S^{1/4} \lambda^{3/4}$	$0.43 C_S \lambda^{1/4} \lambda^{3/4}$
Optimum defocus	$-1.155\ (C_S \lambda)^{1/2}$	$-(C_S \lambda)^{1/2}$
Optimum aperture	$1.515 (\lambda / C_S)^{1/4}$	$1.414\ (\lambda / C_S)^{1/4}$

reversals with specimen thickness, as the phases between various interfering Bragg reflections change. These effects are also nonlocal, so that atom columns adjacent to interfaces will show different intensity from columns of the same composition far from the interface. Along with the fact that atom columns can also change from black to white by changing the objective lens focus, these effects make intuitive interpretation of coherent images rather difficult. Of course, one usually has a known material next to an interface that can help to establish the microscope conditions and local sample thickness, but these characteristics of coherent imaging make it very difficult to invert an image of an interface without significant prior knowledge of its likely structure. Typically, one must severely limit the number of possible interface structures that are considered, e.g., by assuming an interface that is atomically abrupt, with the last monolayer on each side having the same structure as the bulk. Although some interfaces do have these characteristics, it is also true that many, and perhaps most, do not have such simple structures. In the case of the $CoSi_2$/Si(111) interface, this procedure gives six possible interface models. At interfaces made by ion implantation, however, different interface structures were seen by *Z*-contrast imaging (Chisholm et al., 1994). There are now a large number of examples where interface structures have proved to be different from previous models: in semiconductors (Jesson et al., 1991; McGibbon et al., 1995; Chisholm and Pennycook, 1997), superconductors (Pennycook et al., 1991; Browning et al., 1993a), and ceramics (McGibbon et al., 1994, 1996; Browning et al., 1995).

If we apply reciprocity to the STEM annular detector, it is clear that an equivalent mode exists in CTEM using high-angle hollow-cone illumination. In practice, it has not proved easy to achieve such optics. In any case, a much higher incident beam current will pass through the specimen than in the equivalent STEM geometry, because the illumination must cover an angular range much greater than a typical Bragg angle, whereas the STEM objective aperture is of comparable size to a Bragg angle. Beam damage to the sample would therefore become a serious concern.

Another advantage of the STEM geometry that should be emphasized is that the detection angle can be increased without limit, up to the maximum of 180° if so desired. Lord Rayleigh's illumination was limited to the maximum aperture of the condenser lens, so that at the limit of resolution (with condenser and objective apertures equal), significant partial coherence exists in an optical microscope. In STEM, by increasing the inner angle of the annular detector, the residual coherence can be reduced well below the limiting resolution imposed by the objective aperture. It is therefore a more perfect form of incoherent imaging than fixed-beam light microscopy. However, increasing detector angles will lead to reduced image intensity due to the rapid fall-off in atomic scattering factor, and a backscattered electron detector is really not practical. With high-energy electrons, the full unscreened Rutherford scattering cross-section is seen above a few degrees scattering angle, so higher scattering angles bring no advantage. A useful criterion for the minimum detector aperture θ_i to achieve incoherent imaging of two objects separated by ΔR is

$$\theta_i = 1.22\lambda/\Delta R \qquad (1)$$

where the detected intensity varies by <5% from the incoherent expectation (Jesson and Pennycook, 1993). In this case the Airy disc coherence envelope of the detector (illumination) aperture is half the width of that of the probe-forming (imaging) aperture. This condition therefore corresponds to separating the coherence envelopes by double the Rayleigh criterion.

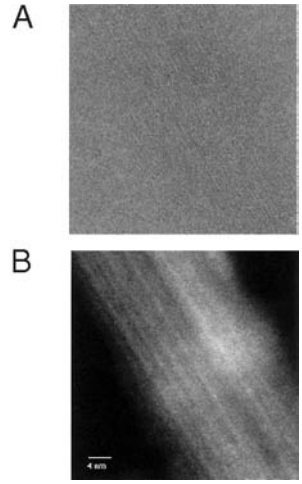


Figure 8. (**A**) Bright-field phase-contrast STEM image of an iodine intercalated bundle of carbon nanotubes showing lattice fringes. (**B**) *Z*-contrast image taken simultaneously showing intercalated iodine.

Another benefit of STEM is that the *Z*-contrast image and the conventional HREM image are available simultaneously. However, phase-contrast images in STEM tend to be substantially noisier than those recorded in TEM using a charge-coupled device (CCD) camera or photographic plate, because STEM is a serial imaging instrument that is far less efficient than the parallel recording capabilities of TEM. An example of simultaneous bright-field and *Z*-contrast imaging of a bundle of iodine-intercalated carbon nanotubes is shown in Figure 8. Because of their cylindrical form, only a few atomic layers are parallel to the electron beam. Nevertheless, lattice fringes from the tubes are seen clearly in the phase-contrast image because it is tuned to the spacing expected and filters out the uniform background near-zero spatial frequency due to all other atoms. On the other hand, the *Z*-contrast image is sensitive to the absolute numbers of atoms under the beam. Here, where there is no significant dynamical diffraction, the *Z*-contrast image can be considered as an image of projected mass thickness. It shows no detectable contrast from the tubes themselves, but the iodine intercalation is now clearly visible. Thus the two kinds of images are highly complementary in this case.

Figure 9 shows a Rh catalyst cluster supported on γ-alumina (Pennycook et al., 1996). In this case, the bright-field image is dominated by phase contrast from

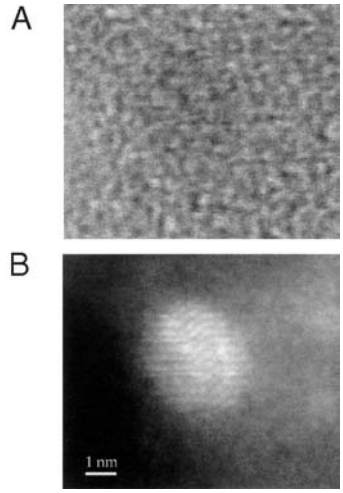


Figure 9. (**A**) Bright-field and (**B**) *Z*-contrast images of an Rh catalyst particle on γ-alumina. The particle is barely detectable in the bright-field image due to the phase contrast of the support, but the dark-field image reveals its atomic structure, internal twins, and external facets.

the carbon film ($Z = 6$) used to support the sample, whereas the Rh particle ($Z = 45$) is clearly visible in the *Z*-contrast image. Of course, if pieces of γ-alumina overlapping holes in the carbon film were chosen, the phase-contrast image would not be so obscured, but nevertheless, bright-field imaging of small metal clusters becomes very difficult for particles $<\sim$1 nm in size, due to the inevitable coherent interference effects from the support and the lack of *Z* contrast (Datye and Smith, 1992). The ultimate example of *Z*-contrast imaging is the detection of single Pt atoms on γ-alumina shown in Figure 10 (Nellist and Pennycook, 1996). Here, again, the two images are very complementary. The orientation of the γ-alumina support can be deduced from the bright-field image, while single atoms, dimers, and trimers are detectable in the *Z*-contrast image. Spacings and angles between the Pt atoms are constrained to match the atomic spacings in the γ-alumina surface, suggesting the possible adsorption sites shown in the schematic.

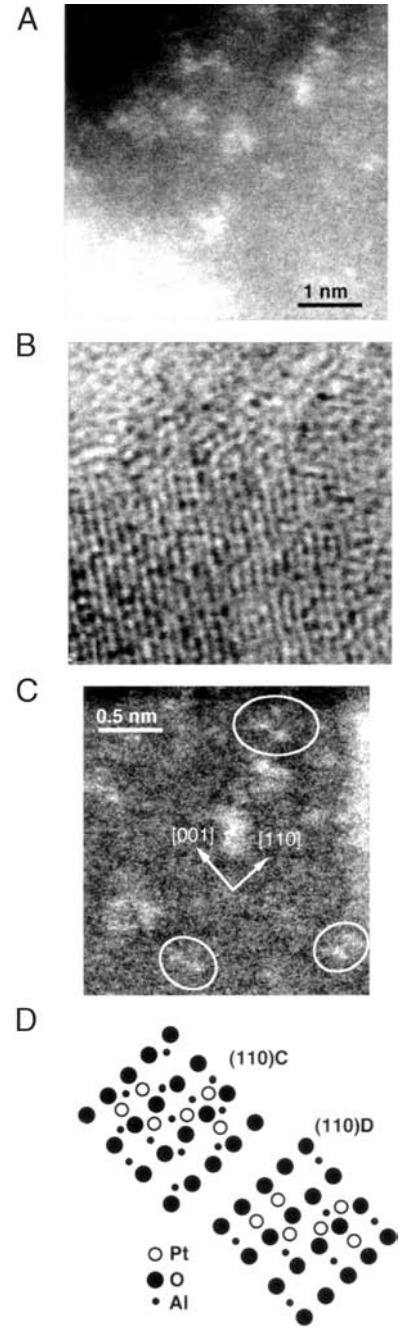


Figure 10. (**A**) *Z*-contrast image of Pt clusters supported on γ-Al_2O_3. (**B**) Bright-field image showing strong {222} fringes of Al_2O_3. (**C**) Higher magnification *Z*-contrast image after high- and low-pass filtering to enhance contrast revealing a Pt trimer and two dimers (circled). (**D**) The spacings and angles between Pt atoms match the orientation of the support, suggesting likely configurations on the two possible (110) surfaces.

Probe Formation

An electron of wavelength λ passing at an angle θ through an objective lens with spherical aberration coefficient C_s set to a defocus of Δf experiences a phase shift γ given by

$$\gamma = \frac{\pi}{\lambda}\left(\Delta f\theta^2 + \frac{1}{2}C_s\theta^4\right) \qquad (2)$$

The amplitude distribution $P(\mathbf{R} - \mathbf{R}_0)$ of the STEM probe at position $\mathbf{R}_0$ on the specimen is obtained by integrating all possible pathways through the objective lens, as defined by the objective aperture,

$$P(\mathbf{R} - \mathbf{R_0}) = \int_{\substack{\text{objective}\\ \text{aperture}}} e^{i[\mathbf{K}\bullet(\mathbf{R}-\mathbf{R_0})+\gamma(\mathbf{K})]} d\mathbf{K} \qquad (3)$$

where $|\mathbf{K}| = \chi\theta$ is the transverse component of the incident electron wave vector $\chi = 2\pi/\lambda$. The intensity distribution is given by $P^2(\mathbf{R} - \mathbf{R}_0)$. A focal series is shown in Figure 11 for a 300-kV STEM instrument with C_s = 1 mm and a Scherzer optimum objective aperture of 9.3 mrad. Notice how the intensity profiles are not symmetric about the Scherzer defocus value of −430 Å. This

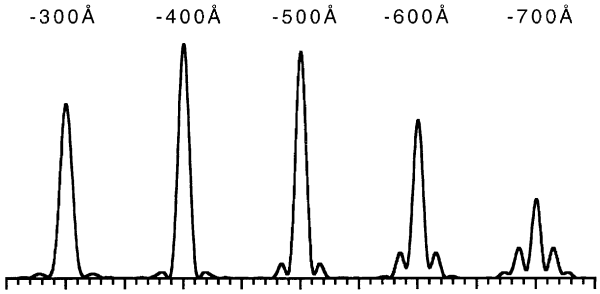


Figure 11. Probe intensity profiles for a 300-kV STEM instrument with $C_s = 1$ mm and a Scherzer optimum objective aperture of 9.4 mrad.

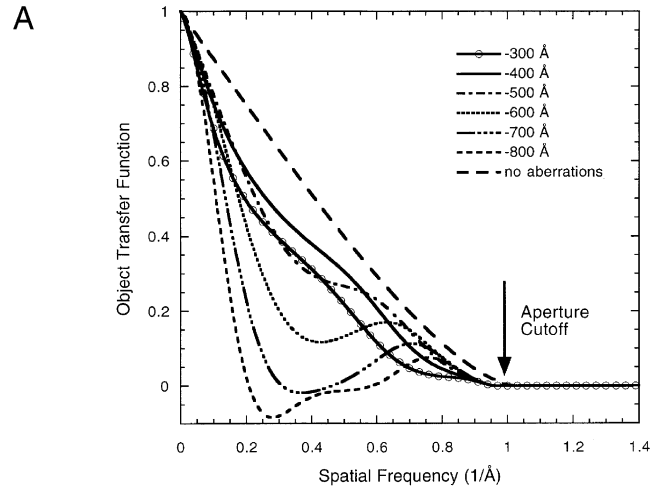
is because the entire probe is (ideally) a single electron wavepacket, all angles fully coherent with each other. The probe is best thought of as a converging, phase-aberrated spherical wave. At a defocus below the Scherzer optimum, the probe is close to Gaussian in nature, while at high defocus values, the probe develops a well-defined "tail," a subsidiary maximum around the central peak. Although the width of the central peak is significantly reduced, actually dropping to below 1 Å at a defocus of −500 Å, now over half the total intensity is in the tails. This gives rise to significant false detail in the image, which makes intuitive interpretation no longer possible, as shown by the corresponding simulated images of Si⟨110⟩ in Figure 12.

The object transfer function for incoherent imaging is the Fourier transform of the probe intensity profile. Figure 13A shows transfer functions corresponding to the focal series of Figures 11 and 12. Also shown is the ideal transfer function without any aberration. These curves show how increasing the defocus enhances the high spatial frequencies but reduces the transfer at lower frequencies. In all cases the transfer reaches zero at the cutoff defined by the aperture. If the aperture size is increased, the transfer function can be extended even further. Figure 13B shows transfer functions obtained with a 13-mrad objective aperture, corresponding to an aperture cutoff of 0.74 Å. At low defocus values there is little transfer at the high spatial frequencies, and an intuitive image is expected but with reduced contrast compared to the optimum aperture.

Experimental verification that the resolution is enhanced under such conditions is seen in Figure 14, which shows images of Si⟨110⟩ obtained with a 17-mrad objective aperture (Nellist and Pennycook, 1998a). Under optimum defocus the dumbbells are well resolved (Fig. 14A) with maxima close to the atomic positions, the expected intuitive image. The Fourier transform of the image intensity, Figure 14B shows the spatial frequencies



Figure 12. Simulated images of Si(110) corresponding to the probe profiles in Figure 11.



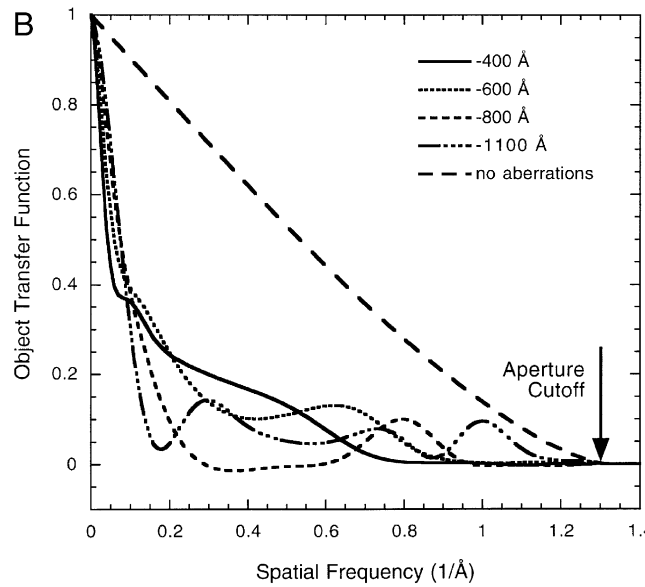


Figure 13. Object transfer functions for (**A**) the optimum condition in Figures 11 and 12; (**B**) extended transfer obtained with an oversized objective aperture of 13 mrad, though contrast at lower spatial frequencies is reduced.

being transferred, which includes the {004} reflection at a spacing of 1.36 Å. On increasing the objective lens defocus (Fig. 14C), the image is no longer intuitive, but many additional spots are seen in the Fourier transform (Fig. 14D). The {444} spot at 0.78 Å is the highest resolution yet achieved in any electron microscope. This result demonstrates very clearly the improved resolution available with incoherent imaging, as first pointed out by Lord Rayleigh.

Incoherent Scattering

Because the probe is necessarily a coherent spherical wave, incoherent imaging must come by ensuring incoherent scattering. Of course, as mentioned earlier, all imaging

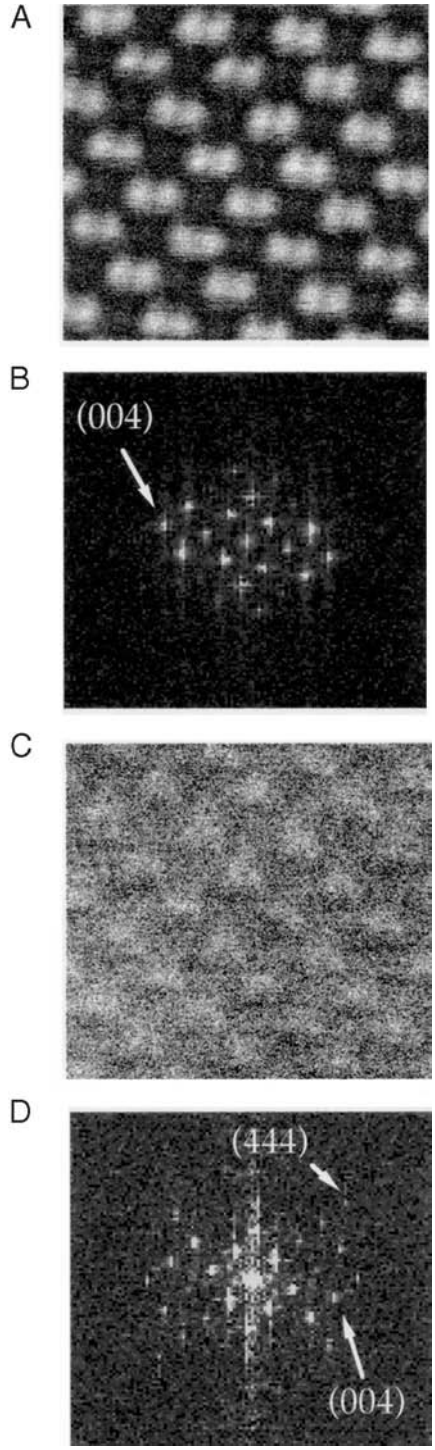


Figure 14. Images and corresponding Fourier transforms of Si<110> obtained with a 17-mrad objective aperture (Nellist and Pennycook, 1998a); (**A, B**) at Scherzer defocus passing the {004} spacing and resolving the dumbbell; (**C, D**) using a high underfocus giving increased transfer at high spatial frequencies, including the {444} spacing at 0.78 Å.

is based on interference, and incoherent characteristics arise when averaging a large number of optical paths of different lengths. For a monolayer raft of weakly scattering atoms, it can be shown rigorously how incoherent imaging can be achieved with a large detector to any desired degree, as shown in Appendix A. A simple justification is that, depending on the projected potential, a monolayer crystal changes only the phase of the electron wave passing through it. Clearly, if only the phase of the spherical wave is altered inside the crystal, the intensity profile will remain the same as that of the incident probe. Each nucleus will scatter in proportion to the intensity of the probe in its vicinity, and incoherent imaging is obtained as shown schematically in Figure 3. The image intensity $I(\mathbf{R})$ is given by a convolution of the probe intensity with the object function $O(\mathbf{R})$,

$$I(\mathbf{R}) = O(\mathbf{R}) * P^2(\mathbf{R}) \tag{4}$$

In the simple case of a monolayer crystal, the object function is just an array of atomic cross-sections that can be approximated as delta functions at the atomic positions.

A typical Bragg angle for a high-energy electron is only $\sim$10 mrad, or $\sim 0.5^\circ$, so that high-angle scattering corresponds to only a few degrees. This means that path differences between atoms separated in the transverse plane are much greater than between atoms separated along the beam direction, as shown in Figure 15. Atoms spaced by x and z in the transverse and longitudinal directions, respectively, when viewed from direction θ, have phase differences of $x \sin\theta$ and $z(1 - \cos\theta)$, respectively. For small scattering angles these are approximately $x\theta$ and $z\theta^2/2$; the latter is much smaller, showing how longitudinal coherence is more difficult to break with the detector alone. Clearly, if the Rayleigh detector is sufficient to ensure incoherence in the x direction, there is no justification for believing that atoms separated by a similar distance in the

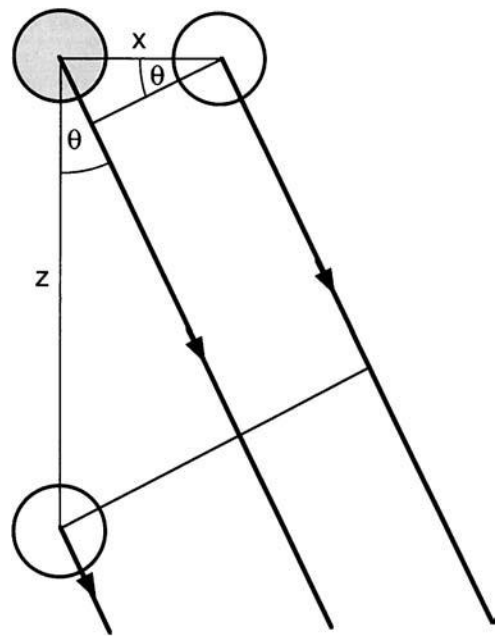


Figure 15. Path differences between scattering from atoms separated by x and z in the transverse and longitudinal directions, respectively.

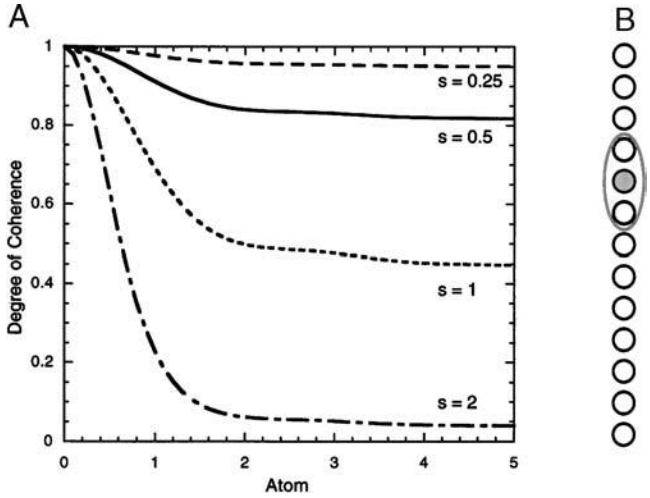


Figure 16. (**A**) Degree of coherence between an atom at the origin and neighboring atoms along a column using a phonon model of thermal vibrations. (**B**) With increasing separation the degree of coherence drops rapidly, defining an effective coherence envelope.

z direction will scatter independently. Indeed, the incoherence in the beam direction comes from another effect, the unavoidable thermal vibrations of the atoms themselves.

Thermal vibrations act to randomize the phase differences between atoms at different heights in the column by inducing transverse atomic displacements. Consider the two atoms separated by z to have an additional instantaneous relative displacement due to thermal motion of u_z and u_x in the longitudinal and transverse directions, respectively. The resulting path differences are $(z + u_z)(1-\cos\theta)$ and $u_x \sin\theta$. Clearly, the longitudinal displacements u_z have relatively little effect as they are much smaller than the atomic separation z, but the transverse displacements u_x are significant. A detailed analysis using a phonon model of thermal vibrations shows that it is indeed transverse modes of phonons with wave vectors in the z direction that are most effective in breaking the coherence along a column (Jesson and Pennycook, 1995). The analysis showed further that the coherence of two atoms separated along the z direction by m unit cells is

$$W_m = \exp\left\{2Ms^2\left[\frac{\mathrm{Si}(\pi m)}{\pi m} - 1\right]\right\} \qquad (5)$$

where Si(x) is the sine integral function, M is the usual Debye-Waller factor, and $s = \theta/2\lambda$ is the scattering vector. Figure 16A shows plots of this function for different values of s showing how the coherence rapidly reduces as the separation of atoms along the column increases. For large separations the degree of coherence approaches the limiting value e^{-2M}, which is the Einstein value for the strength of coherent reflections in his model of independently vibrating atoms. This model, though convenient, does ignore the correlation between near neighbors seen in the figure. The phonon model shows clearly that atoms close together scatter with greater coherence than those far apart, leading to the concept of a longitudinal coherence volume, shown schematically in Figure 16B.

Therefore, the need for large detector angles to ensure *inter*column incoherence (transverse incoherence) will automatically break the *intra*column coherence leading to longitudinal incoherence also. The change in the thickness dependence of the image intensity with increasing annular detector angle is largely governed by longitudinal coherence, changing from predominantly coherent at low angles to predominantly incoherent at large angles. This is illustrated in Figure 17 for a column of Rh atoms 2.7 Å apart illuminated by 300-kV electrons. With low detection angles the scattering is almost entirely coherent, becoming almost entirely incoherent at high detector angles. Intermediate angles exhibit an initial coherent dependence with thickness, changing to an incoherent dependence as the column becomes significantly longer than the correlation length.

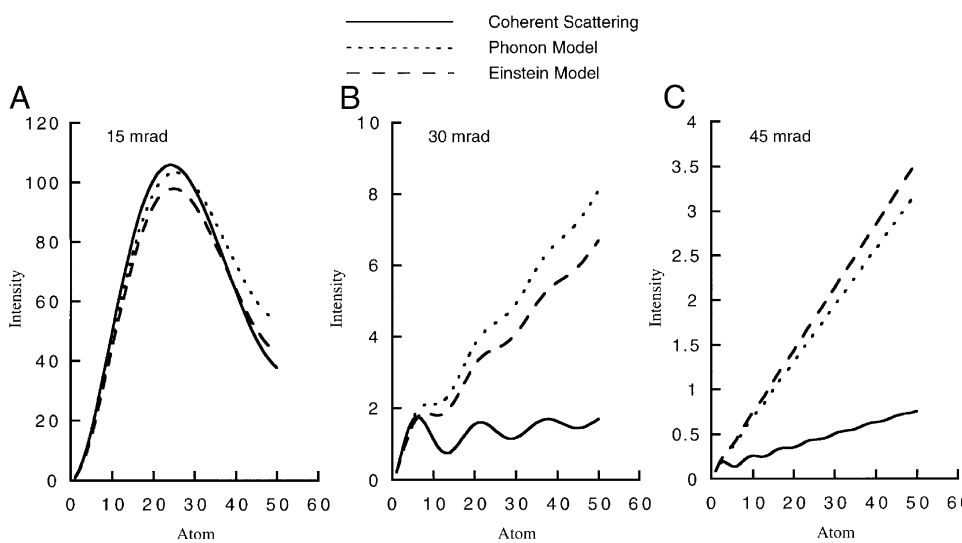


Figure 17. Thickness dependence of the scattering from a column of Rh atoms spaced 2.7 Å apart along the beam direction for (**A**) low, (**B**) medium, and (**C**) high scattering angles. The coherent thickness oscillations at low angles are almost completely suppressed at high angles.

Finally, it should be mentioned that a significant fraction of the thermal displacements of the atoms are due to zero-point fluctuations that will not disappear on cooling the sample. Thus, it should not be assumed that the scattering will become coherent on cooling, although the required detection angles will be increased somewhat.

Dynamical Diffraction

The above analysis makes clear how at high scattering angles thermal vibrations effectively destroy the coherence between atoms in a column. Given the transverse incoherence due to the annular detector, we may therefore think of each atom as scattering independently with an atomic cross-section σ, the electron equivalent of the self-luminous object. Now all that remains is to determine how the probe propagates through the thick crystals typically encountered in materials science studies, i.e., what is the effective illumination of each atom in a column? Electrons interact strongly with matter, undergoing multiple scattering, which is referred to as dynamical diffraction. With amorphous materials, or crystals in a random orientation, the multiple scattering is uncorrelated and leads to beam broadening. In a crystal oriented to a low-order zone axis, the situation is very different since the scattering is strongly correlated, often leading to a periodic behavior with increasing sample thickness. Phase-contrast images usually exhibit periodic contrast changes or reversals. One of the most surprising features of *Z*-contrast images from zone axis crystals is the lack of any apparent change in the form of the image with increasing sample thickness (Pennycook and Jesson, 1990, 1991). The *Z*-contrast image shows no strong oscillatory behavior, just reducing contrast with increasing thickness. The 300-kV STEM instrument at Oak Ridge National Laboratory still resolves the dumbbell spacing of 1.36 Å in a Si⟨110⟩ crystal 1000 Å thick.

The explanation of this remarkable behavior requires a quantum mechanical analysis of the probe propagation. Dynamical diffraction effects can be conveniently described in a Bloch wave formulation, where an incident plane-wave electron is described as a superposition of Bloch states, each propagating along the zone axis with a different wave vector. It is the interference between these Bloch states that leads to dynamical diffraction effects, thickness fringes in a diffraction contrast image, and contrast reversals in a phase-contrast lattice image. Figure 18 shows the first six Bloch states for Si⟨110⟩, which are seen to take the form of molecular orbitals about the atomic strings. Usually, the wave function inside the crystal can be well represented with just a few strongly excited Bloch states, and it is their propagation with different wave vectors through the crystal that leads to depth-dependent dynamical diffraction effects. Note in particular the 1*s* states, which are located over the atomic columns, and overlap little with neighboring columns. They are the most localized states, and we will find they are responsible for the *Z*-contrast image. These 1*s* states exist around every atomic column, and it is because of their localization that we can think of the *Z*-contrast image as a column-by-column image of the crystal. They are also the basis for atomic column spectroscopy, as described under Atomic Resolution Spectroscopy, below. All the other states are much less localized and overlap significantly with neighboring columns, as is clear from the figure. These delocalized states are responsible for the nonlocal effects in phase-contrast imaging, which makes it necessary to simulate images with supercells. Such states are not specific to a particular type of column but are strongly dependent on surrounding columns. They are the reason that atomic columns at an interface in a phase-contrast image may appear different to those in the bulk.

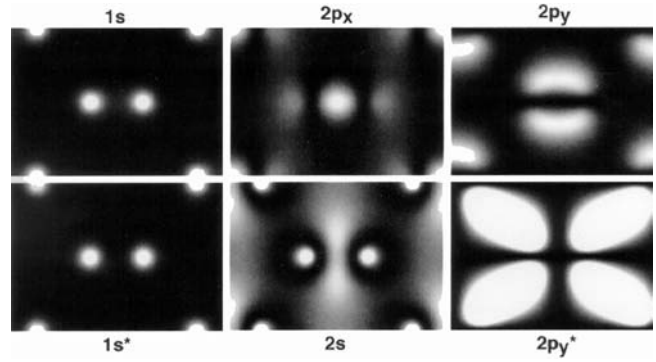


Figure 18. Intensities of the first six Bloch states in Si(110) with their molecular orbital assignments. The 1*s* states are located around the Si atomic columns.

We can now better appreciate the action of the high-angle detector. The 1*s* Bloch states are the most highly *localized* in real space, and therefore the most *extended* in reciprocal space. Only these states have transverse momentum components that extend to the high-angle detector. So the detector acts as an efficient Bloch state filter, with only one type of state per column involved in *Z*-contrast imaging. All states have low-order Fourier components, and so depending on their excitation, they all interfere in a low-angle detector such as used for axial bright-field imaging. This is how the Z-contrast image can show no interference phenomena while simultaneously the bright-field image shows strong interference effects, such as thickness fringes or contrast reversals.

The high-angle scattering comes entirely from the 1*s* states, specifically from their sharp peaks at the atomic columns. The detector therefore acts rather like a detector placed inside the sample at each atomic nucleus. In fact, the high-angle components of 1*s* states scale as Z^2, as would be anticipated on the basis of Rutherford scattering. However, dynamical diffraction does make its presence felt because not all the incident beam excites 1*s* states, and the total amplitude inside the crystal must be described as a sum over Bloch states j propagating with different wave vectors k_z^j along the z direction. The 1*s* states are located over the deepest part of the projected potential and so have the highest kinetic energy and the largest k_z^j. All the other less localized states have very similar k_z^j values, and so in thin crystals they all propagate approximately in phase through the thickness z. Therefore, it is a good

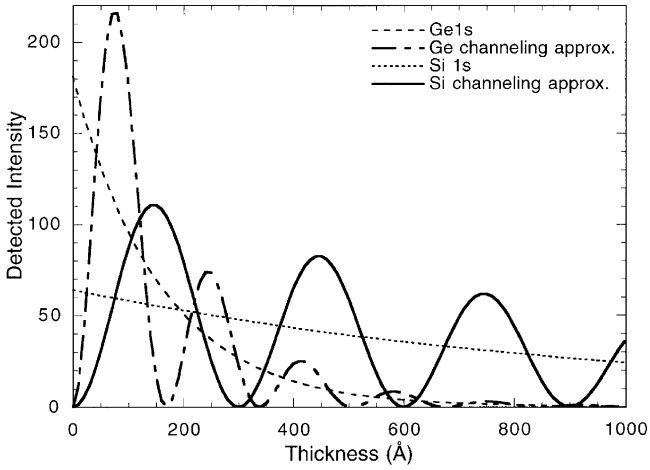


Figure 19. Intensity of coherent scattering reaching the annular detector from Si and Ge⟨110⟩ using the channeling approximation, Equation 6. Parameters: Si, $\xi = 300$ Å, $\mu^{1s} = 0.00048$; Ge, $\xi = 169$ Å, $\mu^{1s} = 0.0032$.

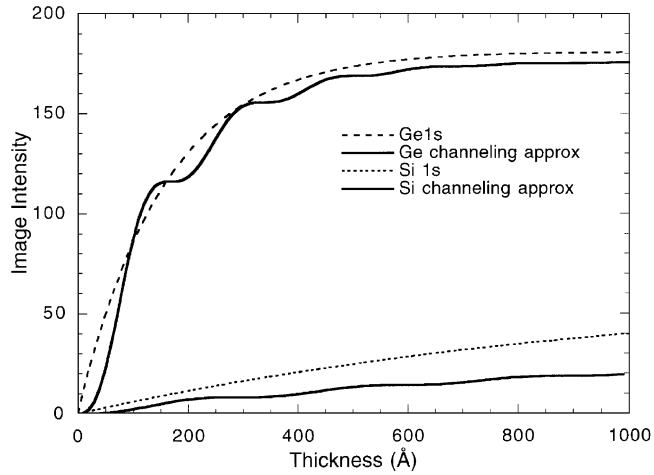


Figure 20. Intensity of incoherent scattering reaching the annular detector from Si and Ge ⟨110⟩ using the channeling approximation, Equation 7. Parameters are the same as Figure 19.

approximation to consider just two components to the electron wave function, the 1*s* state ϕ^{1s} propagating with wave vector k_z^{1s} and a term $1 - \phi^{1s}$ propagating at an average k_z^0. This is referred to as a channeling approximation (Pennycook and Jesson, 1992). The beating between these two components occurs with an extinction distance $\xi = 2\pi/(k_z^{1s} - k_z^0)$, and the total intensity scattered to high angles from a depth z in the dynamical wave field is given by

$$I \propto Z^2 \varepsilon^{av^2} [1 - \cos(2\pi z/\xi)] e^{-2\mu^{1s} z} \tag{6}$$

Here, Z^2 represents the high-angle scattering cross-section of the column; ε^{av} is the 1*s* state excitation averaged over the angular range of the objective aperture. The term in square brackets represents the dynamical oscillations with depth z, and μ^{1s} represents the 1*s* state absorption coefficient that damps the oscillations. Figure 19 shows a plot of this function for Si and Ge in the ⟨110⟩ orientation, compared to the 1*s* state intensities alone.

Experimental images do not show these strong depth oscillations because, as mentioned earlier, most of the scattering reaching the detector is thermal diffuse scattering. This is generated incoherently from the dynamical wave field described in Equation 6 and is therefore given by the integral of Equation 6 over the crystal thickness t. This results in a thickness-dependent columnar object function (Nellist and Pennycook, 1999)

$$O(R,t) \propto \frac{Z^2 \varepsilon^{av^2}}{2\mu^{1s}(\xi^2 \mu^{1s^2} + \pi^2)} \times \left[\pi^2 (1 - e^{-2\mu^{1s} t}) - \xi^2 \mu^{1s^2} e^{-2\mu_i^{1s}} \left(1 - \cos\frac{2\pi t}{\xi}\right) - \pi\xi\mu^{1s} e^{-2\mu^{1s} t} \sin\frac{2\pi t}{\xi} \right] \tag{7}$$

As shown in Figure 20, the thickness integration has removed the strong dynamical oscillations. Again we see that the thermal vibrations are important in breaking the coherence through the thickness t and giving an image intensity that is increasing with thickness. The limiting value is again proportional to Z^2, allowing very simple image quantification. The form of the curve is in good agreement with both experimental observations and multislice simulations for a high detection angle (Loane et al., 1992; Hillyard et al., 1993; Anderson et al., 1997; Hartel et al., 1996). Under these conditions, this channeling approximation is very useful for simulating images and saves enormously on computer time.

For lower detection angles the other states become more important, and a fraction of the coherent exit-face wave function given by Equation 6 must be added. The image contrast begins to show more oscillatory dependence on thickness, and the incoherent characteristics are progressively lost. Other states also become important if the atomic columns are no longer straight but are bent due to the presence of a defect or impurity. Then transitions occur between Bloch states, which is the origin of diffraction contrast imaging in CTEM, and strain contrast effects will also be seen in the annular detector signal, as discussed under Strain Contrast, below. A similar case is that of a single heavy impurity atom at a depth z, which will sample the oscillating wave field of Figure 19 and show depth-dependent contrast (Loane et al., 1988; Nakamura et al., 1997).

Nevertheless, in most cases of high-resolution imaging, the *Z*-contrast image represents a thickness integrated signal, and the oscillatory behavior is strongly suppressed. This is in marked contrast to a phase-contrast image that reflects the exit-face wave function and therefore does show an oscillatory dependence on sample thickness. The *Z*-contrast image is a near-perfect approximation to an ideal incoherent image. It removes the contrast reversals with objective lens focus, the proximity effects due to nonlocal Bloch states, and finally, the depth-dependent oscillations due to dynamical diffraction. The Bloch state description of the imaging process shows clearly the

physical origin of the incoherent characteristics, despite the presence of strong dynamical scattering.

Atomic Resolution Spectroscopy

Scanning transmission electron microscopy provides a sound basis for atomic resolution analysis using x-ray fluorescence or EELS. To date, atomic resolution has been demonstrated only for EELS because of the much lower detection efficiency for x rays (Browning and Pennycook, 1995; Pennycook et al., 1997; Duscher et al., 1998a). Using the incoherent *Z*-contrast image, a stationary probe can be centered over a chosen column by maximizing the annular detector intensity. Provided that incoherent optics are also used for the EELS, i.e., a large collector aperture, the EELS signal from that column will also be maximized at the same probe position, giving an atom column resolved analysis.

Because it is an inner shell electron that is the scatterer rather than the nucleus, the spatial resolution of the spectral information may not be quite as high as that of the *Z*-contrast image, an effect referred to as delocalization. For inner shell excitations, however, the width of the object function is less than an angstrom for most elements, as shown in Figure 21 (Rafferty and Pennycook, 1999). Core loss EELS therefore allows individual atomic columns to be analyzed, as shown in Figure 22. The *Z*-contrast image and spectra from specific atomic columns show the variation in both concentration and valence of Mn in a doped $SrTiO_3$ grain boundary (Duscher et al., 1998a). The highest peak intensity comes from the pair of columns labeled 2, which in the undoped boundary are Ti sites. A much lower concentration is seen at position 3, which represents the Sr sites. The ratio of the two Mn peaks, or white lines, indicates valence and suggests a change from 4+ in the bulk to 3+ at the boundary. Data of this nature are particularly valuable for linking the structure of the grain boundary to its electrical activity.

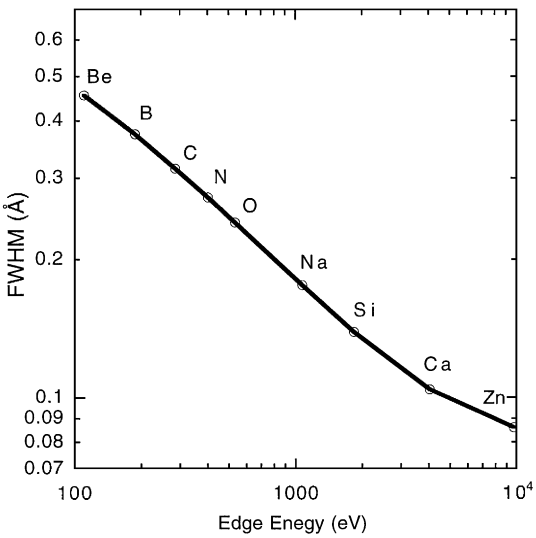


Figure 21. Full-width at half-maxima of EELS object functions for *K*-shell excitations by 300-kV incident electrons.

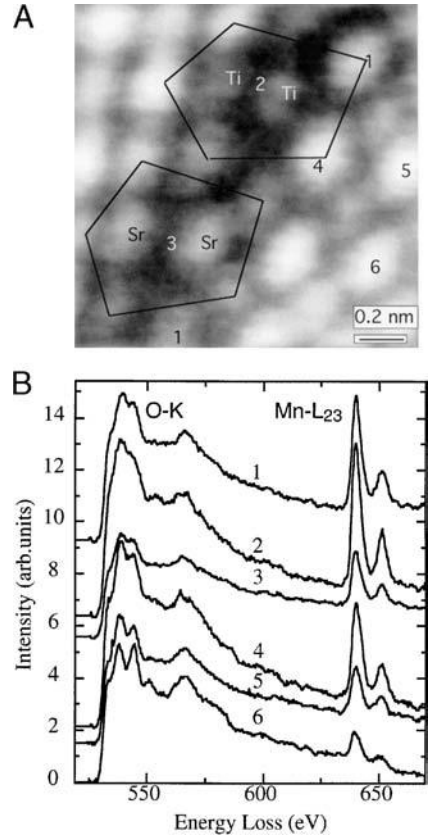


Figure 22. (**A**) *Z*-contrast image from a Mn-doped $SrTiO_3$ grain boundary with EELS spectra from individual atomic columns; (**B**), showing differences in composition and Mn valence at the grain boundary.

With the phase-contrast image of CTEM, it is not so simple to accurately illuminate an individual column. Neither is it practical to form an energy-filtered image at atomic resolution except from low-loss electrons that are intrinsically delocalized, in which case the contrast is due to elastic scattering, but the chemical information is nonlocal. In CTEM the objective lens is behind the specimen (Fig. 5A), so that the collection angle for EELS imaging is restricted to the objective aperture size. Furthermore, the energy loss electrons will suffer chromatic aberration on passing through the lens to form an image. For the energy-filtered image to show atomic resolution, the energy window selected by the filter must be kept small. Core loss cross-sections are orders of magnitude lower than elastic scattering cross-sections that determine the signal-to-noise ratio of the *Z*-contrast image. Attempting to form an image from core loss electrons in CTEM will therefore result in a very noisy image, probably with insufficient statistics to show atomic resolution. Increasing the incident beam current to compensate will increase the chances of beam damage. The information would be more efficiently gathered in the STEM mode by illuminating the chosen column, using the high-intensity *Z*-contrast image as a reference, and collecting the transmission EELS spectrum column by column. The objective lens is before the specimen in STEM, so that

chromatic aberration no longer degrades the spatial resolution, and a large collector aperture can be used to collect a large fraction of the inelastic scattering and partially compensate for the low cross-section.

This combination of the *Z*-contrast image and EELS at atomic resolution is very powerful. It is particularly useful for analyzing elements such as oxygen, which are too light to be seen directly in the *Z*-contrast image (McGibbon et al., 1994; Dickey et al., 1998). In the high-temperature superconductors, a pre-edge feature on the oxygen *K* edge is directly related to the concentration of holes that are responsible for superconductivity. This feature can be used to map hole concentrations at a spatial resolution below the superconducting coherence length (Browning et al., 1993b).

PRACTICAL ASPECTS OF THE METHOD

Besides the need for a high-resolution objective lens, a critical requirement for a probe of atomic dimensions is a field emission source. Without the high brightness such sources offer, there is insufficient current in the probe to focus and stigmate the image and insufficient current for an atomic resolution analysis. The cold field emission source is preferred for EELS because of its smaller energy spread compared to the Schottky-type source, whereas the latter is preferred for general TEM work because it is capable of a higher current mode and therefore gives less noisy images at low magnification when searching the specimen. It is easier to search the sample in the CTEM mode, and once an interesting region is found, the microscope can be switched to the STEM mode for *Z*-contrast imaging.

Formally, atomic-sized probes were only available in dedicated STEM instruments, but in recent years, conventional TEM has dramatically improved its probe-forming capabilities. It it is now possible to obtain true atomic resolution *Z*-contrast capability in a conventional TEM column (James et al., 1998; James and Browning, 1999).

To adjust the STEM instrument, a shadow image or ronchigram is very useful (Cowley 1979, 1986b; Lin and Cowley, 1986). Formed by stopping the STEM probe and observing the intensity in the diffraction plane, it is a powerful means to align the microscope, correct astigmatism, and tilt a specific area precisely on axis. The objective lens strength is used as a nonlinear magnification control. With the probe focused away from the specimen plane, a broad area of sample is illuminated, and the diffraction screen shows a shadow image. This will also contain diffraction information in the form of Kikuchi lines, as different parts of the image are illuminated by electrons at different angles in the defocused probe (see Fig. 23). As the probe is focused on the specimen, the magnification increases, until at focus (infinite magnification) a ronchigram is formed that is an angular map of the objective lens aberrations. A very thin region of sample is required, ideally an amorphous edge or glue, and the optical axis can be seen clearly. Astigmatism can be adjusted by making the phase-contrast features circular, and stability of all lenses and alignments can be checked by wobbling and counting fringes as the probe is displaced. This procedure

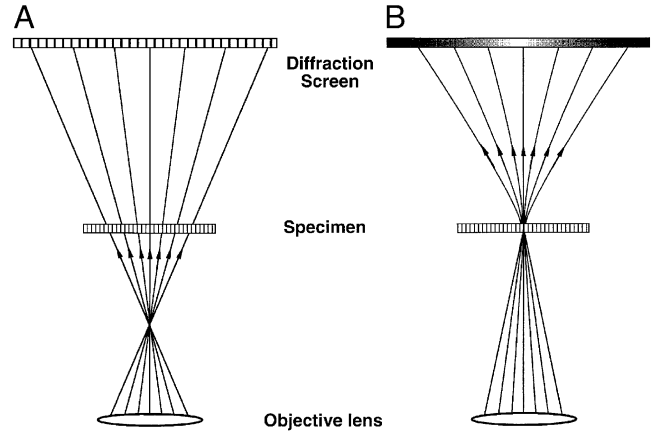


Figure 23. Schematic showing the formation of a shadow image or ronchigram by a stationary probe in the STEM instrument. (**A**) A defocused probe gives a shadow image containing simultaneous image and diffraction information. (**B**) focused probe produces a ronchigram.

gives an accurate alignment and precise position of the objective lens optic axis about which the objective aperture and all detectors are centered. If the microscope has a stage capable of adjusting the sample height, then the shadow image is a useful method to set each sample at the same height and therefore at the same objective lens excitation. In this way all alignments and astigmatism settings are constant and all that remains is to tilt the sample on axis.

With the microscope now adjusted, the scan is turned on and *Z*-contrast images can be collected. Unless there is a large change in height from area to area, the astigmatism should not need adjustment, and any changes in the image will most likely reflect changes in sample tilt. The *Z*-contrast image loses contrast more rapidly than the phase-contrast image with specimen tilt. The first indication is an asymmetry in the image, progressing to a pronounced streaking along the direction of tilt. This can be distinguished from astigmatism because it will not change direction when the objective lens is changed from underfocus to overfocus. Fine focus can be judged by eye, remembering the general characteristics shown in Figure 12, that on weakening the lens (underfocusing), the image contrast first increases, then goes through a maximum at the optimum Scherzer focus, and finally reduces but acquires increasing fine detail, until the image disappears into noise. It should also be remembered that slowing the scan will reduce the image noise quite significantly, although specimen drift becomes a limitation. One can of course also use the conventional phase-contrast image to stigmate and focus, although the noise in a scanned image makes it more difficult to judge minimum image contrast than with conventional TEM.

It should also be realized that in extremely thick regions of sample, the *Z*-contrast image may show contrast reversals, not in the atomic resolution image (usually the region is so thick that an atomic resolution image is no longer possible), but in the relative brightness of materials of different *Z*. In a thick region, the bright-field image has long disappeared, and this effect can be confusing when

searching the sample for the interface of interest. It is caused by multiple elastic scattering that takes electrons off the detector to even higher angles. In the limit there is no more transmission, and obviously the higher Z material will be the first to approach this limit and therefore appear less bright than a material of lower Z.

Finally, the EELS detector needs to be of the highest efficiency possible, which means it should be a cooled CCD parallel detection system (see, e.g., McMullan et al., 1990). Because damage is a more serious concern in EELS, it is also essential to have a time-resolved capability in the system, to allow a series of spectra to be recorded in succession so that they can be compared for signs of damage. If none are seen, they can subsequently be summed to improve statistics. Usually it is in the spectral fine structure that damage effects are first observed, long before the image shows signs of amorphization.

DATA ANALYSIS AND INITIAL INTERPRETATION

Retrieval of the Object Function

The only specimen information lost in recording an incoherent image is at high spatial frequencies, due to the convolution with a probe of finite width. No phase information is lost, as in the recording of a coherent image, because there is no phase information in an incoherent image. It might therefore be assumed that, to retrieve the object, all that would be required would be to deconvolve the probe from the image. Unfortunately this is not useful, as illustrated in Figure 24 (Nellist and Pennycook, 1998b), because there is insufficient information at high frequencies. Although the raw image data in Figure 24A can be Fourier filtered to remove the noise (Fig. 24B), the result of the deconvolution in part C is to produce artifacts between the columns. These arise because, lacking any information at high spatial frequencies, one is forced to cut off the transfer abruptly near the maximum spatial frequency present in the original image. To avoid such artifacts, a slower cutoff can be imposed on the image, but obviously this would then degrade the resolution. In fact, it is not useful to attempt to improve upon the incoherent transfer function in this way, and an alternative method is required to reconstruct the missing high-frequency information.

For atoms far apart, it is reasonable to locate the maximum of each image feature, but this procedure does not work near the limit of resolution because it does not take any account of the probe profile. In Si⟨110⟩, for example, pairs of columns are spaced by distances comparable to the probe size, and the peak image intensity is displaced outward by a few tenths of an angstrom depending on defocus. Maximum entropy (Gull and Skilling, 1984) is a method to accurately account for the effects of the convolution. No prior knowledge concerning the nature of the image is assumed, except that it is incoherent. It does not assume that the object is comprised of atoms, but, given a specified probe profile, it reconstructs the most likely object function consistent with the image data. It produces an object with the least amount of structure needed to account for the image intensity, an object with maximum entropy. Figure 24D shows the reconstructed Si image, obtained by convolving the maximum-entropy object with the probe profile, where the artifacts between the columns are no longer present.

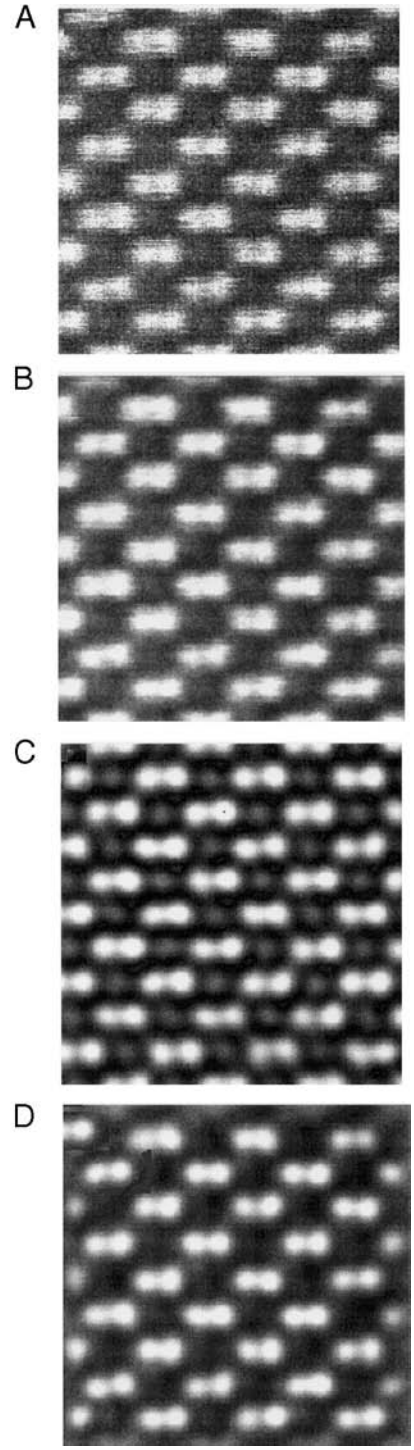


Figure 24. (**A**) Raw Z-contrast image of Si ⟨110⟩; (**B**) low-pass filtered image to reduce noise; (**C**) deconvolution of the probe function leads to artifacts between the columns; (**D**) maximum entropy retrieves the correct object, giving a reconstructed image free of artifacts.

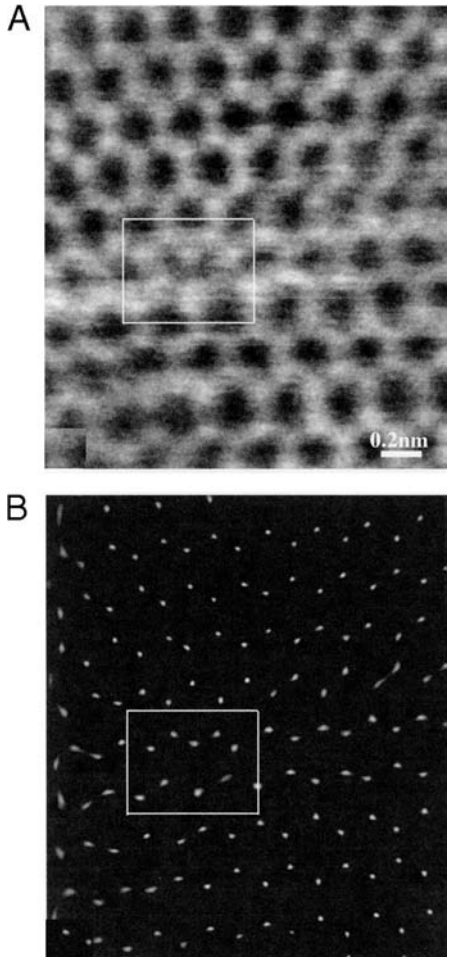


Figure 25. (**A**) *Z*-contrast image of a threading dislocation in a GaN thin film grown on sapphire, viewed along the ⟨0001⟩ direction. (**B**) The 8-fold structure of the core is clear from the maximum-entropy object.

An example of this method applied to a dislocation core in GaN is shown in Figure 25 (Xin et al., 1998). The maximum-entropy object function consists of fine points, the best fit atomic positions, and clearly reveals the core structure of the threading dislocation. The accuracy of the method is easily checked by measuring spacings far from the dislocation core, here, as found typically, $\sim\pm 0.1$ Å for individual columns. More detailed descriptions of the maximum-entropy approach applied to *Z*-contrast images and its accuracy and comparison with alternative schemes are given elsewhere (McGibbon et al., 1999; Nellist and Pennycook, 1998b).

Strain Contrast

Elastic strain fields due to impurities, point defects, or extended defects will affect the image by disturbing the channeling of the probe, even if they do not significantly affect the columnar scattering cross-section. Dislocations are visible because they induce transitions between Bloch states. This is the usual mechanism of diffraction contrast

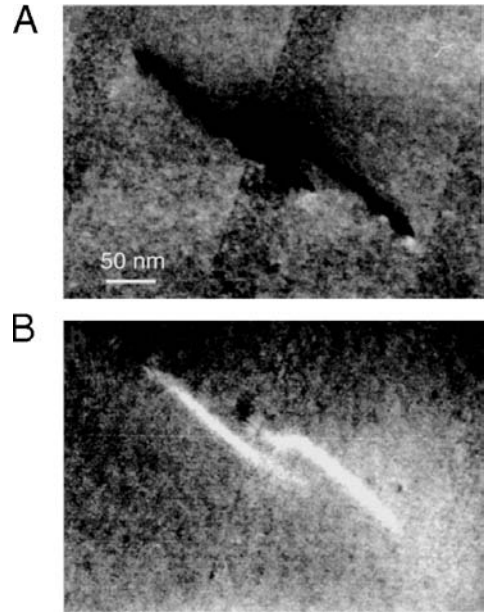


Figure 26. (**A**) STEM bright-field and (**B**) annular dark-field diffraction contrast images of inclined dislocations in a thick Si/Si(B) superlattice.

in CTEM images, but clearly, if the 1*s*-type Bloch states are involved, dislocations will also be visible in the annular detector image. In general, transitions may occur into or out of the *s* states, depending on the depth in the crystal, giving oscillatory contrast from inclined dislocations, as seen in Figure 26 (Perovic et al., 1993a,b). Even dislocations that are viewed end-on may show strain contrast due to the transverse relaxations of the atomic positions that occur near the sample surface. For this reason, grain boundaries, which are closely spaced arrays of dislocations, often appear brighter or darker than the matrix. Strain contrast is relatively long range compared to the lattice parameter and can be removed by Fourier filtering if desired. As shown below, strain contrast depends strongly on detector angle and can be distinguished from compositional changes (*Z* contrast) by comparing images taken with different inner detector angles.

From the discussion of the loss of coherence through atomic vibrations, it is clear that in a zone axis crystal transverse displacements comparable to the atomic vibration amplitude will significantly affect the scattering at high angles. The diffuse scattering cross-section per atom σ depends on scattering angle and temperature through the atomic form factor f and the Debye-Waller factor M^T:

$$\sigma = f^2[1 - \exp(-2M^{\mathrm{T}}s^2)] \qquad (8)$$

where s is the scattering angle, $M^{\mathrm{T}} = 8\pi^2 u_{\mathrm{T}}^2$, and u_{T}^2 is the mean-square thermal vibration amplitude of the atom (Hall and Hirsch, 1965). In the presence of static random atomic displacements, assuming a Gaussian distribution

of strain with a mean-square static displacement of u_s^2, the atomic scattering cross-section will be modified to (Hall et al., 1966)

$$\sigma^s = f^2\{1 - \exp[-2(M^T + M^S)s]\} \quad (9)$$

where $M^s = 8\pi u_s^2$. It is clear from the form of these expressions that both tend to the full atomic scattering cross-section $\sigma = f^2$ at a sufficiently high scattering angle. At lower angles where the Debye-Waller factor is significant, static strains comparable to the thermal vibration amplitude may lead to a significantly enhanced scattering cross-section. Figure 27 shows images of a thermally grown Si-SiO_2 interface. The bright-field phase-contrast image (Fig. 27A) shows dark contrast that could be due to a number of effects, such as strain, thickness variation, bending of the crystal, or a combination of these mechanisms. Figure 27B shows an incoherent dark-field image collected simultaneously using a low (25-mrad) inner radius for the annular detector. Now there is a bright band near the interface indicating additional scattering. With this image alone, this additional scattering could be due either to strain or to the presence of some heavy impurity atoms. However, when the inner detector angle is increased further, to 45 mrad, the bright line disappears (Fig. 27C), showing that the contrast cannot be due to the presence of heavy impurity atoms that would still give increased scattering. The contrast must therefore be due to a static strain effect. Intensity profiles across the two dark-field images are shown in Figure 28A. Taking their ratio normalizes the channeling effect of each column and gives the additional atomic displacement due to strain, as shown in Figure 28B (Duscher et al., 1998b). This is seen to decrease exponentially from the incoherent interface, as would be expected for a uniform array of misfit dislocations.

Strain effects are also commonly seen at grain boundaries and can be distinguished from true *Z* contrast in the same way, by comparing images at different detector

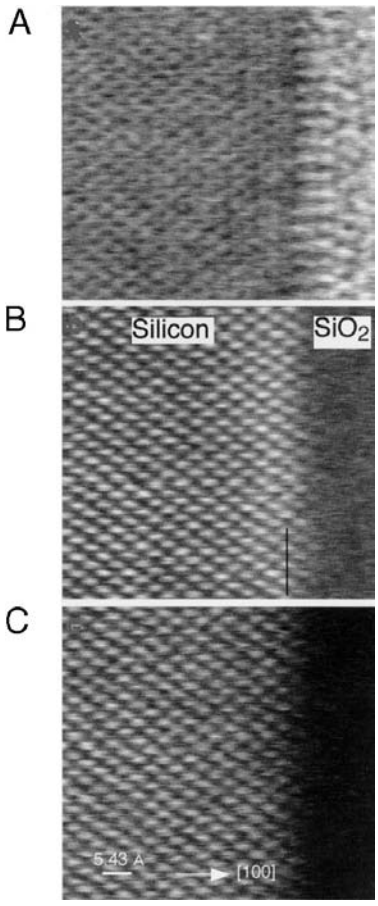


Figure 27. STEM images of a Si-SiO_2 interface: (**A**) bright-field phase-contrast image; (**B**) *Z*-contrast image with 25 mrad inner detector angle showing strain contrast. The vertical line marks the last Si plane used for strain profiling. (**C**) *Z*-contrast image with 45 mrad inner detector angle.

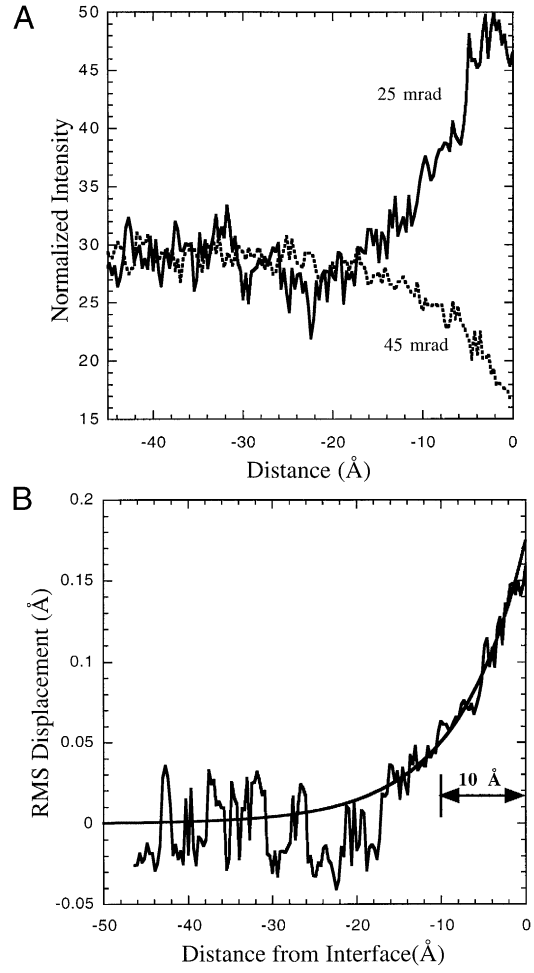


Figure 28. (**A**) Intensity profiles across the *Z*-contrast images of Figure 27. The high-angle profile shows the dechanneling effect near the interface, which can be used to normalize the profile obtained with a lower detector angle. (**B**) Root-mean-squared atomic displacement due to static strain induced by the Si-SiO_2 interface and an exponential decay fitted to the data. Profile ends at the vertical line shown in Figure 27.

angles. A strain effect will appear bright at low detector angles, changing to dark at high detector angles. True segregation of heavy impurities will appear bright at both detector angles.

SAMPLE PREPARATION

Specimen preparation requirements for STEM are essentially the same as for HREM—a similar sample thickness, but generally the thinner the better to avoid problems with projection. However, some differences are required for working with small probes. In particular, the crystal surface must be as free of damage as possible. Ion milling leaves a thin amorphous layer, which broadens the probe before it reaches the crystal surface below. As the resolution is directly determined by the probe size, this leads to loss of contrast and perhaps no atomic resolution at all, whereas it might hardly be noticed in a phase-contrast image. Low-voltage final milling is essential for STEM, and samples made by cleaving or chemical means will give the best results. Nevertheless, most of the images shown in this unit were obtained from samples that had been ion milled. The amorphous layer increases the intensity fluctuations between identical columns as the probe profile is affected by the randomly located atoms above each column.

SPECIMEN MODIFICATION

Damage is usually expected to be a major concern with the high brightness of a small probe, but it should be remembered that for a coherent probe at the electron optical limit, a high demagnification is used and probe currents are in the region of 10 to 100 pA, small compared with the nanoampere currents used typically in microanalysis. There is only one electron in the column at a time, and some excitations, such as electron-hole pairs, may have time to decay before the next electron arrives. For this reason, many damage processes have a threshold beam current, and generally damage by an atomic resolution probe is not a limiting factor for imaging. Nevertheless, when imaging grain boundaries, it is sometimes observed that the structure at the boundary becomes amorphous after a few scans (Maiti et al., 1999). In STEM, because the only area illuminated is that scanned by the beam, it is possible to "walk" the probe down the boundary, taking images from a fresh area each time. In EELS studies damage is a more serious concern, because of the longer exposure times needed for data collection and also because the light elements are the first to suffer displacement damage.

PROBLEMS

Perhaps the main problem with the *Z*-contrast technique has been the lack of available instruments except in selected centers. Now that dedicated instruments are no longer required and the technique is being incorporated into the conventional TEM column, it should see more widespread use.

Because the STEM instrument is serial, with data recorded pixel by pixel, those familiar with CTEM will find a number of differences. One difference is certainly the noisy appearance of the raw *Z*-contrast image, although the atomic positions can be determined through maximum-entropy analysis to an accuracy of ± 0.1 Å, comparable to that of the best conventional methods. Reconstructed images are less noisy, although the noise in the original data now appears as random displacements of the "noise free" atomic columns.

Sample drift also appears in a very different manner than in CTEM. Drift blurs the image in CTEM, whereas in STEM it distorts the image. One can view this as an advantage or disadvantage, but the result is that STEM images often show angular distortion or wavy atomic planes due to specimen drift. This makes it difficult to determine rigid shifts across grain boundaries by the usual methods. One would usually rotate the scan direction so that the fast scan is across the boundary, giving the best measurement of expansion or contraction. One can also compare images taken at different rotations.

Vacuum is a critical requirement for STEM studies, as hydrocarbons are attracted to the small probe and can build up rapidly during high-magnification viewing or analysis when the probe scans only a small area. Sometimes this problem can be alleviated by "flooding," irradiating the sample at the maximum available current by removing all apertures and reducing the demagnification. Obviously, besides polymerizing the hydrocarbons, this may also damage the sample. Usually the vacuum in modern instruments is sufficiently good that the specimen itself becomes the major source of carbon contamination. Use of oil-free specimen preparation techniques is useful, particularly if samples are sensitive to carbon, as in the case of the high-temperature superconductors. Plasma cleaning can be useful, or even a gentle bake in air using an oven or by placing the specimen on a light bulb can be very effective in eliminating contamination. Only those hydrocarbons that adsorb strongly to the sample at room temperature are a problem; lighter species evaporate in the vacuum while heavier species are immobile. Hence, only a moderate temperature is required.

ACKNOWLEDGMENTS

The author is grateful to their colleagues P. D. Nellist, D. E. Jesson, M. F. Chisholm, N. D. Browning, Y. Yan, Y. Xin, B. Rafferty, G. Duscher, and E. C. Dickey for research collaborations. This research was supported by Lockheed Martin Energy Research Corp., under Department of Energy Contract No. DE-AC05-96OR22464, and by appointments to the ORNL Postdoctoral Research Associates Program administered jointly by ORNL and ORISE.

LITERATURE CITED

Ade, G. 1977. On the incoherent imaging in the scanning transmission electron microsocpe. *Optik* 49:113–116.

Anderson, S. C., Birkeland, C. R., Anstis, G. R., and Cockayne, D. J. H. 1997. An approach to quantitative compositional profiling at near atomic resolution using high-angle annular dark-field imaging. *Ultramicroscopy* 69:83–103.

Brown, L. M. 1981. Scanning transmission electron microscopy: Microanalysis for the microelectronic age. *J. Phys. F* 11:1–26.

Browning, N. D., Buban, J. P., Nellist, P. D., Norton, D. P., Chisholm, M. F., and Pennycook, S. J. 1998. The atomic origins of reduced critical currents at [001] tilt grain boundaries in $YBa_2Cu_3O_{7-\delta}$ thin films. *Physica C* 294:183–193.

Browning, N. D., Chisholm, M. F., Nellist, P. D., Pennycook, S. J., Norton, D. P., and Lowndes, D. H. 1993b. Correlation between hole depletion and atomic structure at high-angle grain boundaries in $YBa_2Cu_3O_{7-\delta}$. *Physica C* 212:185–190.

Browning, N. D., Chisholm, M. F., and Pennycook, S. J. 1993a. Atomic-resolution chemical analysis using a scanning transmission electron microscope. *Nature* 366:143–146.

Browning, N. D. and Pennycook, S. J. 1995. Atomic-resolution electron energy loss spectroscopy in the scanning transmission electron microscope. *J. Microsc.* 180:230–237.

Browning, N. D., Pennycook, S. J., Chisholm, M. F., McGibbon, M. M., and McGibbon, A. J. 1995. Observation of structural units at symmetric [001] tilt boundaries in $SrTiO_3$. *Interface Sci.*2:397–423.

Chisholm, M. F., Maiti, A., Pennycook, S. J., and Pantelides, S. T. 1998. Atomic configurations and energetics of arsenic impurities in a silicon grain boundary. *Phys. Rev. Lett.* 81:132–135.

Chisholm, M. F. and Pennycook, S. J. 1997. *Z*-contrast imaging of grain-boundary core structures in semiconductors. *Mater. Res. Soc. Bull.* 22:53–57.

Chisholm, M. F., Pennycook, S. J., Jebasinski, R., and Mantl, S. 1994. New interface structure for A-type $CoSi_2$/Si(111). *Appl. Phys. Lett.* 64:2409–2411.

Coene, W., Janssen, G., Op de Beeck, M., and Van Dyck, D. 1992. Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Phys. Rev. Lett.* 69:3743–3746.

Colliex, C. and Mory, C. 1984. Quantitative aspects of scanning transmission electron microscopy. *In* Quantitative Electron Microscopy (J. N. Chapman and A. J. Craven, eds.). pp. 149–216. Scottish Universities Summer School in Physics, Edinburgh.

Cowley, J. M. 1969. Image contrast in a transmission scanning electron microscope. *Appl. Phys. Lett.* 15:58–59.

Cowley, J .M. 1976. Scanning transmission electron microscopy of thin specimens. *Ultramicroscopy* 2:3–16.

Cowley, J. M. 1979. Adjustment of a STEM by use of shadow images. *Ultramicroscopy* 4:413–418.

Cowley, J. M. 1986a. Principles of image formation. *In* Principles of Analytical Electron Microscopy (J. J. Hren, J. I. Goldstein, and D. C. Joy, eds.). pp. 343–368. Plenum, New York.

Cowley, J. M. 1986b. Electron diffraction phenomena observed with a high resolution STEM instrument. *J. Electron Microsc. Techne*. 3:25–44.

Cowley, J. M. 1997. Scanning transmission electron microscopy. *In* Handbook of Microscopy (S. Amelinckx, D. van Dyck, J. van Landuyt, and G. van Tendeloo, eds.). pp. 563–594. VCH Publishers, Weinheim, Germany.

Crewe, A. V., Wall, J., and Langmore, J. 1970. Visibility of single atoms. *Science* 168:1338–1340.

Datye, A. K. and Smith, D. J. 1992. The study of heterogeneous catalysts by high-resolution transmission electron microscopy. *Catal. Rev. Sci. Eng.* 34:129–178.

Dickey, E. C., Dravid, V. P., Nellist, P. D., Wallis, D. J., and Pennycook, S. J. 1998. Three-dimensional atomic structure of NiO-ZrO_2(cubic) interfaces. *Acta Metall. Mater*. 46:1801–1816.

Duscher, G., Browning, N. D., and Pennycook, S. J. 1998a. Atom column resolved electron energy loss spectroscopy. *Phys. Stat. Solidi (a)* 166:327–342.

Duscher, G., Pennycook, S. J., Browning, N. D., Rupangudi, R., Takoudis, C., Gao, H-J., and Singh, R. 1998b. Structure, composition and strain profiling of Si/SiO_2 interfaces. *In* Characterization and Metrology for ULSI Technology, Conf. Proc. 449 (D. G. Seiler, A. C. Diebold, W. M. Bullis, T. J. Shaffer, R. McDonald, and E. J. Walters, eds.). pp. 191–195. American Institute of Physics, Woodbury, N.Y.

Engel, A., Wiggins, J. W., and Woodruff, D. C. 1974. A comparison of calculated images generated by six modes of transmission electron microscopy. *J. Appl. Phys.* 45:2739–2747.

Fathy, D. and Pennycook, S. J. 1981. STEM Microanalysis of Impurity Precipitates and Doping in Semiconductor Devices, Conf. Ser. 60 (M. J. Goringe, ed.) pp. 243–248. Institute of Physics, Bristol and London.

Gull, S. F. and Skilling, J. 1984. Maximum entropy methods in image processing. *IEE Proc.* 131F:646–659.

Hall, C. R. and Hirsch, P. B. 1965. Effect of thermal diffuse scattering on propagation of high energy electrons through crystals. *Proc. R. Soc. A* 286:158–177.

Hall, C. R., Hirsch, P. B., and Booker. 1966. Effects of point defects on absorption of high energy electrons passing through crystals. *Philos. Mag.* 14:979–989.

Hartel, P., Rose, H., and Dinges, C. 1996. Conditions and reasons for incoherent imaging in STEM. *Ultramicroscopy* 63:93–114.

Hillyard, S., Loane, R. F. and Silcox, J. 1993. Annular dark-field imaging: Resolution and thickness effects. *Ultramicroscopy* 49:14–25.

Isaacson, M. S., Ohtusuki, M., and Utlaut, M. 1979. Electron microscopy of individual atoms. *In* Introduction to Analytical Electron Microscopy (J. J. Hren, J. I. Goldstein, and D. C. Joy, eds.). pp. 343–368. Plenum, New York.

James, E. M. and Browning, N. D. 1999. Practical aspects of atomic resolution imaging and analysis in STEM. *Ultramicroscopy*. 78:125–139.

James, E. M., Browning, N. D., Nicholls, A. W., Kawasaki, M., Xin, Y., and Stemmer, S. 1998. Demonstration of atomic resolution Z-contrast imaging in a JEOL-2010F scanning transmission electron microscope. *J. Electron Microsc.* 47:561–574.

Jesson, D. E. and Pennycook, S. J. 1993. Incoherent imaging of thin specimens using coherently scattered electrons. *Proc. R. Soc. (London) A* 441:261–281.

Jesson, D. E. and Pennycook, S. J. 1995. Incoherent imaging of crystals using thermally scattered electrons. *Proc. R. Soc. (London) A* 449:273–293.

Jesson, D. E., Pennycook, S. J., and Baribeau, J.-M. 1991. Direct imaging of interfacial ordering in ultrathin $(Si_mGe_n)_p$ superlattices. *Phys. Rev. Lett.* 66:750–753.

Lin, J. A. and Cowley, J. M. 1986. Calibration of the operating parameters for an HB5 STEM instrument. *Ultramicroscopy* 19:31–42.

Lipson, S. G. and Lipson, H. 1982. Optical Physics, 2nd ed. Cambridge University Press, Cambridge.

Loane, R. F., Kirkland, E. J., and Silcox, J. 1988. Visibility of single heavy atoms on thin crystalline silicon in simulated annular dark-field STEM images. *Acta Crystallogr.* A44:912–927.

Loane, R. F., Xu, P., and Silcox, J. 1992. Incoherent imaging of zone axis crystals with ADF STEM. *Ultramicroscopy* 40:121–138.

Maiti, A., Chisholm, M. F., Pennycook, S. J., and Pantelides, S. T. 1999. Vacancy-induced stuctural transformation at a Si grain boundary. *Appl. Phys. Lett.* 75:2380–2382.

McGibbon, M. M., Browning, N. D., Chisholm, M. F., McGibbon, A. J., Pennycook, S. J., Ravikumar, V., and Dravid, V. P. 1994. Direct determination of grain boundary atomic structure in $SrTiO_3$. *Science* 266:102–104.

McGibbon, M. M., Browning, N. D., McGibbon, A. J., and Pennycook, S. J. 1996. The atomic structure of asymmetric [001] tilt boundaries in $SrTiO_3$. *Philos. Mag. A* 73:625–641.

McGibbon, A. J., Pennycook, S. J., and Angelo, J. E. 1995. Direct observation of dislocation core structures in CdTe/GaAs(001). *Science* 269:519–521.

McGibbon, A. J., Pennycook, S. J., and Jesson, D. E. 1999. Crystal structure retrieval by maximum entropy analysis of atomic resolution incoherent images. *J. Microsc.* 195:44–57.

McMullan, D., Rodenburg, J. M., Murooka, Y., and McGibbon, A. J. 1990. Parallel EELS CCD detector for a VG HB501 STEM. *Inst. Phys. Conf. Ser* 98:55–58.

Möbus, G. 1996. Retrieval of crystal defect structures from HREM images by simulated evolution. I. Basic technique. *Ultramicroscopy* 65:205–216.

Möbus, G. and Dehm, G. 1996. Retrieval of crystal defect structures from HREM images by simulated evolution. II. Experimental image evaluation. *Ultramicroscopy* 65:217–228.

Nakamura, K., Kakibayashi, H., Kanehori, K., and Tanaka, N., 1997. Position dependence of the visibility of a single gold atom in HAADF-STEM image simulation. *J. Electron Microsc.* 1:33–43.

Nellist, P. D. and Pennycook, S. J. 1996. Direct imaging of the atomic configuration of ultra-dispersed catalysts. *Science* 274:413–415.

Nellist, P. D. and Pennycook, S. J. 1998a. Sub-Ângstrom resolution TEM through under-focussed incoherent imaging. *Phys. Rev. Lett.* 81:4156–4159.

Nellist, P. D. and Pennycook, S. J. 1998b. Accurate structure determination from image reconstruction in ADF STEM. *J. Microsc.* 190:159–170.

Nellist, P. D. and Pennycook, S. J. 1999. Incoherent imaging using dynamically scattered coherent electrons. *Ultramicroscopy.* 78:111–124.

Orchowski, A., Rau, W. D., and Lichte, H. 1995. Electron holography surmounts resolution limit of electron microscopy. *Phys. Rev. Lett.* 74:399–401.

Pennycook, S. J. 1982. High resolution electron microscopy and microanalysis. *Contemporary Phys.* 23:371–400.

Pennycook, S. J. 1995. *Z*-contrast electron microscopy for materials science. *In* Encyclopedia of Advanced Materials (D. Bloor, R. J. Brook, M. C. Flemings, and S. Mahajan, eds.). pp. 2961–2965. Pergamon Press, Oxford.

Pennycook, S. J. and Boatner, L. A. 1988. Chemically sensitive structure imaging with a scanning transmission electron microscope. *Nature* 336:565–567.

Pennycook, S. J., Brown, L. M., and Craven, A. J. 1980. Observation of cathodoluminescence at single dislocations by STEM. *Philos. Mag.* 41:589–600.

Pennycook, S. J., Chisholm, M. F., Jesson, D. E., Norton, D. P., Lowndes, D. H., Feenstra, R., Kerchner, H. R., and Thomson, J. O. 1991. Interdiffusion, growth mechanisms, and critical currents in $YBa_2Cu_3O_{7-x}/PrBa_2Cu_3O_{7-x}$ superlattices. *Phys. Rev. Lett.* 67:765–768.

Pennycook, S. J. and Howie, A. 1980. Study of single electron excitations by electron microscopy; II cathodoluminescence image contrast from localized energy transfers. *Philos. Mag.* 41:809–827.

Pennycook, S. J. and Jesson, D. E. 1990. High-resolution incoherent imaging of crystals. *Phys. Rev. Lett.* 64:938–941.

Pennycook, S. J. and Jesson, D. E. 1991. High-resolution *Z*-contrast imaging of crystals. *Ultramicroscopy* 37:14–38.

Pennycook, S. J. and Jesson, D. E. 1992. Atomic resolution *Z*-contrast imaging of interfaces. *Acta Metall. Mater.* 40:S149–S159.

Pennycook, S. J., Jesson, D. E., and Nellist, P. D. 1996. High angle dark field STEM for advanced materials. *J. Electron Microsc.* 45:36–43.

Pennycook, S. J., Jesson, D. E., Nellist, P. D., Chisholm, M. F., and Browning, N. D. 1997. Scanning transmission electron microscopy: *Z*–contrast. *In* Handbook of Microscopy (S. Amelinckx, D. van Dyck, J. van Landuyt, and G. van Tendeloo, eds) pp. 595–620. VCH Publishers, Weinheim, Germany.

Perovic, D. D., Howie, A., and Rossouw, C. J. 1993a. On the image contrast from dislocations in high-angle annular dark-field scanning transmission electron microscopy. *Philos. Mag. Lett.* 67: 261–277.

Perovic, D. D., Rossouw, C. J. and Howie, A. 1993b. Imaging inelastic strains in high-angle annular dark field scanning transmission electron microscopy. *Ultramicroscopy* 52:353–359.

Rafferty, B. and Pennycook, S. J. 1999. Towards atomic column-by-column spectroscopy. *Ultramicroscopy* 78:141–152.

Lord Rayleigh 1896. On the theory of optical images with special reference to the microscope. *Philos. Mag.* 42(5):167–195.

Scherzer, O. 1949. The theoretical resolution limit of the electron microscope. *J. Appl. Phys.* 20:20–29.

Spence, J. C. H. and Cowley, J. M. 1978. Lattice imaging in STEM. *Optik* 50:129–142.

Wallis, D. J., Browning, N., Sivananthan, S., Nellist, P. D., and Pennycook, S. J. 1997. Atomic layer graphoepitaxy for single crystal heterostructures. *Appl. Phys. Lett.* 70:3113–3115.

Xin, Y., Pennycook, S. J., Browning, N. D., Nellist, P. D., Sivananthan, S., Omnès, F., Beaumont, B., Faurie, J.-P., and Gibart, P. 1998. Direct observation of the core structures of threading dislocations in GaN. *Appl. Phys. Lett.* 72:2680–2682.

Yan, Y., Chisholm, M. F., Duscher, G., Maiti, A., Pennycook, S. J., and Pantelides, S. T. 1998. Impurity induced structural transformation of a MgO grain boundary. *Phys. Rev. Lett.* 81:3675–3678.

Zeitler, E. and Thomson, M. G. R. 1970. Scanning transmission electron microscopy. *Optik* 31:258–280, 359–366.

KEY REFERENCES

Bird, D. M. 1989. Theory of zone axis electron diffraction. *J. Electron. Microsc. Tech.* 13:77–97.

A good introduction to the theory of dynamical diffraction using Bloch states.

Keyse, R. J., Garret-Redd, A. J., Goohhew, P. J., and Lorimer, G. W. 1998. Introduction to Scanning Transmission Electron Microscopy. Bios Scientific Publishers, Oxford.

A general introduction to STEM.

Pennycook, S. J. 1992. *Z*-contrast transmission electron microscopy: Direct atomic imaging of materials. *Annu. Rev. Mater. Sci.* 22:171–195.

A simple description of Z-contrast imaging and a number of early applications.

Pennycook, S. J. and Jesson, D. E. 1992. High-resolution imaging in the scanning transmission electron microscope. *In* Proceedings of the International School on Electron Microscopy in Materials Science (P. G. Merli and M. Vittori Antisari, eds.). pp. 333–362. World Scientific, Singapore.

Includes a detailed discussion on breaking the coherence of the imaging process through detector geometry and thermal vibrations.

Pennycook, S. J., Jesson, D. E., Chisholm, M. F., Browning, N. D., McGibbon, A. J., and McGibbon, M. M. 1995. *Z*-contrast imaging in the scanning transmission electron microscope. *J. Microsc. Soc. Am.* 1:231–251.

A nonmathematical description of Z-contrast imaging and a number of applications to semiconductors and superconductors.

Pennycook et al., 1997. See above.

Provides a more detailed theoretical treatment of incoherent imaging with elastic, thermal, and inelastic scattering and a survey of applications to materials.

APPENDIX A: INCOHERENT IMAGING OF WEAKLY SCATTERING OBJECTS

For a very thin specimen, effects of probe dispersion and absorption may be ignored, and the scattered amplitude y_s in the direction $\mathbf{K}_f$ is obtained immediately from the first Born approximation (Pennycook et al., 1997),

$$\psi_s(\mathbf{K}_f) = \frac{m}{2\pi\hbar^2}\int e^{-i\mathbf{K}_f\bullet\mathbf{R}}V(\mathbf{R})P(\mathbf{R}-\mathbf{R_0})d\mathbf{R} \qquad (10)$$

where $V(\mathbf{R})$ is the specimen potential integrated along the beam direction, called the projected potential. Integrating the scattered intensity $|\psi_s|^2$ over all final states $\mathbf{K}_f$ and using the identity

$$\int e^{-i\mathbf{K}_f\bullet(\mathbf{R}-\mathbf{R}')}d\mathbf{K}_f = (2\pi)^2\delta(\mathbf{R}-\mathbf{R}') \qquad (11)$$

the total scattered intensity is given as

$$I(\mathbf{R_0}) = \int O(\mathbf{R})P^2(\mathbf{R}-\mathbf{R_0})d\mathbf{R} \qquad (12)$$

$$= O(\mathbf{R}) * P^2(\mathbf{R}_0) \qquad (13)$$

a convolution of the probe intensity profile $P^2(\mathbf{R})$ with an object function $O(\mathbf{R})$ given by

$$O(\mathbf{R}) = \left(\frac{\chi}{2E}\right)^2 V^2(\mathbf{R}) \qquad (14)$$

where $\sigma = \chi/2E$ is the interaction constant and E is the beam energy. Therefore, provided *all* scattered electrons could be collected, we see immediately that incoherent imaging would be obtained with a resolution controlled by the incident probe intensity profile.

For many years it was considered that incoherent imaging at atomic resolution was impossible in principle because much of the total scattered intensity occurs at small angles, in the same angular range as the unscattered beam (e.g., Ade, 1977). There is no way to distinguish unscattered electrons from elastically scattered electrons, which was referred to as the hole-in-the-detector problem. However, realizing that the Rayleigh bright-field detector is equivalent to the annular dark-field detector in Figure 5, it is clear that the problem does not arise in the dark-field case and perfect incoherent imaging may be achieved.

APPENDIX B: STEM MANUFACTURERS

Dedicated STEM Instruments

Conventional phase-contrast and diffraction contrast modes are available, but all images are recorded in the scanning mode. Instruments have cold field emission guns and are typically equipped with EELS and x-ray analysis capabilities.

VG Microscopes: No longer commercially available, but often can be found in user centers. The entire microscope is bakeable, resulting in a clean vacuum.

Hitachi: http://www.hii.hitachi.com/prdem.htm

TEM/STEM Instruments

JEOL: http://www.jeol.com/navbar/tembar.html

Hitachi: http://www.hii.hitachi.com/prdem.htm

Philips: http://www.feic.com/tecnai/main.htm

The Philips and JEOL instruments have a Schottky source. Hitachi instruments have a cold field emission source. Note that manufacturers specify resolution in different ways. The best comparison is to find the objective lens spherical aberration coefficient and use the definitions of resolution given in Table 1.

S. J. PENNYCOOK
Oak Ridge National Laboratory
Oak Ridge Tennsessee

SCANNING TUNNELING MICROSCOPY

INTRODUCTION

Scanning tunneling microscopy (STM) is a technique for the determination of the structure of surfaces, with spatial resolution on the angstrom scale. In an oversimplified description, STM allows imaging of the surface of conductive materials down to the atomic level.

The technique is based on the measurement and control of the current of electrons tunneling between a sharp stylus (hereafter called the tip) and the sample surface. During imaging, the tip is separated a few angstroms (3 to 10 Å) from the surface, while it is being rastered with the help of piezoelectric transducers. The area scanned is tens to thousands of angstroms on a side. STM has the advantage over other microscopies that its operation does not require any special environmental conditions. Imaging can be performed in vacuum, at high pressures (including air), and in liquids. The choice of the environment is determined mostly by the requirements of the sample and its surface condition.

Acquisition of STM images is relatively simple with modern instruments. Most of the experimental time is spent on sample preparation and tip conditioning in the appropriate environment (ultrahigh vacuum, electrochemical cell, or controlled atmosphere). Tip conditioning and noise reduction can sometimes add considerable time, depending on factors such as the nature of the experiment and reactivity of the sample. This time depends also on the type of information being sought. A medium-resolution (0.5 to 1 Å in z, i.e., normal to the surface) large-area image of a few hundred angstroms is relatively easy to obtain. In such a case, however, similar information can be obtained with atomic force microscopy (AFM), which is easier to use. It is in the atomic-resolution area that STM excels over any other surface microscopy, because atom-to-atom corrugations <0.5 Å—and sometimes <0.1 Å, as in the case of close-packed metal surfaces—need to be resolved. Therefore, the requirements for both instrument and sample preparation are substantially more stringent.

STM was discovered at the beginning of the 1980s and has since experienced an explosive development in both instrumentation and applications (Olgetree and Salmeron, 1990; Wiesendanger, 1994). It has given rise to numerous related techniques generally referred to as scanning probe microscopies (SPM). The most popular of these is AFM. The related technique known as scanning electrochemical microscopy is covered in detail elsewhere (SCANNING ELECTROCHEMICAL MICROSCOPY).

In considering the ultimate resolution obtainable with the STM, it is important to consider the role of tip geometry. The schematic drawing at the bottom of Figure 1 illustrates the basic point: on flat parts of the sample, where one tip atom is responsible for the tunneling, STM reaches its maximum performance. On rough surfaces, however, the tunneling point on the tip apex changes as it moves over sharp corners and changing slopes. The image is then a convolution of the tip and surface shapes, and the final resolution is determined by the tip radius and profile. STM is thus not the best technique to probe, for example, highly corrugated surfaces, deep-patterned surfaces, or powdered and porous materials.

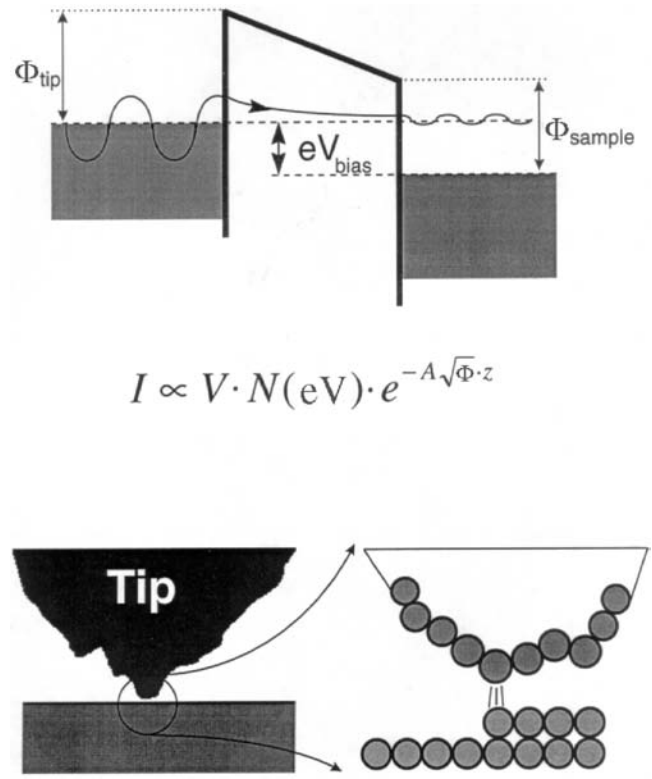


Figure 1. Top: Schematic energy diagram of the tip and sample in an STM experiment. Electrons tunnel from the negatively-biased tip (left) to the sample (right). Only the electrons within the eV range of states in the tip participate in the tunneling. The tunnel barrier is related to the work function Φ of the material involved. Bottom: Schematic of the shape of a tip near a flat surface. At the atomic scale shown on the right, the most protruding atom concentrates the tunnel current. Secondary images can be obtained if another protruding atom can enter within tunneling range in surface irregularities.

Complementary and Competitive Techniques

Atomic force microscopy (AFM) is a technique that has many characteristics in common with STM, although the physical processes involved are quite different. In AFM, there is also a sharp tip but it is mounted on a very flexible cantilever. As in the STM, it is rastered over the surface by means of piezoelectric transducers. Tip-surface interaction forces are sensed in AFM by the deflection of the lever, instead of the tunneling current. AFM is therefore ideal for studies of insulating materials that are not directly accessible to STM imaging.

The resolution of AFM in its most usual operation modes (contact, friction, or tapping) is not truly atomic, as is the case with STM. The forces of interaction produce a contact spot that is several tens of angstroms in diameter, depending on the applied load. Thus atomic-size point defects are not observed in AFM. However, new non-contact imaging techniques using resonant-lever modulation are being developed, and true atomic resolution, i.e., the visualization of point defects, has been demonstrated (Giessibl, 1995; Kitamura and Iwatsuki, 1995; Sugawara et al., 1995). These developments should be incorporated in commercial instruments soon.

Because contact with the (inert) tip is part of the standard operation of the AFM, the technique is less sensitive to chemical instabilities, and to some forms of mechanical vibration noise, than STM. For the routine examination of samples with a resolution of a few nanometers in x-y and angstroms in z, AFM is clearly advantageous over STM.

Electrons in the far field regime, as opposed to tunneling in the near field, are used in transmission electron microscopy (TEM) to provide angstrom-resolved projections of bulk atomic rows in crystalline samples (see SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). TEM has been used on several occasions in an edge-on configuration to resolve surface atomic features. The sample requirements for TEM are quite strict, as samples must be thinned down to a few hundred angstroms to allow transmission of the electron beam. TEM is excellent, for example, for use in studies of bulk phases of crystalline

Table 1. Comparison of Various Microscopies

Technique[a]	Probe/Principle	Environment	Sample Requirements	Resolution (Å)	
				xy	z
STM	Electron tunneling	Air, vacuum, liquid	Flat, conductor	~1–2	$\sim 10^{-2}$
AFM	Forces: attractive, repulsive contact, noncontact	Air, vacuum, liquid	Flat, insulator, conductor	~10	0.1
TEM	Electron transmission	High vacuum	Solid, ~100 Å thin	~2	—
SEM	Secondary electrons	High vacuum	Solid, conductive	20	—
SAM	Auger emission	Ultrahigh vacuum	Conductive	100	—
SIMS	Ion sputtering	Ultrahigh vacuum	Most samples	1000	—

[a]Abbreviations: STM, scanning tunneling microscopy; AFM, atomic force microscopy; TEM, transmission electron microscopy; SEM, scanning electron microscopy; SAM, scanning Auger microscopy; SIMS, scanning secondary ion mass spectroscopy.

precipitates, in metallurgy or in small particles supported on amorphous or crystalline materials. It is also used extensively in studies of biomaterials supported on carbon films. A thin coating (~50 Å) of heavy metal is usually necessary to provide contrast.

Scanning electron microscopy (SEM) is based on the measurement of the secondary electron yield of conductive substrates (see SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING). This yield changes both as a function of composition and local surface slope. The spatial resolution of SEM is determined by the spot size of the electron beam (~20 Å in the best instruments), and by the diffusion of the secondary electrons before exiting the sample. The SEM operates in vacuum and the best results are obtained with conductive substrates. Combined with x-ray emission detection for elemental composition determination, it is a powerful technique for material studies.

At lower resolutions, a number of other techniques should be considered: scanning Auger microscopy (SAM; see AUGER ELECTRON SPECTROSCOPY), scanning secondary ion mass spectroscopy (SIMS), and small-spot x-ray photoelectron spectroscopy (XPS). These techniques offer the great advantage of spectroscopic information in addition to imaging. Table 1 is a compendium of several commonly used techniques for real-space imaging.

PRINCIPLES OF THE METHOD

A schematic view of the operation of the STM is shown in Figure 1. On top is an energy diagram of the sample and tip. The wavefunction of a tunneling electron is represented by the wavy lines inside the solids connected by a decaying line in the intervening gap. The negative bias applied to the tip causes a net flow of electrons to the sample. The electrons leave occupied levels in an energy band of width eV between the Fermi levels of tip and surface, with V being the bias voltage, and tunnel through the potential barrier separating the two materials to occupy empty levels at the sample side. The reverse situation occurs if the bias is positive. Both the height of the tunneling barrier and the separation between the two surfaces enter into the exponential dependence of the current

$$I \propto V \cdot N(\mathrm{eV}) \cdot e^{-A\sqrt{\Phi}\cdot z} \tag{1}$$

where A is a number of order 1 with units $(\mathrm{eV})^{-1/2}$ Å^{-1}, Φ the tunneling barrier height in eV and z the separation in Å. The strong z dependence of I is at the origin of the very high z-distance resolution of the STM. In properly designed microscopes, where the noise level is determined by the Johnson noise of the preamplifier gain resistance, the z resolution can be 1/1000 Å. The in-plane resolution is determined, as illustrated at the bottom of Figure 1, by the atomic arrangement at the tip apex. In the optimal case of a single-atom-terminated tip, it is the electronic "size" of this atom (~1 to 2 Å) that determines the x-y resolution. The nature of the tip (geometry and chemical composition) is paramount in determining the quality of the images. Except in special cases, the nature of the tip is unknown, and the operation of the STM to date is still subject to some degree of empiricism. This empiricism is the art of tip preparation, for which no well-proven recipes exist yet. In practice, however, atomically resolved images are often obtained due to the exponential dependence of I on z, which naturally selects the most protruding atom as the imaging element.

Equation 1 is the well-known quantum mechanical result that relates the tunneling probability of electrons through a one-dimensional potential energy barrier of width z (the tip-sample separation) and height Φ. The value of Φ is related to the work functions of the tip and sample materials and is due to the overlap of the potential energy curves of the electrons beyond the surface atoms. Its value is therefore on the order of a few electron volts.

Typical values of I, V, z and Φ are $I = \sim 1$ nA, $V = \sim 10$ mV to 3 V, $z = \sim 5$ to 15 Å, and $\Phi = \sim 4$ to 5 eV. Equation 1 is valid in the limit of low applied voltages ($V \ll \Phi$), and so it can be viewed as the relation between bias current and gap resistance R. The value of R is the most practical way to characterize the tunneling gap, since it can be directly measured ($R = V/I$).

STM images are typically acquired in a wide range of R values, from MΩ to GΩ and even TΩ. To put the value of R in perspective, it should be compared with the quantum resistance unit $2e^2/h$, which is 12.9 kΩ and observed in single-atom contacts. Between this value and a few tens of MΩs, the risk of strong tip-surface interactions, leading to such problems as sample indentation or tip and surface modifications through atom transfer, is considerable. In surfaces that have low chemical activity or that are covered with passivating layers, contact imaging is sometimes possible in a friction-like regime without permanent modification of the tip. This type of imaging, usually by interposed atoms or molecules between the tip and surface, is not uncommon, even if often not explicitly known to many experimenters.

A clear signature of contact STM imaging is the value of the barrier height Φ, obtained in I versus z measurements. In these, the tip is advanced toward or away from the surface while the current I is measured. The slope of log I versus z should be ~ 2 Å^{-1}, corresponding to a value of $\Phi = \sim 4-5$ eV for clean, true tunneling gaps. This corresponds to a change of about one decade of current per angstrom. Values in the range of a few tenths of an electron volt and below are signatures of "contact" imaging. An example of these two situations is shown in Figure 2, which corresponds to a nominally clean Pt tip and a Pd substrate (Behler et al., 1997a). The explanation for the low barrier height is the elastic deformation of tip and sample due to the contact forces through interposed material in the gap. Because of these elastic deformations, the change in gap width is always less than the change in tip or sample position calculated from the voltages applied to the piezo transducers.

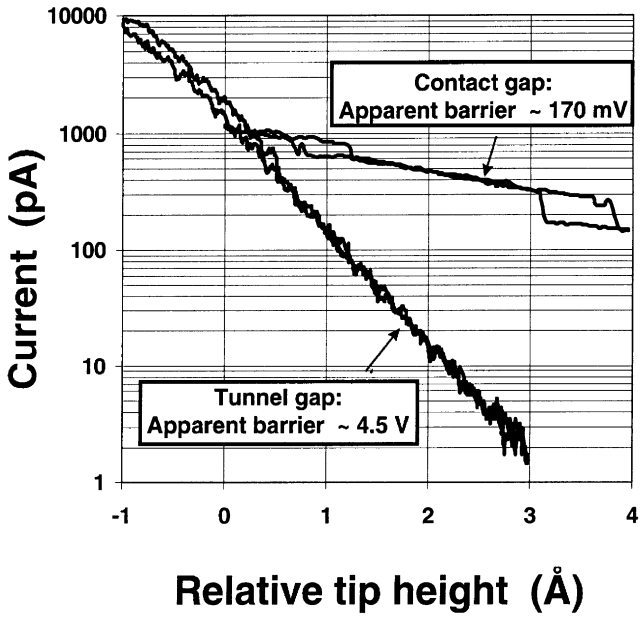


Figure 2. Current versus distance for a tunneling junction of a Pt tip on a Pd(111) crystal surface in ultrahigh vacuum. A true vacuum tunneling produces a change of current of roughly a decade per Å, and corresponds to an apparent barrier height of 4.5 eV. A contact gap, characteristic of dirty junctions, gives a much smaller current decay and apparent barrier heights of less than 1 eV. From Behler et al. (1997a).

The discussion below (see Data Analysis and Initial Interpretation) contains additional information on the physical nature of the tunneling process and its relation to the actual image contrast.

PRACTICAL ASPECTS OF THE METHOD

Acquisition of STM Images

As mentioned above, the tip (or sample) scans over areas that range from a few Å to several μm in x and y, while an electronic feedback control keeps the tunnel current constant. The acquisition of images is done electronically by means of computers interfaced with an STM controller using data conversion cards. The number of data points is variable, but 128×128, 512×512, or 1024×1024 pixels per image are common. On each pixel, the value of z is stored as a 12- or 16-bit number. The resulting data files are 200 kb to 1 Mb in size.

Many variant modes of operation are used, depending on the experiment. Besides the "topographic" mode in which $z(x,y)$ profiles are acquired at constant I, one can also acquire $I(x,y)$ profiles at constant z. The latter is the so-called "current" mode and is useful in fast scanning of atomically flat surfaces. Here the feedback control response time is slow relative to the scanning time, so that only the average value of I is kept constant.

More specialized operation modes include multiple bias imaging, in which interlaced curves at different preset values of V are acquired, and the current image tunneling spectroscopy (CITS) mode, in which the current corresponding to a few selected V values is collected at each pixel during a brief interruption of the feedback control (Hamers, 1992). These two imaging modes have been used principally in investigations of semiconductor surfaces. At selected points of the surface, complete $I(V)$ curves can be obtained for a more detailed study of the band structure (Feenstra et al., 1987). This type of experiment is fraught with difficulties, however, since at the interesting voltages (those of a few volts) that are necessary to access states in the conduction and valence bands, the strong electric field at the tip can easily modify its structure, making reproducibility problematic in many cases. The large information content of the images acquired in these modes can only be displayed in multiple images (one for each value of V) or by post-analysis of the $I(V)$ curves to extract the density-of-states information.

Types of Materials and Problems

The study of semiconductor surfaces is one area where STM has been applied most intensely. Once clean and well-prepared, semiconductor surfaces are relatively easy to image. Large surface reconstructions occur in many of these materials, due to the strong directionality of the tetrahedral covalent bonds, coupled with the tendency to minimize the number of broken bonds. This produces rather large corrugations, from 0.2 to 1 Å in z, and relatively large unit cells in the x-y plane, which makes it

possible to obtain good atomic resolution images even in instruments with modest stability performance. In addition, semiconductor surfaces tend to be less reactive than metals.

Metal surfaces are another important class of materials studied with STM. The applications include surface chemistry, i.e., the study of chemisorbed structures, and electrochemistry. These topics are covered in detail in many recent reviews (Ogletree and Salmeron, 1990). Here we will comment only on some peculiarities in imaging these materials, which should be kept in mind when considering and planning an STM experiment. The most important characteristic of metal surfaces is that their corrugation is substantially smaller than that of semiconductors. This is due to the delocalized nature of the metallic bond. The conduction electrons that are responsible for the tunneling signal are spread over the unit cell, rather than spatially concentrated in covalent bonds. The resulting electronic corrugation is always small, particularly in the close-packed surfaces—i.e., the (111) planes of fcc metals, the (110) planes in bcc metals, and (0001) or basal planes in hcp metals, which are also the most widely studied because of their stability. Corrugations ranging from 0.01 to 0.2 Å are typical. The more open surfaces tend to have more corrugation and are then easier to resolve atomically.

Adsorbates can change the corrugation substantially, since in many cases they give rise to localized, covalent-type bonds. The apparent height of isolated atoms or molecules can vary from a few tenths of an angstrom to several angstroms. Imaging of isolated adsorbates, however, is not an easy task even for strongly bound elements. The reason is the high diffusion rate, at room temperature, of the adatoms. For example, the binding energy of sulfur on Re is very high, ~4.5 eV for the (0001) plane (Dunphy et al., 1993). However, its diffusion energy barrier is ~0.7 eV, which implies a residence time of 100 ms on a site at room temperature. Thus, during scanning, the S adatoms are changing positions continuously, making imaging problematic. More weakly bound adsorbates, e.g., CO molecules, might therefore appear invisible in the STM images because of rapid diffusion. In all these situations, sample cooling is necessary, sometimes to very low temperatures. The pioneering work of Eigler and his collaborators demonstrated the great advantages of low temperature STM for fundamental studies (Eigler and Schweizer, 1990; Zeppenfeld et al., 1992; Heller et al., 1994). As a result, the design and construction of cryogenic STMs has recently increased.

Imaging while the sample is at high temperature imposes strict demands on the design of the microscope head, because of large thermal drifts (Curtis et al., 1997; Kuipers et al., 1993; Kuipers et al., 1995). In vacuum, good shielding, together with careful choice of construction materials with the appropriate expansion coefficients, have made it possible to obtain images up to several hundreds of degrees Celsius (McIntyre et al., 1993). Under high pressures (a few torr to several atmospheres), convection in the gas phase adds to the problem of thermal drifts, and limitations in the highest temperatures are to be expected.

Similar considerations apply to metals in an electrochemical cell, although heating and cooling of the sample is not feasible except in a very narrow temperature window. Therefore, imaging of diluted or isolated adsorbates is difficult or impossible.

Oxides tend to be insulating, and therefore the surfaces of bulk oxide materials cannot be easily imaged by STM. Exceptions include the cases of small bandgap oxides and of oxides that can be made conductive by defect doping. Such is the case for TiO_2 (Rohrer et al., 1990; Onishi and Iwasawa, 1994; Sander and Engel, 1994; Murray et al., 1995), UO_2 (Castell et al., 1996), Fe_3O_4 (Jansen et al., 1996), and others. Nonconductive oxides, in the form of very thin layers, have also been studied by deposition or growth onto metal substrates.

Biological material can, in principle, be imaged with STM. However, the normally poor electrical conductivity of these materials, and the need to adsorb them on flat conductive substrates, impose severe constrictions. As a result, today these materials are preferentially studied with AFM techniques.

Other applications of STM include the study of epitaxial growth of metals and oxides on metals and the mapping of superconducting and normal-conducting areas in superconductors by using a bias voltage above or below the superconducting gap (Hess et al., 1989; Hess et al., 1990a,b). Recently, there has been increasing use of STM as a microfabrication (Quate, 1989) or atomic manipulation tool (Eigler and Schweizer, 1990; Stipe et al., 1997; Bartels et al., 1998).

METHOD AUTOMATION

A major component of any STM instrument is the software for data acquisition and analysis. With the help of such software, color-coded, picture-like graphic displays of the images can be produced. Top views, as well as three-dimensional representations, including artificial light shading, are commonly used to enhance contrast, as in the example of Figure 3. The selection of the color or gray-scale mapping is arbitrary and left to the subjective taste of the operator. Image processing features (e.g., smoothing, filtering, and flattening) are common, as well as analysis capabilities for extracting cursor profiles along selected paths over the image. Other features included in most software packages are Fourier analysis and height histograms, among others. Today's sophisticated programs have rendered the task of analysis and data display relatively easy, and have made possible the beautiful images of atomic-surface landscapes (see the examples in Figure 3) that are currently produced in many STM laboratories.

Complete automation of an STM experiment is not common for either commercial or home-made instruments. The nature of the STM experiment itself is not easily conducive to automation. This is because the atomic structure of the surfaces under study is a delicate function of preparation and environmental conditions. Exceptions are the most chemically inert surfaces such as graphite, MoS_2, and others of limited intrinsic interest, except

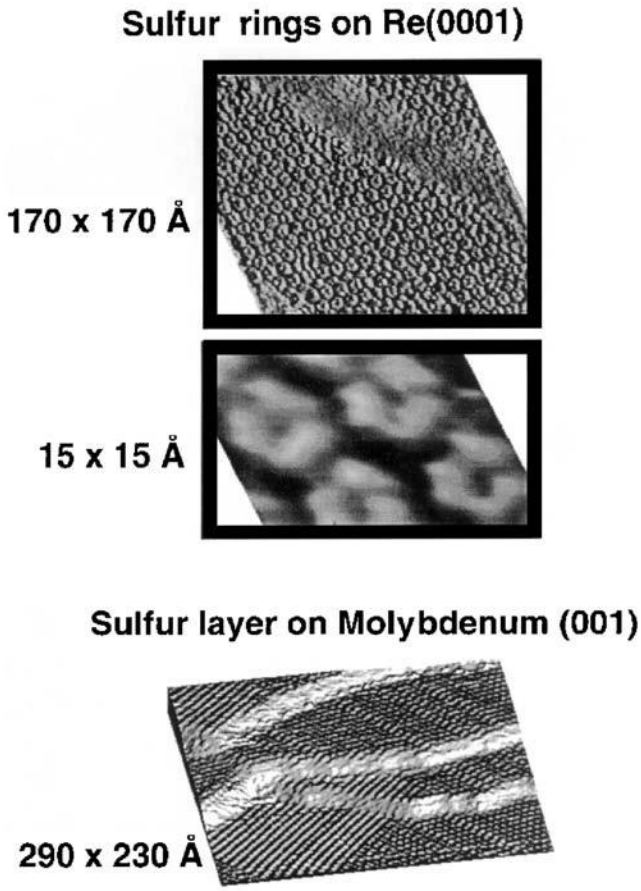


Figure 3. Examples of STM images obtained in vacuum. The top image was obtained on a Re(0001) surface covered by a saturation layer of sulfur. Sulfur is seen forming nearly hexagonal rings. Notice the blurred area corresponding to a monoatomic step. The radius of the tip determines the resolution in this case. The middle image shows a magnified view of the atomic arrangement. The bottom image was obtained on a Mo(001) surface, again with a saturation layer of sulfur. Sulfur atoms are seen forming two domains at a 90°angle. Steps 2-, 3- and 4-atoms high separate the terraces.

when used as substrates for other adsorbed material. Thus, it is not possible in general to simply insert a sample into a sample holder and have the instrument position the sample and acquire atomically resolved images. The approach of the tip from a large distance (centimeters or millimeters) down to a few angstroms from the surface is risky and often accompanied by some instability when nearing the "in range" or tunneling position that leads to atom exchanges and tip or surface modification. STM is thus an operator-intensive technique, as it requires considerable experience and exercise of scientific judgment for its successful realization.

DATA ANALYSIS AND INITIAL INTERPRETATION

The fundamental question in image interpretation is, of course, the relation between the measured topography, i.e., the $z(x,y)$ profile and the actual "height" or position of the atomic features. Contrast and symmetry are largely determined by the bonding orbitals of the atoms or molecules being imaged, and by the chemical nature of the tip termination. In its simplest interpretation, the $z(x,y)$ profile corresponds to the electronic corrugation of the surface at a constant integrated density of states between the Fermi levels of the tip and sample (see diagram in Figure 1). The latent capability of the STM to provide both topographic and electronic maps of the surface cannot therefore be fully realized without a theory of the tunneling process.

While an exact theory does not yet exist, several approximations have been useful in interpreting the images. The first one came from Tersoff and Hamann (1985), who assumed no interaction between the tip and surface. The tunneling current in this model, for a tip with electron states of s symmetry, is proportional to the density of states near the Fermi edge at the position of the tip center (considered to be terminated in a sphere). Other theoretical treatments have also been developed and can be found in recent reviews (Güntherodt and Wiesendanger, 1993). A recent one, called electron scattering quantum chemistry (ESQC), deserves special mention because of its relative simplicity and good results (Sautet and Joachim, 1988; Cerda et al., 1997). It is based on the combination of an exact treatment of the electron transport through the tunneling gap with approximate quantum chemistry methods (e.g., extended Hückel) to calculate the Hamiltonian matrix elements. It does not require the neglect of tip-surface interactions, and the structure of the tip is explicitly taken into account in the calculation of the image. It has been used in chemisorbed systems with remarkable success.

The strong electronic contribution to the image contrast sometimes gives rise to counterintuitive observations. For example, O atoms are usually imaged as holes (lower

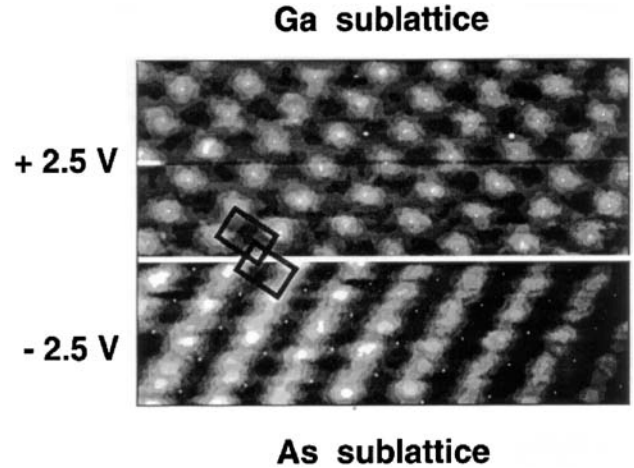


Figure 4. STM image of the cleavage face of an n-type GaAs sample. The bias applied to the sample was switched from $+2.5$ to -2.5 V near the center of the image. At positive bias, electrons tunnel from the tip states to empty states of the GaAs. These states are due to empty Ga dangling bonds, which have a lobe localized nearly atop the Ga atoms. At negative bias, electrons leave As dangling bonds, which are also localized near the As atoms. The two images represent the sublattices of these two atoms, which are offset in the manner indicated by the two boxes representing the surface unit cell.

tunneling probability) rather than protrusions (Kopatzki and Behm, 1991; Wintterlin et al., 1988; Sautet, 1997). Interference, when the tip overlaps several adsorbate atoms, can give rise to strong effects. These effects are manifested sometimes in a contrast distribution inside the unit cell that does not reflect the number of maxima expected from the atomic composition. For example, S on Mo(001) forms a c(2×4) structure, with 3 S atoms in the unit cell and yet, only one maximum is observed in the images (see Fig. 3), giving the apparent impression that only one S atom is present (Sautet et al., 1996).

Electronic effects in the STM images are particularly strong for semiconductor surfaces, due to the existence in these materials of energy gaps and surface states close to the Fermi level that are filled or empty. This is the case of GaAs(110), for example, where a filled dangling bond located on the As atom is imaged at negative sample bias while an empty dangling bond, located at the Ga atom position, is imaged at the opposite polarity (Feenstra and Stroscio, 1987). This is shown in Figure 4.

SAMPLE PREPARATION

The preparation of samples for STM is not different from that in other microscopies. In UHV-STM (ultrahigh vacuum STM), samples are prepared following well established surface-science methods. These include such techniques as sputtering, annealing, characterization by Auger electron spectroscopy (AES; see AUGER ELECTRON SPECTROSCOPY), or low-energy electron diffraction (LEED; see LOW-ENERGY ELECTRON DIFFRACTION), whenever appropriate. Air or high-pressure environments require special procedures, sometimes in combination with sample preparation in environmental chambers. In electrochemical cells equipped with STM, cleaning is part of the oxidation-reduction cycles and can be tested by well-established electrochemical voltammograms (see SCANNING ELECTROCHEMICAL MICROSCOPY).

It is the preparation of tips that is a distinct procedure of STM. Ideally single-atom-terminated tips are desired. However, short of performing field ion microscopy examination prior to or after imaging, no technique or procedure has been developed that meets universal standards of characterization. Tip preparation is thus, regrettably, a technique that is still largely empirical and varies from laboratory to laboratory.

The most widely used tip materials are W and Pt. The first is easily prepared using electrochemical etching methods with alkaline solutions (KOH or NaOH) to thin, cut, and sharpen a wire. Tips with radii from a few tens to a few hundreds of angstroms can thus be produced. If the wire is a single crystal of (111) orientation, single-atom-terminated tips of defined structure can, in principle, be prepared (Kuk, 1992). Pt alloyed with Ir or Rh is also a popular tip material. Such Pt alloys might be preferable in applications with reactive atmospheres and in air, where W tips can oxidize and more easily produce insulating layers. Pt tips can be prepared by electrochemical etching using molten ($\sim$300°C) mixtures of 3:1 $NaNO_3$:NaCl as the electrolyte bath. The sharpened tips usually are subjected to additional conditioning in vacuum, through heating by resistive or electron-beam methods and/or electron field emission on a sacrificial substrate. The purpose of these final treatments is to remove contaminant layers that can diffuse to the tip apex during imaging.

Unfortunately, even these procedures are no guarantee of a "good" tip, which is defined as one that produces stable, well resolved images. Deterioration and accidental changes occur by brief contacts (induced by vibrational noise) with the sample, by capture of contaminant atoms or molecules that form unstable bonds with the tip, and by loss of material. The cleaner and better controlled the sample and environment, the higher is the success rate of tip preparation and utilization. Often, lost performance can be restored by application of voltage pulses to the tip (>1 to 2 V), or by intentional contact with the sample. In the end, as previously pointed out, tip preparation is still an empirical procedure in most STM operations, and performance is decided in a somewhat subjective manner by the operator. For example, tips may be considered "good" and one-atom terminated when sharp and stable images with symmetric "atoms" are obtained. Satisfaction of that criterion, however, does not guarantee that "good" images acquired with different single-atom-terminated tips will have reproducible heights of the atoms, with z variations of a few tenths of an angstrom possible.

PROBLEMS

The immediate signature of a defective STM is the observation of an unstable tunneling gap. This is visible in the high noise level in the z height or the tunneling current signals. The cause may vary from the trivial—such as broken wires, cracked piezos, depolarization, or poor mechanical isolation between head and/or sample and the rest of the apparatus—to the more subtle and problematic, related to uncontrolled chemical processes occurring at the tip or sample. The tunneling gap, which must be stable within a small fraction of an angstrom, is subject to several instabilities. These instabilities can excite mechanical resonances in the head and sample support, which in turn can introduce a phase change in the feedback response. When this change becomes larger than 90°, positive feedback occurs and the tip oscillation is then driven instead of suppressed by the electronic control.

A chemically unstable gap can occur as a result of the continuous transfer of atoms or molecules at the surface or at the tip apex in dirty samples or tips. Sample and tip cleaning, as well as improvement of the environmental conditions, are then in order.

An excellent test of gap stability and a good troubleshooting exercise that should be common practice with students and technicians using STM is the analysis of the harmonic content of the current signal. This analysis gives a good deal of information about the sources and origin of the noise. It is performed by the application of fast Fourier transform (FFT) techniques, which are (or should be) a common feature of most data acquisition and software systems. The feedback control signal (usually the z height, in the form of a voltage applied to the z piezo scanner) can also be analyzed in this way, but it responds only within the frequency range of the feedback bandwidth, which is

rather small. Thus, at low gain, the z response of an unstable gap has the appearance of a saw tooth—the large error signal charges the feedback capacitor in the integrator at a nearly linear rate until saturation. At high gain, the feedback control–gap circuit can enter into resonance and cause the tip to oscillate against the surface. If the frequency of the offending noise is within the feedback bandwidth, the gain can be adjusted to suppress the noise in either z or I, but not both, since this does not solve the problem.

The best way to analyze the noise is to look at the FFT of the tunnel current, with the feedback at low gain or disabled. A clean, stable and well behaved gap should show no peaks in the crucial region of 1 to 500 Hz, since this is the region covered by the feedback response time and is also the region of typical scanning frequencies. An example of FFT analysis is shown in Figure 5, obtained with a beetle-type STM head developed by Besocke and Wolf (Frohn et al., 1989). This head consists of three piezo tubes attached perpendicularly to a disk, forming an equilateral triangle. The free end of each of these piezos moves when voltages are applied, allowing it to displace and rotate. A fourth piezo tube, at the center of the disk, supports the tip and can be used for scanning. The top curve corresponds to a noisy gap, with a z noise of a few angstroms, and is due to a loose sample holder. Loose parts tend to enhance the noise in the low-frequency region, below a few tens of hertz. Other poor conditions can derive from such factors as wires touching the chamber or a cracked piezo. The bottom curve shows the FFT of the same head after fastening all the parts that were loose. Peaks at the frequency of the power line (60 Hz or 50 Hz) are due mostly to ground loops and can be eliminated. The FFT spectrum will also show high frequency peaks due to the natural resonances of the STM scanner head, which should be greater than 1 kHz (the higher the better.) In the example of Figure 5, they are above 4.5 kHz. There are several studies published about this subject that the reader can consult for additional details (Behler et al., 1997b).

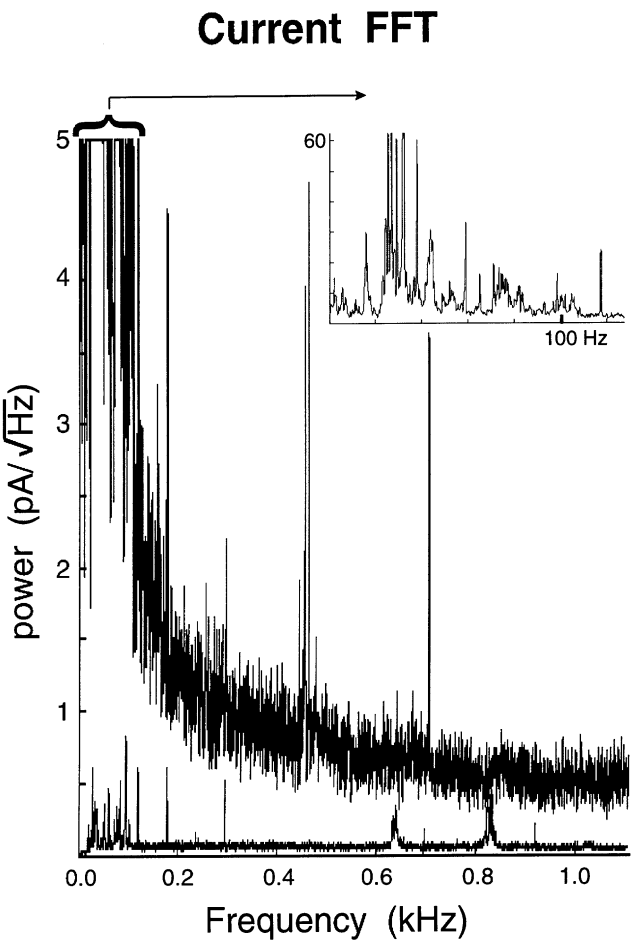


Figure 5. Example of Fourier transform analysis of the tunneling current for a beetle-type STM tip. The top curve was obtained in a highly noisy situation. In this case, the noise was due to the loose condition of the sample holder. Low-frequency noise is dominant, in the region 0 to 50 Hz (see top right inset). This noise couples strongly with the scanning rate at similar frequencies. Rattling of the entire head, which is simply resting on the sample holder in this design, is also excited and produces the peaks at ~500 and ~700 Hz. The spectrum below this was obtained after tightly securing the sample holder. The overall noise decreased by almost 3 orders of magnitude.

ACKNOWLEDGMENTS

The author acknowledges the support and help of various members of the STM group at LBNL, especially M. Rose and D.F. Ogletree for discussions and the data used in this part. This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Science Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

LITERATURE CITED

Bartels, L., Meyer, G., Rieder, K.-H., Velic, D., et al. 1998. Dynamics of electron-induced manipulations of individual CO molecules on Cu(111). *Phys. Rev. Lett.* 80:2004–2007.

Behler, S., Rose, M. K., Dunphy, J. C., Olgetree, D. F., Salmeron, M., and Chapelier, C. 1997a. A scanning tunneling microscope with continuous flow cryostat sample cooling. *Rev. Sci. Instrum.* 68:2479–2485.

Behler, S., Rose, M. K., Ogletree, D. F., and Salmeron, M. 1997b. Method to characterize the vibrational response of a beetle type scanning tunneling microscope. *Rev. Sci. Instrum.* 68:124–128.

Castell, M. R., Muggelberg, C., Briggs, G. A. D., and Goddard, D. T. 1996. Scanning tunneling microscopy of the UO_2/(111) surface. *J. Vac. Sci. Technol. B* 14:1173–1175.

Cerda, J., Van Hove, M. A., Sautet, P. and Salmeron, M. 1997. Efficient method for the simulation of STM images. I. Generalized Green-function formalism. *Phys. Rev. B* 56:15885–15889.

Curtis, R., Mitsui, T., and Ganz, E. 1997. An ultrahigh vacuum high speed scanning tunneling microscope. *Rev. Sci. Instrum.* 68:2790–2796.

Dunphy, J. C., Sautet, P., Ogletree, D. F., Daboussi, O., and Salmeron, M. 1993. Scanning-tunneling-microscopy study of the surface diffusion of sulfur on Re(0001). *Phys. Rev. B* 47:2320–2328.

Eigler, D. M. and Schweizer, E. K. 1990. Positioning single atoms with a scanning tunneling microscope. *Nature* 344:524–526.

Feenstra, R. M. and Stroscio, J. A. 1987. Tunneling spectroscopy of the GaAs(110) surface. *J. Vac. Sci. Technol. B* 5:923–929.

Feenstra, R. M., Stroscio, J. A., and Fein, A. P. 1987. Tunneling spectroscopy of the Si(111)2*1 surface. *Surf. Sci.* 181:295–306.

Frohn, J., Wolf, J. F., Besocke, K. H., and Teske, M. 1989. Coarse tip distance adjustment and positioner for a scanning tunneling microscope. *Rev. Sci. Instrum.* 60:1200.

Giessibl, F. J. 1995. Atomic resolution of the silicon (111)-(7*7) surface by atomic force microscopy. *Science* 267:68–71.

Güntherodt, H.-J. and Wiesendanger, R. (eds.).. 1993. Scanning Tunneling Microscopy, Vol III. Springer-Verlag, Heidelberg.

Hamers, R. J. 1992. STM on semiconductors. *In* Scanning Tunneling Microscopy, Vol. I (H.-J. Güntherodt and R. Wiesendanger, eds.).. pp. 83–129. Springer-Verlag, Heidelberg.

Heller, E. J., Crommie, M. F., Lutz, C. P., and Eigler, D. M. 1994. Scattering and absorption of surface electron waves in quantum corrals. *Nature* 369:464–466.

Hess, H. F., Robinson, R. B., Dynes, R. C., Valles, J. M., and Waszczak, J. V. 1989. Scanning-tunneling-microscope observation of the Abrikosov flux lattice and the density of states near and inside a fluxoid. *Phys. Rev. Lett.* 62:214–216.

Hess, H. F., Robinson, R. B., Dynes, R. C., Valles, J. M., and Waszczak, J. V. 1990a. Spectroscopic and spatial characterization of superconducting vortex core states with a scanning tunneling microscope. *J. Vac. Sci. Technol. A* 8:450–454.

Hess, H. F., Robinson, R. B., and Waszczak, J. V. 1990b. Vortex-core structure observed with a scanning tunneling microscope. *Phys. Rev. Lett.* 64:2711–2714.

Jansen, R., van Kempen, H., and Wolf, R. M. 1996. Scanning tunneling microscopy and spectroscopy on thin Fe_3O_4(110) films on MgO. *J. Vac. Sci. Technol. B* 14:1173–1175.

Kitamura, S. and Iwatsuki, M. 1995. Observation of 7*7 reconstructed structure on the silicon (111) surface using ultrahigh vacuum noncontact atomic force microscopy. *Jpn J. Appl. Phys.* 34:L145–L148.

Kopatzki, E. and Behm, R. J. 1991. STM imaging and local order of oxygen adlayers on Ni(100). *Surf. Sci.* 245:255–262.

Kuipers, L., Hoogeman, M. S., and Frenken, J. W. M. 1993. Step dynamics on Au(110) studied with a high-temperature, high-speed scanning tunneling microscope. *Phys. Rev. Lett.* 71:3517–3520.

Kuipers, L., Loos, R. W. N., Neerings, H., ter Horst, J. et al. 1995. Design and performance of a high-temperature, high-speed scanning tunneling microscope. *Rev. Sci. Instrum.* 66:4557–4565.

Kuk, Y. 1992. STM on metals. *In* Scanning Tunneling Microscopy, Vol. I (H.-J. Güntherodt and R. Wiesendanger, eds.).. pp.17–37. Springer-Verlag, Berlin.

McIntyre, B. J., Salmeron, M., and Somorjai, G. A. 1993. A variable pressure/temperature scanning tunneling microscope for surface science and catalysis studies. *Rev. Sci. Instrum.* 64:687–691.

Murray, P. W., Condon, N. G., and Thornton, G. 1995. Na absorption sites on TiO_2/(110)-1*2 and its 2*2 superlattice. *Surf. Sci.* 323:L281–L286.

Ogletree, D. F. and Salmeron, M. 1990. Scanning tunneling microscopy and the atomic structure of solid surfaces. *Prog. Solid State Chem.* 20:235–303.

Onishi, H. and Iwasawa, Y. 1994. Reconstruction of TiO_2/(110) surface: STM study with atomic-scale resolution. *Surf. Sci.* 313:L783–L789.

Quate, C. F. 1989. Surface modification with the STM and the AFM. *In* Scanning Tunneling Microscopy and Related Methods (R. J. Behm, N. Garcia, and H. Rohrer, eds.).. p. 281. Kluwer Academic Publishers, Norwell, Mass.

Rohrer, G., Henrich, V. E., and Bonnell, D. A. 1990. Structure of the reduced TiO_2/(110) surface determined by scanning tunneling microscopy. *Science* 250:1239–1241.

Sander, M. and Engel, T. 1994. Atomic level structure of TiO_2[110] as a function of surface oxygen coverage. *Surf. Sci.* 302:L263–268.

Sautet, P. 1997. Atomic absorbate identification with the STM: A theoretical approach, *Surf. Sci.* 374:406–417.

Sautet, P., Dunphy, J. C., and Salmeron, M. 1996. The Origin of STM contrast differences for inequivalent S atoms on a Mo(100) surface. *Surf. Sci.* 364:335–344.

Sautet, P. and Joachim, C. 1988. Electronic transmission coefficient for the single-impurity problem in the scattering-matrix approach. *Phys. Rev. B* 38:12238–12247.

Stipe, B. C. Rezaei, M. A., Ho, W., Gao, S., Persson, M., and Lundqvist, B. I. 1997. Single-molecule dissociation by tunneling electrons. *Phys. Rev. Lett.* 78:4410–4413.

Sugawara, Y., Ohta, M., Ueyama, H., and Morita, S. 1995. Defect motion on an InP(110) surface observed with noncontact atomic force microscopy. *Science* 270:1646–1648.

Tersoff, J. and Hamann, D. R. 1985. Theory of the scanning tunneling microscope. *Phys. Rev. B* 31:805–813.

Wiesendanger, R. 1994. Scanning Probe Microscopy and Spectroscopy: Methods and Applications. Cambridge University Press, New York.

Wintterlin, J., Brune, H., Hofer, H., and Behm, R. J. 1988. Atomic scale characterization of oxygen absorbates on Al(111) by scanning tunneling microscopy. *Appl. Phys. A* 47:99–102.

Zeppenfeld, P., Lutz, C. P., and Eigler, D. M. 1992. Manipulating atoms and molecules with a scanning tunneling microscope. *Ultramicroscopy* 42–44:128–133.

KEY REFERENCES

Cerda et al. 1997. See above.

A modern implementation of the ESQC is described.

Güntherodt, H.-J. and Wiesendanger, R. 1994. Scanning Tunneling Microscopy, Vols. I and II. Springer-Verlag, Heidelberg.

Contains an excellent collection of review papers covering most aspects of STM.

Wiesdanger, 1994. See above.

Contains a good description of the state of the STM field at the time of its publication.

INTERNET RESOURCES

http://www.omicron-instruments.com

http://www.RHK-TECH.com

Web sites of two STM manufacturers, describing scanning microscopes and containing links.

MIQUEL SALMERON
Lawrence Berkeley National Laboratory
Berkeley, California

LOW-ENERGY ELECTRON DIFFRACTION

INTRODUCTION

Low-energy electron diffraction (LEED) is a powerful method for determining the geometric structure of solid surfaces. The phenomenon was first observed by Davisson and Germer in 1927 and provided the earliest direct experimental proof of the wavelike properties of electrons. LEED has since evolved into one of the most powerful and widespread tools of the surface scientist. It is similar to x-ray diffraction (XRD) in the type of information that it provides, and XRD will be frequently referred to for analogy throughout this unit. The most obvious difference is that a beam of electrons, rather than x rays, is used. Since electrons have a mean free path (or attenuation length) measured in angstroms (Å) as opposed to microns (μ) for x rays, LEED is particularly sensitive to surface geometry.

LEED is a structural technique that can provide essentially two levels of information. In both cases, it is most commonly used to determine the structure of a solid surface when the bulk structure of the material is already known by other means (e.g., XRD). It is possible, and indeed straightforwardly simple, to use LEED to determine both the absolute dimensions of the surface unit cell and the unit cell symmetry. One can easily deduce that, for example, a surface has a two-dimensional (2D) unit cell twice as large as that of the bulk; this often allows a reasonable guess at the true structure. Sometimes it is sufficient to know that a particular phase is present without knowing the details of the structure. For example, the (2×4) reconstruction of GaAs (nomenclature explained below) has been found to be optimal for device growth (Farrow, 1995). Determination of the exact atomic positions is more difficult, but quantitative experiments can elucidate this second level of information. Sophisticated calculations, generally run on a workstation, can provide atomic coordinates with a typical precision of ± 0.05 Å, which is generally more than adequate to determine the adsorption site of a molecule or the atomic positions in a reconstructed surface.

Many different surfaces can be examined by LEED, and their terminology must be understood. We assume familiarity with the basic crystal structures [face-centered cubic (fcc), zincblende, etc.]. A particular crystal face is specified by its Miller indices. These are of the form (hkl), and the ideal bulk termination can be determined in the following way: place one corner of the unit cell at the origin of a Cartesian coordinate system and construct a plane that intersects the axes at the reciprocal of the respective Miller index (see SYMMETRY IN CRYSTALLOGRAPHY). In the general case, this plane would pass through the points $(1/h, 0, 0)$, $(0, 1/k, 0)$, and $(0, 0, 1/l)$. A Miller index of zero indicates that the plane is parallel to that axis. Further examples can be found in any solid-state textbook (e.g., Ashcroft and Mermin, 1976; Kittel, 1986). It is conceptually useful to regard a surface as being composed of stacked planes of atoms, each parallel to the topmost surface plane (see SYMMETRY IN CRYSTALLOGRAPHY).

The structure of an actual surface is usually very different from an ideal termination of the bulk. The low-Miller-index surfaces—(100), (110), (111)—are relatively flat, but higher-index surfaces generally have a large number of steps and often kinks as well. Adatoms and adsorbed molecules also appear in many systems. Figure 1 illustrates some of these features. LEED is sensitive to any such structures that recur at regular intervals, that is, with measurable periodicity.

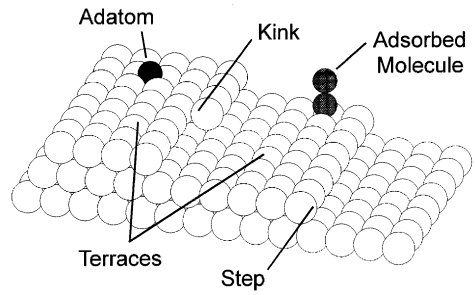


Figure 1. A model of the most common surface features.

LEED is a mature area of surface science and it is useful to consider the state of the art in this field. A great deal of work has been done on clean metal and semiconductor surfaces, and much of this is considered "done"; current work in this area consists mostly of refinements. Most low-index clean metal surfaces are found to merely relax the spacing between the topmost layers. Higher-index metal surfaces tend to be less closely packed and more prone to reconstruction. Clean semiconductor surfaces almost always reconstruct. Adsorbed atoms and molecules are sometimes found to lift reconstructions, and they often modify the dimensions and orientation of the surface unit cell themselves. Several atoms and simple molecules (H, O, S, CO, C_2H_4) have been investigated on low-Miller-index surfaces of semiconductors (Si, GaAs) and most transition metals. An important direction of current research is the growth of ordered crystalline epitaxial films for a variety of purposes. They serve as models for catalytic systems (Boffa et al., 1995), model magnetic structures, or avenues for studying otherwise intractable systems (Roberts et al; 1998), to name a few examples. Other recent advances seek to extend the power of the technique to systems with disorder in geometry (Döll et al., 1997; Starke et al., 1996) or composition (Gauthier et al., 1985; Crampin et al., 1991). Because of the time-consuming trial-and-error computations involved in a structural determination, there is also interest in developing a LEED-based holographic technique for rapid adsorption site determination (Saldin, 1997).

Complementary and Related Techniques

Information similar or complementary to that provided by LEED can be obtained by other techniques. For example, XRD provides the same information as LEED, but without surface selectivity. Similar surface-specific diffraction information can be obtained from such methods as photoelectron diffraction (Woodruff and Bradshaw, 1994; Fadley et al., 1995), small-angle x-ray diffraction (SAXD; SURFACE X-RAY DIFFRACTION), x-ray standing waves (XSW;

Woodruff and Delchar, 1994, pp. 97-104), Rutherford backscattering spectroscopy (RBS; Van Der Veen, 1985), and reflection high-energy electron diffraction (RHEED). The great advantage of LEED over these other diffraction methods is its ease of use and the simplicity of the instrumentation required. X-ray standing wave experiments require a monochromatic x-ray source of a brilliance only to be found at a synchrotron, and small-angle x-ray scattering requires very precise control over sample alignment. RBS needs a mega-electron-volt (MeV) Van de Graaff generator or an ion implantation machine, equipment only found at specialized facilities, to create the needed beam of high-energy ions (50 keV to 2 MeV H^+ or He^+) to probe the surface. In contrast, the majority of all ultrahigh-vacuum (UHV) surface analysis chambers are typically fitted with a relatively simple set of LEED optics, and sample alignment of $\sim 0.5^\circ$ precision is easily attained with standard equipment. RHEED is relatively simple to implement, but it suffers from the disadvantage that full structural determinations are far from routine, and it is generally applicable only in a qualitative mode. Nevertheless, it finds wide application particularly in the study of semiconductor growth because it allows for easy *in situ* study of MBE growth processes. Under some circumstances, RBS, also known as high-energy ion scattering (HEIS), can be used to infer surface structure. This is generally done by channeling ions between the aligned rows of atoms along major crystal axes in the substrate lattice and looking for scattering from displaced surface atoms that block those channels. By channeling along several axes, displaced surface and often second larger atoms can be located by triangulation. This has the advantage of being a real-space technique, rather than a diffraction-based reciprocal-space technique and can distinguish foreign adatoms under available conditions.

Methods that provide information complementary to LEED are of greater interest. The chief weakness of LEED is shared with all other diffraction techniques: it requires single-crystal samples and selectively views only the well-ordered portions of the sample. Neither disordered regions nor intimate details of atomic motion can generally be examined, except in an average sense. For a complete structural picture, the combination of LEED with microscopy techniques is especially powerful; scanning tunneling microscopy (STM; SCANNING TUNNELING MICROSCOPY) and atomic force microscopy (AFM) are especially useful in this regard. This has been addressed in a recent review (Van Hove et al., 1997). Another useful technique is SEXAFS, which is merely surface extended x-ray absorption fine structure (EXAFS) analysis (Koningsberger and Prins, 1988). This technique is a local probe of structure and provides a radial distribution function rather than crystallographic data. It is possible to use SEXAFS to determine structures just as in LEED, but its real utility is its applicability to disordered systems where diffraction cannot be done. SEXAFS also suffers from reliance on synchrotron light sources. Electron microscopy (EM; SCANNING ELECTRON MICROSCOPY, TRANSMISSION ELECTRON MICROSCOPY, SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING) is useful for investigating structures on larger scales, and sophisticated EM instruments can provide atomic resolution as well. Unlike LEED, neither SEXAFS nor EM requires a flat sample in order to work. However, it should be stressed that an inherent problem in all microscopy techniques is the question of representative sampling. LEED rapidly examines a sample of macroscopic proportions ($\sim$1 to 10 mm^2, the spot size of the electron beam).

Finally, LEED is a purely structural technique. Information about bond energies or strengths is simply not available, except insofar as correlations can be drawn with changes in bond lengths. In general, dynamic information is also not available. Temperature-dependent measurements can determine the mean amplitudes of isotropic thermal vibrations, and some efforts have been made to investigate anisotropic vibrations, but this work is in its early stages (Löffler et al., 1995). Certainly, none of this can compete with, for example, femtosecond laser spectroscopic methods for true molecular dynamic information.

PRINCIPLES OF THE METHOD

As the name implies, LEED operates on the basis of diffraction of low-energy electrons, typically in the range of 20 to 400 eV (1 eV = 1.602×10^{-19} J) from solid surfaces. Electrons in this energy range have wavelengths close to an angstrom and thus diffract from crystalline arrays of atoms. Electrons in this energy range are also very surface sensitive. Electrons can act as quantum mechanical waves with a De Broglie wavelength given by $\lambda(\text{Å}) = \sim \sqrt{150.4/E(\text{eV})}$, so energies in the range 20 to 400 eV are typically used to diffract from crystalline solids with lattice constants of typically several angstroms ($1 = 10^{-10}$m). This technique is surface sensitive because electrons in this energy range have a short inelastic mean free path ($\sim$3 to 20 Å) in all known materials. Seah and Dench (1979) have fit an analytic expression to the available experimental data that is plotted in Figure 2. This "universal curve" can be qualitatively understood by recognizing that the dominant mechanism for inelastic scattering is loss of energy to a plasmon. At lower energies, insufficient

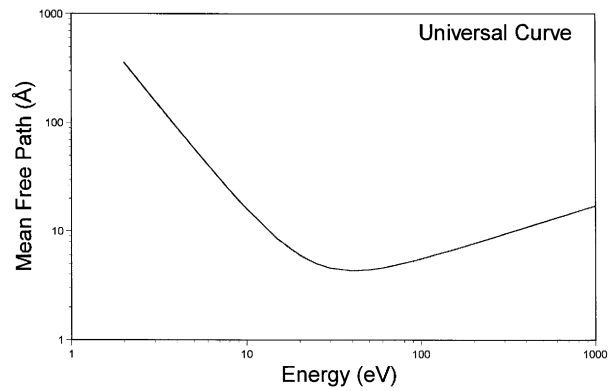


Figure 2. The "universal curve" of the mean free path of electrons in a solid, as a function of electron kinetic energy. (After Seah and Dench, 1979.) The mean free path represents a distance over which the flux of electrons drops to $1/e$ of its original value. The kinetic energy is that measured in vacuum.

energy exists to efficiently excite a plasmon, and at higher energies, the electron may well escape the solid before any excitation occurs. Once the electrons escape the solid, their mean free path in air requires pressures below $\sim 10^{-4}$ Torr for typical operating distances. Lower pressures are generally required for most experiments because of the need to keep samples clean; ultrahigh-vacuum conditions ($\leq 10^{-9}$ Torr) are generally recommended.

LEED is distinguished from more conventional diffraction techniques in that only the surface layers contribute to the diffraction pattern. This has consequences in the types of pattern that are seen. In conventional x-ray diffraction, for example, the real-space three-dimensional (3D) crystalline lattice generates a 3D lattice in reciprocal space (cf. X-RAY TECHNIQUES and SYMMETRY IN CRYSTALLOGRAPHY). The Laue condition for diffraction gives rise to the construction of an Ewald sphere that may or may not intersect another point on the reciprocal lattice. Put plainly, for arbitrary conditions of wavelength (or energy) and incident angle, diffraction will not necessarily occur from a 3D crystal. In practice, this means it is necessary to examine the reflection of x rays from a single crystal at many different angles in order to observe all possible diffraction spots. (Those familiar with typical experiments might notice that x-ray diffraction from a single crystal commonly is seen at all orientations; however, all spots are not seen at the same time. The common appearance of several spots under arbitrary conditions is usually due to deviations of the experiment from "ideal" conditions—finite divergence of the x-ray beam, polychromatic x-ray source.) Diffraction from a rigorously 2D crystal, however, guarantees that the diffraction condition will be met (above some critical energy). To see how this is so, examine the reciprocal lattice vectors, $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ of a real-space lattice with lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. The relation is

$$\mathbf{b}_1 = \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \bullet (\mathbf{a}_2 \times \mathbf{a}_3)} \qquad \mathbf{b}_2 = \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \bullet (\mathbf{a}_2 \times \mathbf{a}_3)}$$
$$\mathbf{b}_3 = \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \bullet (\mathbf{a}_2 \times \mathbf{a}_3)} \qquad (1)$$

The periodicity of a 2D lattice in the direction normal to the lattice plane can be considered infinite. Thus, let $\mathbf{a}_3 = d\mathbf{n}$, where $\mathbf{n}$ is a unit vector normal to the surface plane. As $d \to \infty$ (the limit of infinite interplanar separation), the magnitude of $\mathbf{b}_3$ approaches zero, resulting in continuous rods, rather than points, in reciprocal space. This means that, once the Ewald sphere is sufficiently large (for sufficiently short wavelength), diffraction will always be observed for arbitrary orientation of the incident angle. That is, one is guaranteed to always see spots on the LEED screen for a sufficiently large energy of the electron beam.

The assumption of a rigorously 2D lattice is somewhat incorrect as a model for real 2D diffraction. Real crystals do have periodicity in the direction normal to the surface, but that periodicity terminates abruptly at the surface. This termination manifests as the rods in reciprocal space becoming "lumpy," and diffraction in certain directions being favored. This is realized in the LEED pattern as variations of the intensities of spots with the kinetic energy of the electrons. It is precisely this variation that allows extraction of detailed crystallographic parameters. Certain spots in the pattern will occasionally vanish to zero intensity, but these are isolated instances of angle and beam energy, and the spot will otherwise be generally visible. It happens to turn out that there are almost always conditions under which all of the spots can be seen.

Qualitative Observations

There are essentially two levels of information that can be extracted from a LEED experiment: the qualitative and the quantitative. Qualitatively, a LEED experiment will determine the symmetry of a surface structure. Consider, e.g., Figure 3. Shown in part A is the diffraction pattern obtained from a square 2D lattice [perhaps the (100) plane of a face- or body-centered cubic (bcc) metal] along with the corresponding real-space structure. Part B shows the diffraction pattern due to a so-called (2×2) pattern of adsorbates on the substrate crystal (notation explained below). One such real-space structure is also illustrated in B; adsorbates have been arbitrarily placed atop the atoms.

Also shown in Figure 3 is the indexing usually applied to the diffracted beams. The specularly reflected beam is placed at the center of the coordinate system, and two of the beams with the lowest value of $|\boldsymbol{k}|$ are selected as the (10) and (01) beams. Other beams are labeled according to the coordinate system generated. Negative indices are generally indicated by a bar over the appropriate number. When an adsorbate or reconstruction exists and induces extra spots, the most common method of indexing assigns fractional values to the new spots.

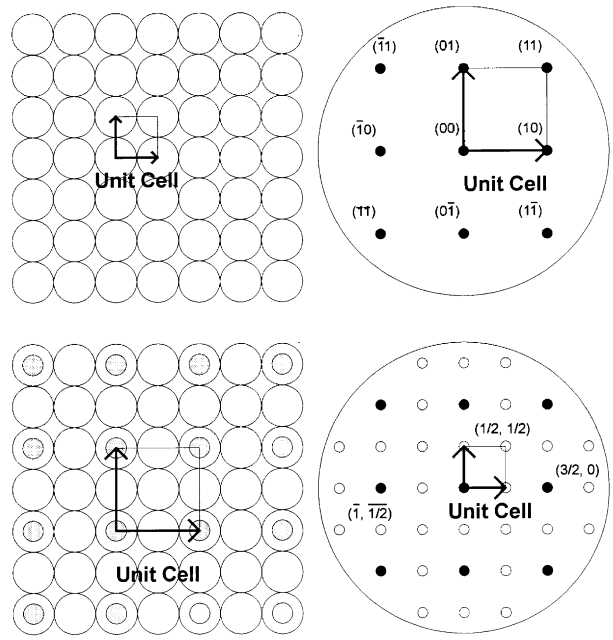


Figure 3. An illustration of the relationship between real and reciprocal space. In (**A**), we see a clean surface of square symmetry, with atoms represented as circles. In (**B**), we see the same surface with a p(2×2) unit cell. The translational symmetry has been lowered by an adsorbate arbitrarily placed atop an atom. Typical spot indexing is indicated by a label above the appropriate beam.

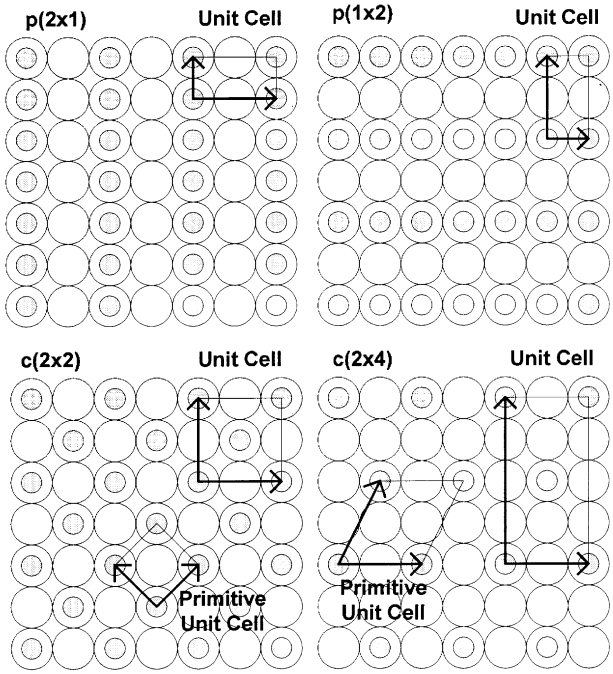


Figure 4. Some simple, single-domain real-space structures on a substrate of square symmetry. The Wood notation for each is given above the structure. These real-space models correspond to the diffraction patterns in Figure 5.

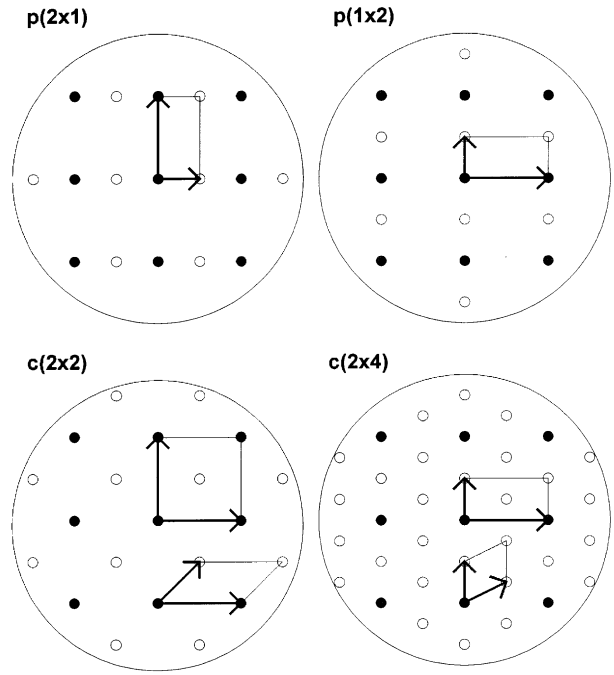


Figure 5. Some simple, single-domain reciprocal space structures on a substrate of square symmetry. These diffraction patterns correspond to the real-space models in Figure 4. Note that the $p(2 \times 1)$ and $p(1 \times 2)$ structures are not incorrectly reversed; the long dimension in real space is the short dimension in reciprocal space. The centered structures, $c(2 \times 2)$ and $c(2 \times 4)$, have two unit cells indicated. The lower one is the primitive unit cell, while the upper one is a larger cell that more obviously preserves the symmetry of the real-space unit cell.

The notation used above, the so-called Wood notation (after Wood et al., 1963), relates the periodicity of the overlayer to that of the substrate. It is commonly found that an overlayer adopts lattice vectors that are integral multiples of the substrate unit vectors. For example, in the (2×2) case above, the overlayer has a unit cell that is twice as large in both dimensions as the substrate unit cell. Figure 4 illustrates, again on a "square" lattice substrate, a $p(2 \times 1)$, a $p(1 \times 2)$, a $c(2 \times 2)$, and a $c(2 \times 4)$ structure. Their diffraction patterns are illustrated in Figure 5. Figures 6 and 7 illustrate the real-space lattice and diffraction patterns, respectively, for a $p(1 \times 2)$, $p(2 \times 2)$, $(\sqrt{3} \times \sqrt{3})R30^\circ$, and $c(2 \times 4)$ overlayer on a surface of hexagonal symmetry [e.g., the (111) plane of an fcc metal]. Notice that all of the "c" structures have an adsorbate at the center of the unit cell as well as at the corners. If one examines the square $c(2 \times 2)$ structure more closely, one can also construct a smaller unit cell that is square, contains only one adsorbate, but is rotated 45° with respect to the substrate. The sides of this square are $\sqrt{2}$ times longer than the dimensions of the substrate unit cell, so this cell can be described as $(\sqrt{2} \times \sqrt{2})R45^\circ$ as well.

More formally, Wood's notation can be described as $j[(b_1/a_1) \times (b_2/a_2)]R\theta$, where j is either p or c, a_n and b_n are substrate and overlayer unit cell lengths, respectively, and θ is the angle through which the cell must be rotated to produce the overlayer unit cell. The notation "$R\theta$" indicates that the unit cell of the substrate must be rotated by an angle θ to give unit cell vectors parallel to those of the overlayer. Typically, $R\theta$ is omitted when θ is zero. The preceding "p" signifies a primitive unit cell, while "c"

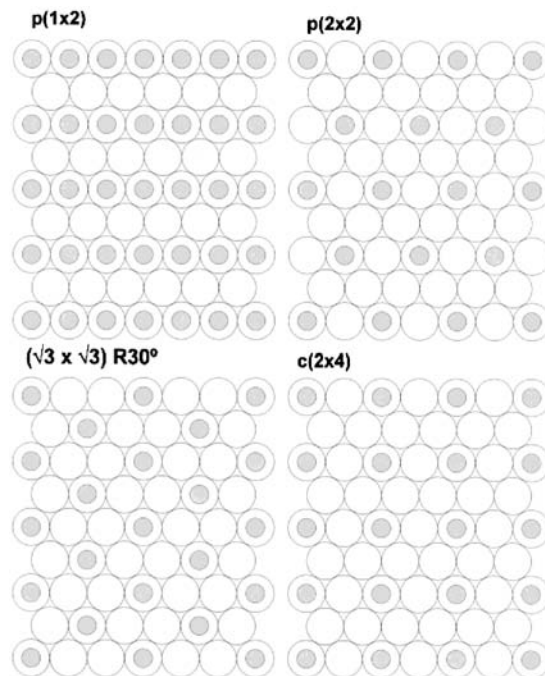


Figure 6. Some simple, single-domain real-space structures on a substrate of pseudo-hexagonal symmetry. The Wood notation for each is given above the structure. These real-space models correspond to the diffraction patterns in Figure 7.

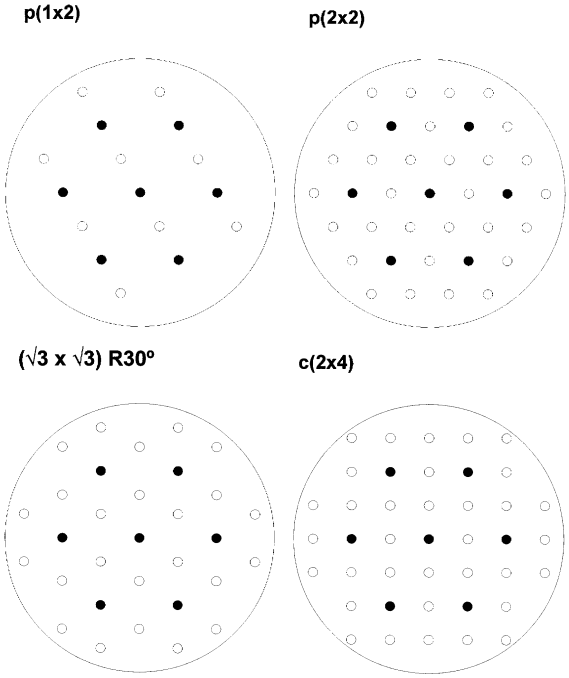


Figure 7. Some simple, single-domain reciprocal space structures on a substrate of square symmetry. These diffraction patterns correspond to the real-space models in Figure 6.

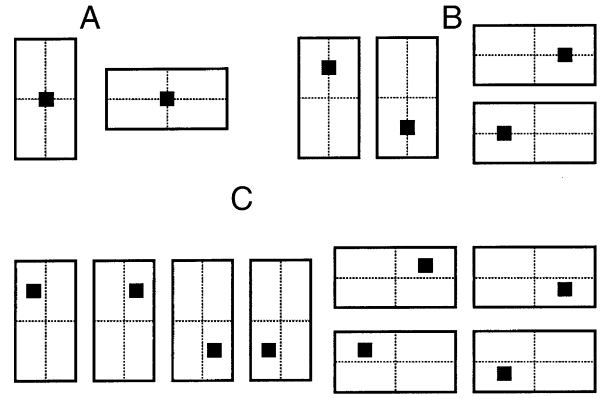


Figure 8. The various possible domains for a (1×2) unit cell on a substrate of square symmetry [e.g., (100) face of a bcc or fcc lattice]. Two domains are possible in (**A**), where the unit cell has the maximum possible symmetry, including two mirror planes and a 2-fold rotation axis. In (**B**), four possible unit cells have only one mirror plane. In (**C**), eight unit cells are possible with no internal symmetry. These unit cells can be distinguished by quantitative LEED structural analysis, but their diffraction patterns are qualitatively identical with that shown in the upper right of Figure 9.

represents a "centered" unit cell with a lattice point at the center of the cell identical to the ones at the corners. The p is sometimes omitted from descriptions of structures as well, although its omission does not always mean that p is implied. It should be apparent that in all cases, a centered cell can be decomposed into smaller primitive cells, but the c usage survives—if only because it is intuitively easier to recognize centered cells than their primitive counterparts.

It should likewise be apparent that structures are possible that cannot be described using Wood's simple notation. When the angle between the unit vectors of the overlayer is different from the angle between the unit vectors of the substrate, the two unit cells cannot be related by a simple expansion and/or rotation. In such cases, matrix notation must be used. The matrix used is the most general relation between the substrate unit vectors, $\mathbf{a}_1$ and $\mathbf{a}_2$, and the overlayer unit vectors, $\mathbf{b}_1$ and $\mathbf{b}_2$ (expressed in terms of a common basis):

$$\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{m}_{11} & \boldsymbol{m}_{12} \\ \boldsymbol{m}_{21} & \boldsymbol{m}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \qquad (2)$$

This has the advantage of always being applicable, but the disadvantage of being more cumbersome.

Another important point concerns the existence of multiple domains of a surface structure. In general, a particular overlayer geometry is energetically equivalent when operated on by one of the symmetry elements of the substrate. If the symmetry of the overlayer is lower than that of the substrate, several domains are possible. For example, on a bcc (100) face, two distinguishable domains of a (1×2) structure exist, which can be termed (1×2) and (2×1), as seen in Figure 4. The bcc (100) surface has a 4-fold rotation axis and four mirror planes, while the (1×2) cell has at most a 2-fold axis and two mirror planes. The importance of this is that at least two domains of the (1×2) will be present on any real sample. Figure 8 illustrates all of the possible domains of a (1×2) superstructure on a bcc (100) substrate. The fewer symmetry elements possessed by the unit cell, the more possible domains will be present. In cases of lower symmetry, the large number of domains will only be observed in the quantitative structure of the $I(E)$ spectra, and the qualitative appearance of the diffraction pattern is indistinguishable from a unit cell with the maximum possible symmetry. Most domains turn out to be larger than the coherence length (see below) but smaller than the typical beam spot size of ~1 mm. This means that all possible domains will be illuminated by the electron beam, and diffraction from all of them will appear on the LEED screen. As a result, the pattern observed from a low-symmetry structure on the LEED screen may not be a Bravais lattice, but a superposition of such lattices. Figure 9 illustrates the patterns typically seen from some common multidomain structures.

Quantitative Measurements

Quantitative crystallographic information is also accessible. Atomic positions can be determined to within several hundredths of an angstrom by modeling the intensities of the diffraction spots observed. These intensity versus beam energy curves are typically called $I(E)$ curves (as well as "*I-V* curves," referring to the intensity and electron beam voltage). In principle, the formalism is similar to

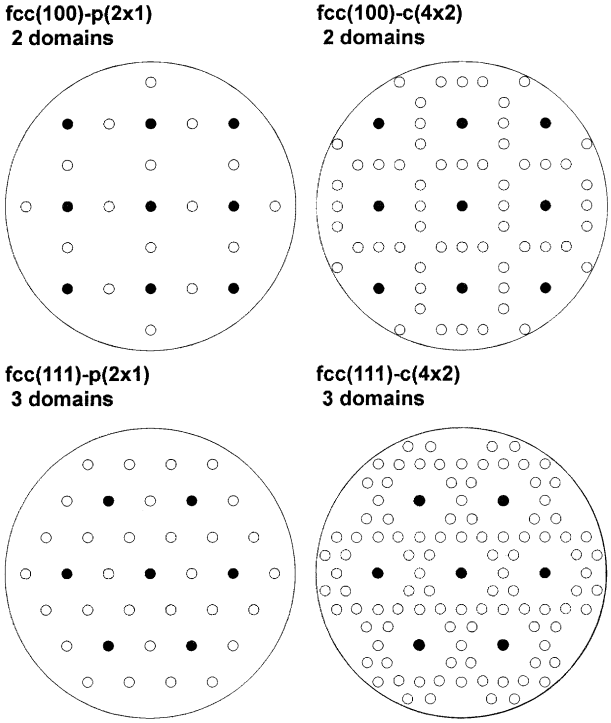


Figure 9. Diffraction patterns of some common multidomain structures. Note that the fcc(111)-p(2×1) structure is indistinguishable from the fcc(111)-p(2×2) structure in Figure 7. However, the fcc(100)-p(2×1) can be distinguished from the p(2×2) structure on the same substrate seen in Figure 3.

x-ray diffraction, but important differences exist owing primarily to the effect of multiple scattering. Because of the low cross-section for x-ray scattering, one may assume that an x ray scatters only once before exiting a crystal, but the cross-section for electron scattering approaches the geometric cross-section of an atom. As a result, multiple scattering is a common occurrence and its effects cannot be ignored. A practical result of this is that an x-ray diffraction pattern can be solved, in principle, by a direct Fourier transform of the pattern. However, a great deal of information is convolved in the process of multiple scattering, and the best that can be achieved is comparison of experimental data with a reference calculation. As an illustration, consider that the wave function of an electron undergoes a phase shift when scattering from an atom, modeled as a square well in Figure 10. This phase shift is indistinguishable from a phase shift due to the electron traversing a longer distance, and this ambiguity denies the possibility of directly determining a lattice spacing as one might do, for example, from x-ray diffraction data. Hence, peaks in the $I(E)$ curves cannot be fit by a simple Bragg law, and sophisticated multiple scattering calculations must be performed.

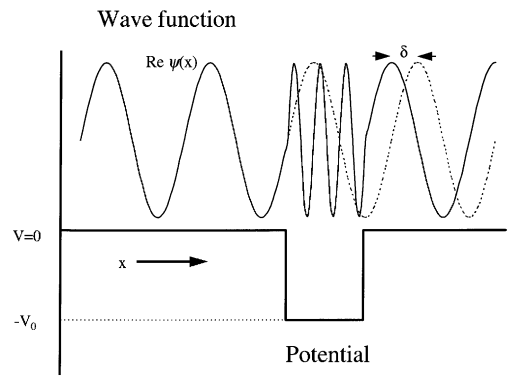


Figure 10. A simple model illustrating the phenomenon of a phase shift. A plane wave passing over a square well will change its wavelength in the region of the well. After exiting the well, its phase will have advanced by an amount indicated by δ.

PRACTICAL ASPECTS OF THE METHOD

Performing the actual LEED experiment is reasonably straightforward. It requires an ultrahigh-vacuum chamber fitted with a single-crystal sample, electron gun, display system, and provisions for intensity measurements. These will be discussed in turn.

The sample is usually a single crystal of some electrically conducting material. Doped semiconductors can also be used easily. The sample must usually be cleaned of surface contaminants prior to use, and one typically needs facilities such as an ion bombardment gun and sample heating capabilities to anneal out defects caused by ion bombardment. Sample cleanliness is usually monitored by such techniques as Auger electron spectroscopy (AES; AUGER ELECTRON SPECTROSCOPY) or x-ray photoelectron spectroscopy (XPS). Since thermal vibrations are known to diminish the intensity of all diffraction peaks, LEED experiments should be done with the sample as cold as is practical. Typically, liquid nitrogen cooling is used during the recording of the pattern, even if higher temperatures are required to prepare the desired surface structure. Another important consideration is that the diffraction pattern is typically viewed from behind the sample, resulting in a need to reduce the bulk of the manipulator as much as possible for maximum viewability (the advent of so-called "rear-view" optics has relaxed this requirement on some systems).

Another requirement for the sample manipulator is the ability to rotate the sample around a minimum of two perpendicular axes. This is necessary to achieve the normal incidence condition useful for quantitative work, although this is not strictly required for qualitative data. The reason for positioning the crystal normal to the incoming electron beam is 2-fold: (1) normal incidence is the easiest angle to specify exactly, so that high angular precision can be ensured; quantitative calculations require a precision of half a degree or better; and (2) normally, incident electron beams also maximize the symmetry of the resulting diffracted beams; making full use of symmetry is a great advantage when doing a quantitative analysis, as will be discussed later. Since finding normal incidence is a nontrivial task, two methods of checking for normal incidence are described briefly. Both depend on the fact that, for a normally incident beam of electrons, the intensities of the diffracted beams have the same symmetries as the crystal plane being investigated. For example, an

fcc(100) face has a 4-fold symmetry axis and four mirror planes, and an fcc(111) plane has a 3-fold symmetry axis and three mirror planes. The simplest method is useful only for highly symmetric surfaces—at normal incidence the intensities of symmetrically equivalent beams are equal. For example, four symmetrically equivalent spots in the diffraction pattern from an fcc(100) surface, say the (10) set of beams [e.g., as distinct from the (11) or (20)], should have equal intensity. If the left pair of spots were brighter than the right pair of spots, some sort of left/right angular adjustment would need to be made. The other method is to allow beams with identical values of $|\Delta\mathbf{k}_{\|}|$ to approach the edge of the screen, which is generally a cylindrically symmetric section of a sphere. These beams should all arrive at the edge of the screen at the same time, as the energy of the incident beam is varied.

The electron guns typically employed are widely used and only a few brief comments are warranted. Since LEED examines back-reflected electrons, the gun is usually mounted inside the display apparatus. As a result, only electrostatic deflection is used to reduce the effect of stray fields on the diffraction pattern. The filament is usually tungsten, although sometimes LaB_6 has been used for its low work function. One problem that can arise is that the light emitted from the hot filament can make the fluorescent screen of the display system difficult to see. For this reason, the filament is typically mounted off of the axis of the gun and the emitted electrons deflected onto the axis. Another consideration is that the beam current is not usually constant as the accelerating voltage is increased. This is important in quantitative measurements because increased beam current will result in increased intensity in the diffracted beams; this can artificially influence the position of peaks in the $I(E)$ curves. There are various ways to control this, but the simplest method is to directly measure the beam current as a function of energy and correct the $I(E)$ curves accordingly.

The display system must provide for measurement of the intensities of diffracted beams. The earliest methods for doing so employed a Faraday cup that could be moved to the proper position. It is currently much more popular to use a fluorescent screen held at high voltage (3 to 7 kV), which glows when struck by electrons with brightness proportional to current. A typical modern apparatus is shown in Figure 11. The hemispherical phosphorescent screen is preceded by a series (two, three, or four) of concentric hemispherical wire meshes or grids, with four grids being most common. The sample is ideally placed at the center of curvature of the instrument. The grid closest to the sample is generally held at ground/earth potential, as is the sample, so that diffracted beams propagate through a field-free region between the two. The second grid is held at a potential a few volts below that of the beam to energy select out all inelastically scattered electrons. In a four-grid system, the second and third grids are held at the same potential to improve the quality of energy selection, and the fourth grid is grounded to isolate the energy selection potential from the high voltage placed on the phosphorescent screen. Two grids are generally sufficient for LEED work, but the four-grid system offers improved energy selection. This is useful because the same apparatus is often also used for AES measurements, for which energy selection is crucial.

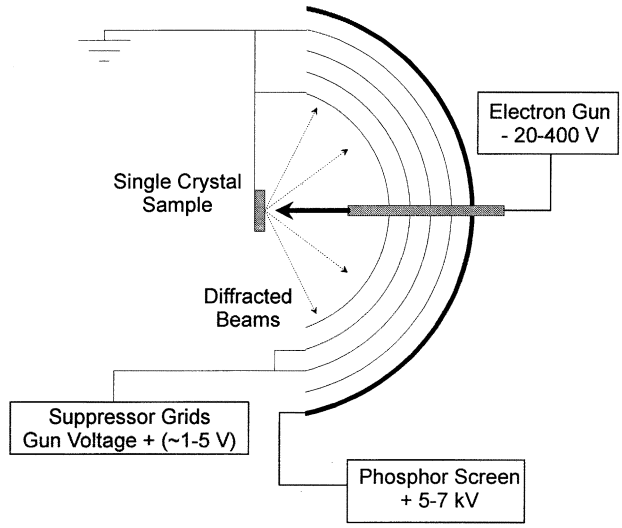


Figure 11. A schematic of the apparatus used in LEED studies. Voltages are indicated on the diagram. Most older instruments view the screen from behind the sample, but newer rear-view optics view the screen from the right side of the diagram.

Several methods have been employed for the direct measurement of spot intensities. The earliest methods involved the use of a Faraday cup that was moved to intersect the diffracted electron beam. This presented difficulties in part because it is difficult to measure currents of the low-energy electrons employed in LEED. Later refinements introduced the postacceleration to a phosphorescent screen described above. A spot photometer could be used to track each diffracted beam and directly measure the $I(E)$ spectrum. However, this process was found to be slow, owing to the need to manually readjust the photometer at each new energy value. Moreover, this method requires that separate acquisitions be made of every spot in the pattern. A method that is currently in wider use involves photographing the screen at regular intervals as the beam energy is increased. This allows for parallel measurement of all visible spots, and delays measurement of intensities until after the rapid data acquisition. This was pioneered by Stair et al. (1975) with a Polaroid camera, but a digital camera interfaced to a computer is more popular currently. In the authors' experience, a common 35-mm camera designed for personal use is not usually sufficient to capture the low light levels emitted by typical fluorescent screens, and a more sensitive device must be employed. The authors successfully used both a manually shuttered Polaroid camera (for rapid development and evaluation) and a computer-controlled CCD camera.

It is important to know the boundaries of LEED's capabilities. As a rule of thumb, LEED interrogates an ordered region of the surface ~100 Å across, and is sensitive to ~5% to 10% of a monolayer. The limitation on lateral resolution is a typical number for most instruments, and it is determined by deviations from the "ideal" beam—a monochromatic plane wave described as A

$\exp(i\mathbf{k} \bullet \mathbf{r})$, with well-defined energy, $\hbar^2|\mathbf{k}|^2/2m_e$, wave vector $\mathbf{k}$ and position in space $\mathbf{r}$. The two important factors characterizing deviations are the finite energy spread of the electron beam, deviating from monochromaticity, and the finite angular divergence of said beam, owing to the filament typically being an imperfect plane wave source. Qualitatively, these imperfections will give rise to unpredictable lateral variations in the phase of the beam; the lateral distance at which the phase relationship becomes arbitrary is generally termed the instrument response function, or coherence length. The coherence length is generally around a few hundred angstroms; an estimate is given in Appendix A.

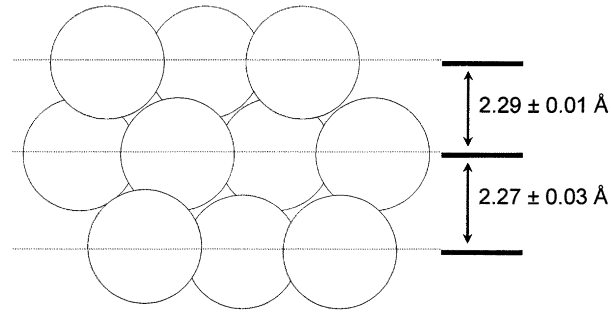


Figure 12. A side view of the Pt(111) surface along the $\{1\bar{1}0\}$ direction. (After Materer et al., 1995.) The bulk interlayer distance is 2.2650 Å.

METHOD AUTOMATION

It is possible and desirable, although not strictly necessary, to acquire LEED data using a computer. The bare minimum requirements for qualitative data can be met with a simple camera. However, as noted above, a common 35-mm camera designed for personal use may not prove sufficient due to the low light level of the phosphorescent screen. Quantitative data requires detailed measurements of spot intensities. It is recommended that a camera interfaced to a computer be capable of at least 8 bits of intensity resolution with adjustable gain and dark level.

Two approaches to automated data acquisition are generally used. One is a procedure whereby a single spot is tracked by a camera while the energy is scanned. This has the advantage of directly measuring the desired $I(E)$ curve in real time. Another popular method is to use a camera to photograph the entire LEED screen at particular energies (e.g., 2-eV steps) and perform an off-line measurement of the intensities later. This method has the advantage of photographing all spots at once for maximum efficiency and not requiring particular effort to track individual spots. Since measuring all of the spots in parallel is also more rapid, instrument drift is minimized. We typically implement the latter method with a CCD camera, a personal computer with a frame-grabber expansion card, and our own software for image acquisition.

DATA ANALYSIS AND INITIAL INTERPRETATION

Two different sorts of data analysis are appropriate to LEED experiments: qualitative and quantitative. Knowledge of the symmetry of the diffraction pattern can yield powerful restrictions on the atomic geometry of the surface. These prove invaluable in reducing the number of candidate structures examined by calculations. The two different forms of analysis will be examined in turn.

Qualitative Analysis

The qualitative features of LEED are the easiest and most commonly applied. As illustrated above, the diffraction pattern reveals a great deal of information about the symmetry of the overlayer structure. This is often enough to make a reasonable guess at a structure or to help deduce a coverage. However, it is dangerous to make blind assumptions about coverage without corroborating information. A less obvious qualitative feature is the $I(E)$ spectra themselves—these can be used as a "fingerprint" of a particular surface structure in much the same way that vibrational spectra are unique to a particular molecule.

This process is perhaps best illustrated by the following examples: Pt(111)-(1×1), Si(100)-(2×1), β-SiC(100)-(2×1), Pt(111)-p(2×2)-C_2H_3, and Ni(100)-(2×2)-2C. These structures can be seen in Figures 12 to 16, respectively. All of these structures have been solved by quantitative LEED techniques, and are examined in light of what qualitative information could be deduced. They also illustrate some of the more important points to bear in mind when investigating a structure.

The Pt(111) surface (Fig. 12) has the simplest possible overlayer structure (Materer et al., 1995). The (1×1) translational symmetry prohibits almost all lateral reconstructions, and the most reasonable structures to consider involve simply relaxations of interlayer spacings. In principle, a shift of the topmost layer of surface atoms to a different registry [at bridge sites or hexagonal close-packed (hcp) hollows instead of fcc hollows] is possible, but it has not been observed in any system examined to date. This structure is also in line with other known metal surface structures; the (111) plane of an fcc metal is the most dense, and lateral reconstructions are rare. Some (100) faces have been known to reconstruct [e.g., the Pt(100)-hex reconstruction; Van Hove et al., 1981], and (110) faces commonly exhibit reconstructions [e.g., Ir(110)-(2×1); Chan and Van Houe, 1986]. For bcc metals, the (110) surface is the most dense and the (111) is the least dense, and the propensity for reconstruction varies accordingly.

The Si(100) structure (Fig. 13; Holland et al., 1984) is simple in that no adsorbate is involved, merely a reconstruction of the substrate. The major question is the geometry of the topmost layers. Semiconductor surface reconstructions can be understood as driven to minimize the number of "dangling bonds." While metals are understood to have nearly isotropic bonding, semiconductors tend to have bonds that are more covalent in nature and highly directional. Consequently, the simple octet rule of "four bonds" from organic chemistry also applies well to silicon surfaces. Si(100) eliminates its "dangling bonds"

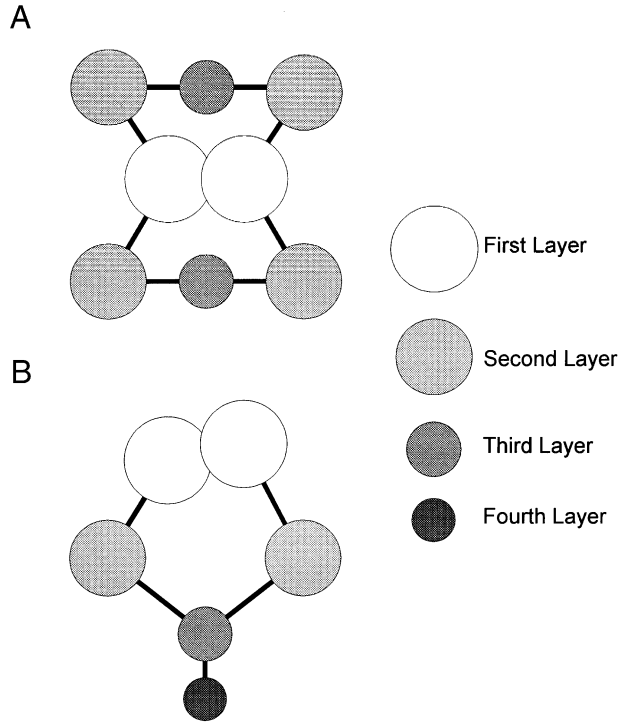


Figure 13. A top (**A**) and side view (**B**) of a single unit cell of the Si(100)-(2 × 1) reconstruction. The details of this structure are still a matter of debate, but the most recent evidence agrees that the surface dimer is tilted. Reconstructions have been found penetrating as far as the fifth atomic layer (Holland et al., 1984.)

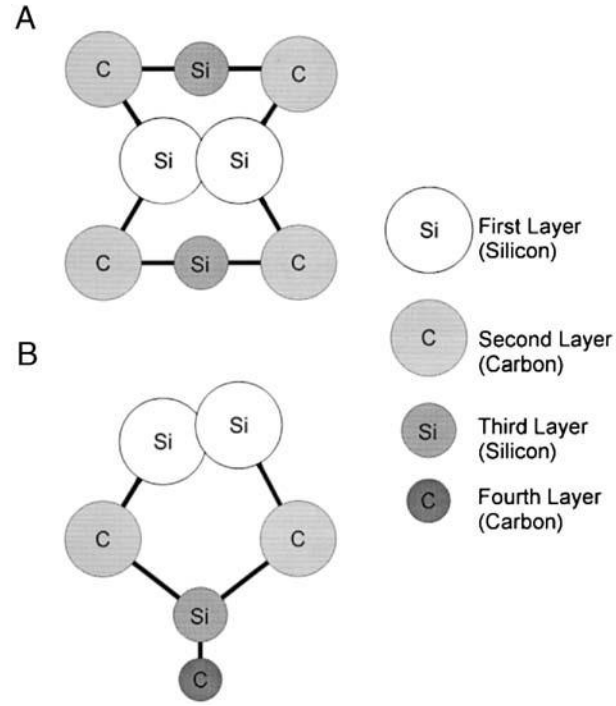


Figure 14. A top (**A**) and side view (**B**) of a single unit cell of the SiC(100)-(2 × 1) reconstruction. (After Powers et al., 1992.) Note the correspondence with Figure 13.

by forming dimers, which halves the number of unoccupied bonds.

The SiC(100) structure (Fig. 14; Powers, et al., 1992) is remarkably similar to the Si(100) structure, which is an example of the extent to which chemical analogy is useful. In the bulk, the (100) planes of SiC are composed of alternating layers of pure Si or pure C composition. However, one should also consider the possibility that chemical mixing will occur at the surface. It turns out that in this case, independent evidence (from Auger spectroscopy) was available to indicate that the face was Si terminated, reducing the number of possible structures to be considered. This illustrates that it is generally advisable to obtain as much information as possible about a surface structure so that some possibilities can be ruled out.

The p(2 × 2) ethylidyne ($CH_3-C\equiv$) structure on Pt(111) (Fig. 15) is representative of adsorbate-induced structures (Starke et al., 1993). The LEED pattern would naively seem to suggest a coverage of 0.25 monolayers (ML), with one molecule adsorbed per unit cell. However, models with 0.75 or 0.50 ML coverage can also be constructed consistent with this symmetry, and one is urged not to leap too rapidly to such a conclusion. Another complication of this system is that a p(2 × 2) structure on an fcc(111) surface is indistinguishable from three rotationally equivalent domains of a p(2 × 1) structure. This ambiguity requires that both models be considered, and in fact, there was a debate in the literature at one time as to which was correct before a general agreement was reached. Part of the data that resolved this ambiguity was a careful calibration of the coverage. The measured coverage of 0.25 ML is consistent with a p(2 × 2) but not a p(1 × 2) structure. Another interesting point is that the calculated structure based on quantitative LEED only indicates the positions of the carbon atoms. Hydrogen atoms are very light, and hence poor scatterers of electrons. It is generally considered impossible to locate hydrogen atoms by LEED, except indirectly by their effects on the positions of heavier atoms.

The case of carbon on Ni(100) (Fig. 16; Gauthier et al., 1991) is interesting because the apparent structure, which resembles a p(2 × 2), leads one to the wrong conclusion upon simplistic assessment. Instead of a coverage of 0.25 or 0.75 ML, it turns out to be 0.50 ML. A critical clue is found in the systematic extinctions of certain spots in the LEED pattern. The (0, n/2) and (n/2, 0) fractional-order spots all vanish when the diffraction pattern is viewed with the electron beam at normal incidence to the crystal. This is characteristic of glide-plane symmetry elements, and these elements place severe restrictions on the possible structures. It is important to note that if the electron beam is not at normal incidence, the symmetry of the beam with respect to the glide planes is broken and the extinguished spots reappear.

Quantitative Analysis

Quantitative structural determinations represent the pinnacle of what LEED can currently achieve. A full tutorial in the theory, programming, and use of LEED computer codes is beyond the scope of this unit, and one is referred to Pendry (1974), Clarke (1985), Van Hove and Tong

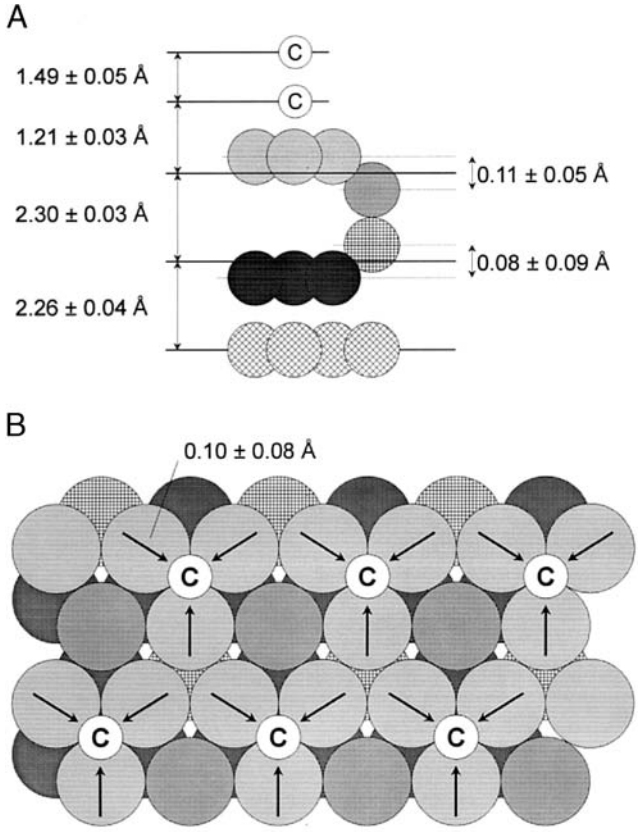


Figure 15. A side (**A**) and top view (**B**) of the structure of ethylidyne on platinum, Pt(111)-p(2 × 2)-C_2H_3. The hydrogen atoms have been omitted since the original investigation (Starke et al., 1993) was not able to locate them; this might have been due to either poor scattering characteristics of the light atoms or dynamic motion that was averaged out in the diffraction experiment. The distances shown are not to scale and are exaggerated to display some of the subtle atomic displacements detected.

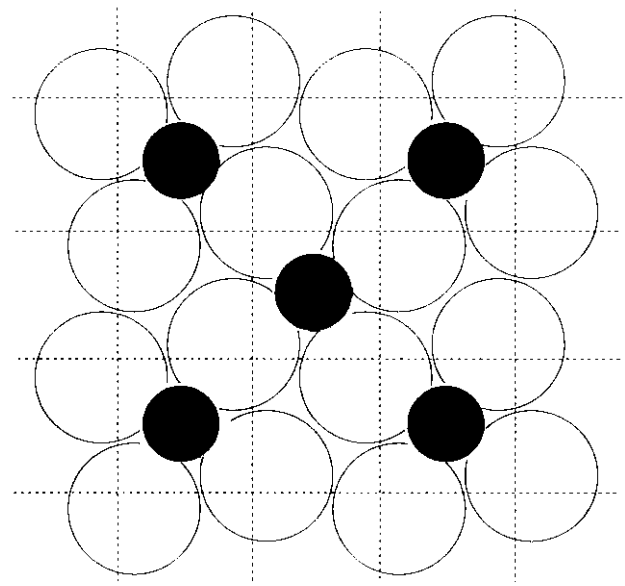


Figure 16. A top view of the Ni(100)-(2 × 2)-2C structure. Carbon atoms are represented as small black circles, and the glide planes are indicated by dotted lines.

(1979), and Van Hove et al. (1986) for the details. However, the outline and important elements of the theory will be described.

The objective of a LEED calculation is to compute the intensities of diffracted beams as a function of the beam energy. These intensities are evaluated for a particular geometry and compared with the experimentally measured data. If the agreement is not satisfactory, the intensities are recomputed for a different trial geometry, and this process of trial and error is repeated until a best agreement is found.

The scattering of LEED electrons is modeled quantum mechanically as reflection from the topmost layers of the surface. The electron beam is generally modeled as a plane wave, $A \exp(i\mathbf{k} \bullet \mathbf{r})$, where $\mathbf{k}$ is complex. The surface is typically modeled with a muffin-tin, one-electron potential. The calculation is typically organized in a hierarchical fashion by determining the scattering of electrons from atoms, organizing those atoms into 2D layers, and finally evaluating the scattering between successive layers. An atomic-like potential for the surface atoms is embedded in and matched to a constant muffin-tin potential, which also includes an imaginary component to simulate the damping of the electron wave by inelastic scattering. The real part of the potential is a calculational device, and it has only limited physical significance, although qualitatively one would expect its value to change for a particular system in correlation with work function changes. The energy of a LEED electron inside the crystal, between the ion cores, is

$$\frac{\hbar^2 k^2}{2m_e} = \frac{\hbar^2}{2m_e}(k_r + ik_i)^2 = \frac{\hbar^2}{2m_e}(k_r^2 - k_i^2) + i\frac{\hbar^2}{m_e}k_r k_i = E + V_{0r} + iV_{0i} \quad (3)$$

thus,

$$\frac{\hbar}{2m_e}(k_r^2 - k_i^2) = E + V_{0r} \quad \text{and} \quad \frac{\hbar^2}{m_e}k_r k_i = V_{0i} \quad (4)$$

where V_{0r} is the real part of the potential, V_{0i} is the imaginary part, k_r is the real part of the complex magnitude of the wave vector of the electron, k_i is the imaginary part, and E is the electron kinetic energy outside of the crystal. Typically, V_{0i} is ~1 to 5 eV, and V_{0r} is generally used as a fitting parameter. An estimate of V_{0i} can be obtained from the width of the observed peaks; approximately, $\Delta E_{\text{FWHM}} \approx V_{0i}$ (Van Hove et al., 1986, p. 119). LEED electrons are not terribly sensitive to valence electrons, so some liberties can be taken with the details of the atomic part of the potential. Most commonly, one uses the results of a band structure calculation as the source of the atomic potentials, such as those of Moruzzi et al. (1978). However, LEED is sufficiently sensitive to valence electrons that

simple free atom calculations do not give good results (Van Hove, 1986, p. 121). Further details of the calculations are given in Appendix B.

In order to actually perform a structural search, some quantitative measure of the goodness of fit of two sets of $I(E)$ curves must be devised. This is typically done by means of an R factor, as is used in x-ray crystallography. The simplest criterion one might imagine is a squared difference, as is employed in a least-squares type of fit. Many other R factors have been proposed and used as well. One particularly popular one was suggested by Pendry (1980), who recognized that peak positions contain most of the important information and chose to emphasize position at the expense of intensity. In part, this is useful because peak positions are determined primarily by geometric factors, but the factors affecting intensity include nonstructural parameters, such as thermal motion, that introduce artificial biases. Pendry's R factor (R_P) is based on the logarithmic derivative of the intensity, $L = (I)^{-1}(dI/dE)$, evaluated as part of a Y function, $Y = L/(1 + V_{0i}^2 L^2)$ so as to avoid singularities near zero intensity. The R factor is defined as

$$R_P = \frac{\int (Y_{expt} - Y_{theo})^2 \, dE}{\int (Y_{expt}^2 + Y_{theo}^2) \, dE} \tag{5}$$

The values for Pendry's R factor range from 0 to 1, with 0 being a perfect correlation between theory and experiment and 1 being no correlation.

Another major advance has been the development by Rous and Pendry (1989) of Tensor LEED, a perturbative method that greatly accelerates structural searches. This method computes a tensor (hence the name) that describes the effect on the LEED matrices of small geometric displacements. Since the full dynamical calculations of $I(E)$ spectra are the most time-consuming step in a fit, this method accelerates a search by allowing a single reference calculation to cover a wide range of parameter space. Instead of requiring an inefficient grid search of parameter space at, say, 0.05-Å intervals, Tensor LEED calculations allow all structures within a certain radius of convergence to be rapidly tested. This radius has been estimated as ~0.3 to 0.4 Å (Rous, 1992), although more conservative estimates are often recommended. In any event, it is wise to check the results of a Tensor LEED calculation by starting a new reference calculation at the last minimum, followed by a new minimization, and perhaps iterating this once or twice.

An important consideration is what kind of experimental data set is required for these calculations. This is usually estimated by examining the total data range for each of the contributing $I(E)$ curves and summing them together to obtain a cumulative energy range. Cumulative ranges of at least 1000 eV are recommended, and more is always better, especially when many parameters are to be fit. A method of estimating the uncertainty in a structural determination was developed by Pendry (1980), and it scales inversely with the cumulative energy range. The sampling of points in experimental spectra should be at least every 3 eV, and 1 to 2 eV is preferable, because peaks in $I(E)$ spectra are generally ~4 to 5 eV wide. This brief, but comprehensive, overview of the theoretical aspects involved in a detailed structural solution provides the reader with adequate insight to critically examine the current LEED literature.

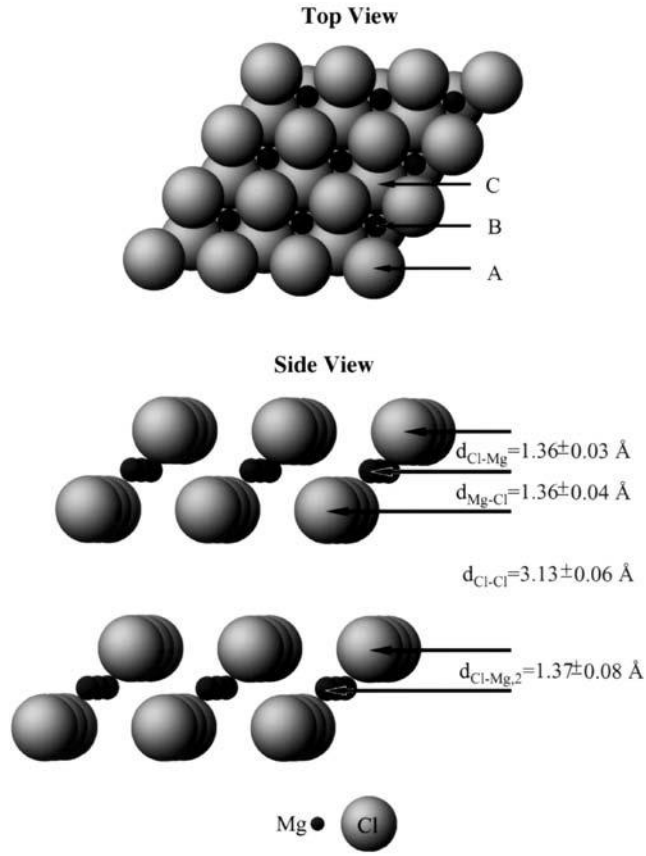


Figure 17. Top view (top) of the α-$MgCl_2$(0001) surface with labels indicating the lateral stacking of the ions. Ionic radii were reduced by 15% for clarity. Side view (bottom) of the α-$MgCl_2$ (0001) surface with the surface termination on top and with the ionic radii reduced by 50%. Refined interlayer spacings are labeled with there respective error bars. Bulk interlayer spacing values are $d_{Mg-Cl,bulk} = d_{Cl-Mg,bulk} = 1.33$ Å and $d_{Cl-Cl,bulk} = 3.22$ Å. The labels A, B, C in the top view indicate the different layers in the side view.

The analysis of the $MgCl_2$ multilayer surface structure (Roberts et al., 1998) will be used as an example of a typical LEED publication. Figure 17 illustrates the objective of a LEED $I(E)$ analysis—the surface structure itself. The use of top and side views familiarizes the reader with the crystallography of $MgCl_2$. In this example, the hexagonal surface symmetry is clearly evident in the top view (Fig. 17 top) along with the relative lateral positions of the atoms. Additionally, it is correctly assumed that the lateral coordinates were held at their bulk values. This is indicated by the absence of any explicit labeling of atomic positions. The side view (Fig. 17 bottom) shows the layered alternating stacking of the Mg and Cl ions along with the interlayer spacings, the only structural parameters varied, in the refined region with their respective error bars. The

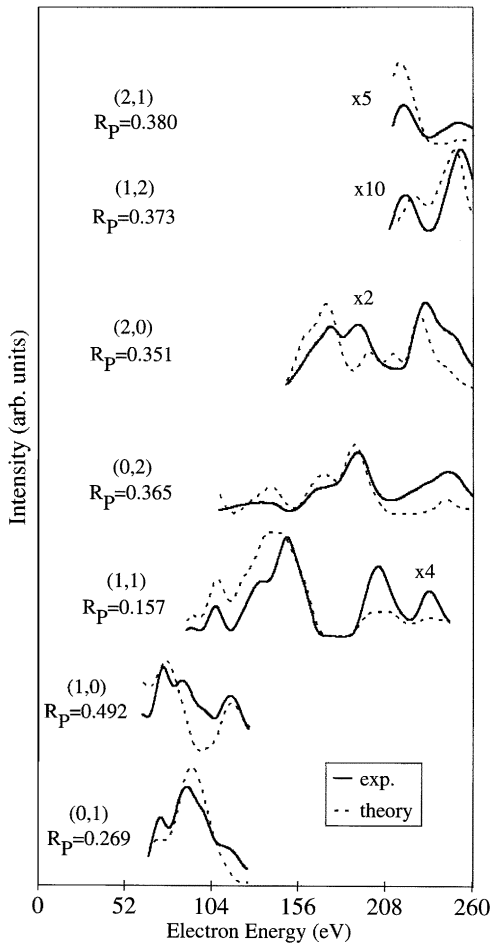


Figure 18. Comparison of the theoretical (dashed lines) and experimental (solid lines) $I(E)$ curves for the fully optimized α-$MgCl_2$(0001) surface. All beams used in the calculation are plotted along with their indices.

spacings below the refined region are again assumed to have bulklike separations because no explicit values are given for these parameters.

The accompanying figure to the structural illustration of all finished calculations is the comparison of the experiment and theoretical $I(E)$ curves (Fig. 18). Although the goodness of fit is quantified by the overall R factor (here $R_P = 0.32$), a visual inspection of the experimental data allows one to determine if the researcher made the correct choice in the number of refinable structural parameters. For $MgCl_2$, the total experimental energy range of 658 eV is smaller than for most data sets, because of the material's extreme sensitivity to the electron beam. This small data set limited the optimization to only four interlayer spacings. Additional structural parameters would have resulted in refined values with very large uncertainties, although the overall R factor would be lowered due to the fact that more variables were used to fit the experimental $I(E)$ curves. When inspecting the $I(E)$ curves, the quality of the theoretical fit to the data should not be judged by how well the theoretical beam intensities correlate to the experimental intensities, since the reason for the inequity between these intensities is directly related to the nature of the Pendry R factor. This R factor is sensitive to the shape of the $I(E)$ curve with respect to energy and not to their intensities. Consequently, a good fit is determined if the minima and maxima of the theoretical curves occur at the same energies as in the experimental curves. The comments and questions that arose from this specific example are common to all LEED-determined structures and provide a more profound understanding of the analysis than could be obtained from just looking at the overall R factor.

Since these calculations are computationally expensive, the tractability of the problem is always a consideration. The difficulty of calculations scales primarily with the number of atoms in a unit cell, usually as the square or cube of that number. Some examples of recent calculations follow. A study of Pt(111)-p(2 × 2)-NO and Ni(111)-c(4 × 2)-2 NO required calculation times of 4 to 5 h for one geometry on an IBM RISC 6000 (Materer et al., 1994). On the same computer, a study of Rh(110)-(2 × 1)-2 O required 1.5 h to calculate a single geometry, and 200 s to complete a search of nearby structures using a Tensor LEED algorithm (Batteas et al., 1995). One benchmark structure is the Si(111)-(7 × 7) reconstruction (Tong et al., 1988), which takes ~1 day on an IBM RISC 6000. A (1 × 1) cell of $MgCl_2$, a simple structure (Roberts et al., 1998), took ~15 min of computational time on a DEC Alpha, using current LEED codes utilizing all benefits afforded by the high symmetry. If symmetry is ignored, the same full dynamical calculation takes ~1 h. For comparison, the DEC Alpha used is observed to be roughly 10 to 20 times faster than the quoted IBM machine. The Si(111)-(7 × 7) structure takes ~1 h to compute on the DEC Alpha, using all available symmetries.

SAMPLE PREPARATION

The basic requirements of a sample are that it be flat, well ordered, and able to reflect diffracted electrons. Samples are typically polished to at least an optically fine finish (1.0 μm or better polishing paste) and mounted in a UHV chamber. Preparation of clean, well-ordered samples for UHV surface science work is somewhat of an art unto itself. Beyond the advice given by Musket et al. (1982), the best that can be said is that one should experiment for oneself. Generally, cycles of noble gas ion bombardment (e.g., 0.5 to 3.0 keV Ar^+), annealing (usually 500 to 2000°C, depending on the material), and gas treatments (usually H_2 or O_2) are used. Ion bombardment can be thought of as "atomic sandblasting" and is good for removing gross contaminants no greater than a few microns in thickness. If contaminants are thick enough to be visible to the naked eye, other methods of cleaning should be used. It is sometimes useful to "momentum match" the ion used to the type of contaminant expected (e.g., Ar^+ for S, Ne^+ for C), but the relative cost of the various rare gases is more often the deciding factor. Annealing heals the damage caused by ion bombardment, and it also allows diffusion of impurities to take place. It is not uncommon for a new crystal to

require several weeks to clean for the first time, while the near-surface region is depleted of contaminants. Subsequent cleaning occurs far quicker, however, and it can usually be accomplished in ~1 h on a daily basis. Chemical treatments are useful in especially difficult cases, if an impurity can be volatilized by reaction (e.g., burning carbon to CO and CO_2). Not all of these steps are necessarily required or even useful in every case.

SPECIMEN MODIFICATION

The only modification to the specimen typically observed is possible electron beam damage. Most of the time, this is not a serious concern, especially when fairly heavy atoms are involved (e.g., reconstructed metals or semiconductors). Damage has more commonly been observed as a result of more energetic beams than those used for LEED (e.g., during Auger electron spectroscopy; see AUGER ELECTRON SPECTROSCOPY), but it is a serious issue for insulating overlayers in some cases. In a recent study of ice multilayers on Pt(111) (Materer et al., 1997), electron beam damage was so severe that a LEED experiment could only be done by use of special Digital LEED equipment (Ogletree et al., 1992).

PROBLEMS

There are several kinds of difficulties that experimenters often face. The most basic is that of finding an ordered structure in a given adsorbate-substrate system. This is the most difficult to address because the specifics vary from one system to another. One must begin with a substrate that is itself well ordered. The spots on the display system should be sharp, nearly as narrow as the incident beam, and the space between the spots should have minimal background intensity. Unless the substrate reconstructs, the pattern seen should form a simple Bravais lattice corresponding to the face of the crystal being studied. If the sample is polycrystalline, one may see either a uniformly bright background or many spots that cannot be reconciled with a single Bravais lattice, perhaps without any obvious pattern at all. The second alternative can usually be distinguished from an unknown substrate reconstruction by the fact that even a complex multiple-domain reconstruction will generally give a pattern with all or most of the bulk rotational and mirror-plane symmetry elements of this crystal plane when a normal incidence electron beam is used. Polycrystalline samples tend to lack recognizable symmetry elements.

Once the substrate crystallinity is assured, a common approach is to adsorb the desired overlayer at as cold a temperature as practical, and begin inspecting the LEED pattern. Frequently, one must warm the sample somewhat to anneal the structure into place. At very cold temperatures, molecules frequently adsorb wherever they strike the surface, and thermal energy must be supplied to overcome kinetic barriers to diffusion so that the well-ordered thermodynamic ground state can be attained. The annealing temperature must be kept low enough that the adsorbate not desorb from the surface. The results of a temperature programmed desorption (TPD) experiment are thus frequently useful in designing a "recipe" for sample preparation (see Woodruff and Delchar, 1994, for more about TPD). Another, more subtle problem is that the adsorbate might itself rearrange on the surface. For example, ethylene adsorbed on Pt(111) at room temperature gives a p(2 × 2) LEED pattern, but it was not obvious until some time later that the ordered species was ethylidyne, $CH_3{-}C{\equiv}$, rather than ethylene itself. Since these transformations are typically thermally driven, one again would wish to use the minimum temperature necessary to anneal an overlayer to avoid side reactions of the adsorbate.

One of the best indicators of reliable data is a close correspondence between the $I(E)$ curves of symmetrically equivalent beams. For example, at normal incidence, the (10), (01), $(\bar{1}0)$, and $(0\bar{1})$ beams are all symmetrically equivalent on an fcc(100) surface. These curves are usually averaged together to reduce noise, but they should look very similar to begin with. Another delicate factor is the integration of intensities. In the photographic method, one integrates the intensity of the spot profile found within a particular window as the total intensity of the spot, since the spot is of finite dimensions. One should take care that the integration windows do not overlap or include intensity from other spots. A less obvious precaution is to normalize the $I(E)$ spectra to the intensity of the primary electron beam. The beam current will generally change as a function of beam voltage, and this can artificially influence the position of peak maxima. Electron beam damage sometimes occurs, as mentioned above. This can be detected by making repeated measurements in ascending and descending order of electron energy. If the two spectra look different, damage has occurred.

Although obtaining an ordered structure represents the greatest challenge to surface crystallographers, finding the sample position where the incident beam is normal to the crystal is always tedious, especially if the overlayer and/or substrate is susceptible to damage from the electron beam. The procedure previously presented for checking normal incidence does have one pitfall—small residual electromagnetic fields inside the chamber can cause deflections in the diffracted beams particularly at low energies. This effect can be circumvented by using higher-order beams to confirm normal incidence of the electron beam on the sample.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the U.S. Department of Energy through the Lawrence Berkeley National Laboratory.

LITERATURE CITED

Ashcroft, N. W. and Mermin, N. D. 1976. Solid State Physics. W. B. Saunders, Philadelphia.

Batteas, J. D., Barbieri, A., Starkey, E. K., Van Hove, M. A., and Somorjai, G. A. 1995. The Rh(110)-p2mg(2 × 1)-2O surface

structure determined by Automated Tensor LEED—structure changes with oxygen coverage. *Surf. Sci.* 339:142–150.

Beeby, J. L. 1968. The diffraction of low-energy electrons by crystals. *J. Phys. C* 1:82–87.

Boffa, A. B., Galloway, H. C., Jacobs, P. W, Benitez, J. J., Batteas, J. D., Salmeron, M., Bell, A. T., and Somorjai, G. A. 1995. The growth and structure of titanium oxide films on Pt(111) investigated by LEED, XPS, ISS, and STM. *Surf. Sci.* 326:80–92.

Chan, C. M. and Van Hove, M. A. 1986. Confirmation of the missing-row model with three-layer relaxations for the reconstructed Ir(110)-(1 × 2) surface. *Surf. Sci.* 171:226–238.

Clarke, L. J. 1985. Surface Crystallography—An Introduction to Low-Energy Electron Diffraction. Wiley, New York.

Crampin, S. and Rous, P. J. 1991. The validity of the average *t*-matrix approximation for low energy electron diffraction from random alloys. *Surf. Sci.* 244:L137–L142.

Davisson, C. and Germer, L. H. 1927. Diffraction of electrons by a crystal of nickel. *Phys. Rev.* 30:705–740.

Döll, R., Gerken, C. A., Van Hove, M. A., and Somorjai, G. A. 1997. Structure of disordered ethylene adsorbed on Pt(111) analyzed by diffuse LEED: Asymmetrical di-sigma bonding favored. *Surf. Sci.* 374:151–161.

Fadley, C. S., Van Hove, M. A., Hussain, Z., and Kaduwela, A. P. 1995. Photoelectron diffraction—New dimensions in space, time and spin. *J. Electron Spectrosc. Relat. Phenom.* 75:273–297.

Farrow, R. F. C. 1995. Molecular Beam Epitaxy. Noyes Publications, Park Ridge, N.J.

Gauthier, Y., Baudoing-Savois, R., Heinz, K., and Landskron, H. 1991. Structure determination of p4g Ni(100)-(2 × 2)-C by LEED. *Surf. Sci.* 251/252:493–497.

Gauthier, Y., Joly, Y., Baudoing, R., and Rundgren, J. 1985. Surface-sandwich segregation on nondilute bimetallic alloys: $Pt_{50}Ni_{50}$ and $Pt_{78}Ni_{22}$ probed by low-energy electron diffraction. *Phys. Rev. B* 31:6216–6218.

Holland, B. W., Duke, C. B., and Paton, A. 1984. The atomic geometry of Si(100)-(2 × 1) revisited. *Surf. Sci. Lett.* 140:L269–L278.

Kittel, C. 1986. Introduction to Solid State Physics, 6th ed. Wiley, New York.

Koningsberger, D. C. and Prins, R. 1988. X-Ray Absorption: Principles, Applications, Techniques of EXAFS, SEXAFS and XANES. Wiley, New York.

Löffler, U., Muschiol, U., Bayer, P., Heinz, K., Fritzsche, V., and Pendry, J. B. 1995. Determination of anisotropic vibrations by Tensor LEED. *Surf. Sci.* 331-333:1435–1440.

Materer, N., Barbieri, A., Gardin, D., Starke, U., Batteas, J. D., Van Hove, M. A., and Somorjai, G. A. 1994. Dynamical LEED analyses of the Pt(111)-p(2 × 2)-NO and the Ni(111)-c(4 × 2)-2NO structures—substrate relaxation and unexpected hollow-site adsorption. *Surf. Sci.* 303:319–332.

Materer, N., Starke, U., Barbieri, A., Döll, R., Heinz, K., Van Hove, M. A., and Somorjai, G. A. 1995. Reliability of detailed LEED structural analyses: Pt(111) and Pt(111)-p(2 × 2)-O. *Surf. Sci.* 325:207–222.

Materer, N., Starke, U., Barbieri, A., Van Hove, M. A., Somorjai, G. A., Kroes, G. J., and Minot, C. 1997. Molecular surface structure of ice(0001): Dynamical low-energy electron diffraction, total-energy calculations and molecular dynamics simulations. *Surf. Sci.* 381:190–210.

Morruzi, V. L., Janak, J. F., and Williams, A. R. 1978. Calculated Electronic Properties of Metals. Pergamon Press, Elmsford, N.Y.

Musket, R. G., McLean, W., Colmenares, C. A., Makowiecki, D. M., and Siekhaus, W. J. 1982. Preparation of atomically clean surfaces of selected elements: A review. *Appl Surf. Sci.* 10:143–207.

Ogletree, D. F., Blackman, G. S., Hwang, R. Q., Starke, U., Somorjai, G. A., and Katz, J. E. 1992. A new pulse counting low-energy electron diffraction system based on a position sensitive detector. *Rev. Sci. Instrum.* 63:104–113.

Pendry, J. B. 1974. Low Energy Electron Diffraction. Academic Press, New York.

Pendry, J. B. 1980. Reliability factors for LEED calculations. *J. Phys. C* 13:937–944.

Powers, J. M., Wander, A., Van Hove, M. A., and Somorjai, G. A. 1992. Structural analysis of the β-SiC(100)-(2 × 1) surface reconstruction by Automated Tensor LEED. *Surf. Sci. Lett.* 260:L7–L10.

Roberts, J. G., Gierer, M., Fairbrother, D. H., Van Hove, M. A., and Somorjai, G. A. 1998. Quantitiative LEED analysis of the surface structure of a $MgCl_2$ thin film grown on Pd(111). *Surf. Sci.* 399:123–128.

Rous, P. J. 1992. The tensor LEED approximation and surface crystallography by low-energy electron diffraction. *Prog. Surf. Sci.* 39:3–63.

Rous, P. J. and Pendry, J. B. 1989. The theory of tensor LEED. *Surf. Sci.* 219:355–372.

Saldin, D. K. 1997. Holographic crystallography for surface studies: A review of the basic principles. *Surf. Rev. Lett.* 4:441–457.

Seah, M. P. and Dench, W. A. 1979. Quantitative electron spectroscopy of surfaces: A standard data base for electron inelastic mean free paths in solids. *Surf. Interface Anal.* 1:2–11.

Stair, P. C., Kaminska, T. J., Kesmodel, L. L., and Somorjai, G. A. 1975. New rapid and accurate method to measure low-energy-electron-diffraction beam intensities: The intensities from the clean Pt(111) crystal face. *Phys. Rev. B.* 11:623–629.

Starke, U., Barbieri, A., Materer, N., Van Hove, M. A., and Somorjai, G. A. 1993. Ethylidyne on Pt(111)—Determination of adsorption site, substrate relaxation, and coverage by automated tensor LEED. *Surf. Sci.* 286:1–14.

Starke, U., Pendry, J. B., and Heinz, K. 1996. Diffuse low-energy electron diffraction. *Prog. Surf. Sci.* 52:53–124.

Tong, S. Y., Huang, H., and Wei, C. M. 1988. Low-energy electron diffraction analysis of the Si(111)-(7 × 7) structure. *J. Vac. Sci. Technol.* 6:615–624.

Tong, S. Y. and Van Hove, M. A. 1977. Unified computation scheme of low-energy electron diffraction—the combined-space method. *Phys. Rev. B* 16:1459–1467.

Van Der Veen, J. F. 1985. Ion beam crystallography of surfaces and interfaces. *Surf. Sci. Rep.* 5:199–288.

Van Hove, M. A., Cerda, J., Sautet, P., Bocquet, M.-L., and Salmeron, M. 1997. Surface structure determination by STM vs. LEED. *Prog. Surf. Sci.* 54:315–329.

Van Hove, M. A., Koestner, R. J., Stair, P. C., Biberian, J. P., Kesmodel, L. L., Bartos, I., and Somorjai, G. A. 1981. The surface reconstructions of the (100) crystal faces of iridium, platinum and gold. I. Experimental observations and possible structural models. *Surf. Sci.* 103:189–217.

Van Hove, M. A. and Tong, S. Y. 1979. Surface Crystallography by LEED—Theory, Computation and Structural Results. Springer-Verlag, Berlin.

Van Hove, M. A., Weinberg, W. H., and Chan, C.-M. 1986. Low Energy Electron Diffraction—Experiment, Theory and Surface Structure Determination. Springer-Verlag, Berlin.

Wood, E. A. 1963. Vocabulary of surface crystallography. *J. Appl. Phys*. 35:1306–1312.

Woodruff, D. P. and Bradshaw, A. M. 1994. Adsorbate structure determination on surfaces using photoelectron diffraction. *Rep. Prog. Phys*. 57:1029–1080.

Woodruff, D. P. and Delchar, T. A. 1994. Modern Techniques of Surface Science, 2nd ed. Cambridge University Press, Cambridge.

KEY REFERENCES

Pendry, 1974. See above.

Clarke, 1985. See above.

Two excellent introductions to the field, both very readable by the beginner. Both emphasize the background necessary to LEED calculations, while still describing other qualitative aspects.

Van Hove et al., 1986. See above.

The more recent of Van Hove's books, which provides a solid exposition of the theory needed to perform structural calculations. The Barbieri/Van Hove software package is available from Michel Van Hove at Lawrence Berkeley Laboratory and is highly recommended, as its use will save a great deal of effort in programming one's own computer codes.

Woodruff and Delchar, 1994. See above.

An excellent overview of all of the major techniques used in modern surface science, with enough information for both the newcomer and the experienced user looking to broaden his or her scope.

Another good book on the same topic as the previous reference.

Rivere, J. C. 1990. Surface Analytical Techniques. Clarendon Press, New York.

APPENDIX A: ESTIMATION OF COHERENCE LENGTH

The following quantitative estimate is after Pendry (1974): variations in the wave vector parallel and perpendicular to the beam can be estimated from the dot- and cross-product, respectively, between $\mathbf{k}$ and $\Delta\mathbf{k}$. The parallel deviations can be described as

$$|\mathbf{k}|^{-2}\overline{(\Delta\mathbf{k}\bullet\mathbf{k})^2} = |\mathbf{k}|^2\,\overline{(\Delta E)^2}/(4E^2) \tag{6}$$

since

$$\hbar|\mathbf{k}|^2 = 2m_eE \tag{7}$$

where $\overline{(\Delta E)^2}$ is the mean-square spread in energy. Perpendicular variations can be described as

$$|\mathbf{k}|^{-2}\overline{|\Delta\mathbf{k}\times\mathbf{k}|^2} = \overline{(\Delta\theta)^2}|\mathbf{k}|^2 \tag{8}$$

where $\overline{(\Delta\theta)^2}$ is the mean-square angular spread of the beam. On the surface, two points separated by a distance $\mathbf{l}$ differ in phase by $\mathbf{l}\bullet\mathbf{k}$, and the mean-square deviation from this phase is given by

$$\overline{(\Delta\phi)^2} = \overline{(\Delta\theta)^2}\frac{1}{2}|\mathbf{k}|^2|\mathbf{l}|^2\sin^2(\alpha)+\overline{(\Delta E)^2}(2E)^{-2}|\mathbf{k}|^2|\mathbf{l}|^2\cos^2(\alpha) \tag{9}$$

where α is the angle between $\mathbf{l}$ and $\mathbf{k}$. When the mean-square phase deviation approaches π^2, all phase relationship has vanished. The coherence length becomes

$$l_c = \frac{2\pi|\mathbf{k}|}{[2\overline{(\Delta\theta)^2}\sin^2(\alpha)+\overline{(\Delta E/E)^2}\cos^2(\alpha)]^{1/2}} \tag{10}$$

Choosing, e.g., $\Delta\theta \cong 0.001$ radians, $\Delta E \cong 0.2$ eV, $E = 150$ eV and $\alpha = 45^\circ$ gives

$$l_c \cong 500\text{Å} \tag{11}$$

APPENDIX B: DETAILS OF CALCULATIONS

The details of the atomic scattering event are only important at the level of a "black box"—knowing what goes in and comes out. Consequently, knowledge of the wave function of the electron beam within the atom is unnecessary; knowledge of the amount by which the phase of the scattered wave was advanced is all that is useful or required, and this describes the scattering completely. These phase shifts, δ_l are components of the atomic scattering t matrix, which has the form

$$t_l = \frac{\hbar^2}{2m_e}\left(\frac{1}{k_0}\right)\sin(\delta_l)\exp(i\delta_l) \tag{12}$$

where $k_0 = |\mathbf{k}|$, $\hbar \equiv h/2\pi$ (where h is Planck's constant), and m_e is the mass of the electron. Note that it has been found convenient to represent scattering from an assumed spherical atom by a spherical wave formalism. This necessitates decomposing the incident plane waves into spherical partial waves for purposes of the scattering event. In a practical calculation, one cannot carry out the spherical wave expansion to infinite orders of precision, so one may approximate the maximum required angular momentum, l by $l_{\max} \approx k_0 \times r_{\text{mt}}$, where r_{mt} is the muffin-tin radius of the atom in question. Of course, this is only a rule of thumb, and one should check that the inclusion of higher-order terms does not change the results of a calculation.

Once the atomic scattering parameters have been determined, an atomic layer is constructed. Usually, it is simple to divide the crystal into layers, although sometimes it is necessary to create a "composite layer" of inequivalent atoms at either the same or similar heights; this is most frequently true for the topmost layer and for closely spaced planes. Scattering matrices are determined for each plane, $M_{\text{gg}}^{\pm\pm}$, which relate the amplitudes of a scattered wave with momentum $k_{\text{g}}^{\pm}$ to that of an incident wave with momentum $k_{\text{g}}^{\pm}$. In the previous notation, "+" is taken to signify a wave propagating into the crystal while "−" denotes an outgoing wave, and k_{g} is the momentum of an electron having experienced parallel momentum transfer of $\mathbf{g}$, one of the reciprocal lattice vectors of the crystal. A matrix inversion formalism can be used for this calculation, using a spherical wave representation

for each of the subplanes of a composite layer. This approach was developed by Beeby (1968) and generalized by Tong and Van Hove (1977) for use with all beams. The following are defined in atomic units $\hbar = m_e = 1$ in a spherical wave representation [$L = (l,m)$] directly after Van Hove and Tong (1979):

$t_l^i = \frac{1}{2k_0}\exp(i\delta_l)\sin(\delta_l)$: Scattering t-matrix for a single atom in subplane i.

$\tau_{LL'}^i$: Scattering matrix that contains all scattering paths within subplane i.

$T_{LL'}^i$: Scattering matrix that includes all those scattering paths within the composite layer that end at subplane i.

$G_{ll'}^{ji}$: Structural propagator describing all unscattered propagations from atoms in subplane i to atoms in subplane j.

The diffraction matrix is defined as

$$M_{gg}^{\pm\pm} = -\frac{16\pi^2 i}{Ak_{g'\perp}}\sum_{LL'} Y_L(k_{g'}^{\pm})\sum_{i=1}^{N}\{\exp[i(\pm k_g^{\pm} \mp k_{g'}^{\pm})\cdot r_i]T_{LL'}^i\} \times Y_{L'}^8(k_{g'}^{\pm}) + \delta_{g'g}\delta_{\pm\pm} \quad (13)$$

with

$$G_{LL'}^{ji} = \hat{G}_{LL'}^{ji}\exp[-ik_g^{\pm}\cdot(r_j - r_i)] \quad (14)$$

$G_{LL'}^{ji}$ may be expressed as a sum over reciprocal lattice points with

$$\hat{G}_{LL'}^{ji} = \frac{16\pi^2 i}{A}\sum_g \frac{\exp[ik_g^{\pm}\cdot(r_j - r_i)]}{k_g^{\pm}} Y_L^*(k_g^{\pm})Y_{L'}(k_g^{\pm}) \quad (15)$$

where Y_{lm} is a spherical harmonic, A is the unit cell area, and r_i and r_j are the positions of arbitrary reference atoms in subplanes i and j, respectively. The subplane τ matrix is

$$\tau_{LL'}^i = [(I - t^i G^{ii})^{-1}]_{LL'} t_l^i \quad (16)$$

where I is the multiplicative identity matrix. According to Beeby (1968), the matrices $T^1, \ldots, T^N$ are obtained directly from the equation

$$\begin{pmatrix} T^1 \\ T^2 \\ \vdots \\ T^N \end{pmatrix} = \begin{pmatrix} I & -\tau^1 G12 & \ldots & -\tau^1 G1N \\ -\tau^2 G21 & I & \ldots & -\tau^2 G2N \\ \vdots & \vdots & & \vdots \\ -\tau^N G^{N1} & -\tau^N G^{N2} & \ldots & I \end{pmatrix}^{-1} \begin{pmatrix} \tau^1 \\ \tau^2 \\ \vdots \\ \tau^N \end{pmatrix} \quad (17)$$

This completes the calculation of $M_{g'g}^{\pm\pm}$.

The final step is to calculate the interlayer scattering, which is typically done using the renormalized forward scattering (RFS) method developed by Pendry (1974). The RFS method takes advantage of the physical fact that forward scattering is much more probable than backscattering (backscattered flux is typically 1% to 5% incident flux), and the reflectivity is expanded in terms of the number of reflections experienced by the electron wave. Note that only an odd number of reflections will contribute to any observed diffraction beam. This method can be implemented by allowing the transmitted (forward scattered) waves to propagate through the crystal until their amplitude decays to a negligible value due to inelastic effects. The backscattered waves are then propagated back through the crystal, and the exiting amplitudes become the first-order result. The second order is obtained by allowing each of the backscattered waves to reflect twice more and exit the crystal, and so on for higher orders. This is continued until convergence is attained, when the reflected amplitude changes negligibly from a previous pass. This method tends to converge because inelastic damping prohibits very high-order waves from surviving. The formalism makes use of two iterative expressions for the interlayer amplitudes, one for penetration

$$a_{(i)g}^{\text{new}} = \sum_{g'}[M_{gg'}^{++}P_{g'}^{+(i-1)}a_{(i-1)g} + M_{gg'}^{+-}P_{g'}^{-(i)}a_{(i)g'}] \quad (18)$$

and one for emergence (Van Hove et al., 1986)

$$a_{(i)g}^{\text{new}} = \sum_{g'}[M_{gg'}^{--}P_{g'}^{-(i+1)}a_{(i+1)g'} + M_{gg'}^{-+}P_{g'}^{+(i)}a_{(i)g}] \quad (19)$$

where $P_g^{\pm -i}$ are plane wave propagators between appropriate reference points on successive layers and $a_{(i)g}^{\text{new}}$ is constantly overwriting $a_{(i)g}$. RFS typically requires 12 to 15 layers and 3 to 4 orders to achieve convergence (Van Hove et al., 1986, p. 174). This method also requires that layer spacings be greater than $\sim$1 Å, or too many plane waves are required for convergence. In these cases, alternative methods such as layer doubling must be employed.

Craig A. Gerken
Gabor A. Somorjai
University of California at Berkeley and Lawrence Berkeley National Laboratory
Berkeley, California

ENERGY-DISPERSIVE SPECTROMETRY

INTRODUCTION

Energy-dispersive x-ray spectrometry (EDS) is a technique for measuring the intensity of x-ray emission as a function of the energy of the x-ray photons (Fitzgerald et al., 1968; Goldstein et al., 1992). The excitation source for the x-rays can be energetic electrons, photons, or ions, and the target can be solid, liquid, or gas, although solid targets are the norm in virtually all materials science applications. The measured x-ray intensity can be related to the concentration (i.e., mass or atomic fraction) for each

element present by performing physical/empirical matrix corrections to account for the interelement modification of the generated x radiation caused by (1) primary radiation stopping power and ionization effects, (2) attenuation of secondary characteristic x-rays during their passage through matter, and (3) any inefficiencies in detection. In this chapter, various aspects of the process of obtaining high-quality x-ray spectra, the interpretation of spectra, and the accurate extraction of x-ray intensities from spectra will be considered.

Materials analysis techniques that employ EDS include the following.

Electron excitation: (1) electron probe x-ray microanalysis (EPMA)/analytical scanning electron microscopy (ASEM), where EDS complements wavelength-dispersive x-ray spectroscopy (WDS); (2) analytical electron microscopy (AEM).

Photon excitation: X-ray fluorescence (XRF).

Ion excitation: particle-induced x-ray emission (PIXE).

This chapter will concentrate on EDS performed in electron-beam instruments (Williams et al., 1995). The general principles of EDS x-ray detection, processing, display, and spectral manipulation are similar for all analytical techniques regardless of the excitation source. The procedures for quantitative analysis depend in detail on the excitation process. Quantitative x-ray microanalysis by electron excitation is considered in detail in the section on Electron Probe Microanalysis. Those aspects of quantitative x-ray microanalysis that are particular to EDS spectral measurement will be considered.

Measurement Challenge. All elements except H and He produce characteristic x rays. Because of self-absorption in the target and absorption in the components of the spectrometer/detector, the practical minimum photon energy that is accessible by any analytical spectrometry technique is ~0.1 keV, which excludes measurement of Li (0.052 keV). Measurement of all elements beginning with the Be K-shell (0.108 keV) is possible, with heavy elements typically detected with their M- or L-shell lines (with an upper bound of ~15 keV for the L-shell). For certain applications, photon energies as high as 100 keV may be of interest (e.g., K-lines of any element excited with the AEM operating at electron energies of 200 keV or higher). The ideal form of x-ray spectrometry would measure photons (1) in the range 100 eV to 100 keV with high-energy resolution to avoid spectral peak interferences and to reduce the contribution of background processes to the peaks; (2) with true parallel detection of different photon energies to maximize measurement efficiency and to improve the utility for qualitative analysis; or (3) if serial detection of photons is the only possibility, then with a short measurement time constant to minimize coincidence events; and (4) with no added artifacts in the spectrum. These factors will be considered for semiconductor detector-based EDS systems in current use, and for other types of EDS detectors now under development for future application to analytical systems.

PRINCIPLES OF THE METHOD

The basic principles of the measurement process for the semiconductor energy-dispersive x-ray spectrometer are illustrated in Figure 1A, and a schematic diagram of a generic semiconductor EDS system is shown in Figure 1B (Goldstein et al., 1992). An x ray with energy E_v is absorbed photoelectrically by an atom in a semiconductor crystal (Si or Ge) creating a photoelectron with energy $E_v - E_c$, where E_c is the critical ionization energy (binding energy) for the shell (e.g., $E_K = 1.84$ keV for the Si K-shell) and leaving the absorbing atom in an excited, ionized state. Consider a detector fabricated with Si. The Si atom will subsequently de-excite through electron shell transitions that will cause the subsequent emission of

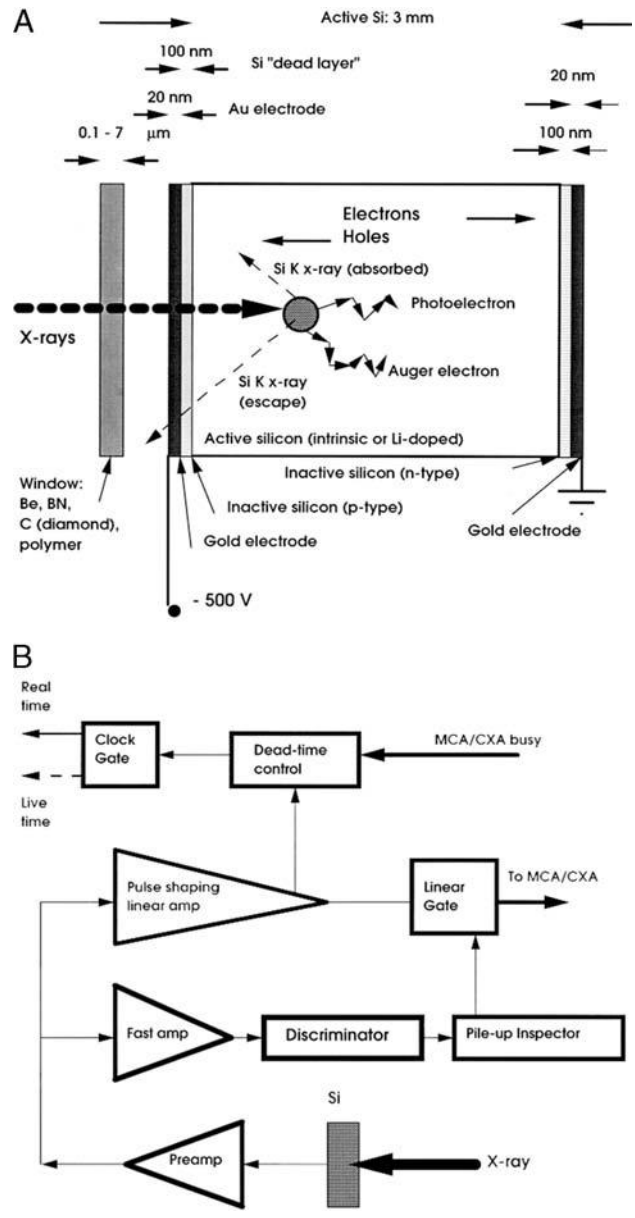


Figure 1. (**A**) The measurement process for the semiconductor energy-dispersive x-ray spectrometer. (**B**) Schematic diagram of a generic semiconductor EDS spectrometry system.

either an Auger electron (e.g., Si KLL) or a characteristic x-ray (Si Kα, $E = 1.74$ keV or Si Kβ, E = 1.83 keV). If an x-ray is emitted it may be reabsorbed in a photoelectric event with a lower binding energy shell (e.g., Si L-shell) with the emission of another photoelectron and subsequently another Auger electron. This cascade of energetic electron generation within the semiconductor is completed within picoseconds. The energetic electrons subsequently scatter inelastically with a very short mean free path (tens of nanometers or less) while traveling in the crystal, giving up their energy to crystal electron and phonon excitations. In particular, electrons from the filled valence level in the semiconductor crystal can be promoted to the empty conduction band where they are free to move under an applied electric field, leaving behind a positive "hole" in the valence band, which can also move. By placing a bias potential (500 to 1000 V) across the faces of the crystal, the free electrons and holes can be separated before they recombine. Creation of an electron-hole pair in silicon requires ~3.6 eV, so that a 3.6-keV photon will create ~1000 electron-hole pairs. The charge carriers drift out of the detector in a time of hundreds of nanoseconds. Measuring the collected charge gives a value proportional to the energy of the original photon. This is the critical measurement in x-ray EDS. A detailed description of the amplification and charge measurement electronics of the EDS system is beyond the scope of this unit. It is critical that a high-gain, low-noise amplification process be employed, since the detection of a photon involves measuring a deposited charge of $\sim 10^{-16}$ C and this measurement must be completed in ~50 μs (Williams et al., 1995).

An EDS system consists of several key components, Figure 1B: (1) A semiconductor crystal detector, typically housed with a field-effect transistor (FET) preamplifier in a cryostat cooled with liquid nitrogen to maintain a temperature of ~100 K or lower at the detector to minimize thermal noise. (Some systems use electrical Peltier cooling and operate at higher temperature and noise to avoid the need for liquefied gas.) (2) A high-voltage power supply to provide the detector bias (~ 500 V). (3) A slow main amplifier, which takes the pulses from the preamplifier and provides further amplification and pulse shaping. (4) A fast pulse inspection function, which monitors the charge state of the detector as a function of time and is used to reduce pileup events. (5) A "deadtime" correction function, which compensates for the time when the detector is processing a pulse and unavailable for other pulses by adding time to the clock to achieve the specified "live" accumulation time. (6) A computer-assisted analysis system [multichannel analyzer (MCA) or computer x-ray analyzer (CXA)] for spectrum display and manipulation. With the rapid evolution of computer control of measurement systems, many of these functions may take place within the computer control system, and in modern systems virtually all are controlled by software commands. Moreover, digital pulse processing, in which the entire preamplifier waveform is digitized by a high-speed circuit for subsequent processing, is emerging as a means to improve pulse handling and deadtime correction (Mott and Friel, 1995).

The EDS x-ray spectrum presented to the analyst by the computer-assisted analyzer consists of a histogram of channels calibrated in energy units containing counts corresponding to the detection of individual photons. An example of an EDS spectrum of a pure element, titanium, excited with an incident beam energy of 20 keV is shown in Figure 2A and expanded vertically in 2B. Consider this spectrum in terms of the measurement challenge posed above. The semiconductor EDS detector fulfills the first requirement that a wide energy range be accessible. Photons in the range 100 to 20 keV can be readily measured with a Si EDS detector, as shown in the logarithmic display in Figure 2C, and with decreasing efficiency, photon energies as high as 30 keV can be measured. The range can be extended to 100 keV with the use of Ge as a higher-density detector. It is this "energy-dispersive" character that provides the greatest value in practical analytical x-ray spectrometry, because it enables access to the entire periodic table (except H, He, and Li) with every spectral measurement. A complete qualitative analysis can (and should!) be performed at every sample location analyzed. With EDS measurements, constituents should not be missed unless they suffer severe interference from another peak or are present at concentrations below the limit of detection for the measurement conditions chosen.

Concerning the second criterion, the resolution (i.e., width) of x-ray peaks measured by EDS is relatively coarse and is limited by the statistical nature of the charge generation and collection process. A typical figure of merit used to define the resolution of a system is the width of the Mn Kα peak (chosen because of the use of the ^{55}Fe radioactive source, which decays by electron capture with subsequent emission of Mn K-radiation). The resolution of a typical Si EDS is measured as the full width of the peak at half the peak intensity (full width half-maximum, FWHM) and is typically ~130 to 140 eV at Mn Kα (5890 eV) for Si detectors operating in the optimum resolution (i.e., long integration time constant) or a fractional width of 2.3%. This compares to the natural peak width of ~ 2 eV; i.e., the line is broadened by a factor of ~75. The Ge EDS can achieve an improved resolution of ~120 eV. The resolution is energy dependent and is described approximately by the following expression (Fiori and Newbury, 1978):

$$\mathrm{FWHM} = [2.5(E - E_{\mathrm{ref}}) + \mathrm{FWHM}^2_{\mathrm{ref}}]^{0.5} \qquad (1)$$

where the reference peak is usually Mn Kα. This relatively poor energy resolution leads to a significant number of peak overlaps in practical analysis situations (e. g., S K, Mo L, and Pb M), and also constrains the limit of detection (concentration) when a significant background process is present, as in the case of the electron-excited bremsstrahlung (continuum) radiation.

Considering the third criterion for parallel collection, when the accumulation of an EDS spectrum is observed, the EDS appears to provide parallel collection across the entire energy range. Actually, the term "energy dispersive" in the name of the technique is traditional and not strictly correct. No dispersion in the classic spectroscopic sense actually occurs in the measurement of the x-ray spectrum with EDS. [In the older literature, the more descriptive term "nondispersive" was often applied

(Heinrich, 1981).] The photons are detected individually, and the EDS detection process can only accommodate one photon at a time since the entire volume of the monolithic semiconductor crystal is involved. While a photon of any energy within the range specified above can be detected at any given time, detection is a serial process, so that if a second photon enters the detector while the first is being processed, the measurement of one or both may be corrupted. Because it takes time for the charge to drift out of the detector and an even longer time to measure that charge accurately, the EDS detection process is paralyzable. As more and more photons enter the detector, the output count rate first increases with increasing input count rate, but eventually reaches a maximum and decreases with any further increase in the input count rate.

The final criterion for defect-free spectra is also not satisfied for semiconductor EDS, because of the formation of certain parasitic peaks during the detection process (escape and sum peaks), the nonuniform efficiency with

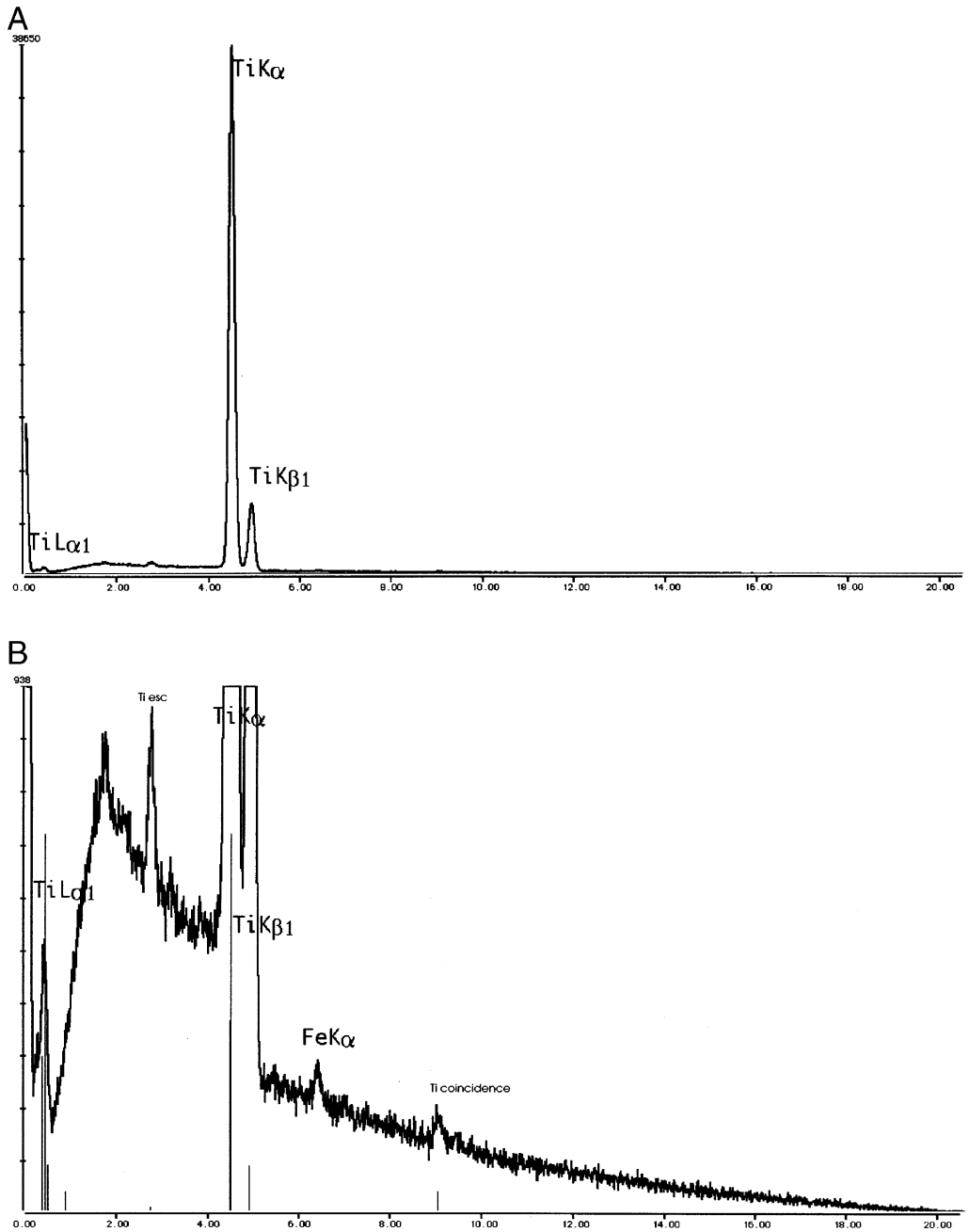


Figure 2. EDS spectrum of a pure element, titanium, excited with an incident beam energy of 20 keV. (**A**) Vertical (linear) scale set to highest peak; (**B**) vertical (linear) scale expanded to show bremsstrahlung background and artifact peaks due to coincidence and escape from the detector; (**C**) logarithmic scale.

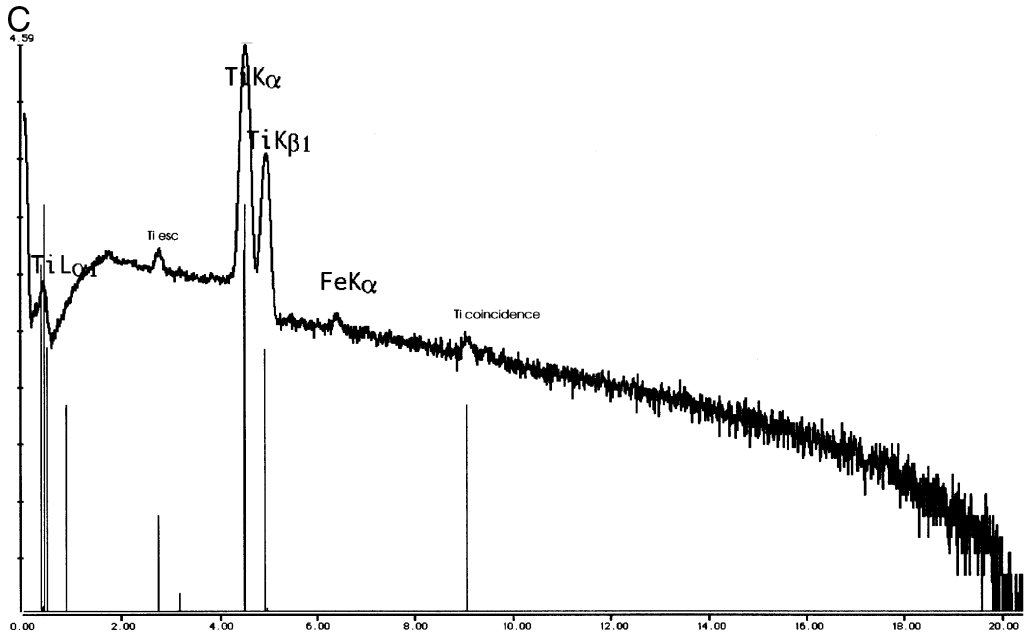


Figure 2 (*Continued*)

photon energy, and the fundamental distortion imposed upon the spectrum by the relatively coarse resolution function.

1. *Escape Peaks.* The escape peak arises because during the photon capture process within the detector, the de-excitation of the Si (or Ge) atom leads to the emission of a Si Kα photon (or Ge Lα and Ge Kα) in ~10% of the events. A material is relatively transparent to its own characteristic x rays, so the range of Si Kα in Si is much greater than that for the Si photoelectrons and Si Auger electrons, and moreover, the Si Kα x-ray does not lose significant energy by inelastic scattering within the detector. While the Si Kα x ray can be absorbed photoelectrically in an interaction with a Si L-shell electron, there is a significant chance that it will escape the detector, which robs the photon being measured of 1.74 keV of energy. These deficient pulses form a parasitic peak displaced by 1.74 keV below the parent peak, as noted in the spectrum of pure titanium shown in Figures 2B and C. The relative size of the escape peak depends on the average depth of capture of the primary photon within the detector crystal, so that peaks with energies just above the silicon K edge (1.84 keV) show the largest escape peaks. For P Kα (2.015 keV), the escape peak is ~2% of the parent peak intensity, while for photons above 5 keV, the escape peak becomes less significant, e.g., 0.5% for Fe Kα.
2. *Sum Peaks.* The pulse-inspection circuitry constantly monitors the rate of change of charge flow from the detector. After a photon arrival is detected and the measurement process is initiated, any subsequent photon detection during the pulse processing time changes the rate of charge flow from the detector and triggers a decision to retain or exclude the initial photon measurement in process. This pulse inspection circuit has a finite time resolution, so that it is possible for two photons to enter the detector so close in time that they cannot be distinguished. In such a case, their energies are added, creating a coincidence or "sum" peak. It is possible for any two photons to coincide, but it is only the high-abundance peaks that can give rise to noticeable spectral defects, as marked in Figures 2B and C. Coincidence peaks become progressively greater in relative height as the input count rate increases. For system deadtime below ~30%, the coincidence effect will be negligible, especially if the target contains several major elements so that the effective peak count rate is reduced compared to that for a pure element. Note that for equivalent counting rates, sum peaks become relatively more significant as the photon energy decreases, because the charge pulse produced by a low-energy photon is closer to the fundamental noise level and the coincidence exclusion circuit is therefore less efficient at distinguishing photon events. For low-energy photons (<1 keV), the pileup inspector may not function at all.
3. *Nonuniform Detection Efficiency.* Any photon in the energy range 100 to ~15 keV that actually enters the active volume of the detector is captured with near unit efficiency. Photons >15 keV in energy are increasingly lost with increasing energy due to penetration through a Si detector, while for a Ge detector, useful efficiency remains out to 100 keV. However, photons below ~4 keV suffer significant loss due to absorption in the spectrometer components: aluminum reflective coatings, window materials, gold surface electrodes, and silicon "dead" (actually,

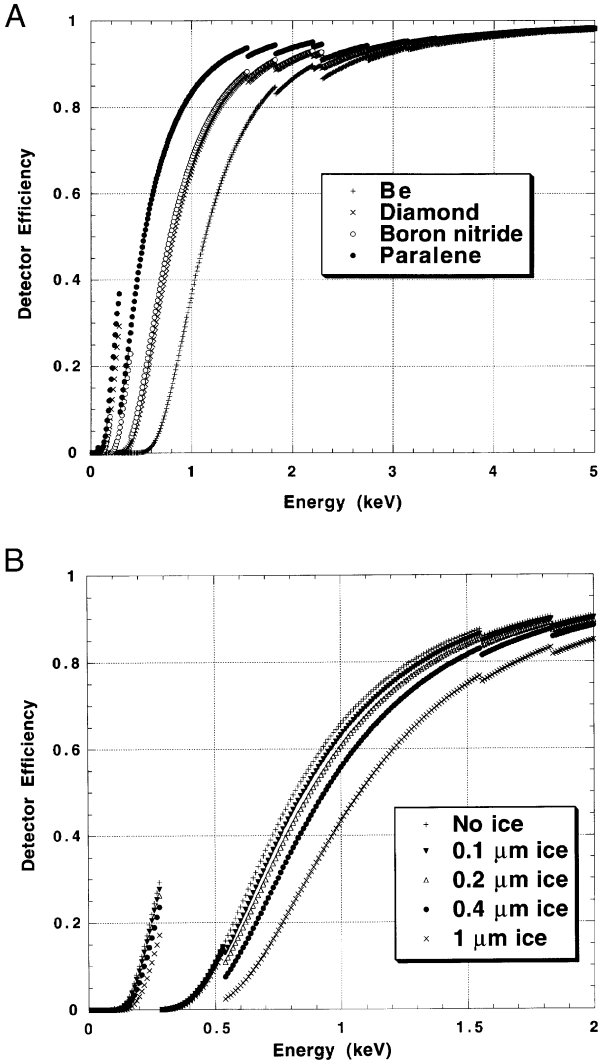


Figure 3. (**A**) Detector efficiency for various windows: boron nitride (0.5 μm), diamond (0.4 μm), beryllium (7.6 μm), and paralene (0.3 μm) (Al window coating = 0.02 μm; Au electrode = 0.01 μm; Si dead layer = 0.03 μm). The absorption edge at 0.28 keV is due to the presence of carbon in some of the windows. Edges due to the Au electrode and the Si dead layer can also be seen in the figure. (**B**) Detector efficiency for a diamond (0.4 μm) thin window showing the effect of ice buildup.

partially active) layers, as well as pathological layers due to possible contamination on the window, and ice on the detector. Efficiency curves are presented in Figure 3A for various windows (Be, diamond, boron nitride, and polymer) on a Si detector with a 20-nm gold electrode, a 30-nm silicon dead layer, and a 20-nm Al reflective coating. The effect of ice buildup on the detector efficiency (diamond window) is shown in Figure 3B.

4. *Spectrum Distortion.* The action of the detector resolution function on the spectrum is illustrated in Figure 4, where the "true" spectrum calculated theoretically (using NIST-NIH Desktop Spectrum Analyzer; Fiori et al., 1992) following absorption in the specimen is compared with the actual EDS spectrum after all stages of the measurement process. The broadening of the peaks from "line" spectra (the characteristic peaks are shown as single lines because the true peak width is actually less than the 10-eV channel width) into Gaussian peaks 10 to 15 channels wide is evident. Less evident is the effect of spectrometer broadening on background distort ion, where the absorption edges are broadened. The parasitic escape and pileup peaks also contribute to spectrum distortion.

PRACTICAL ASPECTS OF THE METHOD

Optimizing EDS Collection

To obtain the best possible x-ray spectra for subsequent analysis, the analyst must carefully choose and/or adjust several aspects of the operation of the EDS system.

Resolution/Count Rate Range. The effective resolution of an EDS system depends on the time spent measuring each pulse (the "shaping time"). A tradeoff can be made between the maximum acceptable count rate and the resolution. Operating at the best resolution limits the maximum input count rate to $\leq$3 kHz. If the x-ray peaks of interest are well separated, it may be acceptable to operate with poorer resolution to obtain a higher acceptable counting rate, 10 kHz or higher. Whatever the choice, it is critical that the analyst be consistent when collecting a series of spectra, especially if measurements are taken over a period of time and compared to digitally archived standards.

Energy Calibration. The charge deposited in the detector must be accurately converted to the equivalent energy in a linear fashion. At the beginning of each measurement campaign, the user should check that the calibration (peak position) is correct within $\pm$10 eV using known x-ray peaks from standards. Accurate calibration is critically important for both qualitative and quantitative analysis. Adjustments of the zero and coarse/fine gain may be necessary via hardware or software settings to achieve accurate calibration. Ideally, the test peaks should span the analytical range, e.g., 100 to 12 keV. A typical first choice for this task is pure copper, which provides peaks at 0.93 keV (Cu Lα), 8.04 keV (Cu Kα), and 8.68 keV (Cu Kβ). Note that a prudent analyst will also check selected lines at intermediate energies [e.g., Si Kα (1.74 keV), Ti Kα (4.51 keV), and Fe Kα (6.40 keV)], as well as higher-energy lines [e.g., Pb Lα (10.549 keV)]. It is commonly observed that even with a linear calibration established above 1 keV, the energy response may be nonlinear for photons below 1 keV. Due to the effects of incomplete charge collection for low-energy photons that are absorbed near the front electrode of the crystal, the peak position (and shape) may deviate significantly. Calibrating with a peak in this range is likely to introduce a nonlinearity for intermediate energies. The best procedure is to calibrate with the low reference peak chosen at 1 keV or above and then to check the position of important low-energy peaks, e.g., C K (0.282 keV) and O K (0.523 keV).

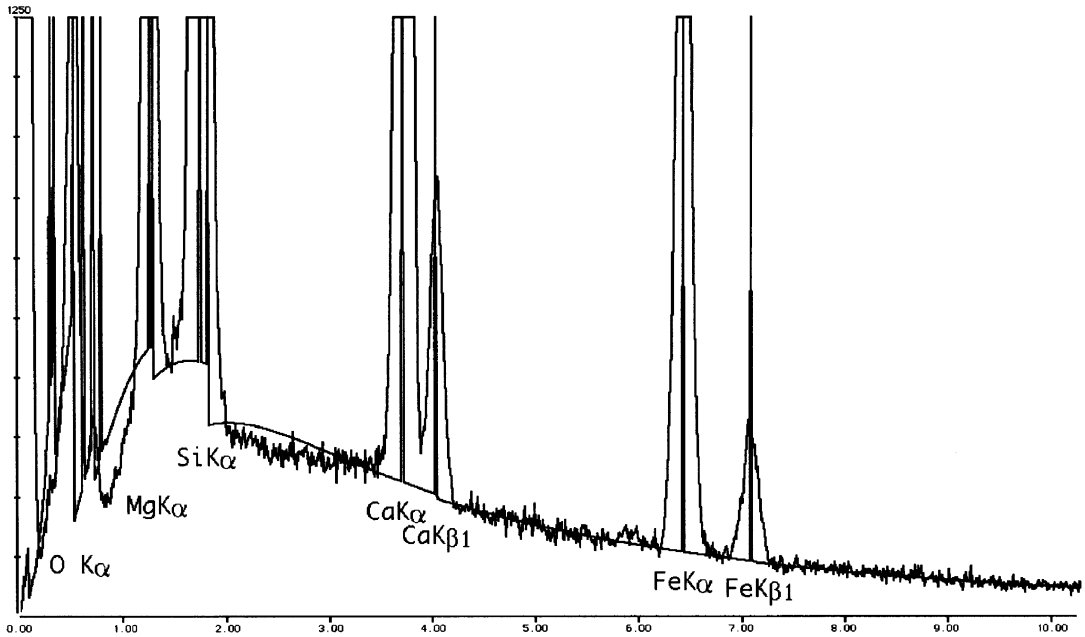


Figure 4. Comparison of experimentally measured spectrum (noisy) with theoretical spectrum as generated in the target (smooth). Specimen, NIST Standard Reference Material K-411 glass [Mg (0.08885); Si (0.254); Ca (0.111); Fe (0.112); O (0.424) mass fraction]; beam energy, 20 keV.

Deadtime Correction. The basis of quantitative analysis methods is the measurement of the x-ray intensity produced with a known dose (i.e., beam current $\times$ integration time). However, because the EDS has a paralyzable deadtime, the effective integration time would depend on the count rate if a simple clock (real) time were used for the measurement. The deadtime correction circuit automatically determines how long the detector is busy processing pulses and therefore unavailable to detect the arrival of another photon, and then adds on additional time units to the specified time of accumulation to compensate for this deadtime. The performance of the deadtime correction function can be checked by noting that all signals scale linearly with the beam current. An accurate current meter should be used to measure the beam current in a Faraday cup. Choosing a fixed "live time" (i.e., deadtime-correction applied), spectra should be collected from a pure-element standard (e.g., Cu or Fe) with progressive increases in the beam current. The x-ray counts in any peak (or background region) should increase linearly with the beam current to indicated system deadtime of $\geq$50%. Note that operation with the dose rate chosen to keep the deadtime $<$30% is advisable, since at higher deadtime artifacts such as coincidence peaks and pileup distortions in the background will occur.

METHOD AUTOMATION

Computers play an extensive role in energy-dispersive x-ray spectrometry, and in fact the advance of laboratory computerization has been a driving force in the incorporation of dedicated automation systems in electron microscope/EDS laboratories throughout the history of the technique (Williams et al., 1995). In current instrumentation, the computer plays a role at all stages of EDS detector control, digital pulse processing, spectrum display, qualitative and quantitative analysis, and archiving of spectral data. The multichannel analyzer (MCA) has given way to the computer-assisted x-ray analyzer (CXA), which is capable of controlling both the full functions of the scanning electron microscope (SEM) or analytical electron microscope (AEM) and the EDS system for unattended, automatic operation. Spectra can be collected from an operator-specified list of predetermined sites on the specimen, or in the most advanced systems, objects of a certain class, such as particles, can be automatically detected and located in electron images and then analyzed according to a specified protocol.

DATA ANALYSIS AND INITIAL INTERPRETATION

Qualitative Analysis

Qualitative analysis is the identification of the spectral peaks present and their assignment to specific elements. While state-of-the-art computer-assisted analysis systems will automatically identify peaks, a prudent analyst will always check on the suggested solutions to see if they agree with the analyst's assessment. This is particularly true for the low intensity peaks, which might correspond to minor (e.g., 1 to 10 wt.%) and trace (e.g., $<$1 weight percent) constituents, but which might also arise as minor family members, escape peaks, or sum peaks of high-concentration constituents. Automatic qualitative analysis usually operates on the basis of a look-up table in which the detected peak is compared with an energy database

and then all possible elements within a specified range, e.g., ±50 eV, are reported. These automatic results should only be considered as a guide, and the analyst must take the responsibility to confirm or deny each suggested possibility with manual qualitative analysis. One possible strategy for manual qualitative analysis is described below (Fiori and Newbury, 1978).

Before attempting manual qualitative analysis, it is critical that the analyst first determine the reliability of the database of x-ray information available in the computer-assisted analysis system in use. Certain minor lines in the L and M families (especially L1 and M_2N_4) produce low-relative-intensity peaks but yet are well separated in energy from the main family lines and therefore are readily visible. Such lines may not be included in the database and consequently will not appear in the "KLM" line markers for the display. If those lines are unavailable in the database, it is almost certain that they will be misidentified and assigned to another element. To evaluate the situation, the analyst should examine spectra measured from pure heavy metals (e.g., Ta, Au, Pb) to test the particular implementation of KLM markers as well as the automatic qualitative-analysis procedure against complex families of x-ray peaks. An example of a complete treatment of the lines is shown for the bismuth L and M families in Figure 5. The relative weights for the L- and M-family lines may also be of questionable value, and should only be considered as a crude guide. In addition, the analyst should test the response of the automatic qualitative analysis procedure against escape peaks and sum peaks with pure-element spectra, e.g., Al, Si, and Ti.

The basis for robust qualitative analysis is the identification, when possible, of multiple lines for each element and the careful location and notation of all minor-family-member and artifact peaks (Fiori and Newbury, 1978). To improve confidence in the identification of light elements that only produce a single line in the low photon-energy range (<3 keV), it is necessary to carefully exclude possible L- and M-line interferences from heavy elements. Finally, the identification of low-intensity peaks as arising from minor and trace constituents can only be made with confidence if all minor family members from high-intensity parent peaks have been located and properly assigned, as well as the corresponding escape and sum peaks. Adequate peak statistics must be obtained, which may require additional spectrum accumulation for identification of minor and trace constituents.

1. First identify the high-intensity peaks that arise from major constituents. The analyst should start at the upper end of the spectrum (>5 keV) and work down in photon energy because the members of each x-ray family are most widely separated above 5 keV and are almost always resolvable by semiconductor EDS, thus giving greater confidence to elemental assignments by providing two or more peaks for identification.
2. Once a tentative assignment has been made to a peak, all other family members must be located, e.g., Kα-Kβ; Lα-Lβα-Lγ-Lη-Ll; Mα-Mβ-Mγ-Mζ-M_2N_4. Both energy position and relative height (weights of lines) are important in making an assignment.
3. For x-ray peaks above 5 keV, if a K family is identified, then a lower-energy L family of the same element will also be present and must be located and identified. Similarly, if an L family above 5 keV is identified, there must be a corresponding M family at lower energy.
4. The escape peaks and sum peaks corresponding to each high-intensity peak must be located and marked. Note that both the escape and sum peaks increase in relative significance as the parent peak energy decreases.

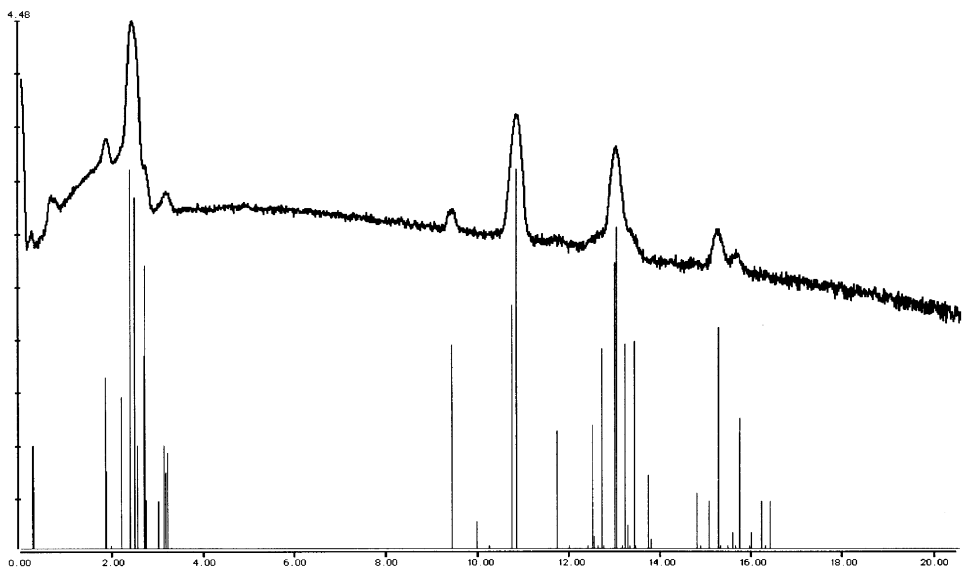


Figure 5. EDS spectrum of bismuth (logarithmic scale), showing KLM markers depicting the full M and L families. Beam energy, 25 keV.

5. Low-atomic-number elements ($Z < 15$) will produce a single unresolved K peak, since the Kα-Kβ peaks become progressively closer in energy and the relative Kα/Kβ ratio increases to more than 50:1 as the atomic number decreases. For fluorine and below, only one K peak exists. Thus, light elements can only be identified with a single peak, which reduces the confidence with which such assignments can be made. This is especially true for photon energies <1 keV, where x-ray absorption in the specimen significantly reduces the measured peak height relative to the higher-energy portion of the spectrum. Thus, care in identifying the higher-atomic number-elements that produce low-energy L- and M-family members is critical.
6. After the spectrum has been examined for major constituents and all possible peaks located and marked, the remaining low-intensity peaks can be assigned to minor and trace constituents following the same strategy. Note that because of the lower intensity of the peaks produced by minor and trace constituents, it may not be possible to locate and identify more than one peak per element, thus reducing the confidence with which an assignment can be made. If it is necessary to increase the confidence level that a minor or trace constituent is present, a longer accumulation time may be required to improve the counting statistics and reveal additional peaks against the statistical fluctuations in the background.
7. A final question should always be considered: What minor/trace elements could be obscured by the high-intensity peaks of the major constituents, e.g., Ti-K and Ba-L, Ta-M, and Si-K?

Quantitative Analysis

The first step in achieving accurate quantitative analysis is the extraction of the characteristic x-ray intensities from the EDS spectrum (Heinrich and Newbury, 1991; Goldstein et al., 1992; Williams et al., 1995). The relatively poor resolution in EDS results in low spectral peak-to-background ratios and frequent severe peak interferences. Careful spectral deconvolution is a critical step in any quantification scheme. Several mathematical tools are used in various combinations in different deconvolution strategies.

Background Removal

Characteristic peaks can be separated from background by two approaches: modeling and filtering.

In background modeling, a physical model for the background parameterized in terms of primary radiation energy, x-ray energy, and specimen composition is used to predict the background intensity as a function of energy, constrained to match the experimental spectrum at two or more energy values. Figure 6 shows the results of a background fit using the background model of Small et al. (1987) to a multielement glass (NIST K309). An excellent fit is found across the full energy range, with the only significant deviation occurring below the oxygen K line (0.523 keV). When background modeling the spectrum from an unknown specimen, the composition is initially estimated as part of a quantitative analysis procedure and an iteration procedure is followed as the specimen composition is refined. An accurate background fit is most critical for minor and trace constituents. High-intensity peaks are less sensitive to the background, so that the estimate/iteration approach is effective for unknowns.

In background filtering, a mathematical algorithm that acts as a frequency filter is applied to the spectrum (Goldstein et al., 1992). Transformed to frequency space, the peaks in the spectrum reside in the high-frequency component, while the background represents the low-frequency component. These components can be separated using a top-hat filter, which acts by transforming the contents of

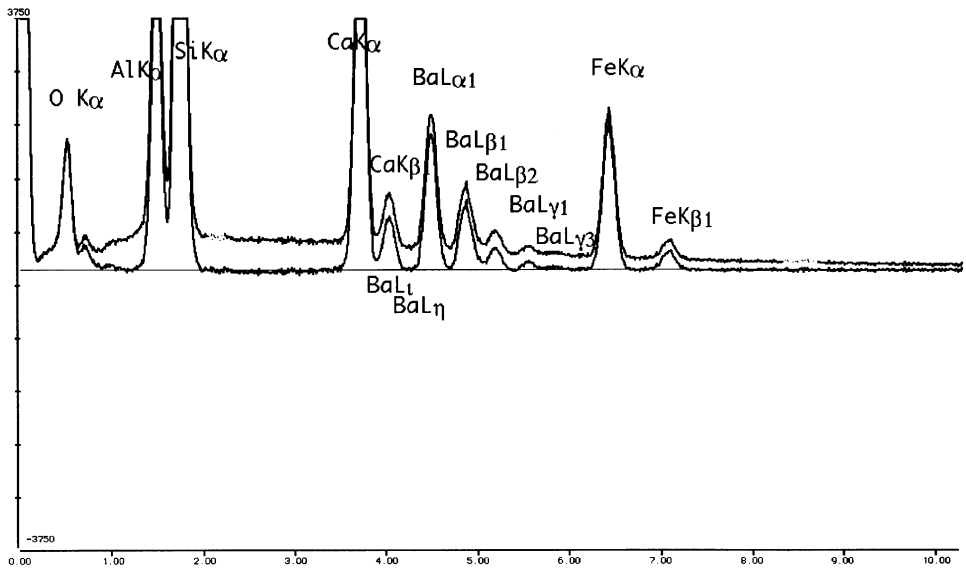


Figure 6. Result of a background fit to a complex specimen. NIST glass K309 [Al (0.079 mass fraction); Si (0.187); Ca (0.107); Fe (0.105); Ba (0.134); O (0.387)]. Beam energy, 20 keV.

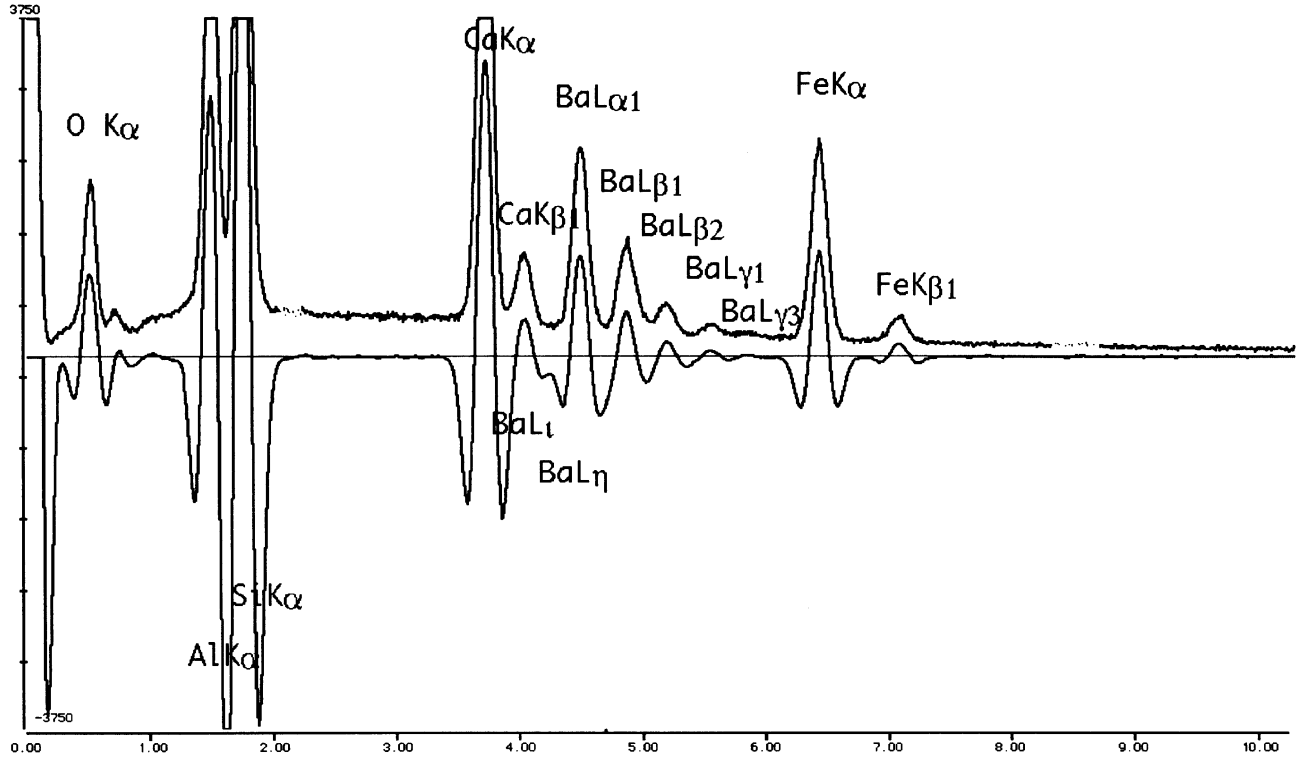


Figure 7. Result of the application of a top-hat digital filter to a complex specimen. NIST glass K309 [Al (0.079 mass fraction); Si (0.187); Ca (0.107); Fe (0.105); Ba (0.134); O (0.387)]. Beam energy, 20 keV.

a channel with a sum of scaled values from adjacent channels. Figure 7 shows the effect of a top-hat filter on the K309 spectrum.

Deconvolution of Peak Overlaps

Most commercial software systems utilize peak fitting by the method of multiple linear least squares (MLLS). In MLLS, a peak reference is first prepared that contains the full peak shape for an element separated from the background and without interference from other elements. Ideally this reference is measured on the same EDS system that is used to measure the unknown. The reference contains the measured response of the particular EDS for all peaks in the family in the energy range of interest. The reference shows the true response of the EDS to each peak, including any deviations from the ideal Gaussian shape, such as incomplete charge collection, which is a function of the particular detector. To deconvolve an unknown, references for all elements expected in a specified energy region of interest are collectively fit to the unknown peak bundle. The goodness of fit between the synthesized spectrum and the real spectrum is determined on a channel-by-channel basis, using a statistical criterion such as the chi-squared value. The only variable is the peak amplitude for each component, so combinations of the references can be synthesized by linear superposition. An example of MLLS fit to the Sr-L, W-M region of a $SrWO_4$ spectrum is shown in Figure 8, which contains both the original spectrum and the background remaining after peak stripping. Because linear mathematics is used, MLLS fitting can be very fast, requiring <1 sec per spectral region of interest. It is critical that the MLLS references be appropriate to the EDS system in use in terms of calibration and peak shape (resolution). Any deviation in system response will impact the accuracy of an MLLS deconvolution. It is extremely useful to be able to inspect the spectrum after MLLS peak stripping to detect failures in the fitting. A fit to the Ti Kα,β peaks in $SrTiO_3$ with a good peak reference derived from TiO_2 measured in the same sequence is shown in Figure 9A. Both the original spectrum and the residuals following peak stripping are shown. The effect of a 10-eV (one-channel) shift in the position of the recorded spectrum relative to the reference is shown in Figure 9B, and a 20-eV difference in resolution (unknown broader than reference) is shown in Figure 9C. The value of examining the residuals following any spectrum modification operation such as stripping is well illustrated by these examples.

Sequential simplex fitting is an alternative approach to peak deconvolution that does not require prior determination of peak references (Fiori et al., 1981). The simplex approach requires an assumption of the peak shape, taken as a Gaussian for EDS x-ray peaks. A test spectrum is synthesized for the elements being fit, and the correspondence to the measured spectrum is compared on a channel-by-channel basis. The simplex is a nonlinear fitting procedure, and can simultaneously be fit for three variables for each peak: position (energy calibration), width (resolution), and amplitude. Thus, simplex fitting can respond

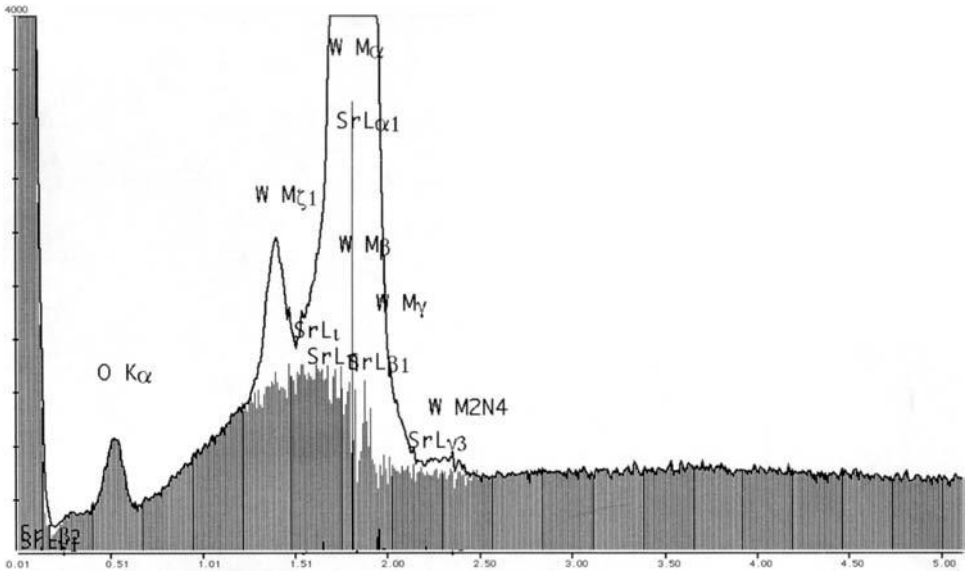


Figure 8. Multiple linear least-squares stripping of the W M family and the Sr L family in SrWO4. Beam energy, 20 keV. Solid black trace, original spectrum; gray spectrum, background remaining after stripping.

to instability in the EDS system. Because of the more complex, nonlinear calculations, the simplex procedure is significantly slower, by a factor of 10 or more, than the MLLS approach.

Matrix Corrections

Once characteristic x-ray intensities have been extracted from spectra of unknowns and standards, quantitative analysis proceeds in the same way for EDS as for wavelength-dispersive x-ray spectrometry (WDS) in electron probe microanalysis. The central paradigm of quantitative electron-excited x-ray microanalysis with WDS or EDS is the measurement of the unknown and standards under identical conditions of beam energy, spectrometer efficiency, known electron dose, and spectral deconvolution. The first step is the determination of the ratio of intensities for the same x-ray peak in the unknown and the standard, which to a first approximation is equal to the ratio of the elemental concentrations in weight fractions (Castaing, 1951):

$$k = I_{\text{unk}}/I_{\text{std}} \sim C_{\text{unk}}/C_{\text{std}} \qquad (2)$$

Note that in this ratio the efficiency of the detector cancels quantitatively because the same x-ray peak is measured for the unknown and for the standard.

The concentration ratio is not exactly equal to the k ratio because of the action of matrix (interelement) effects (Goldstein et al., 1992). Matrix effects arise because of compositionally dependent differences between the sample and standard in the following processes.

1. ***Electron Scattering.*** The backscattering of high-energy electrons increases with the atomic number of the target (and therefore varies depending on composition). Backscattering reduces the ionization power of the beam electrons.
2. ***Electron Stopping.*** The loss of energy from the beam electrons due to inelastic scattering depends upon the composition. The efficiency of the ionization depends upon the ratio of the electron energy to the critical excitation energy of the atomic shell of interest. Differences in electron stopping with composition modify the ionization power.
3. ***X-ray Absorption.*** Characteristic x rays are produced over a range of depth in the target and must travel through the solid to reach the detector. X rays are subject to photoelectric absorption, and this absorption depends on the path length and the mass-absorption coefficients, which are compositionally dependent.
4. ***Secondary X-Ray Fluorescence.*** X rays that are higher in energy than the excitation energy of a particular atomic shell can be photoelectrically absorbed through interaction with that shell, leading to subsequent emission of the shell's characteristic x ray. This secondary fluorescence can be induced by higher-energy characteristic x rays or by continuum (bremsstrahlung) x rays.

These four effects can be expressed as a series of multiplicative correction factors that convert the k ratio of intensities to the equivalent ratio of weight concentrations.

$$C_{\text{unk}}/C_{\text{std}} = kZAF \qquad (3)$$

The Z, A, and F factors can be derived from first-principles physical calculations (such as Monte Carlo electron-trajectory simulations), but because of the difficulty in determining certain critical parameters with adequate

accuracy, the matrix factors are often based at least partially upon experimental measurements of the required data.

Since the Z, A, and F factors depend upon the composition, which is initially unknown, a starting estimate of the composition is obtained by normalizing the set of measured k values for all elements in the specimen:

$$C_{i,n} = i_i/\Sigma k \tag{4}$$

From this initial estimate of composition, a set of correction factors is calculated and the corresponding k values are calculated and compared with the measured k values. This procedure is iterated until convergence is achieved, which is generally within three iterations. Note that for oxidized systems, the low energy of the oxygen characteristic x ray (0.523 keV) leads to severe absorption in the target and in the detector components, and consequently a large uncertainty in the calculated absorption correction. It is thus common practice to include oxygen by the method of assumed stoichiometry, where based upon the analysis of the cations, an appropriate amount of oxygen is added according to the assumed cation valences.

An example of a quantitative analysis of a microhomogeneous standard reference material (SRM 482; NIST) is listed in Table 1. The SRM values are listed along with the value determined by EDS analysis. The standards were pure elements, and matrix corrections were calculated with the NIST ZAF procedure (Myklebust et al., 1979; Fiori et al., 1992). Note that the relative errors are defined as

$$\text{rel err } (\%) = 100\% \times (\text{measured} - \text{true})/\text{true} \tag{5}$$

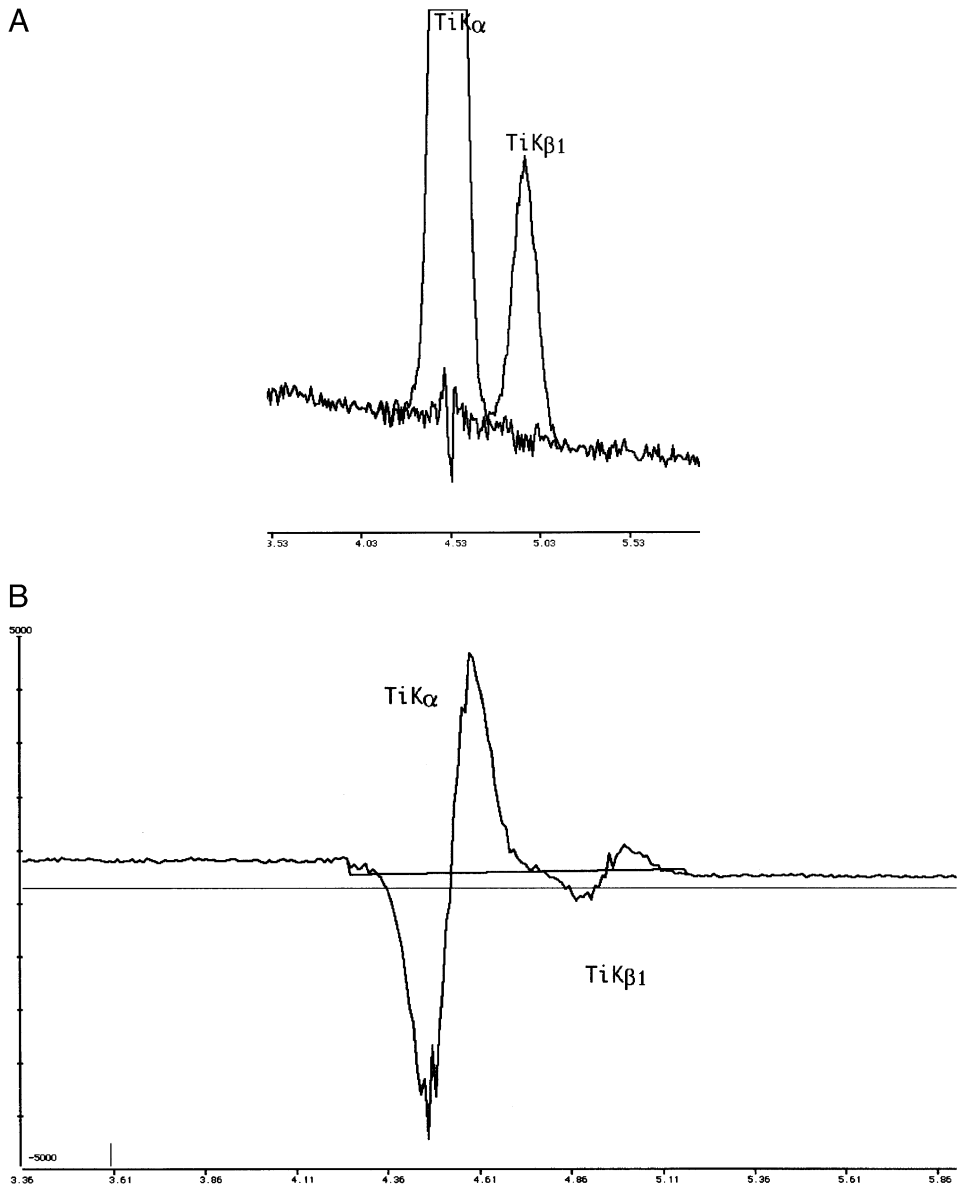


Figure 9. (**A**) Multiple linear least-squares fit to the titanium Kα-Kβ peaks with good peak references. (**B**) Effect of a one-channel (10-eV) shift between the reference and the experimental spectrum on multiple linear least-squares fitting. Specimen, Ti; beam energy, 20 keV. (**C**) Effect of a 20-eV difference in resolution between the reference and the experimental spectrum on multiple linear least-squares fitting. Specimen, Ti; beam energy, 20 keV.

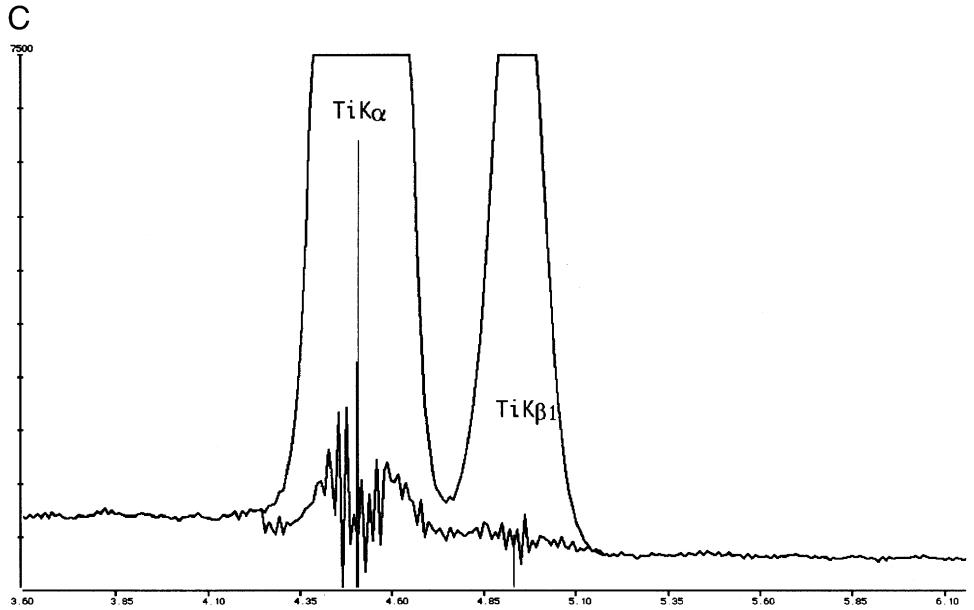


Figure 9 (*Continued*)

In no case is the error >1.6% relative. This is comparable to the errors that are encountered with WDS analysis of the same materials.

When large numbers of analyses of known materials spanning a large fraction of the periodic table are performed, the distribution of analytical errors can be determined. For such studies, spectra are recorded with a large number of counts to reduce the statistical variation to a negligible level compared to the systematic errors. An example of this distribution for the case of binary metallic alloys analyzed with pure-element standards is shown in Figure 13A. This distribution predicts that in 95% of the analyses performed, the relative errors are less than ±5%. For EDS analysis, this distribution is expected for concentrations greater than 0.05 mass fraction (or 5% by weight) and with minimal peak interference. As the concentration is reduced below a mass fraction of 0.05, the uncertainty in the background correction becomes more significant, and the level of error will depend strongly on the photon energy.

Standardless Analysis

An increasing trend in recent years in performing quantitative electron probe x-ray microanalysis with energy-dispersive x-ray spectrometry has been the substitution of "standardless" methods in place of the traditional approach of measuring standards containing the elements of interest on the same analytical instrument under the same excitation and detection conditions. In standardless methods, the standard intensities necessary for quantification are either calculated from first principles, considering all aspects of x-ray generation, propagation through the solid target, and detection ("true standardless"); or else they are derived from a suite of experimental measurements performed remotely and adjusted for the characteristics of the local instrument actually used to measure the unknowns ("fitted standards"). With either route to obtaining "standard intensities," the resulting k values (intensity of the unknown/intensity of the standard) are then subjected to matrix corrections with one of the usual approaches—e.g., ZAF or $\phi(\rho z)$.

The apparent advantages of standardless analysis are considerable. Instrument operation can be extremely simple. There is no need to know the beam current, and indeed, it is not even necessary for the beam current to be stable during the spectrum accumulation, a real asset for instruments such as the cold-field emission gun SEM where the beam current can be strongly time dependent. Moreover, the detector solid angle is of no consequence to

Table 1. Electron-Excited EDS X-Ray Microanalysis of SRM 482 (Au-Cu)[a]

Cu SRM	Cu Conc	Rel Err %	Au SRM	Au Conc	Rel Err %	Total
0.198	0.198	0	0.801	0.790	−1.4	0.988
0.396	0.399	0	0.603	0.594	−1.6	0.993
0.599	0.605	1.0	0.401	0.402	0.1	1.007
0.798	0.797	−0.1	0.200	0.199	−1.2	0.996

[a] Values in mass fractions. Conc,, concentration; rel err, relative error; SRM, standard reference material.

the quantitative procedure. Both the beam-current uncertainty and the detector solid-angle uncertainty are effectively hidden by normalizing the analytical total to a predetermined value, e.g., unity when all constituents are measured. When the spectrum has been accumulated to the desired level of statistical precision, the analyst needs to specify only the beam energy and x-ray takeoff angle, and list of elements to be quantified (or this list can be derived from an automatic qualitative analysis). The software then proceeds to calculate the composition directly from the spectrum, as described below, and in the resulting output report the concentration is often specified to 3 or 4 significant figures. The measurement precision for each element, calculated from the integrated peak and background counts, is also reported. The precision is really only limited by the patience of the analyst and the stability of the specimen under electron bombardment, so that relative precision values below 1% (1 σ) can be readily achieved, even for minor constituents. While such excellent measurement precision is invaluable when sequentially comparing different locations on the same specimen, the precision is no indication of accuracy. When standardless analysis procedures are tested to produce error histograms like that in Figure 13A, the errors are found to be substantially larger (Newbury et al., 1995; Newbury, 1999).

First-Principles Standardless Analysis

Calculating an equivalent standard intensity from first principles requires solution of Equation 6 (Goldstein et al., 1992):

$$I_{\text{std}} = \left[(\rho N_0/A)\omega \int_{E_0}^{E_c} \frac{Q}{(dE/ds)} dE\right] Rf(\chi)\varepsilon(E)_{\text{v}} \qquad (6)$$

where the terms in brackets represent the excitation function: ρ is the density, N_0 is Avogadro's number, A is the atomic weight, ω is the fluorescence yield, Q is the ionization cross-section, dE/ds is the stopping power, E is the electron energy, E_0 is the incident-beam energy, and E_c is the critical excitation energy. The other terms correct for the loss of x-ray production due to electron backscattering (R), the self-absorption of x rays propagating through the solid [$f(\chi)$], and the efficiency of the detector, $\varepsilon(E_{\text{v}})$, as a function of photon energy, E_{v}. It is useful to consider the confidence with which each of these terms can be calculated.

Excitation. There are three critical terms in the excitation function: the ionization cross-section, fluorescence yield, and stopping power.

Ionization cross-section. Several parameterizations of the K-shell ionization cross-section are plotted in Figure 10. The variation among these choices exceeds 25%. While it is not possible to say that any one of these is correct, it is certain that they cannot all be correct. Moreover, because of the continuous energy loss in a solid target, the cross-section must be integrated from E_0 to E_c, through the peak in Q and the rapid decrease to $U = 1$ (where U is overvoltage $U = E_0/E_c$). This region of the cross-section is poorly characterized, so that it is difficult to choose among the cross section formulations based on experimental measurements. The situation for L- and M-shell cross-sections is even more unsatisfactory.

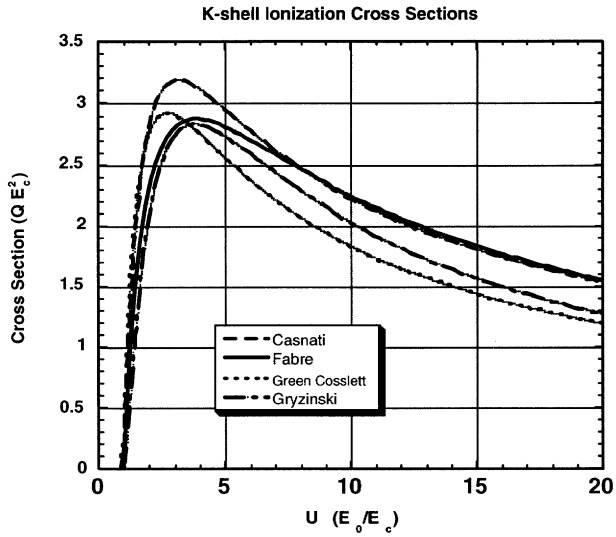


Figure 10. Ionization cross-sections (Casnati, S4; Fabre, S5; Green-Cosslett, S6; and Gryzinski, S7) as a function of overvoltage U, as formulated by different authors.

Fluorescence yield. Figure 11 plots various experimental determinations of the K-shell fluorescence yield. Again, a variation of >25% exists for many elements. The situation for L- and M-shell transitions is even less certain.

Stopping power. The classic Bethe formulation of the stopping power becomes inaccurate at low beam energies (<5 keV), and eventually with decreasing energy becomes

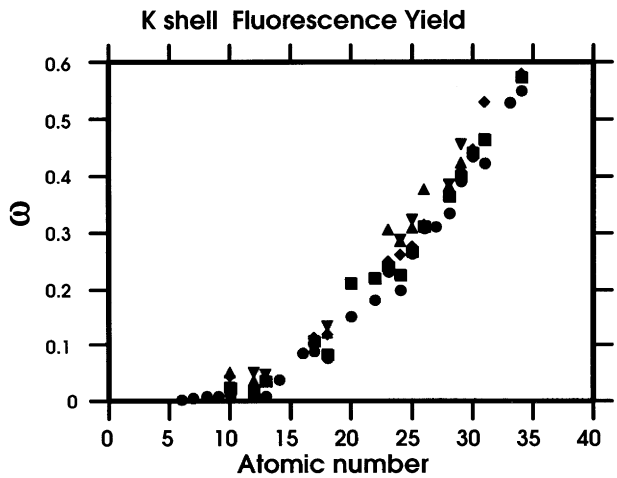


Figure 11. K-shell fluorescence yield as a function of atomic number. Symbols indicate measurements by different workers, as derived from the compilation in Fink et al. (1966).

physically unrealistic with a sign change. The accuracy of the stopping power matters for calculating Equation 5, because the cross-section must be integrated down to E_c, which for low-energy x-rays (e.g., C, N, O, F) involve electron energies in this regime. Several authors (e.g., Heinrich and Newbury, 1991) have suggested modifications to the Bethe formulation to correct for the low-beam-energy regime. Unfortunately, the base of experimental measurements necessary to select the best choice is just being developed.

Backscatter Loss. The backscatter loss correction factor R was initially formulated based upon experimental measurements of the total backscatter coefficient and the differential backscatter coefficient with energy (Heinrich, 1981). While careful, extensive measurements of total backscatter were available in the literature, the database of the differential backscatter coefficient with energy, a much more difficult experimental measurement, was limited to a few elements and was available at only one emergence angle. The development of advanced Monte Carlo simulations has permitted the rigorous calculation of R over all scattering angles and energy losses so that this factor is now probably known to an accuracy within a few percent across the periodic table and the energy range of interest (Heinrich and Newbury, 1991).

X-Ray Self-Absorption. The self-absorption of x rays in the hypothetical standard is calculated based upon the distribution in depth of x-ray production, a parameter that has been extensively studied by experimental measurement of layered targets and by Monte Carlo electron-trajectory simulation. The formulation of the absorption factor due to Heinrich and Yakowitz, as used in the NIST ZAF matrix correction, is employed (Heinrich, 1981). Fortunately, the absorption correction is generally small for the x rays of a pure element, so that at least for higher-energy characteristic x rays, e.g., photon energies >3 keV, there is little error in this factor. However, the self-absorption increases both as photon energy decreases and as incident electron energy increases, so that the error in calculating the intensity emitted from an element emitting low-energy photons, such as carbon, can be significant.

Detector Efficiency. The last term in Equation 5 is one of the most difficult with which to deal. In the traditional k value approach, the detector efficiency cancels quantitatively in the intensity ratio and can be ignored because the same x-ray peak is measured for the unknown and the standard under identical (or at least accurately reproducible) spectrometer conditions. When standardless analysis is performed, this cancellation can not occur because x-ray peaks of different energies are effectively being compared, and therefore accurate knowledge of the detector efficiency becomes critical. Detector efficiency is mainly controlled by absorption losses in the window(s) and detector structure. The expression for detector efficiency, Equation 7, consists of a multiplicative series of absorption terms for each component: detector window (denoted "win" in the subscripted terms in Equation 6), gold surface electrode ("Au"), semiconductor ("Si" or "Ge") dead layer ("DL"; actually a partially active layer below the electrode and the source of incomplete charge phenomena), and a transmission term for the detector thickness. Additionally, for most practical measurement situations there may be pathological absorption contributions from contamination on the detector crystal, usually arising from ice buildup due to pinhole leaks in the window or support ("ice"), and from contamination on the detector window usually deposited as hydrocarbons from the microscope environment ("con").

$$\begin{aligned}\varepsilon(E_v) = \exp[&(-\mu/\rho)_{win}\rho_{win}\mathrm{win} + (-\mu/\rho)_{Au}\rho_{Au}\mathrm{Au} \\ &+ (-\mu/\rho)_{Si}\rho_{Si}\mathrm{SiDL} + (-\mu/\rho)_{con}\rho_{con}t_{con} \\ &+ (-\mu/\rho)_{ice}\rho_{ice}t_{ice}\}]^*[1 - \exp(\mu/\rho)_{Si}\rho_{Si}t_{Si}] \quad (7)\end{aligned}$$

In Equation 7, the mass absorption coefficients are those appropriate to the photon energy, Ev, of interest. An example of the detector efficiency as a function of photon energy for several window materials is shown in Figure 3A. The choice of the window material has a strong effect on the detector efficiency for photon energies <3 keV, and accurate knowledge of the window and detector parameters is vital for accurate interelement efficiency correction across the working range of the detector, typically 100 to 12 keV. The pathological change in the detector efficiency with the accumulation of ice is illustrated in Figure 13B. The buildup of ice and other contaminants and the resulting loss in efficiency is referred to as "detector aging." Detector aging can result in significant loss of low-energy photons (<3 keV) relative to higher-energy photons (3 to 12 keV). Detector aging due to ice buildup can often be reversed by following the manufacturer's recommended procedure for warming the detector.

Fitted-Standards Standardless Analysis

The fitted-standards technique is the more widely used approach for implementing standardless analysis on commercial computer-assisted EDS analyzer systems. In the fitted standards technique, a suite of pure element standards covering K-, L-, and M-family x-rays is measured at one or more beam energies on an electron-beam instrument equipped with an EDS detector whose efficiency is known from independent measurements or at least can be estimated from knowledge of the detector construction. An example of such a measurement for a portion of the K-series from pure elements measured at 20 keV is shown in Figure 12A. In the fitted standards technique, missing elements can be calculated by mathematically fitting the available peaks and interpolating (e.g., in Fig. 12A, the intensity for gallium could be estimated by fitting the smoothly varying data and interpolating). From the smooth change in peak height with atomic number seen in Figure 12A, such an interpolation should be possible with reasonable accuracy. The situation is not as satisfactory in the L and M families, as illustrated in Figures 12B and 12C, because the cross-section/fluorescence yield

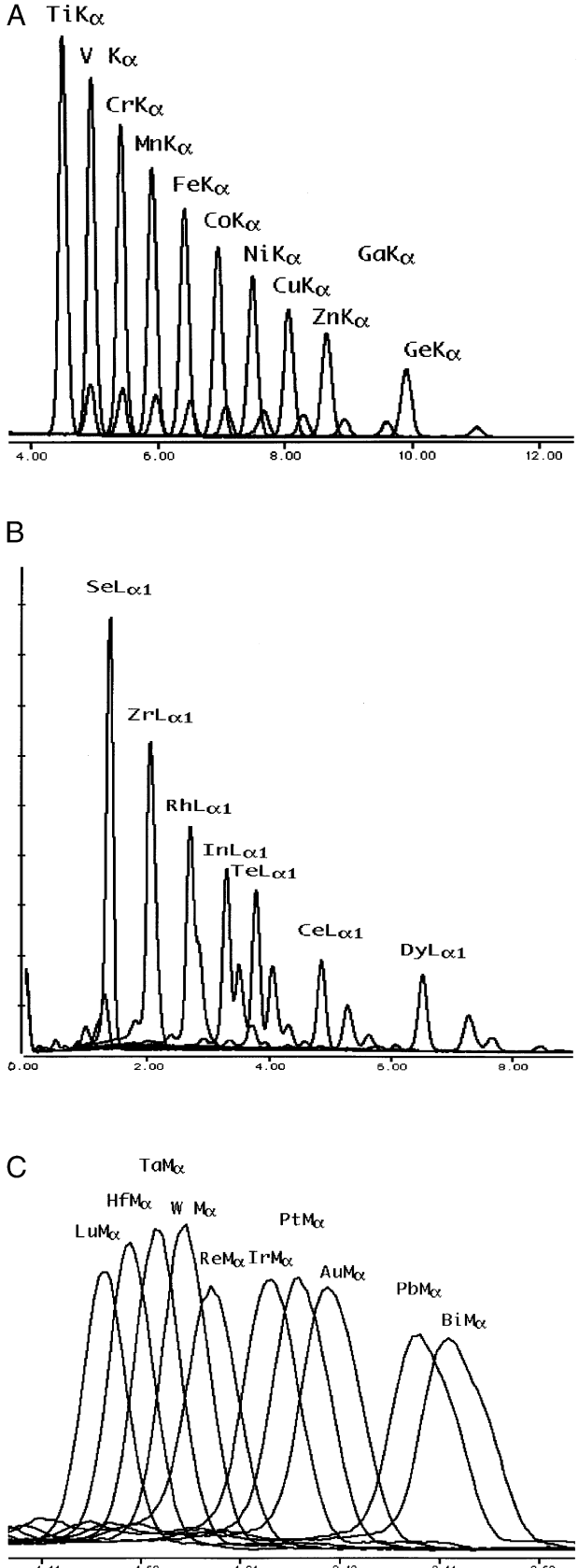


Figure 12. (**A**) K-family peaks from transition elements; E_0, 20 keV. (**B**) L-family peaks; E_0, 20 keV. (**C**) M-family peaks; E_0, 20 keV.

product is a much more complicated function of the atomic number.

If the analysis must be performed at a beam energy other than that of the spectral database, then Equation 6 must be used to shift the fitted standards intensities appropriately. Similarly, if a different EDS detector is used, the detector efficiency must be corrected for the differences in efficiency between the two detectors using Equation 7. The fitted-standards standardless procedure is expected to be more accurate than the "first-principles" standardless procedure because it is tied to actual experimental measurements that directly incorporate the effects of the cross-section/fluorescence yield product, at least over the range of the elements actually measured.

Testing the Accuracy of Standardless Analysis

The accuracy of standardless analysis procedures has been tested by carrying out analyses on microhomogeneous materials of known composition (Newbury et al., 1995): NIST Microanalysis Standard Reference Materials (SRM; Newbury, 1999), NIST Microanalysis Research Materials (RM), stoichiometric binary compounds (e.g., III-V compounds such as GaAs and II-VI compounds such as SrTe), and other materials such as ceramics, alloys, and minerals for which compositions were available from independent chemical analysis and for which microhomogeneity could be established (Newbury et al., 1995). Compositions were carefully chosen to avoid serious spectral overlaps (as in, e.g., PbS and MoS_2). Light elements such as boron, carbon, nitrogen, oxygen, and fluorine were also eliminated from consideration because of large errors due to uncertainties in mass absorption coefficients. In oxidized systems, the oxygen was calculated by means of assumed stoichiometry, but the resulting oxygen values were not included in the error histograms because of their dependence on the cation determinations.

Figure 13B shows an error histogram for the first-principles standardless analysis procedure embedded in the National Institute of Standards and Technology–National Institutes of Health Desktop Spectrum Analyzer (DTSA) x-ray spectrometry software engine. The error distribution shows symmetry around 0% error, but in comparing this distribution with that for the conventional standards/ZAF procedure shown in Figure 13A, the striking fact is that the error bins are 10 times wider for first-principles standardless analysis. Thus, the 95% error range is $\sim \pm 50\%$ relative rather than $\pm 5\%$ relative.

The error distribution for a commercial standardless procedure based upon the fitted-standards approach is shown in Figure 13C. This distribution is narrower than the first principles standardless approach, but the error bins are still 5 times wider than those of the conventional standards/ZAF procedure, so that the 95% error range is $\pm 25\%$ compared to $\pm 5\%$. It must be emphasized that this distribution represents a test of only one of the many implementations of standardless analysis in commercial software systems and that more extensive testing is needed.

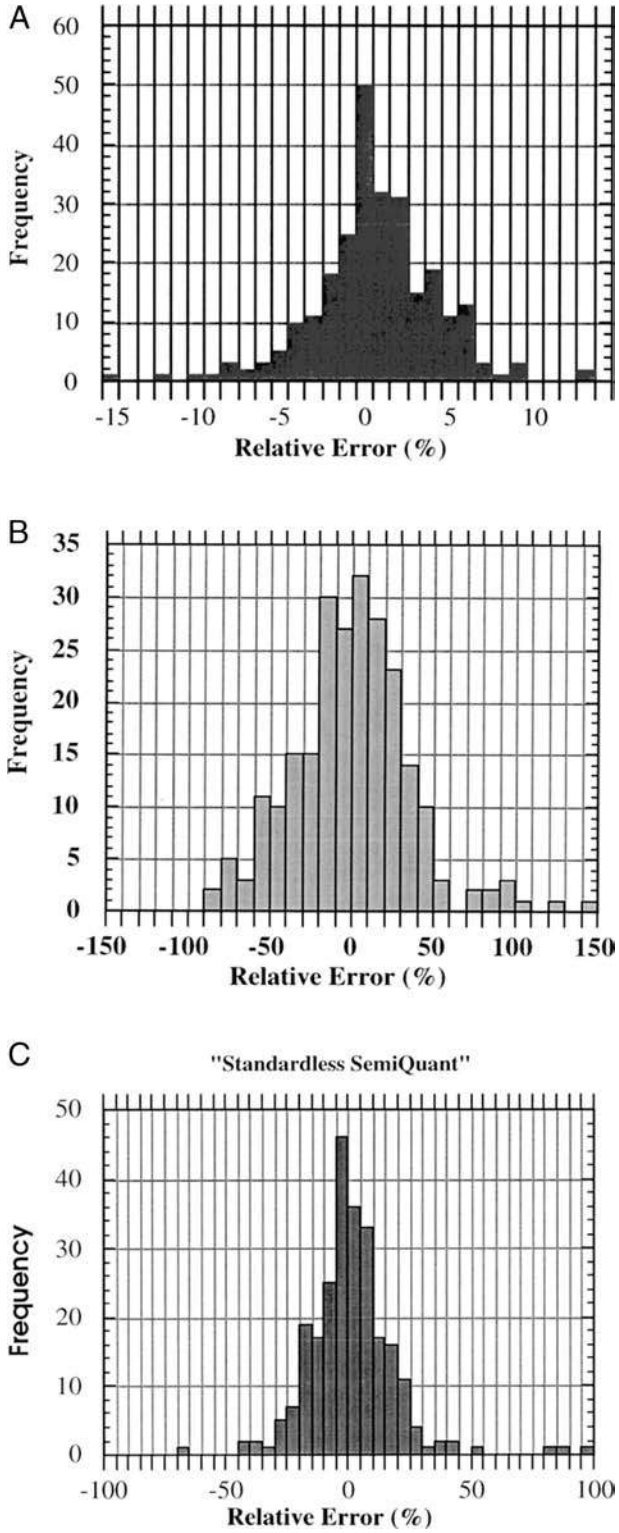


Figure 13. (**A**) Distribution of errors observed in the analysis of binary metal alloys against pure element standards (Heinrich-Yakowitz binary data ZAF; Heinrich, 1981). Beam energy, 20 keV. (**B**) Error distribution for the first-principles standardless analysis procedure embedded in the NIST-NIH Desktop Spectrum Analyzer x-ray spectrometry software system. Beam energy, 20 keV. (**C**) Error distribution for a fitted-standards standardless analysis procedure embedded in a commercial x-ray spectrometry software system (beam energy, 20 keV).

Using Standardless Analysis

Given the broad width of these error distributions for standardless analysis, it is clear that reporting composition values to 3 or 4 significant figures can be very misleading in the general case. At the same time, the error distributions show that there are significant numbers of analyses for which the relative errors are acceptably small, <10%. Usually, these analyses involve elements of similar atomic number (e.g., Cr, Fe, and Ni in stainless steel) for which the x-ray peaks are of similar energy and are therefore excited and measured with similar efficiency. Users of standardless analysis must be wary that any confidence obtained in such analyses does not extend beyond those particular compositions. The most significant errors are usually found when elements must be measured involving a mix of K-, L-, and M-shell x rays. The errors in the tails of the distributions are so large that the utility of concentration values obtained in this fashion is very limited.

Standardless analysis does have a legitimate value. If a microhomogeneous material with a known composition similar to the unknown specimen of interest is available, then the errors due to standardless analysis can be assessed and included with the report of analysis. For example, if the Fe-S system is to be studied, then analyzing the minerals pyrite (FeS_2) and troilite (FeS) would serve as a good test to challenge the standardless analysis procedure.

The analyst should never attempt to estimate relative concentrations merely by inspecting a spectrum. There are simply too many complicated physical effects of relative excitation, absorption, and detection efficiency to allow such a quantitative detail to be obtained from a casual inspection of a spectrum. Standardless analysis incorporates enough of the corrections to allow a sensible classification of the constituents of the specimen into broad categories, e.g.:

major: (>10 wt.%)
minor: (1 to 10 wt.%)
trace: (<1 wt.%)

In the absence of a known material to test a standardless analysis procedure, it is recommended that these broad classification categories be used instead of numerical concentration values, which may imply far more apparent accuracy than is justified and which may lead to a loss of confidence in quantitative electron-probe microanalysis when an independent test is conducted.

Limits of Detection with EDS X-Ray Microanalysis. The ultimate limit of detection is determined by statistical variance of the background counts within the peak region of interest. As such, the concentration limit of detection (C_{DL}) for a peak that does not suffer from interference from a nearby high-intensity peak depends on the peak-to-background ratio (Ziebold, 1967):

$$C_{DL} \geq 3.3a/[n\tau P(P/B)]^{0.5} \quad (8)$$

where a is a factor relating the concentration to the measured intensity ratio and is generally close to unity, n is

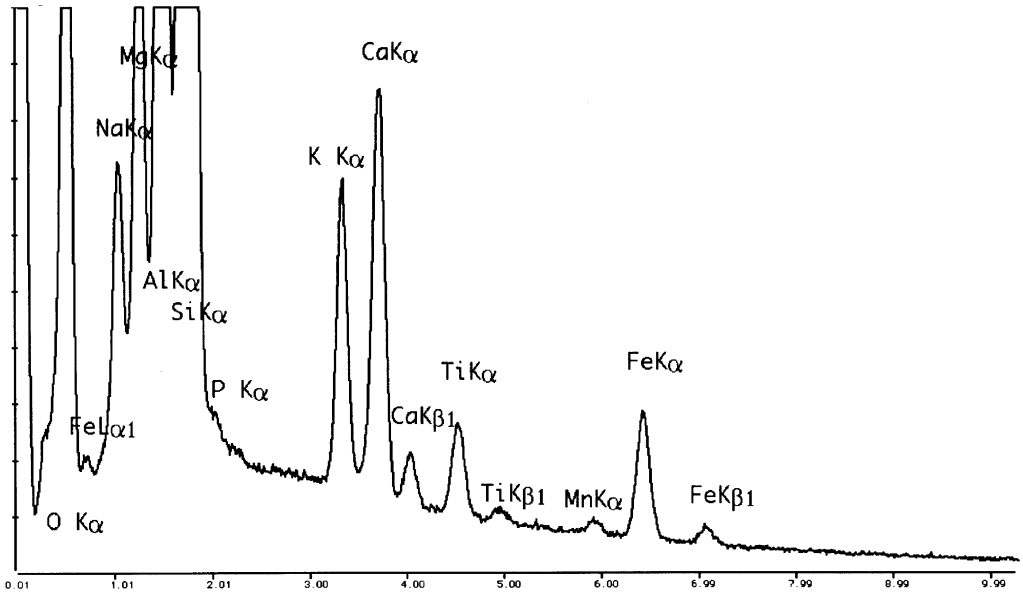


Figure 14. EDS spectrum of NIST glass K961; beam energy, 15 keV. Composition: O (0.470); Na (0.0297); Mg (0.0301); Al (0.0582); P (0.0022); K (0.0249); Ca (0.0357); Ti (0.012); Mn (0.0032); Fe (0.035) (mass fraction).

the number of replicate measurements, τ is the measurement time, and P and B are the peak and background count rates, respectively. While the detection limit can be lowered by using longer counting times, the appearance of this term within the square root limits the improvement. For practical counting conditions ($\tau < 1000$ s) and deadtime ($< 30\%$), the limit of detection for EDS varies from concentration levels of 0.0005 to 0.002 (mass fraction) depending on the analytical line and the matrix. An example of a spectrum showing minor and trace elements is presented in Figure 14. Note the obvious detection of Mn (0.0032 mass fraction), while P (0.0022) suffers severe interference from the major Si peak.

SAMPLE PREPARATION

The ideal form of the specimen for x-ray analysis with any type of primary excitation (electrons, photons, or ions) is a flat, mechanically polished surface with topography reduced below 50 nm (Goldstein et al., 1992). The requirement for flatness arises because of the strong absorption of x rays, particularly for photon energies <4 keV and increasing in severity as the photon energy decreases. Low-energy x rays, e.g., C-K, N-K, and O-K, are particularly strongly absorbed. Even with a flat surface, 90% or more of the x rays generated within the target are absorbed along the exit path toward the EDS detector. The effect of surface topography is to introduce an uncontrolled geometric variable that can modify the x-ray spectrum independent of the composition, which in conventional analysis is assumed to be the only variable. An example of the magnitude of the effect of anomalous absorption is shown in Figure 15, which shows two spectra obtained from a "123" $YBa_2Cu_3O_7$ high-T_c superconductor. The spectrum obtained with the beam placed on the polished surface shows significant x-ray counts at all photon energies, while the spectrum obtained in the crack shows significant attenuation of the low-energy photons, which is caused by the additional path length through the solid. Note that if only the "in-crack" spectrum were available, it would not be readily apparent that the anomalous absorption situation existed. The low-energy photon peaks from Y-L, Cu-L, and O-K can easily be detected, but their intensities are not representative of the polished material, and if quantification were attempted, the concentrations calculated from these peaks would be below the correct value by a large factor.

SPECIMEN MODIFICATION

Generally, EDS performed with electron bombardment does not lead to significant changes for most specimen compositions, such as metal alloys, ceramics, glasses, and minerals. Important exceptions occur, however, where charged ionic species can move under the influence of charge injected by the beam and for certain classes of specimens for which radiation damage can result in the generation and loss of volatile components (Joy et al., 1986; Goldstein et al., 1992).

Ionic Migration

Insulating materials can accumulate charge under electron bombardment, leading to the development of local fields strong enough to deflect the incident beam. To minimize the effects of charging on the primary beam, a thin ($\sim$ 10- to 20-nm) conducting coating such as carbon is typically applied to the surface of the specimen and this layer is connected to an electrical ground. However, beam electrons passing through this layer can accumulate within

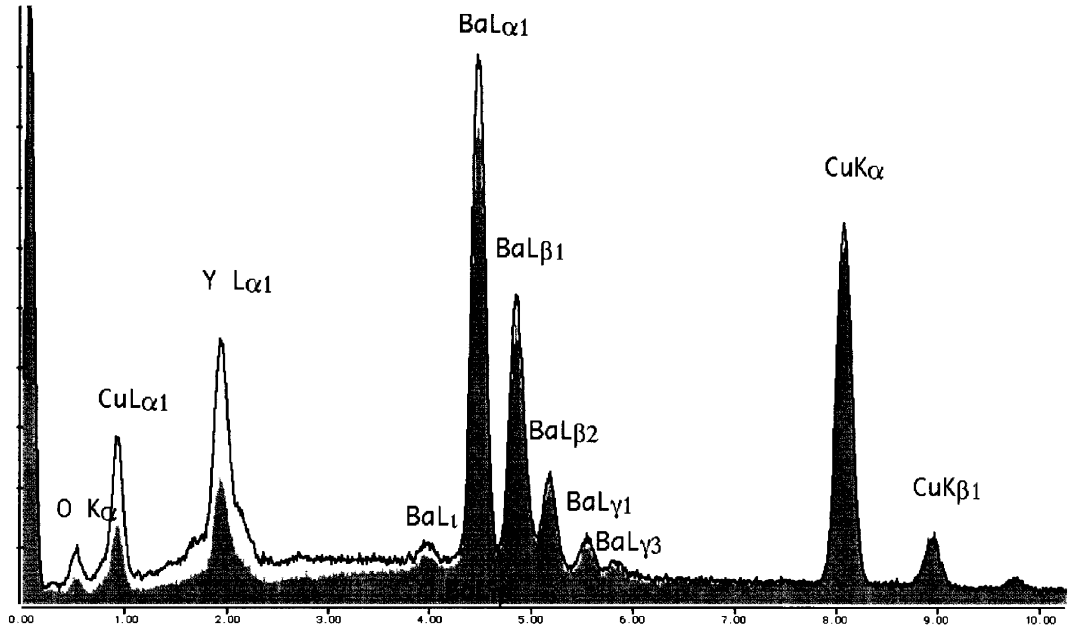


Figure 15. Comparison of spectra obtained on polished surface (trace) and in crack (solid gray). Specimen: $YBa_2Cu_3O_7$ high-T_c superconductor. Beam energy, 25 keV.

the specimen and create a significant internal electric field. Sodium, for example, is known to be capable of motion under such fields, which can be detected as a time dependency in the x-ray signal. The degree of sodium migration depends on the specimen composition.

Loss of Volatile Components

Some specimen compositions are sensitive to radiation damage, which can produce broken chemical bonds or free radicals, for example. Biological and organic specimens are particularly prone to the loss of water and other volatile components under electron bombardment. Even after chemical substitution to stabilize such specimens, significant modification may occur during electron bombardment. To correct for such mass loss in the quantification of minor and trace constituents, Hall (1968) developed a method of normalizing the characteristic peaks to a band of high-energy bremsstrahlung radiation that scaled with the total mass of the electron-excited volume.

PROBLEMS

Energy-dispersive x-ray spectrometry systems are capable of stable, long-term operation with a high degree of reproducibility. In the next section operational quality-assurance protocols will be described. However, occasional deviations will occur, even in an otherwise properly operating EDS system, and it is the responsibility of the user to recognize such pathological conditions. The careful user of EDS will develop the ability to recognizing such deviations. One of the best ways to do this is to record and archive spectra from high-purity elements and simple binary compounds and carefully study the forms of the spectra. Modern computer-assisted x-ray analyzers provide ready means for comparing spectra both visually and mathematically. An archive of reference spectra obtained on the local instrument is invaluable for rapid comparison with spectra obtained under current operating conditions.

Pathological Conditions

Ground Loops. The EDS detector-amplifier chain operates at extremely high gain and may therefore be sensitive to other sources of electromagnetic radiation. Improperly shielded cables may act as antennae, introducing spectral degradation through resolution loss, background distortion, false peaks, and other artifacts. Another source of interference arises from "ground loops"—AC currents flowing between two points at nominal ground potential. It is critical to prevent electrical contact between the EDS detector snout and the microscope column. A high value of electrical resistance should be maintained to isolate the EDS from the microscope. The EDS (as well as microscope power supplies) should all be connected to a common high-quality ground. An example of spectra obtained with and without a ground loop operating is shown in Figure 16. Note the high and complex background when the ground loop is present. It is highly recommended that when a new EDS system is being installed, the EDS performance first be checked on the bench (i.e, with the EDS system isolated from the microscope) with a radioactive source (e.g., ^{55}Fe). Upon installation on the microscope, the EDS system should be checked again with the radioactive source placed in the microscope specimen position. Ideally, these two spectra should be identical. Any degradation in EDS performance (resolution, deadtime, etc.) is probably due to ground loops and should be corrected by careful cable shielding and routing to avoid proximity to power supplies, etc. If no degradation is found, the microscope should then be fully activated with the vacuum system in full operation and the electron beam on but blanked to prevent it from hitting the specimen (the

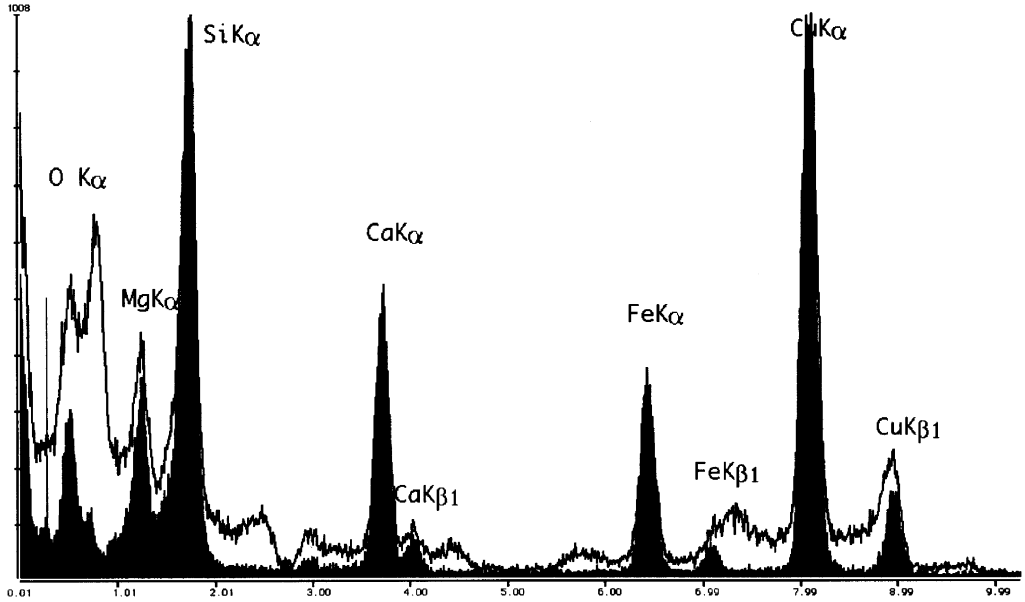


Figure 16. Pathological artifact: Action of ground loop in a low-resistance path between the EDS and the microscope chassis distorting the spectrum. Trace, spectrum with ground-loop active; solid filled areas, spectrum with ground loop eliminated.

^{55}Fe source). A third radioactive source spectrum should be obtained. If this spectrum is identical to the previous two conditions, then the EDS system is free of ground loops.

Light Leakage. Under electron bombardment, certain specimens (especially minerals, ceramics, and some semiconductors) emit light, a process called cathodoluminescence. The vacuum-isolation window of the EDS is coated with a thin (~20-nm) aluminum layer to reflect visible light. This layer may not provide sufficient attenuation of the light, or there may be leakage through pinholes in the coating. Depending on the light intensity, the spectrum may undergo energy shift (miscalibration) or resolution degradation, and in extreme cases the peaks may undergo enormous distortion, as illustrated in Figure 17 for ZnS. Other sources of light can include inspection lights and the infrared source chamber TV camera. These sources should be shut off during EDS operation. A symptom of light leakage into the EDS detector is anomalously high deadtime, which can reach 100% (no x-ray photons processed) if the light source is strong enough. Note that after the EDS detector is exposed to a high level of light, it may take several minutes to return to stable operation

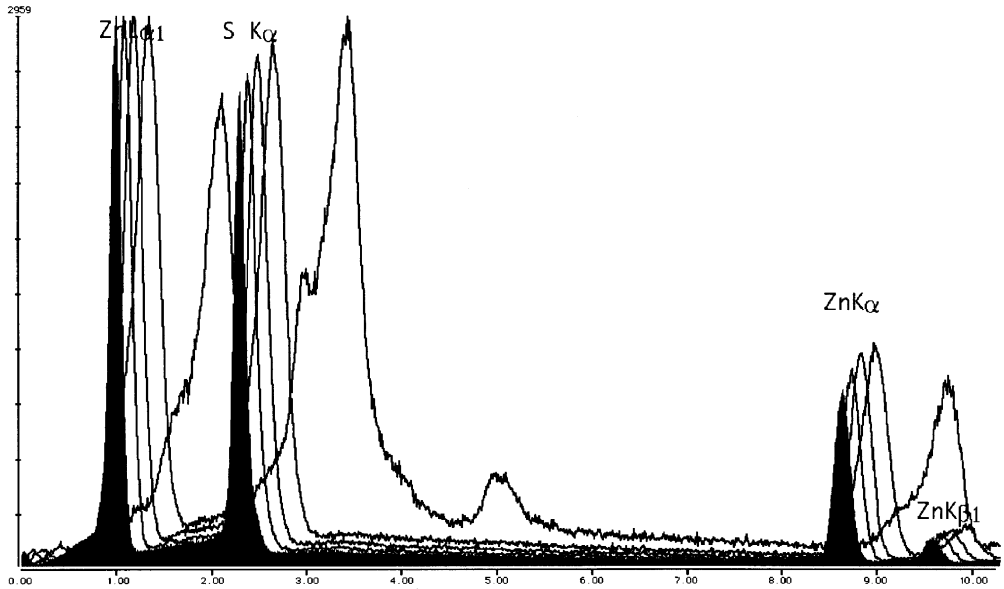


Figure 17. Pathological artifact: Effect of cathodoluminescent light leakage into detector. Specimen, ZnS; beam energy, 20 keV. Solid filled areas, reference ZnS; traces, increasing cathodoluminescence output achieved by increasing scanned area at fixed current.

and can be subject to severe spectral artifacts during this recovery.

Ice Accumulation. Figure 3B shows the calculated decrease in low-energy detection efficiency with the build-up of ice. To monitor this phenomenon in practice, a pure nickel target can provide a spectrum in which the Ni Kα radiation (7.477 keV) is sufficiently energetic that it is unaffected by detector windows and ice layers, while the low-energy Ni Lα (0.849 keV) suffers severe absorption passing through ice, since it is located above the O K edge (0.531 keV). A plot of the Ni Lα/Ni Kα ratio is an effective monitor of detector performance. When the Ni Lα/Ni Kα ratio has fallen by a significant percentage (e.g., arbitrarily 10%, to be determined by the user's established quality-assurance plan) relative to the performance of the new detector, the EDS manufacturer's protocol for conditioning the detector should be followed scrupulously.

Distortions of Low-Energy Peaks. Low-energy photons, less than 1 keV, are measured with charge pulses that are very close in magnitude to the system noise. Several artifacts are observed in this energy range even when the detector is performing optimally for photons above 1 keV in energy. Peaks are distorted from the ideal Gaussian shape by the effects of incomplete charge collection on the low- energy side. This effect depends entirely on the detector construction and cannot be modified. The pulse-pileup inspector is virtually ineffective in this energy range. Pulse pileup effects can be severe, with the result that the peak position and shape may change with input count rate, which can be controlled by the analyst. Both peak shift and peak distortion at high deadtime (50% versus 8%) can be seen in the carbon spectra shown in Figure 18.

Stray Radiation. The operating environment of the EDS system on an electron microscope can be severe. Beam

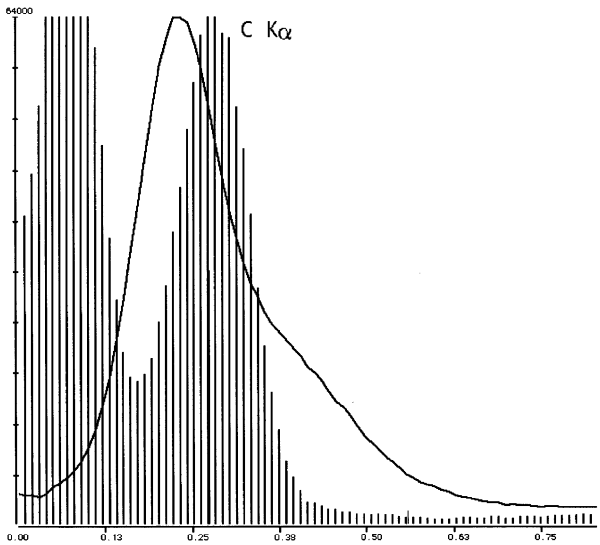


Figure 18. Pathological artifact: Peak shift induced by count-rate effects. Barred spectrum taken at low deadtime (8%); line spectrum taken at high deadtime (50%). Beam energy, 15 keV.

electrons will backscatter from the specimen, and additionally may scatter off microscope components such as apertures and stage materials, especially in a high-energy analytical electron microscope. The EDS is equipped with an "electron trap" (actually a permanent magnet deflector) to prevent energetic electrons (whose prevalence actually exceeds the rate of x-ray production by a large factor) from reaching the detector and dominating or distorting the spectrum. However, these remote sources of electrons can create anomalous contributions to the x-ray spectrum. The EDS is also commonly equipped with a collimator to restrict the contributions of x rays from the remote electron scattering, but typically such collimation is limited to an area with linear dimensions of millimeters centered around the on-axis beam impact point on the specimen. Electrons not in the focused beam that strike the specimen within this collimation acceptance disk will contribute to the spectrum. To assess the magnitude of stray electron contributions, it is highly recommended that an "in-hole" spectrum be obtained. For the solid specimen SEM, a "Faraday cup" should be constructed using materials not present in the microscope construction. A polished block of titanium is a good choice. Into this metal block should be drilled a blind hole and over this hole a microscope aperture with a small opening (<100 μm) should be placed. If possible, the aperture metal should be different from that actually used for apertures in the microscope. Three spectra should then be recorded for the same live time: (1) beam placed on the titanium; (2) beam placed on the aperture; and (3) beam placed in the blind hole. Ideally, the in-hole spectrum should have no counts for either characteristic or bremsstrahlung x rays. It is likely that a small component of x rays representing nonbeam electrons striking the aperture and/or metal block will be detected. This contribution should be $<1\%$ of the intensity recorded when the beam is placed on these materials.

Note that this in-hole procedure detects only electrons not in the focused beam. In normal operation with the beam striking the specimen, significant numbers of backscattered electrons will be generated that will strike the polepiece above the specimen and can rescatter as an uncollimated radiation source raining down on the specimen. To examine this effect, a second specimen should be fabricated consisting of microscopic particles or fibers embedded in a silver-doped conducting epoxy and polished. Choose a particle or fiber whose diameter exceeds the expected interaction volume by at least a factor of 4: e.g., at least 20-μm diameter if the electron x-ray range is 5 μm. With the beam placed on the center on the particle, which should have a composition that does not contain silver, examine a spectrum accumulated with at least 200,000 counts in the full energy range. Any silver radiation must arise from rescattered backscattered electrons. This rescattering component might be reduced if the polepiece can be shielded with a plate consisting of a light element, such as carbon. It is very important never to use carbon paint.

Electron Penetration. Energetic electrons that penetrate the detector window and strike the detector will be measured as if they were x-ray photons, leading to anomalous

spectral background shape (Fiori and Newbury, 1978). To avoid or at least minimize this problem, a permanent-magnet deflector is usually placed in the collimator assembly. During very-high-beam-energy operation in the AEM (100 keV or more), electron penetration into the detector may still occur, especially during surges in beam current that can happen during changes in the operating parameters of the electron lenses, especially under computer automation. Such electron penetration may actually damage the EDS detector.

PROTOCOLS

Measurement protocols have been partially described as part of the previous sections, e.g., the recommendations for testing a new EDS detector system for ground loops, stray radiation, etc., and the recommendations for testing qualitative-analysis procedures. In this section, protocols for quality assurance in operating an EDS system will be summarized.

Quality-assurance operations for an EDS system operating on an electron microscope takes place at several time intervals: once per session, weekly, and long term.

Session

At the start of each measurement session, the analyst must go through a checklist of operational parameters to ensure that the system is operating in the configuration that is required. This is especially true if the facility has multiple users. In the following discussion, please note that terms are used which are as generic as possible to describe the functions. There is no standardization for these terms, and when the historical range of EDS systems is considered, these functions may be implemented with controls ranging from hardware to full software selection and control.

The key elements on this checklist are as follows.

Resolution (Count-Rate Range). The longer the time spent processing a pulse, the more precisely the energy can be determined, and thus the narrower the peak and the better the resolution. Selection of this parameter (e.g., specified as "shaping time," "amplifier time constant," "count rate range," and "resolution") is made depending on the analytical problem. Generally, one will wish to accumulate as many x-ray counts as possible in the available counting time, so it is desirable to operate at the highest allowable count rate and then select a beam current/detector solid angle to achieve that count rate, recognizing that the EDS will operate at the poorest resolution. However, if the elements of interest produce interfering peaks, especially if an element of interest is present as a minor or trace constituent whose peak is close to the intense peak of a major constituent, then the poor-resolution performance at high count-rate range may preclude achieving an adequate detection limit, and it may be necessary to operate at high resolution and a lower limiting count rate. Whatever the choice, the analyst must be consistent when quantitative procedures are followed with archived standards.

EDS Spectrum Channel Width and Number. The width of the individual channel of the EDS histogram can be selected, as well as the number of channels. Typical width choices are 5, 10, 20 and 40 eV. It is important to have as many channels as possible to define the x-ray peak, so a choice of 10 or 5 eV is desirable. If possible, it is also important to record the entire x-ray spectrum up to the beam energy, at least in the SEM case where E_0 is typically between 5 and 30 keV. Thus, it is desirable to record 2048 or 4096 channels. Given the continuing drastic decreases in the cost of computer memory and mass storage, it is advisable to record and archive the complete spectrum from every measurement location.

Calibration. The calibration should be checked on a specimen with characteristic peaks in the 1- to 2-keV and 8- to 10-keV ranges, e.g., Cu Kα (8.04 keV) and Cu Lα (0.928 keV). The calibration should be established within 10 eV of the reference value. After these endpoints have been fixed, several intermediate peaks should be checked, e.g., Si (1.740 keV), Ti (4.508 keV), and Fe (6.400 keV). Noting that the low-energy photon peaks (<1 keV) are likely to be out of calibration, it is useful to record pure carbon (0.282 keV) and an oxygen-containing target, e.g., Al_2O_3 (O = 0.523 keV).

Specimen Position. The first condition to define the detector geometric efficiency (solid angle) is the specimen position. In electron beam systems with an optical microscope, the specimen can be accurately positioned because of the shallow depth of focus of the optical microscope. If only SEM imaging is available, then to position the specimen relative to the EDS in terms of the working distance, the objective lens current can be used to select a particular lens strength. By using a large final aperture to minimize the depth of focus, the specimen position can then be selected with the z motion (along the beam) of the stage.

Detector Solid Angle. The absolute x-ray intensity depends upon the electron dose (current × live time), the detector quantum efficiency, and the detector solid angle. Many detectors are capable of mechanical motion along the axis of the takeoff angle. The solid angle Ω of collection depends on the detector area A (as collimated) and the detector-to-specimen distance, r:

$$\Omega = A/r^2 \tag{9}$$

Because of the strong dependence of the solid angle on the detector-to-specimen distance, it is critical to have a method to accurately reset the detector position. Typically a ruler scale is provided for coarse adjustments, but this may not be adequate for the most accurate work. If the beam current and specimen position can be set accurately and reproducibly, fine adjustment of the solid angle can be based upon the absolute intensity measured in a high-energy x-ray peak, e.g., Cu Kα.

Weekly

The performance of the deadtime correction circuit should be checked by using the beam current (as measured in a

Faraday cup) to independently control the input count rate. The output count in a defined live time should scale with the electron dose (beam current) at least up to an indicated deadtime of 50%, and preferably to even higher deadtime. Detector aging should be monitored with a spectrum of Ni.

Long Term

EDS systems are capable of a high degree of stability provided they are continuously maintained under power. When powered up from a cold start, several hours are typically needed to establish thermal equilibrium in all electronic circuits and thus achieve a stable operating condition. It is therefore not advisable to power the EDS system down as part of routine operation but rather to maintain power at all times. When such a shutdown must be carried out, it is important to reestablish thermal equilibrium before resuming measurement operations. All parameters then should be checked, and adjusted if necessary: resolution, calibration, and deadtime correction.

LITERATURE CITED

Castaing, R. 1951. Application of Electron Probes to Metallographic Analysis. Ph.D. dissertation, University of Paris.

Fink, R. W., Jopson, R. C., Mark, H., and Swift, C. D. 1966. Atomic fluorescence yields. *Rev. Mod. Phys.* 38:513–540.

Fiori, C. E., Myklebust, R. L., and Gorlen, K. E. 1981. Sequential simplex: A procedure for resolving spectral interference in energy dispersive X-ray spectrometry. *In* Energy Dispersive X-ray Spectrometry (K. F. J. Heinrich, D. E. Newbury, R. L. Myklebust, and C. E. Fiori, eds.). pp. 233–272. National Bureau of Standards Publication 604, U.S. Government Printing Office, Washington, D.C.

Fiori, C. E. and Newbury, D. E. 1978. Energy dispersive x-ray spectrometry in the scanning electron microscope. *Scanning Electron Microsc.* 1:401–422.

Fiori, C. E., Swyt, C. R., and Myklebust, R. L. 1992. Desktop Spectrum Analyzer. Office of Standard Reference Data, National Institute of Standards and Technology, Gaithersburg, Md.

Fitzgerald, R., Keil, K., and Heinrich, K. F. J. 1968. Solid-state energy-dispersion spectrometer for electron-microprobe X-ray analysis. *Science* 528:159–160.

Goldstein, J. I., Newbury, D. E., Echlin, P., Joy, D. C., Romig, A. D. Jr., Lyman, C. E., Fiori, C., and Lifshin, E. 1992. Scanning Electron Microscopy and X-ray Microanalysis. Plenum, New York.

Hall, T. 1968. Some aspects of the microprobe analysis of biological specimens. *In* Quantitative Electron Probe Microanalysis (K. F. J. Heinrich, ed.) pp. 269–299. U.S. Dept. of Commerce, Washington, D.C.

Heinrich, K. F. J. 1981. Electron Beam X-ray Microanalysis. Van Nostrand–Reinhold, New York.

Heinrich, K. F. J. and Newbury, D. E. (eds.). 1991. Electron Probe Quantitation. Plenum, New York.

Joy, D. C., Romig, A. D. Jr., and Goldstein, J. I. 1986. Principles of Analytical Electron Microscopy. Plenum, New York.

Mott, R. B. and Friel, J. J. 1995. Improved EDS performance with digital pulse processing. *In* X-ray Spectrometry in Electron Beam Instruments (D. B. Williams, J. I. Goldstein, and D. E. Newbury, eds.). pp. 127–157. Plenum, New York.

Myklebust, R. L., Fiori, C. E., and Heinrich, K. F. J. 1979. FRAME C: A Compact Procedure for Quantitative Energy-Dispersive Electron Probe X-ray Analysis. National Bureau of Standards Technical Note 1106, U.S. Dept. of Commerce, Washington, D.C.

Newbury, D. E. 1999."Standardless" quantitative electron-excited x-ray microanalysis by energy-dispersive spectrometry: what is its proper role? *Microscop. Microanalys.* 4:585–597.

Newbury, D. E., Swyt, C. R., and Myklebust, R. L. 1995. 'Standardless" quantitative electron probe microanalysis with energy-dispersive x-ray spectrometry: Is it worth the risk? *Anal. Chem.* 67:1866–1871.

Small, J. A., Leigh, S. D., Newbury, D. E., and Myklebust, R. L. 1987. Modeling of the bremsstrahlung radiation produced in pure-element targets by 10-40 keV electrons. *J. Appl. Phys.* 61:459–469.

Williams, D. B., Goldstein, J. I., and Newbury, D. E. (eds.). 1995. X-ray Spectrometry in Electron Beam Instruments. Plenum, New York.

Ziebold, T. O. 1967. Precision and sensitivity in electron microprobe analysis. *Anal. Chem.* 39:858–861.

KEY REFERENCES

Goldstein et al., 1992. See above.

Comprehensive reference to x-ray microanalysis procedures for SEM.

Heinrich and Newbury, 1991. See above.

Comprehensive treatment of leading topics in microanalysis research.

Joy et al., 1986. See above.

Comprehensive reference to x-ray microanalysis procedures for AEM.

Williams et al., 1995. See above.

Comprehensive treatment of leading topics in x-ray spectrometry.

DALE E. NEWBURY
National Institute of Standards and Technology
Gaithersburg, Maryland

AUGER ELECTRON SPECTROSCOPY

INTRODUCTION

Auger electron spectroscopy (AES) is a powerful technique for determining the elemental composition of the few outermost atomic layers of materials. Surface layers often have a composition that is quite different from the bulk material, due to contamination, oxidation, or processing. In the most commonly used form of AES, a specimen is bombarded with electrons having an energy between 3 and 30 keV, resulting in the ejection of core-level electrons from atoms to a depth up to $\sim$1 μm. The resulting core vacancy can be filled by an outer-level electron, with the excess energy being used to emit an x ray (the principle of electron probe microanalysis) or another electron (the principle of Auger electron spectroscopy) from the atom. This emitted electron is called an Auger electron and is

named after Pierre Auger who first observed such events in a cloud chamber in the 1920s (Goto, 1995). AES is a surface-sensitive technique, due to the strong inelastic scattering of low-energy electrons in specimens. Auger electrons from only the outermost few atomic layers are emitted from the specimen without energy loss, and contribute to the peaks in a spectrum. Auger electrons that have lost energy in escaping from the specimen will appear as an additional background signal at lower kinetic energies. Obviously, hydrogen and helium cannot be detected, as three electrons are needed for the Auger process. The Auger electron kinetic energies are characteristic of the material, and the measurement of their kinetic energies is used to identify the elements that produce these Auger electrons. As the Auger electron kinetic energies depend on the binding energies of the electron levels involved, changes in surface chemistry (such as oxidation) can produce detectable shifts in Auger kinetic energies, thereby providing useful information about the surface chemistry as well. Auger electrons emitted from the specimen will appear as peaks on a continuous background of secondary electrons and backscattered electrons. Secondary electrons are usually defined as the electron background below 50 eV, and backscattered electrons are those in the background with energies from 50 eV up to the incident beam energy. The concentrations of elements detected can be determined from the intensities of the Auger peaks. An Auger spectrum is usually a plot of the number of electrons detected as a function of kinetic energy, but sometimes it is displayed as the first derivative of the number of electrons emitted, to enhance the visibility of the Auger electrons and to suppress the continuous background of secondary and backscattered electrons. Electron beams can be focused to small diameters, thereby allowing the composition of small areas ($\sim$50 nm and below) on a surface to be determined. This is often called point analysis. The electron beam can be defocused or rastered over small areas to reduce possible electron beam damage where this might be a problem. Analysis can also be performed at preselected points or areas on the specimen. The electron beam can alternatively be scanned in a straight line across part of the specimen surface and Auger data can be acquired as a function of beam position, resulting in what is called an Auger line scan. Auger maps can also be measured, showing the variation in elemental composition (and concentration) across a region of the surface; this is referred to as scanning Auger microscopy (SAM). Variation in composition with depth can be determined by depth profiling, which is usually accomplished by continuously removing atomic layers by sputtering with inert gas ions while monitoring the Auger signals from the newly created surfaces. For most elements, the detection limit with AES is between 0.1 and 1 atom %.

Competitive and Related Techniques

Methods for surface analysis that are alternatives to AES include x-ray photoelectron spectroscopy (XPS), which is also often called electron spectroscopy for chemical analysis (ESCA); low-energy ion scattering (LEIS), which is also called ion scattering spectroscopy (ISS); and secondary ion mass spectrometry (SIMS). XPS/ESCA uses soft x rays to eject electrons from atoms. Compared with AES, the interpretation of XPS/ESCA spectra is generally simpler; it is more easily quantifiable and easier to use when studying insulators, and produces less beam damage. The detection limit of XPS/ESCA is similar to that of AES, but the highest spatial resolution is a few micrometers (as opposed to $\sim$50 nm with AES). Auger peaks will also appear in XPS spectra. In some cases, energetic ions can also produce Auger electrons. ISS is the most surface sensitive of these techniques; low-energy inert gas ion beams measure just the outermost atomic layer; but it is not generally useful for identifying unknown elements at the surface due to its relatively poor specificity. Of these techniques, SIMS has the highest detectability as well as high elemental specificity; it is also capable of detecting atomic parts per million and below. In SIMS, the ions that are sputtered from the surface are detected in a mass spectrometer. Static SIMS is used for identifying the surface composition, whereas dynamic SIMS is used for depth profiling. SIMS can detect hydrogen as well as isotopes of the elements, which AES cannot. SIMS can be quantified for simple, well-characterized specimens, but is not routinely quantifiable for complex surface compositions. A variation of SIMS is sputtered neutral mass spectrometry (SNMS), where the neutral species that are emitted by sputtering are post-ionized, and then measured with a mass spectrometer.

Surface analysis systems cost several hundred thousand U.S. dollars. Systems can be stand-alone (e.g., AES only) or can incorporate more than one technique (such as AES and XPS). Analysis is performed in a vacuum chamber at pressures typically of the order of 10^{-18} Pa. A vacuum is needed since electron beams are involved, and ultrahigh vacuum ($< 10^{-7}$ Pa) is required so the surface composition does not change during analysis. There are several manufacturers of Auger systems, and manufacturers usually offer more than one model. Manufacturers include JEOL (*http://www.jeol.com*), Omicron Associates (*http://www.omicron-instruments.com*), Physical Electronics (*http://www.phi.com*), Staib Instruments (*http://www.staib-instruments.com)*, and VG Scientific (*http://www.vgscientific.com*). These companies, and many others, also manufacture components such as electron sources, electron kinetic energy analyzers, vacuum systems, and systems for computerized data acquisition and processing, which can be assembled to make a customized system. There are several commercial companies that provide analytical services using AES, and costs are typically 1000 to 3000 U.S. dollars per day. Some commercial companies in the United States are Anderson Materials Evaluation (*http://www.andersonmaterials.com*), Charles Evans & Associates (*http:// www.cea.com*), Geller Micro Analytical Lab (*http:// www.gellermicro.com*), and Physical Electronics (*http://www.phi.com*). Samples can usually be mounted and inserted into the analysis chamber in $\sim$30 min. Simple analysis can be performed within another 30 min, whereas depth profiling and Auger mapping can take longer.

Surface topography can be measured using optical microscopy (OPTICAL MICROSCOPY and REFLECTED-LIGHT

OPTICAL MICROSCOPY); scanning electron microscopy (SEM; SCANNING ELECTRON MICROSCOPY); scanning tunneling microscopy (STM; SCANNING TUNNELING MICROSCOPY); and atomic force microscopy (AFM). Scanning Auger microscopy systems have SEM capability, and this is used for locating regions on the surface of the specimen to be studied.

PRINCIPLES OF THE METHOD

An example of the Auger process is shown in Figure 1. With AES, the electron energy levels are denoted by the x-ray notation, K, L, M, etc., corresponding to the principal quantum number 1, 2, 3, etc. Subscripts are based on the various combinations of the orbital angular momentum 0, 1, 2, 3, etc. and the electron spin $+1/2$ or $-1/2$, and have the values 1, 2, 3, 4, 5, etc. The K level is not given a subscript. The L levels are L_1, L_2, and L_3, the M levels are M_1, M_2, M_3, M_4, and M_5, etc. Sometimes the energy levels are close in energy and not resolvable in the measurement, and are then designated as $L_{2,3}$, $M_{2,3}$, $M_{4.5}$, etc. In the example shown in Figure 1, the K and L levels in the atom are shown to be fully occupied by electrons (Fig. 1, panel A). In this example, the initial ionization occurs in the K level (Fig. 1, panel B). Following relaxation and the emission of an Auger electron, the atom has two vacancies in outer levels, the L_2 and the L_3 levels in this example (Fig. 1, panel C). This process can be thought of as consisting of the following processes. (1) An electron drops down from the L_2 level to fill the K level vacancy. (2) Energy equal to the difference in binding energies between the K and the L_2 levels is produced. (3) This energy is sufficient to remove an electron from the L_3 level of the atom, with the excess energy being the energy of the Auger electron. The three atomic levels involved in Auger electron production are used to designate the Auger transition. The first letter designates the shell containing the initial vacancy and the last two letters designate the shells containing electron vacancies created by Auger emission. The Auger electron produced in the example shown in Figure 1 is therefore referred to as a KL_2L_3 Auger electron. When a bonding electron is involved, the letter V is sometimes used (e.g., KVV and LMV). Coupling terms for the electronic orbital angular momentum and spin momentum may also be added where known, e.g., $L_3M_{4,5}M_{4,5}$; ^1D. More complicated Auger processes can also occur; e.g., if an atom is doubly ionized before Auger emission, it will be triply charged after emission (Grant and Haas, 1970). In such cases, the Auger transition can be designated by separating the initial and final states by a dash, e.g., LL-VV would be used for an atom that is doubly ionized in the L shell before Auger emission, and that uses two electrons from the valence shell for the Auger emission. When an Auger relaxation process involves an electron from the same principal shell as the initial vacancy, e.g., L_1L_2M, it is sometimes referred to as a Coster-Kronig transition. If both electrons involved in relaxation are from the same principal shell as the initial vacancy (e.g., $N_5N_7N_7$), it can be called a super Coster-Kronig transition. In some cases, such as MgO, electrons from neighboring atoms can also be involved in the Auger process and are referred to as cross transitions (Janssen et al., 1974). After the initial Auger electron is produced, further Auger relaxations will occur to fill the electron vacancies, resulting in a cascade of Auger electron emissions as vacancies are filled by outer electrons.

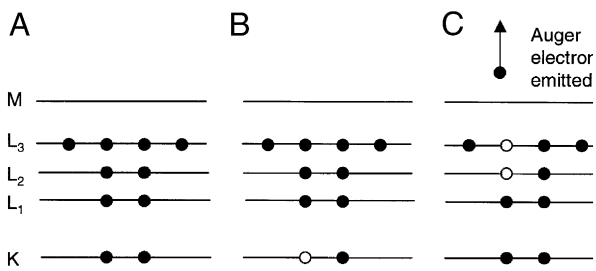


Figure 1. Schematic diagram illustrating the KL_2L_3 Auger process. Atom showing: (**A**) electrons present in filled K and L levels before an electron is removed from the K level; (**B**) after removal of an electron from the K level; and (**C**) following the Auger process, where a KL_2L_3 Auger electron is emitted. In (c), one L-level electron fills the K vacancy and the other L-level electron is ejected due to the energy available on filling the K level.

The energy of the Auger electron will depend on the binding energies, E_b, of the levels involved (and relaxation effects) and not on the energy of the incident beam. In its simplest form, the kinetic energy, E_k, of the Auger electron KL_2L_3 is given approximately by

$$E_k(KL_2L_3) \approx E_b(K) - E_b(L_2) - E_b(L_3) \qquad (1)$$

This equation neglects contributions from the interaction energy between the holes in the L_2 and L_3 levels, as well as intra-atomic and extra-atomic relaxation. Therefore, each element (except hydrogen and helium) will have a characteristic spectrum of Auger electrons, and this forms the basis for qualitative analysis. The intensity of the Auger electrons emitted forms the basis for quantitative analysis. Since binding energies (and relaxation effects) depend on the chemical environment of atoms, information about the chemical environment can often be obtained by studying changes in the kinetic energies of the Auger electrons, which are typically a few electron volts (Haas and Grant, 1969; Haas et al., 1972a,b). Auger transitions involving valence electrons can also exhibit changes in line shape (Haas and Grant, 1970; Haas et al., 1972b). In solids, after an Auger electron is produced, it has to escape from the specimen and pass through the energy analyzer to be measured. All elements except hydrogen and helium can produce Auger electrons, and each has at least one major Auger transition with a kinetic energy below 2000 eV. In AES systems, the electron energy analyzers typically measure electron kinetic energies up to 2000 or 3000 eV. The dominant Auger transitions are KLL, LMM, MNN, etc., with combinations of the outermost levels producing the most intense peaks. These Auger electrons have inelastic mean free paths of only a few atomic layers (Seah and Dench, 1979), and make AES a surface-sensitive technique. Auger electrons produced deeper in

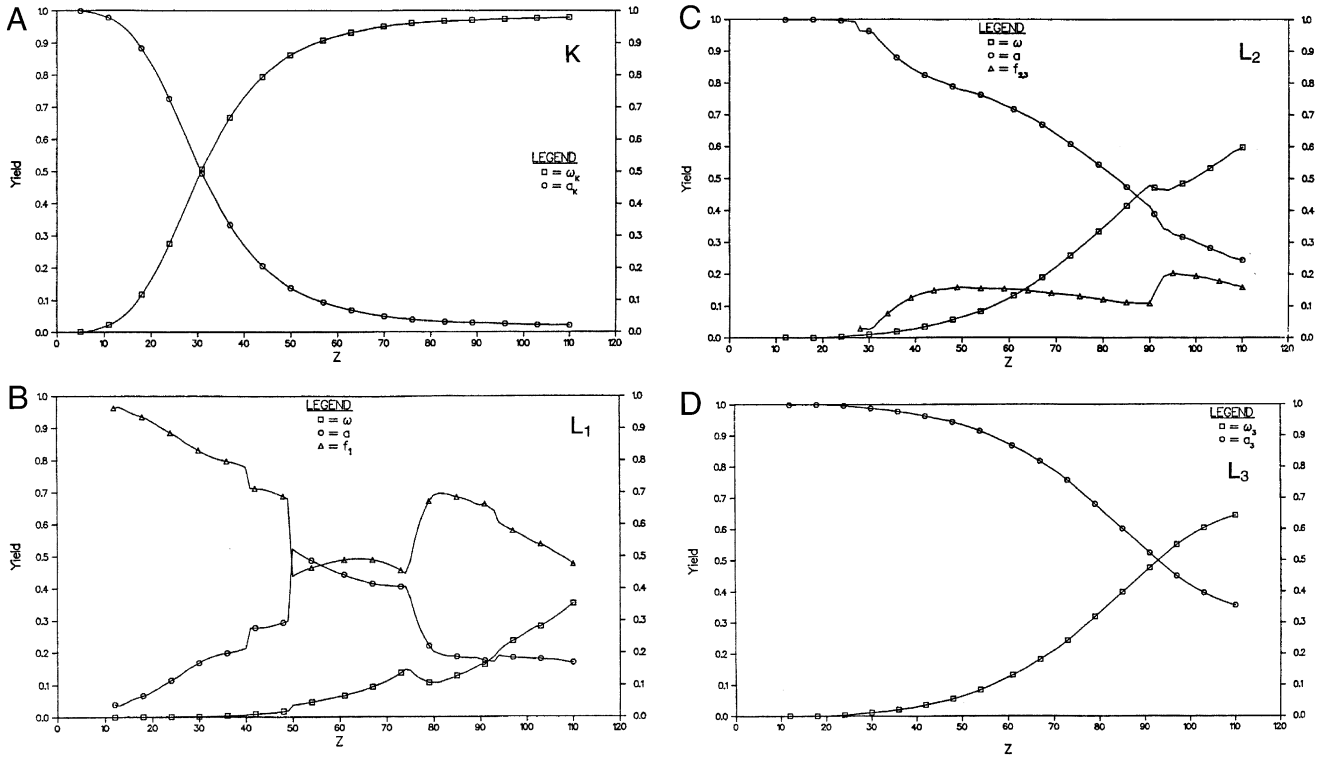


Figure 2. Examples of fluorescent yield and Auger yield as a function of atomic number. Fluorescent yield (boxes), non-Coster-Kronig Auger yield (circles), and Coster-Kronig Auger yield (triangles) for: (**A**) K, (**B**) L_1, (**C**) L_2, and (**D**) L_3 levels as a function of atomic number, Z. Auger electron energy analyzers usually measure kinetic energies up to ~3 keV. For this energy range, K levels in atoms up to atomic number 19 and $L_{2,3}$ levels up to atomic number 50 can be ionized; in these atoms, the fluorescent yields for these levels are much lower than the combined Auger yields. Data from Krause, 1979. (Reprinted with permission.)

the specimen will lose energy in traveling to the surface and appear as a background at lower kinetic energy.

A competing process for Auger electron emission from atoms is x-ray emission. Following the creation of the initial core-level vacancy, a characteristic x ray can be emitted (instead of an Auger electron) when an outer level electron fills this vacancy. The probability for relaxation by Auger emission (called the Auger yield) is much higher than for x-ray emission (called the fluorescent yield) for the energies usually measured in AES (Krause, 1979). Examples of fluorescent yield and Auger yield as a function of atomic number for the K, L_1, L_2, and L_3 levels are shown in Figure 2 (note that Coster-Kronig Auger yields are identified separately from the non-Coster-Kronig Auger yields in this figure).

PRACTICAL ASPECTS OF THE METHOD

As mentioned earlier, most AES work is done with electron-beam excitation. Tungsten filaments, lanthanum hexaboride cathodes, and field emitter tips are all available as electron sources. Lanthanum hexaboride and field emitters have high brightness and are used for high-spatial-resolution instruments. Electron beams can be focused to small areas for point analysis, and rastered in one direction for Auger line scans or in two directions for Auger mapping. Some materials are susceptible to electron beam damage, and care must be exercised in studying them (see Specimen Modification). Insulating specimens can also charge electrically during analysis, resulting in a shift in the measured Auger kinetic energy. Such charging can usually be reduced and/or stabilized using low-energy electron flood guns to replace the lost charge.

A number of electron energy analyzers can be used for AES, the two most popular being the cylindrical mirror analyzer and the spherical sector analyzer. Specimens must be positioned accurately, particularly for single-pass cylindrical mirror analyzers where the measured kinetic energy depends somewhat on specimen position. Positioning is also important in sputter depth profiling where the electron beam, the ion beam, and the specimen must all be aligned at the focus of the analyzer. The electrons transmitted through the analyzer are usually amplified by one or more channel electron multipliers (or microchannel plates) and measured using pulse counting, voltage-to-frequency conversion, or phase-sensitive detection coupled with electron beam or analyzer modulation.

Because of the surface sensitivity of AES, measurements should be made in an ultra-high-vacuum system ($\sim 10^{-7}$ Pa and below) to prevent contamination of the specimens under study. Sputter depth profiling at higher background pressures can be tolerated if the sputter rate is sufficiently high that contamination from the

background gases does not occur. Commercial systems usually include devices for specimen manipulation and treatment (e.g., specimen heating, cooling, and fracture). For many specimens, some treatment, such as inert gas sputtering to remove surface contaminants, might be necessary before meaningful analysis can be made. Some specimens, even though they might appear clean to the eye, may have a thin contaminant layer precluding the detection of any Auger signal from the material of interest. Specimen handling is very important to avoid any contamination of the specimen. Specimens are usually inserted into the analysis chamber through a vacuum interlock.

Auger line scans give the distribution of elements in a straight line across the surface, and Auger maps provide the distribution of elements within a selected area on the surface. The distribution of elements with depth can be obtained by depth profiling. Analysis is rapid, from a few minutes to a few hours depending on the information required, and several specimens can be examined in one day.

Types of Auger Spectra

Auger spectra can be acquired in several ways, but the methods fall into two main categories, namely, (1) the measurement of the total electron signal including Auger electrons and the background from secondary and backscattered electrons, or (2) the measurement of the derivative of this signal using a lock-in amplifier in order to suppress the slowly varying background. Both types of spectra are measured as a function of the kinetic energy of the electrons leaving the sample. Backgrounds can also be suppressed by taking a derivative of the total electron signal obtained in (1) above with a computer. Examples of such spectra are shown in Figure 3, and were taken from a used Cu gasket. Figure 3A is often referred to as a "direct spectrum," and the Auger electrons appear superimposed on the continuous background of secondary and backscattered electrons. This spectrum was obtained in the pulse-counting mode, and the Auger peaks from Cu, Cl, C, N, and O are identified. A large part of the increasing background with increasing kinetic energy is due to the way Auger spectra are usually measured—i.e., with the analyzer energy resolution set at a constant percentage of the measured kinetic energy of the emitted electrons. This results in a continuous, linear change in the measured energy resolution across the spectrum, and therefore the enhancement of the kinetic energy distribution of electrons entering the analyzer, $N(E)$, by a multiple of the kinetic energy, E. This is why the ordinate is usually labeled $E^*N(E)$ in Auger spectra. Figure 3B is referred to as a "derivative spectrum", where the background is reduced and the Auger features are enhanced; note especially the improved N signal. This spectrum was obtained by differentiation of the spectrum in Figure 3A with a computer, and then cutting off some of the large signal at low kinetic energy. The ordinate is then labeled $d[E^*N(E)]/dE$.

Specimen Alignment

The specimen is usually moved to the analyzer focus and approximate analysis position using an optical method

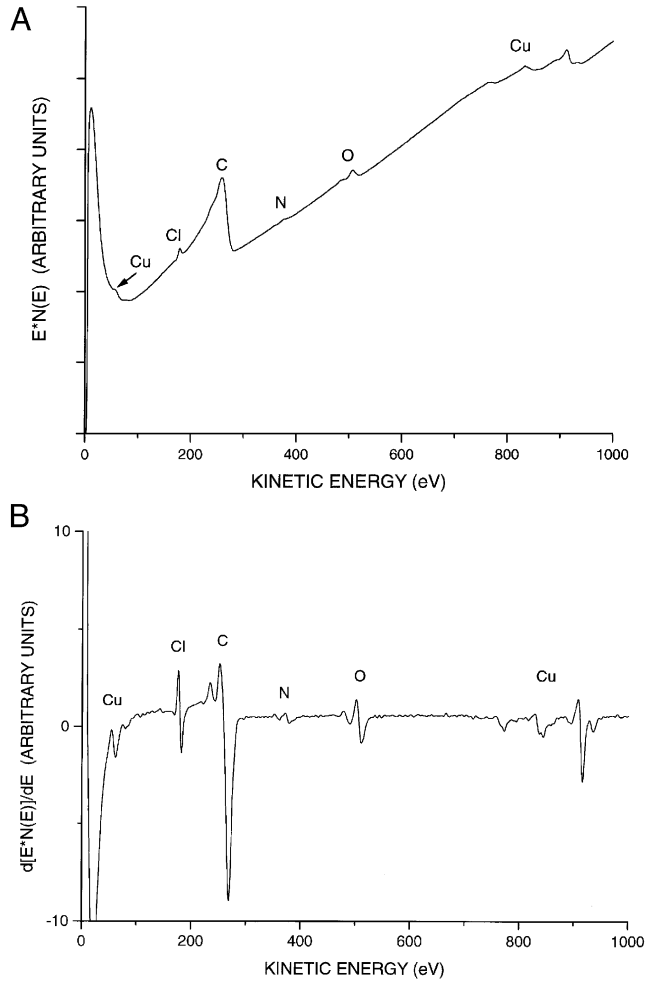


Figure 3. Auger spectra of a used Cu gasket showing: (**A**) the direct spectrum and (**B**) the derivative spectrum. Spectra were obtained using a 40-nA, 5-keV electron beam for excitation, and a hemispherical analyzer with a 1-eV step interval for analysis. The spectrum took 18 min to acquire.

such as a previously calibrated video camera. The electron gun is then switched on, and precise positioning of the specimen is accomplished by using the scanning electron microscopy (SEM) capability of the system. The principles of SEM are discussed in SCANNING ELECTRON MICROSCOPY, and images are obtained from the signal variation with beam position produced by either secondary electrons, backscattered electrons, or specimen current to ground. The analyzer focus can also be checked by maximizing the signal detected by the electron energy analyzer at the same energy as the electron beam, typically 1, 2, or 3 keV. This signal arises from elastically backscattered primary electrons.

Qualitative Analysis

Elements present at the surface are identified from the energies of the Auger peaks in the spectra. An element (or compound) can be considered positively identified if the peak energies, peak shapes, and the relative signal strengths (for multiple peaks) coincide with a standard

reference spectrum of the element (or compound). The procedure starts with the most intense peak in a spectrum and is repeated until all peaks have been identified (Fig. 3). Peak identification depends on an increase in signal intensity above the slowly varying background, and if a computer program is used for peak identification it compares the kinetic energy of the peak with a reference library of associated peaks. If no element lies within a certain energy range from the measured peak, it is not identified by the program. Such problems can arise if the kinetic energy scale is not calibrated, or if the measured kinetic energies shift due to specimen charging. In some cases, peaks can also be misidentified with such software, particularly when different elements have Auger peaks with similar kinetic energies, and the user needs to be aware of this limitation in qualitative analysis. Handbooks of Auger spectra are available and are very useful for helping the novice identify (or confirm) peaks in spectra (Sekine et al., 1982; Childs et al., 1995). The reported Auger peak energies vary by a few electron volts between handbooks, but such variations between handbooks or laboratories should be eliminated in the future, now that energy calibration procedures for AES have been provided for both direct and derivative spectra (Seah et al., 1990). The relative intensities of peaks at widely different kinetic energies will also vary between instruments (and the handbooks) due to instrument design and operating conditions. Therefore, reference spectra should be measured in the same instrument used for the unknown. This problem of intensity variation with kinetic energy can also be reduced by measuring direct spectra of Cu, Ag, and Au in a given instrument, allowing intensity-energy relationships to be derived by comparing such spectra with standard reference AES spectra of Cu, Ag, and Au (Smith and Seah, 1990).

Sets of Auger spectra from some of the elements (Mn, Fe, Co, Ni, and Cu) in the first row of transition elements are shown in direct form in Figure 4, and in derivative form in Figure 5. These spectra were taken from argon-ion sputtered metal foils. The dominant peaks at higher kinetic energy (500 to 1000 eV) are the $L_{2,3}MM$ series, and those at low kinetic energy (40 to 70 eV) are from $M_{2,3}VV$ transitions. The peaks near 100 eV are from M_1VV transitions. Note the rich structure in the $L_{2,3}MM$ series due to the various final M shell vacancies, and how these sets of peaks are similar to each other but move to higher kinetic energy with increasing atomic number. Spectra from other elements will show similar trends in their LMM Auger spectra, as will KLL and MNN transitions. The low-kinetic-energy $M_{2,3}VV$ Auger peaks from the transition metals in Figures 4 and 5 shift by smaller amounts than the $L_{2,3}MM$ peaks, and are not usually used for elemental analysis. The types of Auger spectra shown in these two figures are often referred to as survey spectra and are taken with relatively poor energy resolution and relatively large energy steps so as to optimize the signal-to-noise for a relatively short data acquisition time. Survey spectra are usually the first spectra taken in an analysis, as they provide an overall indication of the surface composition, and are typically taken over an energy range from a few electron volts to 1000 or 2000 eV.

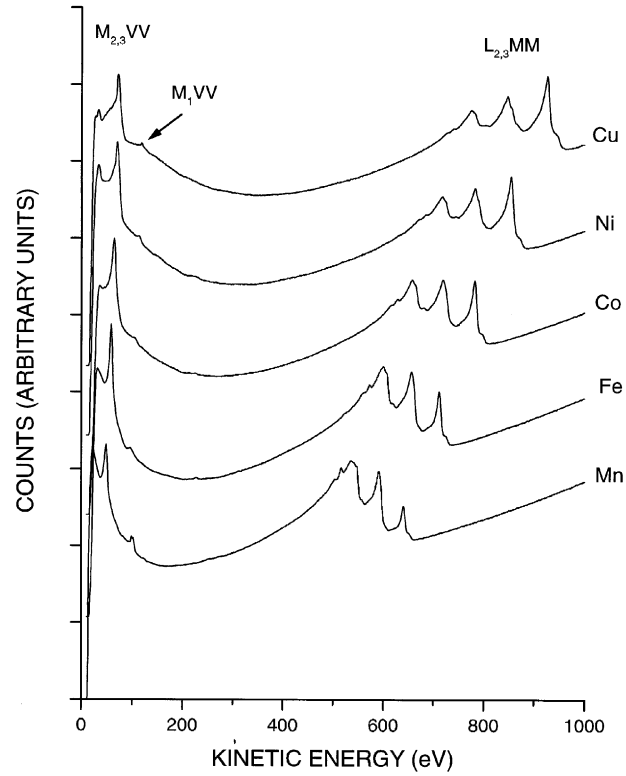


Figure 4. Auger spectra of Mn, Fe, Co, Ni, and Cu taken from argon-sputtered foils, shown in the direct mode. These spectra were obtained with a 40-nA, 5-keV electron beam, and a hemispherical analyzer with a 1-eV step interval. Each spectrum took <2 min to acquire. The Fe, Co, Ni, and Cu spectra are offset for clarity.

The derivative spectra are normally used for peak identification because the Auger features are enhanced in these spectra. The data can also be taken over a longer time to increase the detectability, and higher energy resolution spectra can be obtained if such information is needed. The Auger spectra displayed in Figures 4 and 5 were taken using a hemispherical analyzer with an energy resolution of 0.6% of the kinetic energy and 1.0-eV energy steps applied to the analyzer; each took less than 2 min to acquire with a 40-nA, 5-keV electron beam.

Chemical Effects

Besides identifying the elements present at surfaces, AES can often provide useful information about the chemical environment of surface atoms. Such chemical effects can affect the measured Auger spectrum in a number of ways, e.g., the kinetic energy at which an Auger peak occurs can change, the energy distribution of the Auger electrons can change, or the energy loss structure (see Problems in reference to plasmon loss peaks) associated with Auger peaks can change.

Energy shifts are expected to occur whenever there is charge transfer from one atom to another. Thus, changes in composition of metallic components in metal alloys would not be expected to produce measurable changes in

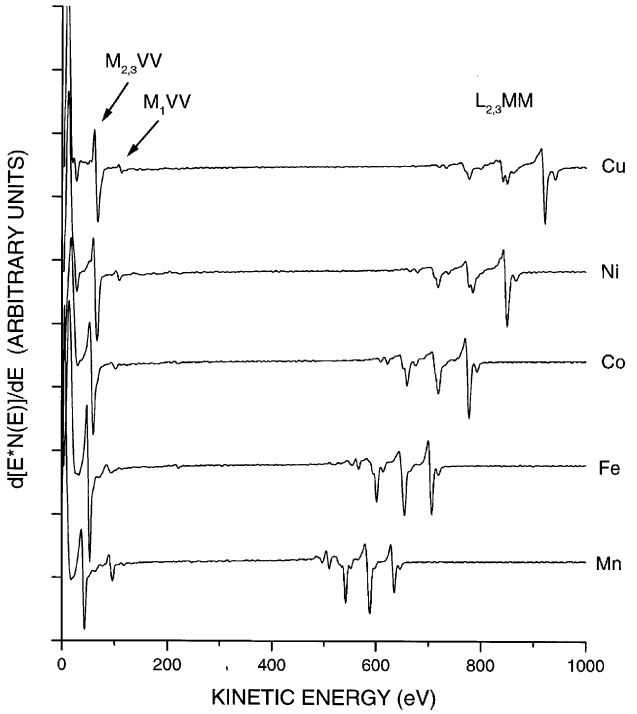


Figure 5. Auger spectra of Mn, Fe, Co, Ni, and Cu taken from argon-sputtered foils, shown in the derivative mode. Derivatives were taken from the direct spectra shown in Figure 4. Note how the spectral features are enhanced in the derivative mode, compared to the direct mode (Fig. 4). The spectra were obtained with a 40-nA, 5-keV electron beam, and each took <2 min to acquire. The spectra are offset for clarity.

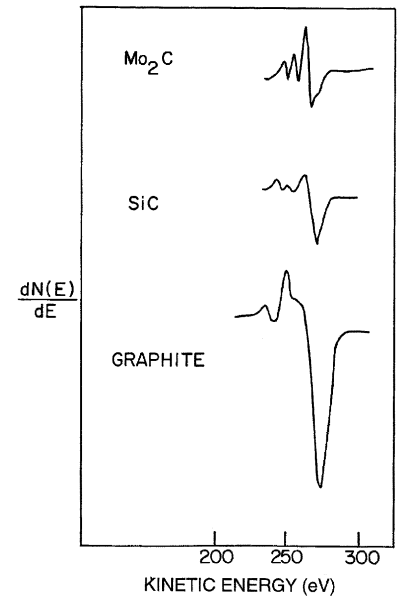


Figure 6. Carbon KVV Auger spectra from molybdeum carbide, silicon carbide, and graphite, showing the different Auger line shapes. Metal carbides usually exhibit the fairly symmetric, three-peak structure of molybdenum carbide, whereas surface contamination from organics usually resembles graphite.

the Auger energies (for core levels) of the components. However, submonolayer quantities of oxygen adsorbed on clean metal surfaces can produce measurable changes in the metal Auger peaks, the shift (from a few tenths of an eV to >1 eV) increasing with coverage (Haas and Grant, 1969). For bulk metal oxides, shifts in the metal Auger peaks are of the order of 10 eV. Changes in the line shape of Auger spectra can also occur due to changes in bonding, particularly when one or two valence electrons are involved in the relaxation process (Haas et al., 1972b). Changes in lineshape of transitions involving valence electrons are usually accompanied by significant shifts in Auger energies as well (Grant and Haas, 1970). Examples of changes in the carbon KVV Auger line shape for different chemical states of carbon are shown in Figure 6, and are due to differences in the density of states in the valence band.

Sensitivity and Detectability

AES measures the composition of the several outermost atomic layers. The sensitivity of the technique should be limited by the shot noise in the electron beam used for Auger electron production. This can be checked by measuring the noise in the direct spectrum, which should follow the square root of the counts detected. Obviously, the longer the time allowed for taking data, the better the detection limit. Also, the sensitivity factor (see Data Analysis) varies for different elements, and for different Auger transitions of the same element. Problems in detectability can also arise due to peak overlap, and this is discussed further below (see Data Analysis). Typical detection limits are in the range of 0.1 to 1 atom %. Remember that H and He are not detected.

Auger Line Scans

With Auger line scans, Auger data are acquired as a function of position across a straight line on the specimen. The maximum length of the line depends on the acceptance area of the analyzer and the deflection capability of the electron gun. The specimen itself is imaged in the SEM mode, and the analysis line is then selected on this image. Also selected is a region (or regions) of the Auger energy spectrum from which to derive the signal that is monitored as the scan progresses.

An SEM image of a Cu grid over a Pd foil, together with a selected line scan position (the white horizontal line near the center), are shown in Figure 7. There is a small vertical displacement between the Cu grid and the Pd foil. The specimen was cleaned by argon-ion bombardment at several azimuths to eliminate shadowing. The Cu and Pd Auger line scans are shown in Figure 8. The Auger line scan signals in this case were the difference between the Auger spectral intensity at the Cu LMM and Pd MNN Auger peak maxima and the backgrounds at the high-kinetic-energy sides of each peak, in the direct spectra. The Cu line scan (solid line) is rather simple, with the signal intensity being high at the Cu grid, as expected. The Pd signal (dashed line) is more complex, being zero at the Cu grid (as

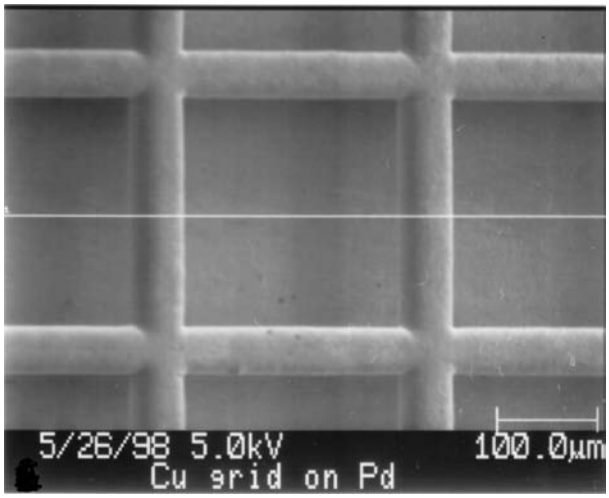


Figure 7. SEM image of a slightly elevated Cu grid over a Pd foil, together with a selected line scan position shown as the white horizontal line near the center. This image was obtained from secondary electrons, using a 5-keV electron beam in the Auger system.

expected), and at a maximum in the region between the Cu grid except for a region to the left of the grid where the signal decreases to near zero. This decrease in signal is due to shadowing of part of the Pd signal by the Cu grid (i.e., blocking Auger electrons from reaching the analyzer), as there is a small gap between the grid and the Pd foil. The specimen was mounted horizontally in the vacuum chamber; the incident electron beam was in the vertical plane that included the line scan, and at an angle of 45° to the horizontal from the left of the image. The analyzer was at an angle of 45° to the horizontal, toward the top of the image.

Several Auger survey spectra are shown in Figure 9 for different positions of the electron beam as it approached

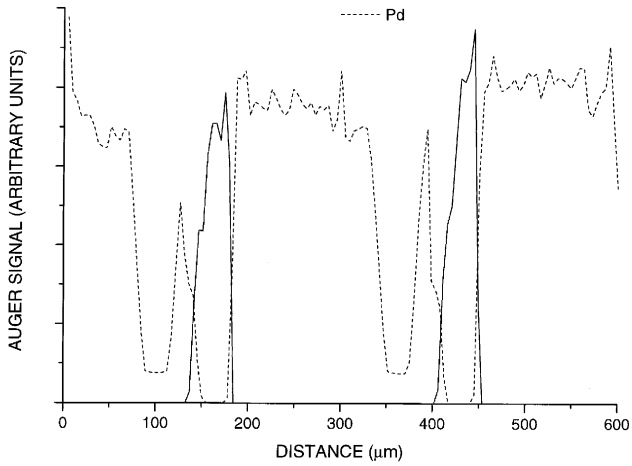


Figure 8. Auger line scans of Cu (solid line) and Pd (dashed line) taken where indicated by the horizontal white line in Figure 7. The Auger line scan signals were, respectively, the difference between the intensity at the Cu LMM or Pd MNN Auger peak maxima and the backgrounds at the high kinetic energy sides of each peak, in the direct spectra. Each scan took 6 min to acquire, using a 40-nA, 5-keV electron beam.

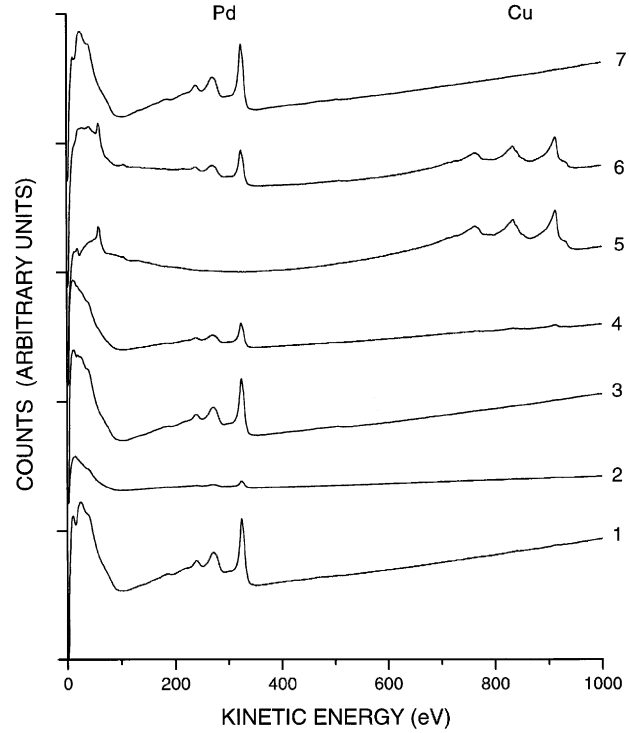


Figure 9. Seven Auger survey spectra in the direct mode, taken at different positions of the electron beam as it approached and crossed the Cu grid at the line scan position shown in Figure 7. The spectra 2 to 7 have been offset for clarity, but were not normalized. Spectra were acquired with a 40-nA, 5-keV electron beam, and each spectrum took ~1 min to acquire.

and crossed the elevated Cu grid. Spectrum number 1 was obtained when the electron beam was on the Pd foil, to the left of the grid as displayed in Figure 7. The Auger peaks between 150 and 350 eV are from Pd. Spectrum 2 was obtained when the electron beam was still on the Pd but underneath the Cu grid, where most of the Pd Auger electrons were blocked by the grid from reaching the analyzer. Spectrum 3 was obtained when the electron beam had passed beneath the elevated Cu grid, but the beam was still not on the grid. Spectrum 4 was obtained when the beam moved partly onto the side of the grid (but shadowing reduced the Cu Auger intensity). Spectrum 5 was obtained when the beam was actually on top of the Cu grid (maximum Cu intensity). Spectrum 6 when obtained when the beam had moved partly off the Cu grid, and spectrum 7 was obtained when the beam was on the Pd foil once more. From these spectra, it can be seen that the initial signal drop in the Pd Auger line scans occurs when the electron beam is not yet on the Cu grid, and is due to shadowing by the grid and not to surface contamination. Other problems in Auger line scans due to topography (not shadowing) and electron backscattering are discussed below (see Data Analysis).

Auger Maps

With Auger maps, Auger data are acquired as a function of position within a defined area on the specimen. The

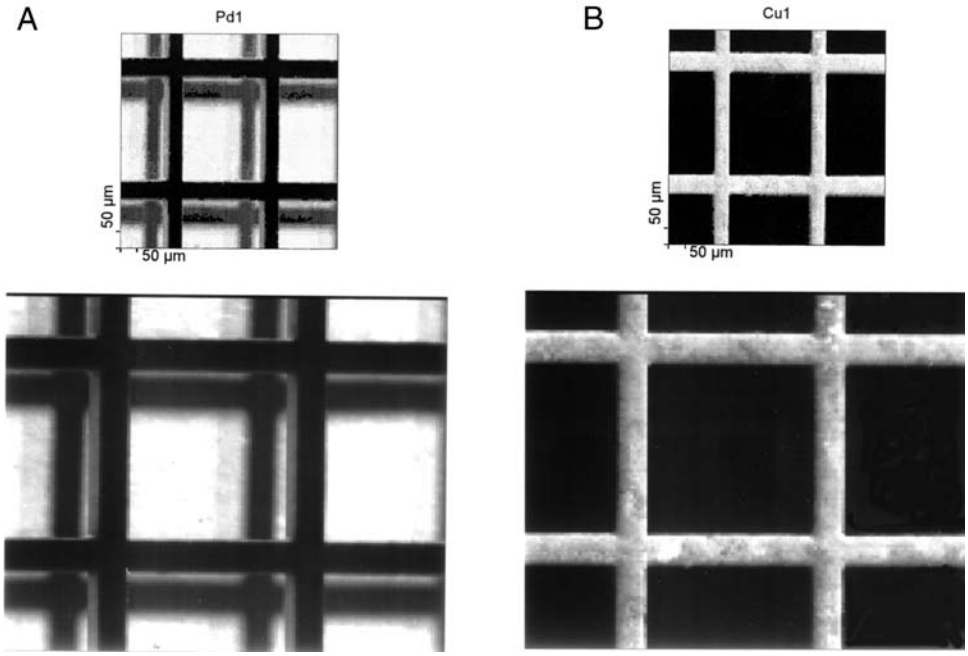


Figure 10. Auger maps of the Cu grid on Pd for, (**A**) Pd, and (**B**) Cu. Bright areas correspond to high signal intensity, and darker areas to lower intensity. These maps were taken with a 40-nA, 5-keV electron beam, with a 128×128 point spatial resolution. Each map took ~25 min to acquire.

maximum dimensions of the area that can be mapped depend on the acceptance area of the analyzer and the deflection capability of the electron gun. The specimen itself is imaged in the SEM mode, and the analysis area is then selected on this image.

Pd and Cu Auger maps of the Cu grid on Pd foil are shown in Figure 10, where the bright areas correspond to higher Auger intensity. The Cu Auger map on the right is rather simple, with the grid being displayed quite clearly. The Pd map is more complex as shadowing also comes into play (see discussion of Auger Line Scans). The region of the Cu grid itself is black, but there is another image of the grid in gray due to the shadowing of this part of the Pd foil by the Cu grid, resulting in a reduction of the Pd Auger signal from this part of the foil.

Auger maps can be displayed in different ways, such as a selected number of gray levels, on a thermal scale, in pseudocolor, or in a selected color for comparing with a map of a different element in the same map area. Auger line scans for each element mapped can also be displayed across a specified line on the appropriate map. As in the case of Auger line scans, corrections of Auger intensity due to topography and electron backscattering are often required (see Data Analysis).

Depth Profiling

In AES, depth profiling is usually accomplished by inert gas ion bombardment to remove successive layers of material from the specimen. The removal of atoms by ion bombardment is also referred to as sputtering. Auger measurements are made either continuously while simultaneously sputtering, or sequentially with alternating cycles of sputtering and analysis. Ar or Xe ions at an energy of a few kilo-electron volts are usually used for sputtering. Typical sputter rates are of the order of 10 nm/min. It is important that the ion beam is correctly aligned to the analysis area on the specimen, and the ion beam is often rastered over a small area of the specimen to ensure that the electron beam used for Auger analysis strikes the flat-bottomed crater formed by the rastering.

Auger survey scans could be taken in depth profiling, but a technique called multiplexing is normally used instead. With multiplexing, several energy spectral windows are preselected, with each one encompassing an Auger peak of the element of interest (Sickafus and Colvin, 1970). This procedure saves time, as only signals from energy regions of interest are measured. This will also provide a higher density of data points in the depth profile, but means that other elements that might be present below the original surface could go undetected.

An Auger sputter depth profile of SiO_2 on Si is shown in Figure 11. The Auger transitions selected were O KLL, Si LMM, and Si KLL. Depth profiles can be plotted with peak height intensities in direct spectra, or with peak-to-peak heights in derivative spectra. The depth profile shown here was obtained using the derivative spectra. The ion beam was rastered over a 2-mm × 2-mm area on the specimen. Auger measurements were made sequentially with sputtering, and the abscissa is plotted as sputter time. Two scans of each Auger transition were obtained before sputtering (one of these is plotted as a negative time). Different compounds sputter at different rates, and it is often not a trivial matter to convert sputter time to depth. Note that near the interface the O signal decreases towards zero, and the Si LMM signal increases

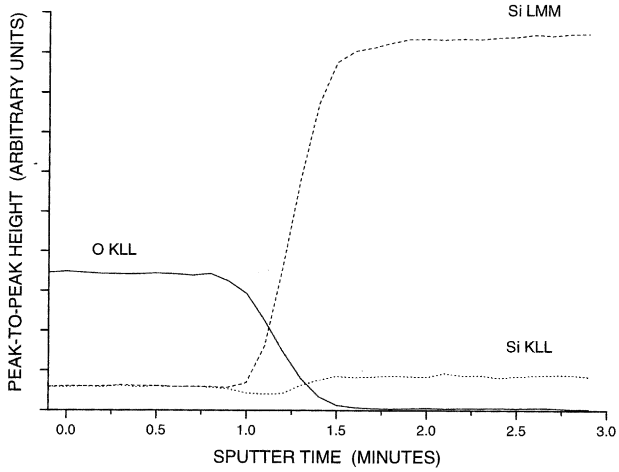


Figure 11. An Auger sputter depth profile of SiO_2 on Si. The Auger transitions selected were O KLL (solid line), Si LMM (dashed line), and Si KLL (dotted line). The profile was obtained by sequentially taking Auger spectra and sputtering. Two scans of each Auger transition were obtained before sputtering (one of these is plotted as a negative time). The sample was sputtered for 0.1 min between measurements. The Auger spectra were taken with a 40-nA, 5-keV electron beam. A 3-keV Ar^+ ion beam, rastered over an area 2 mm×2 mm, was used for sputtering.

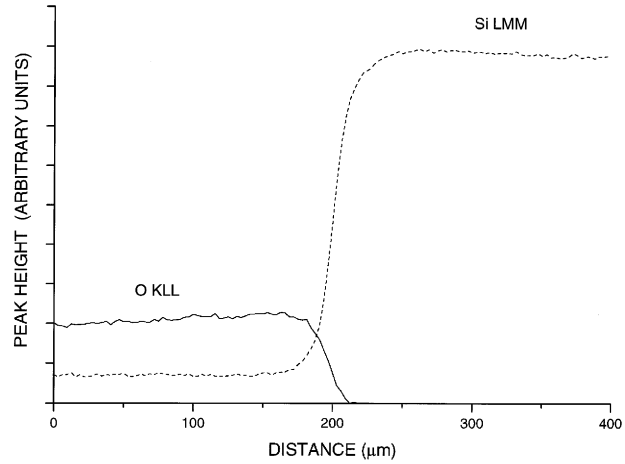


Figure 13. Auger line scans for the O KLL (solid line) and Si LMM (dashed line) Auger transitions taken across the crater edge shown in Figure 12b, using Auger peak heights above background in the direct spectra. The line scans were taken with a 40-nA, 5-keV electron beam, and each scan took ~1 min to acquire.

significantly. The Si KLL signal undergoes a different behavior with an apparent dip at the interface and a small increase in intensity in the pure Si substrate. For discussion of these signal variations, see Data Analysis.

Depth profiles can also be obtained by conducting an Auger line scan across the "crater wall" formed by sputtering (Taylor et al., 1976). An SEM image of the sputtered area for the SiO_2 layer on Si is shown in Figure 12A, with the left-hand crater wall shown expanded in Figure 12B. The Auger line scan was obtained across 400 μm of the left-hand crater wall, as shown in Figure 12B. The Auger line scans for the O KLL and Si LMM Auger transitions are shown in Figure 13 using peak height above background in the direct spectra. Line scans across crater walls are also useful if there was a problem with the alignment of the electron and ion beams, since the Auger line scans are performed on the crater wall after sputtering is finished. In the case of very thin films (up to ~10 nm thickness) where Auger peaks from the substrate can also be detected without sputtering, depth information can be obtained by angle-resolved measurements (the specimen is tilted in relation to the analyzer to change the angle of emission). Other variations on depth profiling include mechanical methods for specimen preparation such as angle lapping (Tarng and Fisher, 1978), where a flat sample is polished at a small angle (~0.5°) to the original surface to expose the depth distribution of elements, and ball cratering (Walls et al., 1979), where a spherical ball grinds a crater into the sample to expose the depth distribution. These two techniques are

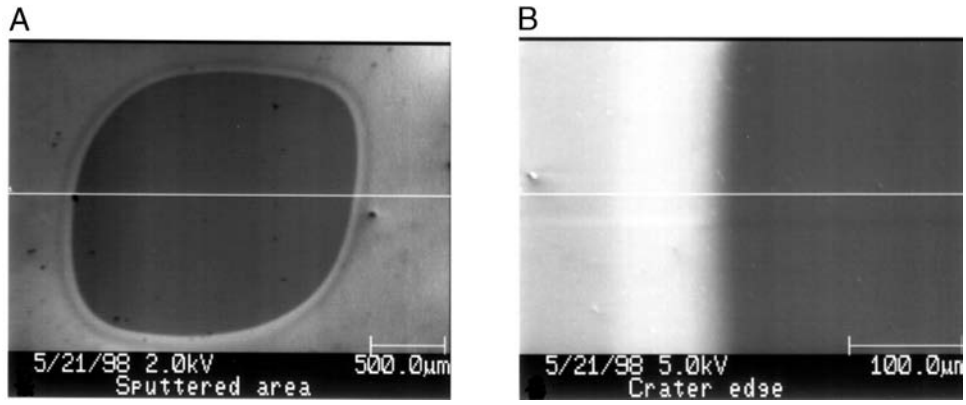


Figure 12. An SEM image of the sputtered area for the SiO_2 layer on Si, following the acquisition of the depth profile, is shown in (**A**). The left-hand crater edge is shown expanded in (**B**), and the horizontal white line in (**B**) is the region that was used for crater edge profiling. The SEM images were taken with (**A**) 2-keV and (**B**) 5-keV electron beams.

typically used where analysis is required to depths >1 μm. The depth resolution in sputter depth profiling is typically several percent of the depth sputtered, but the depth resolution can be improved by rotating the specimen during sputtering (Zalar, 1985).

METHOD AUTOMATION

All modern AES systems operate under computer control, and many repetitive functions can be carried out by the computer.

Specimen introduction is semiautomatic: the specimen insertion chamber can be evacuated automatically, the gate valve between the insertion chamber and the analysis chamber can be pneumatically opened and closed, and the specimen manipulator in the analysis chamber can be parked in the transfer position by computer control. Some AES systems also allow transferability of specimen location information, on a platen obtained prior to insertion, to the specimen manipulator in the analysis chamber.

The electron gun used for producing Auger electrons in the specimen is also controlled by the computer. As electron beam energy is changed, the lens parameters required for focusing, beam alignment in the column, astigmatism correction, etc., are all set to predetermined values. The beam current can also be set to preselected values. Different regions of interest on the specimen surface can also be selected and stored in the computer, so these different regions can be examined in turn automatically during an analysis, e.g., survey spectra could be obtained from the different regions, or individual depth profiles could be obtained from each region during one sputtering operation.

The electron energy analyzer also operates under computer control, although any physical apertures in the analyzer are usually selected manually. The energy resolution of the analyzer is usually controlled by computer, as is the collection and display of the output signal. The kinetic energy range for acquiring data, the kinetic energy step interval, the dwell time at each energy step for acquiring data, and the number of sweeps, among other settings, are all controllable with the computer. The computer also controls the acquisition of Auger line scans and Auger maps.

The computer is also used for data processing. This ranges from simple procedures such as differentiation of spectra to rather complex procedures such as linear least-squares fitting and factor analysis of data. This will be discussed further in the Data Analysis and Initial Interpretation section.

DATA ANALYSIS AND INITIAL INTERPRETATION

Qualitative analysis was discussed earlier (see Practical Aspects of the Method); there are, however, some very useful data analysis methods that need to be mentioned in this section, namely, spectrum subtraction, linear least-squares fitting, and factor analysis. Also, these techniques have all been used to enhance the analysis and interpretation of spectra and to improve detectability.

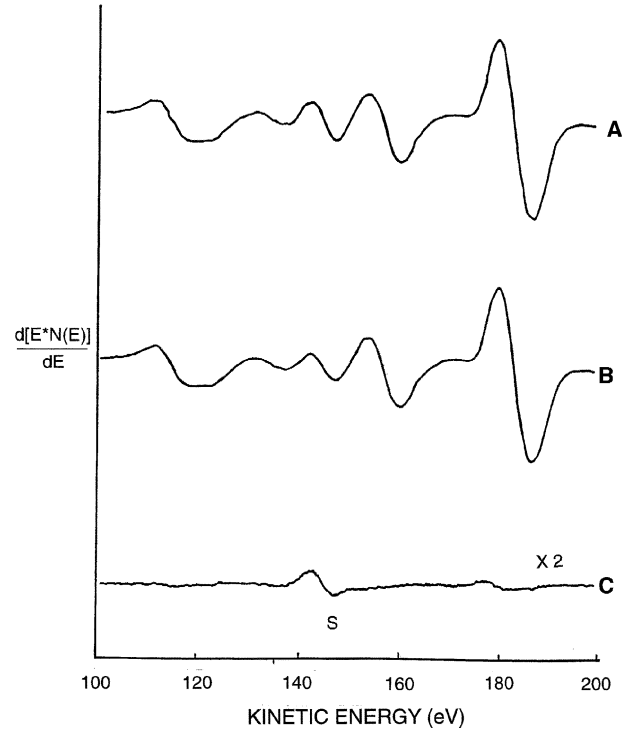


Figure 14. An example of spectrum subtraction: (**A**) the derivative Auger spectrum from a Mo surface, (**B**) the spectrum from pure Mo, and (**C**) the result after subtracting (**B**) from (**A**) to minimize the Mo peak near 180 eV. Note the recovery of an S Auger signal after subtraction.

Spectrum subtraction is a rather simple procedure whereby a spectrum from one element is subtracted from that of another to eliminate overlap problems (Grant et al., 1975a). In this procedure, the spectra must be aligned to allow for any shifts due to chemical effects or specimen charging. If the line shapes are different due to chemical effects, the subtraction will not completely remove the features that are due to the subtracted element. An example of spectrum subtraction is shown in Figure 14, where the S Auger signal (C) is recovered from an Auger spectrum of a Mo surface (A) by subtracting the spectrum from a clean Mo surface (B). The S is then more easily quantified as well.

The different behavior of the Si LMM and Si KLL intensities in the depth profile of SiO_2 on Si (Fig. 11) is due to chemical effects that occur in Si Auger peaks in SiO_2 when compared with pure Si. This depth profile was obtained using derivative Auger data, and the actual Si LMM and KLL spectra used are shown in Figure 15, panels A and B, respectively. Note that in each case the spectra cluster around two different line shapes, which are due to the different Si chemistries in Si and SiO_2. The LMM spectrum from SiO_2 has its main structure ~65 to 75 eV, whereas that from Si is ~80 to 100 eV. Also, the peak-to-peak heights of these LMM transitions increase quite markedly in going from the oxide to the Si substrate. Both of these large differences are due to

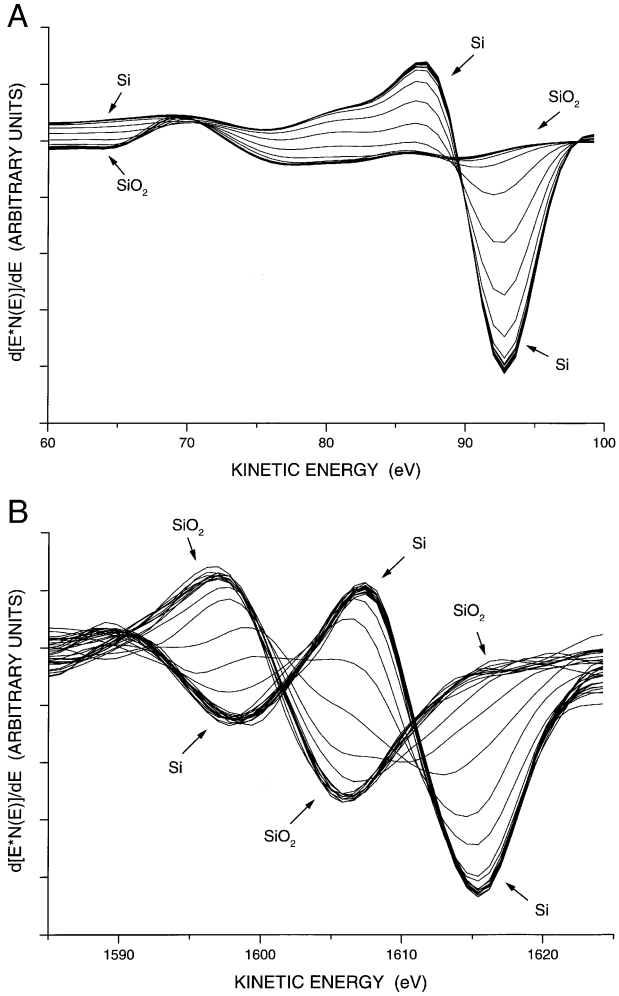


Figure 15. The Si (**A**) LMM, and (**B**) KLL derivative Auger spectra obtained at different depths during the depth profile of SiO_2 on Si shown in Figure 11. Note how in each case the spectra cluster around two different line shapes, corresponding to the Auger transitions in SiO_2 and Si. The few intermediate spectra are from the region near the SiO_2/Si interface.

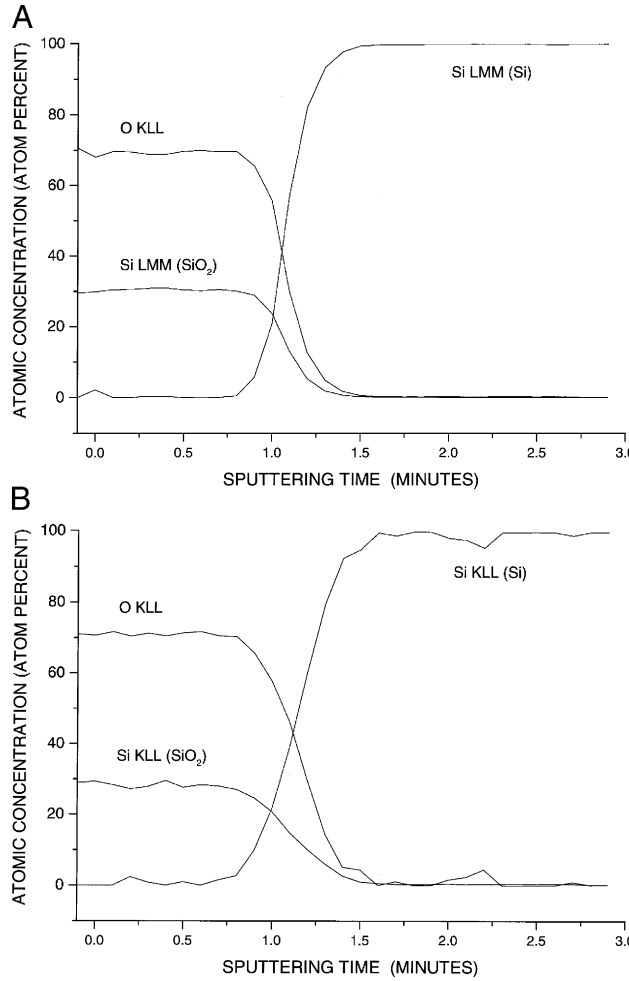


Figure 16. The sputter depth profiles of SiO_2 on Si obtained after applying linear least-squares processing to (**A**) the derivative Si LMM Auger spectra, and (**B**) the derivative Si KLL spectra to separate the Si signals from the oxide and the substrate. The concentrations were calculated using the manufacturer-supplied elemental sensitivity factors for Si and O.

involvement of valence electrons in the LMM transitions. For the KLL transitions, a well defined chemical shift occurs between the oxide and the substrate (the oxide being at lower kinetic energy), and the main peak intensity is more symmetrical in the oxide. Also present in these spectra are a few spectra from the interfacial region, but it is not obvious whether these spectra are linear combinations of the spectra from the oxide and from pure Si, or are line shapes corresponding to a different chemical state of Si. However, it can be seen that the reason for the dip in the Si KLL profile at the interface is a reduction in the peak-to-peak height of the Si KLL signal near the interface since both KLL peaks will be present side-by-side in this region, and not necessarily a decrease in Si concentration. Numerical methods such as linear least-squares fitting (Stickle and Watson, 1992) and factor analysis (Gaarenstroom, 1981) can be used to separate overlapping spectra, or the different chemical components in spectra such as in the SiO_2 on Si sputter depth profile in Figure 11. These processing programs are incorporated into numerical software packages, and have been included in some surface analysis processing software. Examples of use of the linear least-squares method to separate the two different chemical states of Si in the sputter depth profile of SiO_2 on Si are shown in Figure 16A for the derivative Si LMM and O KLL peaks, and in Figure 16B using the derivative Si KLL and O KLL peaks. Note in both examples how the Si Auger signal has been separated into the two different chemical states of Si, namely SiO_2 and pure Si, and that there is no unusual decrease in total silicon concentration at the interface, as indicated from the raw KLL data in Figure 11. The atomic concentrations were calculated using the manufacturer-supplied elemental sensitivity factors for Si and O (see discussion of Quantitative Analysis). Similar results were obtained by selecting two-component fits in factor analysis.

Quantitative Analysis

The most commonly used method for quantitative analysis of Auger spectra is that of sensitivity factors. With sensitivity factors, the atomic fractional concentration of element i, C_i, is often calculated from

$$C_i = \frac{I_i/I_i^\infty}{\sum_j I_j/I_j^\infty} \quad (2)$$

where I_i is the measured Auger intensity of element i in the unknown, I_i^∞ is the corresponding intensity from the pure element i, and the summation is over one term for each element in the unknown (remember that H and He are not detected, and are not included in this calculation). The Auger intensities should all be measured under the same experimental conditions to achieve reliable results. Sometimes, sensitivity factors are measured in the same instrument relative to a particular element such as silver, and the method is then referred to as using relative sensitivity factors. Unfortunately, most people do not measure sensitivity factors in their own instruments but rely on published values, which can differ by more than a factor of 2. Sensitivity factors between elements can vary by as much as a factor of 100 across the periodic table (Sekine et al., 1982; Childs et al., 1995).

Further improvements to Equation 2 can also be made by including corrections for atom size, electron inelastic mean free path, and electron backscatter (Seah, 1983). The atom size adjusts for density, the inelastic mean free path adjusts for the analysis depth, and backscattering adjusts for the increase in Auger intensity that occurs when energetic electrons are scattered back toward the surface and have another chance to produce Auger electrons in the atoms that are closer to the surface. In a homogeneous binary alloy of elements A and B, the atomic concentration ratio of A (C_A) to B (C_B) is then given by

$$\frac{C_A}{C_B} = F_{AB}\frac{I_A/I_A^\infty}{I_B/I_B^\infty} \quad (3)$$

where

$$F_{AB} = \left(\frac{1 + r_A(E_A)}{1 + r_B(E_A)}\right)\left(\frac{a_B}{a_A}\right)^{3/2} \quad (4)$$

and F_{AB} is nearly constant over the entire AB compositional range. $r_i(E_i)$ is the backscattering factor for element i for an electron with kinetic energy E_i, and a_i is the size of atoms in element i (Seah, 1983). These corrections are straightforward and can improve quantitative analysis, but unfortunately are rarely made.

Auger currents cannot be measured directly as they are superimposed on background currents due to secondary and backscattered electrons. Also, there is current on the low kinetic energy side of Auger peaks due to inelastic losses suffered by Auger electrons in escaping from the specimen being analyzed, and this background must be removed in the measurement. Some work has been conducted to measure the areas under Auger peaks, but this is not done routinely in quantitative analysis due to the difficulty in removing the background. Peak height measurements (direct spectra) or peak-to-peak height measurements (derivative spectra) are normally used for quantitative analysis, and each method has its own sets of sensitivity factors.

Peak height measurements are often used for Auger line scans and Auger maps, as data acquisition times can be minimized by collecting data at just two or three energy values for each element. The sloping background above the peaks should be extrapolated to the energy of the peak maximum to measure peak height, although this is not always done. This can be done using a simple algorithm if measurements are made at three energies, namely at the peak maximum, at the background just above the peak, and at a higher background energy equally spaced from the central energy. Problems in quantitative analysis for Auger line scans or maps, due to specimen topography and changes in electron backscattering coefficient, can be satisfactorily corrected in peak height data by dividing the Auger intensity by the background on the high-energy side of the peak (Prutton et al., 1983).

Peak-to-peak height measurements from derivative spectra are the most commonly used method for quantitative analysis in AES, and this is the method normally used for survey scans and for depth profiles. The peak-to-peak height method has met with a lot of success, particularly in studying metals and their alloys. The accuracy of such measurements is typically $\pm 20\%$. More accurate quantitative analysis can be achieved if standards with concentrations that are close to those of the actual specimens are used, thereby reducing errors due to variations in electron backscatter and inelastic mean free path. Large errors can occur if the standards are not measured using the same analysis equipment as for the unknown. For example, different instruments from the same manufacturer have shown a fivefold variation in the intensity ratios for high- and low-energy Auger peaks (Baer and Thomas, 1986).

Problems also arise in quantitative analysis, particularly when using peak-to-peak heights, if Auger lineshape changes occur due to a change in chemistry. For example, for oxygen adsorption on nickel, the concentration of oxygen can be in error by a factor of two if peak-to-peak height measurements are used, even though the oxygen Auger line shape changes are subtle (Hooker et al., 1976).

Improvements in quantitative analysis can be made by measuring the intensity/energy dependency of each instrument for Cu, Ag, and Au for the set of conditions normally used for analysis (e.g., different analyzer energy resolutions), and comparing the results with standard reference spectra (see NPL under Standards Sources section). This will provide the intensity response function for each experimental condition, allow accurate transferability of results between instruments, and allow the performance of a particular instrument to be monitored over time.

Standards Sources

There are several standards for AES that have been developed by national standards organizations. Standards for

depth profiling are available from the United States National Institute of Standards and Technology (NIST), and from the National Physical Laboratory (NPL) in the United Kingdom. Software for calibration of Auger electron spectrometers is available from NPL. Standard terminology, practices, and guides have been developed by the American Society for Testing and Materials (ASTM). Details are listed in the Appendix.

SAMPLE PREPARATION

Since AES is a surface-sensitive technique, the handling, preparation, mounting, and analysis of a sample is of paramount importance. For example, a fingerprint on a specimen can be sufficient contamination to mask any signal from the actual surface of interest. If the specimen has been contaminated in some way, cleaning prior to mounting might be required. Cleaning with acetone, followed by rinsing in methanol, is often used to remove soluble contaminants. It must be remembered that any solvent used for cleaning might also leave a contaminant residue. Loose contaminating particles can be removed using a jet of nitrogen gas. If the contaminant is a thin layer, it might be possible to remove the contamination by low-energy inert gas ion sputtering in the analysis chamber. Sputtering can also cause its own problems, as elements sputter at different rates and their relative concentrations at the surface can therefore change. Sample handling is described in detail in ASTM Guides E1078 and E1829 (see the Appendix).

Size limitations on specimens depend on the particular analysis chamber. Specimen diameters up to several centimeters can usually be accommodated, but height limitations are often ~1 cm when vacuum-interlocked sample insertion systems are used. Some AES systems are designed to handle 300-mm wafers. If larger samples are to be routinely studied, customized insertion systems can be designed. Samples might need to be cut from a larger piece or sectioned for analysis. If this is done, care should be exercised to minimize possible contamination. If information is needed to a depth greater than ~1 μm, samples might be sectioned or ball cratered (see discussion of Depth Profiling).

Powder-free gloves should always be used when mounting specimens, and the tools used should be clean and nonmagnetic. Samples are usually mounted on Al or stainless steel holders, and are held in place with small screws or clips. Sometimes, samples are mounted with colloidal Ag or colloidal graphite. Powders are often mounted on conducting adhesive tape.

Samples prone to outgassing can be prepumped in an auxiliary vacuum system, or in the introduction chamber, to remove volatile species. However, cross-contamination between samples might occur.

When in the analysis chamber, specimens can sometimes be prepared by fracture, cleavage, or scribing. In some cases, it may be possible to remove contaminants by heating if subsequent studies on clean surfaces are required, such as on how specific gases react with a clean surface.

Once in the analysis chamber, specimens can be moved by a manipulator and imaged with a microscope or video camera (initially) and then with the electron beam (as in an SEM), to locate the regions to be analyzed.

If samples are electrically insulating, they are sometimes mounted beneath a metal mask with a small opening or beneath a metal grid to minimize charging by placing the incident electron beam close to the metal. Tilting the sample to change the angle of incidence of the electron beam can help control charging. Changing the beam energy, reducing the beam current density, and using a low-energy electron flood gun are also effective in controlling charging.

SPECIMEN MODIFICATION

Since an energetic electron beam is used for Auger analysis, and since it is a surface-analysis technique, specimens can sometimes be modified by the beam. A manifestation of this would be that the surface composition varies with time of exposure to the beam (Pantano and Madey, 1981).

Specimen modification can occur from heating by the beam, where species might diffuse to the surface from the bulk, diffuse across the surface, or desorb from the surface. The electron beam itself can also cause desorption (e.g., fluorine desorption), it can assist the adsorption of molecules from the vacuum environment onto the surface,

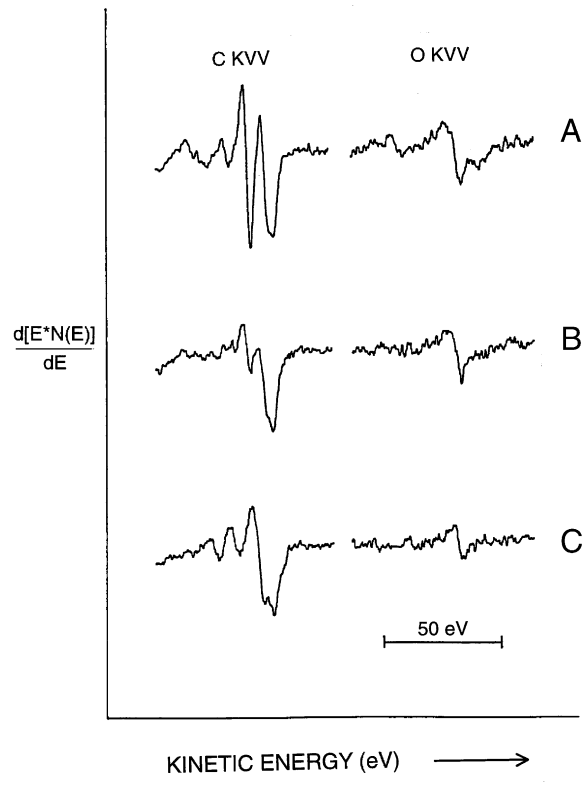


Figure 17. C and O KVV derivative Auger spectra from (**A**) CO molecularly adsorbed on Ni, (**B**) after exposure to a 1.5-keV electron beam with a current density of ~1 μA/mm^2 for 10 min, and (**C**) after exposure to the beam for 40 min.

and it can decompose species at the surface. Electron beam modification can be reduced by decreasing the beam energy, by decreasing the current density at the surface by defocusing the beam, or by rastering the beam over a certain area. If the specimen is laterally homogeneous, electron beam rastering can always be used. An example of electron beam decomposition is shown in Figure 17, where it can be seen that a low-energy beam (1 keV) with a current density of $\sim 1\ \mu A/mm^2$ produces marked changes, in a few minutes, in the C KVV and O KVV Auger spectra for CO molecularly adsorbed on Ni. In fact, some C and O is desorbed from the surface, and the remaining species are converted into nickel carbide and oxide (Hooker and Grant, 1976).

PROBLEMS

Problems in conducting an Auger analysis can occur from many sources. These include specimen modification from the electron beam and electrical charging of insulators. Problems can also be encountered in data reduction, such as peak overlap, misidentification of Auger peaks, and the possible presence of ionization loss peaks, plasmon loss peaks, and ion-excited Auger peaks in the spectra. Since ionization loss peaks, plasmon loss peaks, and ion-excited Auger peaks have not been discussed earlier, their effects on the interpretation of spectra will be discussed here.

Ionization Loss Peaks

When an electron is ejected from an atomic shell (having a binding energy E_b) by an electron from the incident beam (having a kinetic energy E_p), the energy remaining will be $(E_p - E_b)$. This remaining energy is shared between the ejected electron and the incident electron, and these electrons will also appear in the measured energy distribution, $N(E)$. This means that there will be a background of electrons in $N(E)$ from this process up to a kinetic energy of $(E_p - E_b)$, resulting in a small reduction in the intensity of $N(E)$ above this energy. This small step in $N(E)$ is sometimes observed in derivative spectra, and is called an ionization loss peak (Bishop and Riviere, 1970). Its source can be readily confirmed by changing the incident beam energy, as the feature will move in the $N(E)$ distribution by the change in beam energy (remember that Auger peaks do not change their kinetic energy when the electron beam energy is changed). These ionization loss peaks can be useful in studying surfaces, for example, where overlap of Auger peaks might make detection difficult. They can also be used to study changes in surface chemistry, as any shift in binding energy of the energy level in the atom will be reflected as a change in kinetic energy of the ionization loss peak. An example of ionization loss peaks is shown in Figure 18, for the Si L_1 and $L_{2,3}$ levels in SiO_2. In the case of conductors, ionization loss peaks might overlap plasmon loss peaks (see Plasmon Loss Peaks section), for levels with low binding energy.

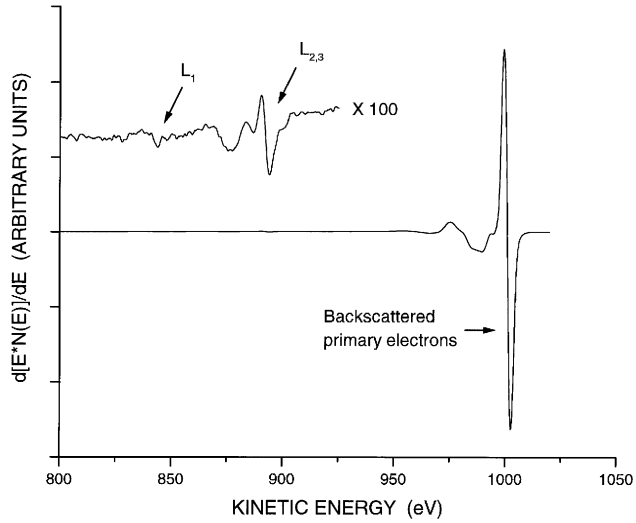


Figure 18. Ionization loss peaks in the derivative mode for the L_1 and $L_{2,3}$ levels in SiO_2. The spectrum was taken with a 100-nA, 1000-eV electron beam in <10 min. The symmetrical peak at 1000 eV is from backscattered primary electrons.

Plasmon Loss Peaks

An electron traveling through a conducting material can interact with the free electrons in the material and lose a specific amount of energy, resulting in an additional peak separated in energy by what is called the plasmon energy (Klemperer and Shepherd, 1963). Several orders of bulk plasmons can occur, and they will appear as a series of equally separated peaks of decreasing intensity on the low-kinetic-energy side of the Auger peak. Besides these bulk plasmons, the presence of the surface results in additional plasmons, called surface plasmons, that are associated with the back-scattered primary electrons and with each bulk plasmon peak. Higher-order surface plasmons are very weak and not observed directly in the loss of spectrum (Pardee et al., 1975). An example of plasmon peaks is shown in Figure 19A for 1-keV backscattered primary electrons from an Al surface, where the bulk (B) and surface (S) plasmons have energies of 15.3 and 10.3 eV, respectively. This backscattered spectrum was taken under high energy resolution, and typical Auger peaks and their corresponding plasmon loss peaks will be broader, as can be seen for the Al KLL Auger spectrum in Figure 19B. The Al $KL_1L_{2,3}$ Auger electrons near 1336 eV have their own plasmons, and the $KL_1L_{2,3}$ Auger peak also overlaps with plasmon peaks from Al $KL_{2,3}L_{2,3}$ Auger electrons, which have a peak near 1386 eV. Note that the plasmon peaks are equally spaced and decrease in intensity as they move away from the Auger peaks. Aluminum oxide would not have any plasmon peaks. The presence of plasmons will change the shape of an Auger peak, and take electrons away from the Auger peak, making quantitative analysis somewhat more difficult.

Ion-Excited Auger Peaks

Core electron levels for some elements can be ionized by low-energy (1 to 5 keV) inert gas ion bombardment,

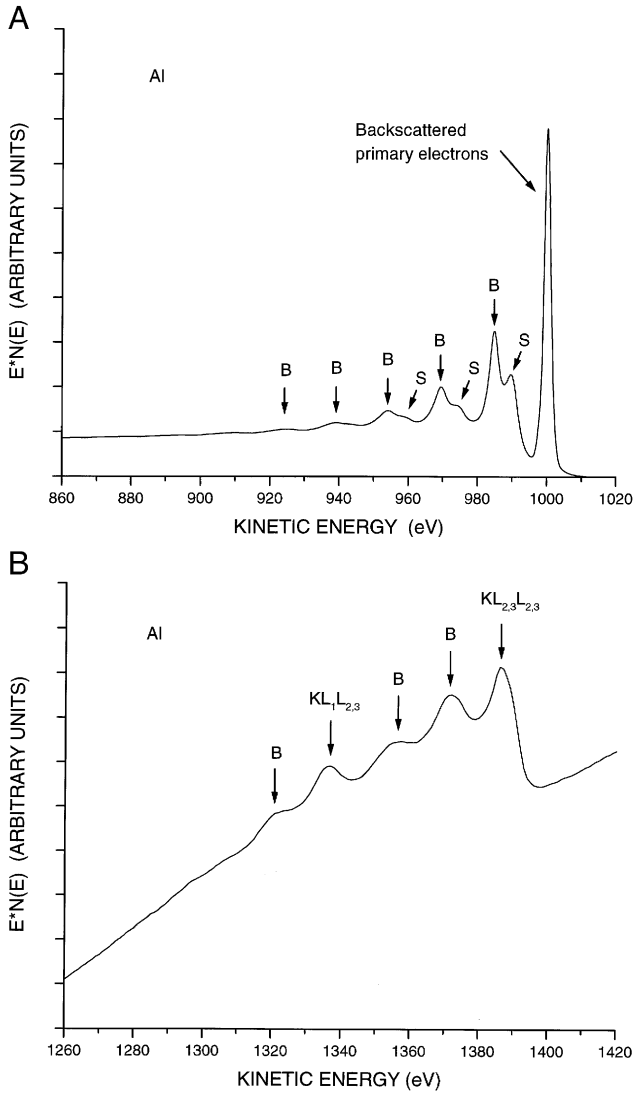


Figure 19. Plasmon peaks (bulk B, surface S) associated with (**A**) a backscattered primary electron beam at 1000 eV from Al, and (**B**) the Al $KL_1L_{2,3}$ (near 1337 eV) and $KL_{2,3}L_{2,3}$ (near 1386-eV) Auger peaks from Al obtained with a 5-keV electron beam. For Al, the bulk and surface plasmons have energies of 15.3 and 10.3 eV, respectively. The surface plasmons in (**B**) are obscured by the broad bulk plasmon peaks (reflecting the broad Auger peaks).

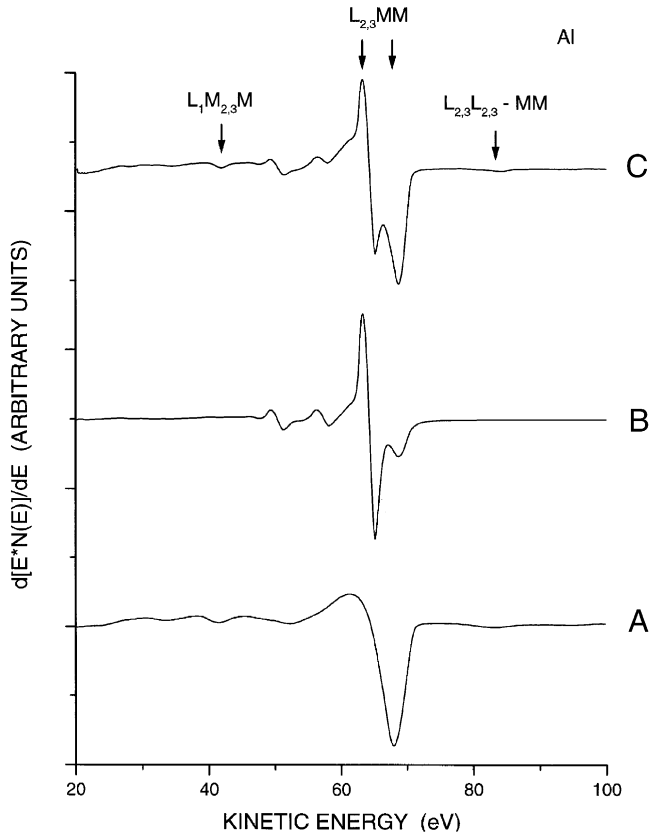


Figure 20. Derivative LMM Auger spectra, (**A**) excited by a 5-keV electron beam, (**B**) excited by a 3-keV Ar^+ ion beam, and (**C**) excited by both electron and ion beams together, as might occur in a sputter depth profile through Al.

resulting in the emission of Auger electrons from these atoms (Haas et al., 1974). Only those energy levels in an atom of the sample that can be promoted to a much lower binding energy by the incident ion during its interaction with the atom can be ionized (the relatively heavy ion cannot transfer enough energy to an electron to ionize core shell electrons). The atom is ionized in a core shell if an electron is transferred to the incident ion from a promoted core shell energy level during impact, and if it remains transferred after the collision. The ionized level then returns to its higher binding energy, and Auger electron emission can then occur. Typical levels that can be ionized by this method are the $L_{2,3}$ levels of Mg, Al, and Si, and additional Auger intensity might occur on depth profiling through layers containing these elements. The L_1 level of Mg, Al, and Si cannot be ionized, as this level is not promoted during collision with the incoming ion. During ion bombardment, Auger emission can occur from sputtered specimen ions as well as from the specimen. The spectrum produced by the ion beam will be different from that produced by an electron beam in two ways (1) the spectrum produced by ion impact will not contain all the transitions produced by electron impact, since not all levels can be promoted (e.g., the L_1 levels in these atoms), and (2) sharp, atomic-like Auger transitions will occur from sputtered ions as well (Grant et al., 1975b). A comparison of electron- and ion-excited Auger emission from Al is shown in Figure 20. The spectrum produced by electron excitation, Figure 20A, shows the main $L_{2,3}MM$ transition at 68 eV, and the $L_1L_{2,3}M$ and $L_{2,3}L_{2,3}$-MM transitions at 41 and 83 eV, respectively. The spectrum produced by Ar^+ excitation, Figure 11d.20B, shows the $L_{2,3}MM$ transition at 68 eV from the bulk Al, together with some overlapping symmetric Auger peaks from sputtered Al ions, but the $L_1L_{2,3}M$ and $L_{2,3}L_{2,3}$-MM transitions are not seen. The L_1 level is not promoted on impact with an Ar^+ ion, so the $L_1L_{2,3}M$ transition cannot occur. Also, double initial ionization of the $L_{2,3}$ level does not occur on ion impact.

Composite spectra, such as the one shown in Figure 20C, can be obtained during simultaneous electron and ion bombardment of Al. Note that the symmetric, ion-excited peaks from sputtered ions are sharper than the electron-excited peaks, and will therefore appear relatively larger in derivative spectra than in direct spectra.

LITERATURE CITED

Baer, D. R. and Thomas, M. T. 1986. A technique for comparing Auger electron spectroscopy signals from different spectrometers using common materials. *J. Vac. Sci. Technol. A* 4:1545–1550.

Bishop, H. E. and Riviere, J. C. 1970. Characteristic ionization losses observed in Auger electron spectroscopy. *Appl. Phys. Lett.* 16:21–23.

Childs, K. D., Carlson, B. A., LaVanier, L. A., Moulder, J. F., Paul, D. F., Stickle, W. F., and Watson, D. G. 1995. Handbook of Auger electron spectroscopy. Physical Electronics, Inc., Eden Prairie, Minn.

Gaarenstroom, S. W. 1981. Principal component analysis of Auger line shapes at solid-solid interfaces. *Appl. Surf. Sci.* 7:7–18.

Goto, K. 1995. Historical Auger electron spectroscopy. I. Works of P. Auger and other pioneers. *J. Surf. Anal.* 1:328–341.

Grant, J. T. and Haas, T. W. 1970. Auger electron spectroscopy of Si. *Surf. Sci.* 23:347–362.

Grant, J. T., Hooker, M. P., and Haas, T. W. 1975a. Spectrum subtraction techniques in Auger electron spectroscopy. *Surf. Sci.* 51:318–322.

Grant, J. T., Hooker, M. P., Springer, R. W., and Haas, T. W. 1975b. Comparison of Auger spectra of Mg, Al, and Si excited by low-energy electron and low-energy argon-ion bombardment. *J. Vac Sci. Technol.* 12:481–484.

Haas, T. W. and Grant, J. T. 1969. Chemical shifts in Auger electron spectroscopy from the initial oxidation of Ta(110). *Phys. Lett.* 30A:272.

Haas, T. W. and Grant, J. T. 1970. Chemical effects on the KLL Auger electron spectrum from surface carbon. *Appl. Phys. Lett.* 16:172–173.

Haas, T. W., Grant, J. T., and Dooley, G. J. 1972a. On the identification of adsorbed species with Auger electron spectroscopy. *In* Adsorption-Desorption Phenomena (F. Ricca, ed.) pp. 359–368. Academic Press, New York.

Haas, T. W., Grant, J. T., and Dooley, G. J. 1972b. Chemical effects in Auger electron spectroscopy. *J. Appl. Phys.* 43:1853–1860.

Haas, T. W., Springer, R. W., Hooker, M. P., and Grant, J. T. 1974. Ion excited Auger spectra of aluminum. *Phys. Lett.* 47A:317–318.

Hooker, M. P. and Grant, J. T. 1976. Auger electron spectroscopy studies of CO on Ni(110)–spectral line shapes and quantitative aspects. *Surf. Sci.* 55:741–746.

Hooker, M. P., Grant, J. T., and Haas, T. W. 1976. Some aspects of an AES and XPS study of the adsorption of O_2 on Ni. *J. Vac. Sci. Technol.* 13:296–300.

Janssen, A. P., Schoonmaker, R. C., Chambers, A., and Prutton, M. 1974. Low energy Auger and loss electron spectra from magnesium and its oxide. *Surf. Sci.* 45:45–60.

Klemperer, O. and Shepherd, J. P. G. 1963. Characteristic energy losses of electrons in solids. *Adv. Phys.* 12:355–390.

Krause, M. O. 1979. Atomic radiative and radiationless yields for K and L shells. *J. Phys. Chem. Ref. Data* 8:307–327.

Pantano, C. G. and Madey, T. E. 1981. Electron beam effects in Auger electron spectroscopy. *Appl. Surf. Sci.* 7:115–141.

Pardee, W. J., Mahan, G. D., Eastman, D. E., Pollak, R. A., Ley, L., McFeely, F. R., Kowalczyk, S. P., and Shirley, D. A. 1975. Analysis of surface- and bulk-plasmon contributions to x-ray photoemission spectra. *Phys. Rev. B* 11:3614–3616.

Prutton, M., Larson, L. A., and Poppa, H. 1983. Techniques for the correction of topographical effects in scanning Auger microscopy. *J. Appl. Phys.* 54:374–381.

Seah, M. P. 1983. Quantification of AES and XPS. *In* Practical Surface Analysis (D. Briggs and M. P. Seah, eds.)., Chapter 5. John Wiley & Sons, Chichester, U.K.

Seah, M. P. and Dench, W. A. 1979. Quantitative electron spectroscopy of surfaces: A standard data base for electron inelastic mean free paths in solids. *Surf. Interface Anal.* 1:2–11.

Seah, M. P., Smith, G. C., and Anthony, M. T. 1990. AES: Energy calibration of electron spectrometers. I. An absolute, traceable energy calibration and the provision of atomic reference line energies. *Surf. Interface Anal.* 15:293–308.

Sekine, T., Nagasawa, Y., Kudoh, M., Sakai, Y., Parkes, A. S., Geller, J. D., Mogami, A., and Hirata, K. 1982. Handbook of Auger Electron Spectroscopy. JEOL Ltd., Tokyo.

Sickafus, E. N. and Colvin, A. D. 1970. A mutichannel monitor for repetitive Auger electron spectroscopy with application to surface composition changes. *Rev. Sci. Instrum.* 41:1349–1354.

Smith, G. C. and Seah, M. P. 1990. Standard reference spectra for XPS and AES: Their derivation, validation and use. *Surf. Interface Anal.* 16:144–148.

Stickle, W. F. and Watson, D. G. 1992. Improving the interpretation of x-ray photoelectron and Auger electron spectra using numerical methods. *J. Vac. Sci. Technol. A* 10:2806–2809.

Tarng, M. L. and Fisher, D. G. 1978. Auger depth profiling of thick insulating films by angle lapping. *J. Vac. Sci. Technol.* 15: 50–53.

Taylor, N. J., Johannessen, J. S., and Spicer, W. E. 1976. Crater-edge profiling in interface analysis employing ion beam etching and AES. *Appl. Phys. Lett.* 29:497–499.

Walls, J. M., Hall, D. D., and Sykes, D. E. 1979. Composition-depth profiling and interface analysis of surface coatings using ball cratering and the scanning Auger microprobe. *Surf. Interface Anal.* 1:204–210.

Zalar, A. 1985. Improved depth resolution by sample rotation during Auger electron spectroscopy depth profiling. *Thin Solid Films* 124:223–230.

KEY REFERENCES

Annual Book of ASTM Standards, Vol. 03.06. American Society for Testing and Materials, West Conshohocken, Pa.

This volume includes the ASTM Standards that have been developed for surface analysis.

Briggs, D. and Seah, M. P. (eds.). 1990. Practical Surface Analysis, 2nd ed., Vol. 1. Auger and X-ray Photoelectron Spectroscopy. John Wiley & Sons, Chichester, U.K.

This is perhaps the most popular book on Auger electron spectroscopy, and is available in paperback.

Ferguson, I. 1989. Auger Microprobe Analysis. Adam Hilger, Bristol, U.K.

This book provides an overall view of Auger electron spectroscopy, with many historical references. It also has a lot of information on the scanning aspects of the measurement.

INTERNET RESOURCES

See the Appendix for additional internet resources.

http://sekimori.nrim.go.jp/

Web site of the Surface Analysis Society of Japan. This site has access to a Common Data Processing System that is being developed, and to Auger spectra from many of the elements in the periodic table. The spectra can be viewed on line, and members of the Society can download files. Membership information is available online.

http://www.vacuum.org/

The American Vacuum Society is the main professional society for those working in Auger electron spectroscopy in the United States. This page has links to its scientific journals, scientific meetings, a buyers guide, and to the International Union for Vacuum Science, Technique, and Applications.

http://www.uwo.ca/ssw/

Surface Science Western home page. A surface science mailing list for users in surface science (mainly AES and XPS) has been set up by Surface Science Western at the University of Western Ontario, and information about the mailing list can be accessed from this page. An archive of messages is also kept.

APPENDIX: SOURCES OF STANDARDS FOR AUGER ELECTRON SPECTROSCOPY

National Institute of Standards and Technology (NIST)

Standard Reference Materials Program, Building 202, Room 204, Gaithersburg, Md., 20899. Telephone: (301) 975-6776; Fax: (301) 948-3730; e-mail: *srminfo@nist.gov*; WWW: *http://ts.nist.gov/ts/htdocs/ 230/232/232.htm*.

The following two standards from NIST are for calibrating equipment used to measure sputtered depth and erosion rates in surface analysis.

SRM 2135c: Nickel/chromium thin-film depth profile standard.

SRM 2136: Chromium/chromium oxide thin-film depth profile standard.

National Physical Laboratory (NPL)

Surfaces and Interfaces Section, CMMT, National Physical Laboratory, Queens Road, Teddington, Middlesex TW11 0LW, U.K. Telephone: +44 181 943 6620; Fax: +44 181 943 6453; e-mail: *sjs@npl.co.uk*; WWW: *http://www.npl.co.uk/npl/cmmt/sis/refmat.html*.

CRM261: Tantalum pentoxide depth profile reference materials, consisting of four 30-nm-thickness samples and four 100-nm-thickness samples of Ta_2O_5 on Ta, each $\sim$10 mm$\times$5 mm in size.

SCAA87: Copper, silver, and gold reference materials for intensity and energy calibration of Auger electron spectrometers.

AES intensity calibration software: Used to calculate the intensity response (or transmission) function of the spectrometer over a kinetic energy range of 20 to 2500 eV.

PC138 software: To check that data files can be saved in compliance with the VAMAS (Versailles Project on Advanced Materials and Standards) Standard Data Transfer Format.

American Society for Testing and Materials (ASTM)

100 Barr Harbor Drive, West Conshohocken, Pa. 19428-2959. Telephone: (610) 832-9500; Fax: (610) 832-9555; e-mail: *gcolling@astm.org*; WWW: *http://www.astm.org/COMMIT/*.

There are currently 13 standards covering terminology, practices and guides published by ASTM, that relate to Auger electron spectroscopy. Some of the most useful documents are as follows.

E673: Standard Terminology Relating to Surface Analysis.

E827: Standard Practice for Elemental Identification by Auger Electron Spectroscopy.

E983: Standard Guide for Minimizing Unwanted Electron Beam Effects in Auger Electron Spectroscopy.

E984: Standard Guide for Identifying Chemical Effects and Matrix Effects in Auger Electron Spectroscopy.

E995: Standard Guide for Background Subtraction Techniques in Auger Electron and X-ray Photoelectron Spectroscopy.

E1078: Standard Guide for Specimen Preparation and Mounting in Surface Analysis.

E1127: Standard Guide for Depth Profiling in Auger Electron Spectroscopy.

E1829: Standard Guide for Handling Specimens Prior to Surface Analysis.

JOHN T. GRANT
University of Dayton
Dayton, Ohio

ION-BEAM TECHNIQUES

INTRODUCTION

Ion Beam Analysis (IBA) involves the use of a well-defined *Beam* of energetic *Ions* for the purpose of *Analyzing* materials. In general, it is the combined attributes of the incident ions' energy, mass, and mode of interaction with the target that distinguish each of the analytic techniques described in this chapter. The range of possible ion-atom interactions that can occur in a target, and the results of that interaction that produces the detected signal, give each technique specific advantages under specific circumstances. In the units of this chapter, IBA techniques are described which utilize ion species spanning nearly the entire periodic table, ion energies which extend from several electron volts to several million electron volts, and detected signals consisting of nuclei, electrons, γ rays, x rays, optical photons, collected electrical charge, and even logic states in digital circuits.

One feature of IBA that distinguishes it from most other analytical techniques is the incredible wealth of signals that are available for detection. A simple beam of protons incident on a target offers the experimenter a "detection menu" consisting of: (1) back-scattered protons, (2) α particles formed by nuclear reactions (if the energy is high enough), (3) γ rays caused by these same reactions or Coulomb excitation, (4) x rays produced by inner atomic shell ionization followed by radiative decay, (5) secondary ions emitted by sputtering, (6) secondary electrons ionized near the surface, (7) photons produced by the ionolumenescence process, (8) electron-hole pairs collected from within biased semiconductor targets, and on and on in glutinous profusion. Of course we are not limited to just protons—one can accelerate deuterons, tritons, ^{3}He, ^{4}He, ^{6}Li, ^{7}Li, etc. Given all of the possible signals from all of the possible beams at all of the possible energies, a virtually mind-boggling range of measurements is possible. Moreover, owing to the interplay of ion, energy, and target composition combined with the capability to finely focus the beam, the experimenter can preferentially probe different regions of a sample at times in three dimensions. In the simplest of cases, for example, low-energy ions interact only with atoms that are near the surface, while high-energy ions permit measurements of the composition deep within a specimen. Thus, a modern IBA lab can usually provide a way to analyze virtually any isotope of any element, at any position, inside any sample.

The connecting thread that binds the broad range of analytical techniques presented in this chapter may not be immediately obvious, due to the widely disparate energies and ion species that are employed to probe for specific elements in specific targets. Yet, the relationships among the different IBA techniques becomes more apparent when each is plotted in a "phase space" defined by the projectile ion, energy, and the electronic and nuclear stopping powers for a specific target (i.e., the rates at which an ion loses energy in a sample due to electronic and nuclear collisions). In such a phase space of IBA techniques, shown in Figure 1, a coordinate system of projectile atomic number (specifically for H, He, C, Si, Mo and Au) is plotted versus ion energy/mass (logarithmic in E/amu) within the framework of a log-log plot of nuclear stopping power [$S(n)$] versus electronic stopping power [$S(e)$], where the stopping power units are MeV/(mg/cm^2). This plot has been generated for the "favorite son" of IBA—silicon—as the target. Aside from the qualitative rendering of the relationships among the IBA techniques, such plots also readily provide quantitative comparisons of the stopping powers of all ions across nearly all energies in a given target.

The different IBA techniques are mapped into this space by assigning each technique that region of the phase space representing its typical ion and energy parameters. Beginning with the least utilized region, i.e., that of relativistic energies, the region labeled "REL" applies to energies greater than $\sim$100 MeV/amu—approximately half the speed of light. The region labeled NRA (nuclear reaction analysis) is next. Its low energy boundary is set by the height of the Coulomb barriers for collisions between ions and Si; above this barrier, the nuclei can collide permitting the possibility of nuclear reactions. Below the Coulomb barrier is the elastic scattering region characterized by RBS (Rutherford back-scattering) for low-Z projectiles, and ERD (elastic recoil detection) for high-Z projectiles. Note that it is in this region that the electronic (and total) stopping power is maximum, and therefore the highest depth resolution IBA depth profiling is done with ions in this category. This is also the region where proton-induced x-ray emission (PIXE) is performed. The low energy boundary for this region occurs where electron screening effects become important. At still lower energies, MEIS (medium energy ion scattering) and HIBS (heavy ion back-scattering spectrometry) come into use. HIBS denotes the use of higher-Z projectiles with larger stopping powers and high sensitivity to trace elements. MEIS also takes advantage of very large stopping powers and the enormous Coulomb scattering cross-sections which tend to vary with the inverse of the projectile energy squared. Because of this, MEIS techniques also have very good depth resolution and, at times, extremely high sensitivity. We have defined this region's lower energy boundary using a velocity equal to approximately one-half the Bohr velocity. Ions with energies less than the Bohr velocity become neutralized very easily because their velocity is almost exactly matched to that of a target's outer-shell electrons which can then be easily captured by the ion. This marks the onset of the LEIS (low energy ion scattering) regime; owing to its low energy, LEIS is highly surface sensitive. When inert-gas ions are used, LEIS can provide an analysis of the outermost atomic layer in a material. Moreover, since in this regime the nuclear stopping power is a maximum, SIMS (secondary ion mass spectrometry) is performed at these energies with heavier ions that have correspondingly large sputtering yields. Finally, for ions with energies of less than 100 eV, we leave the realm of

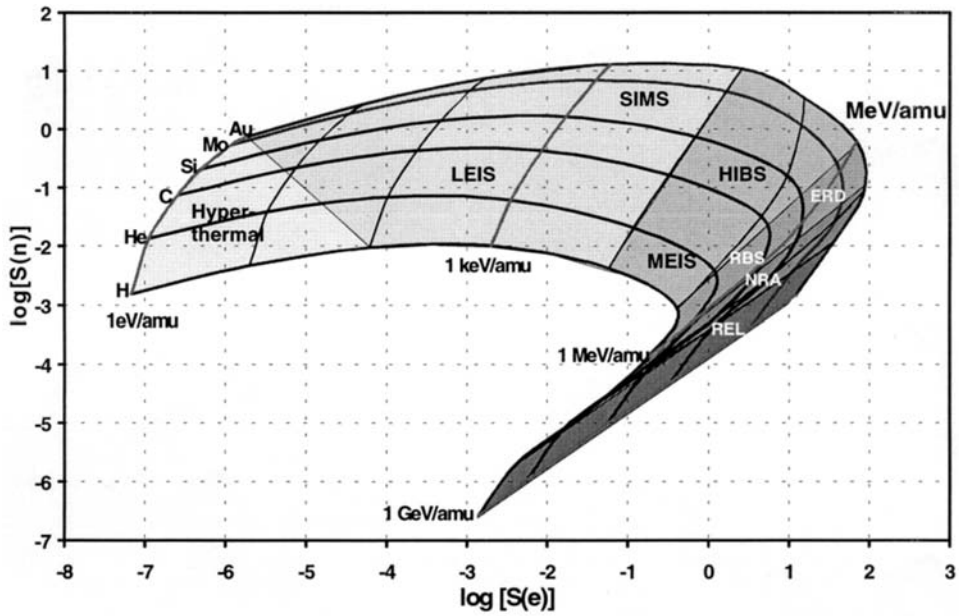


Figure 1. IBA Regimes for Si Target Incident Ions H-Au Energies from 1 eV to 1 GeV/amu.

binary collisions and enter a regime called hyperthermal, where ion-solid interactions begin to include both multibody kinematic and chemical interactions.

Common to each of the IBA techniques described in this chapter is the use of the particle accelerator—a device that the public most commonly associates with the "atom smashers" used in the early part of the 20th century to probe the structure of the atom, and later, the nucleus. Our dependence on this tool is somehow fitting given that virtually all of the techniques discussed in the following units can trace their development to the pioneering work performed by the great physicists of that era—Rutherford, Compton, Bethe, Bohr, etc. This legacy becomes apparent by simply noting that in performing any of these analyses, we are usually repeating a classic experiment, albeit with a different emphasis and better equipment, that was once performed by one of the giants of modern physics. Today's IBA techniques are the direct descendants of investigations into the nature of the atom by Lord Rutherford using back-scattered α particles, the structure of the electronic orbitals by Niels Bohr using characteristic x-ray spectra, and even the nature of stellar energy generation by Hans Bethe using the cross-sections of ion-induced nuclear reactions. When performing IBA, it is sometimes hard not to actually feel the presence of one of these legends on whose shoulders we stand.

As the reader delves into each of the following chapters, the underlying similarities in these techniques will repeatedly appear. This is not surprising, since the basis of all ion-atom scattering kinematics is contained in the simple central force problem first solved by Kepler, (and perfected for IBA purposes by Minnesota Fats and Willie Mosconi!). Likewise, in a large fraction of these techniques, the scattering cross-sections are given by the famous Rutherford equation. Thus, while it is the unique nature of the collision physics within each ion/energy regime that lends each IBA technique its particular utility for a given measurement, it is the shared traits of ion-solid interactions which unite these techniques into a single scientific field which has been the useful servant of much of today's science and technology.

BARNEY L. DOYLE
KEVIN M. HORN

HIGH-ENERGY ION BEAM ANALYSIS

INTRODUCTION

It is very hard to identify the earliest application of high-energy ion beam analysis (IBA), which was a natural outgrowth of accelerator-based atomic and nuclear physics that started in the 1940s with the development of particle accelerators. In these early experiments, the composition of the target to be studied was well known, and it was the nuclear scattering or reaction cross-section that was measured. It would not be surprising if one of the very first nuclear physics pioneers turned this emphasis around and used their beams to measure the compositional uniformity of an unknown sample. Perhaps the earliest known proposal for IBA was that of E. Rubin in 1949 (Rubin, 1949, 1950) at the Stanford Research Institute, which spelled out the principle of Rutherford backscattering spectrometry (RBS). Another very early application of IBA was performed by Harald Enge at MIT (Enge et al., 1952). Enge used mega-electron-volt deuterons from one of Van de Graaff's first accelerators with a magnetic spectrometer to measure the constituents of aluminized Formvar in perhaps the earliest materials analysis application of nuclear reaction analysis (NRA). It is interesting to note that the

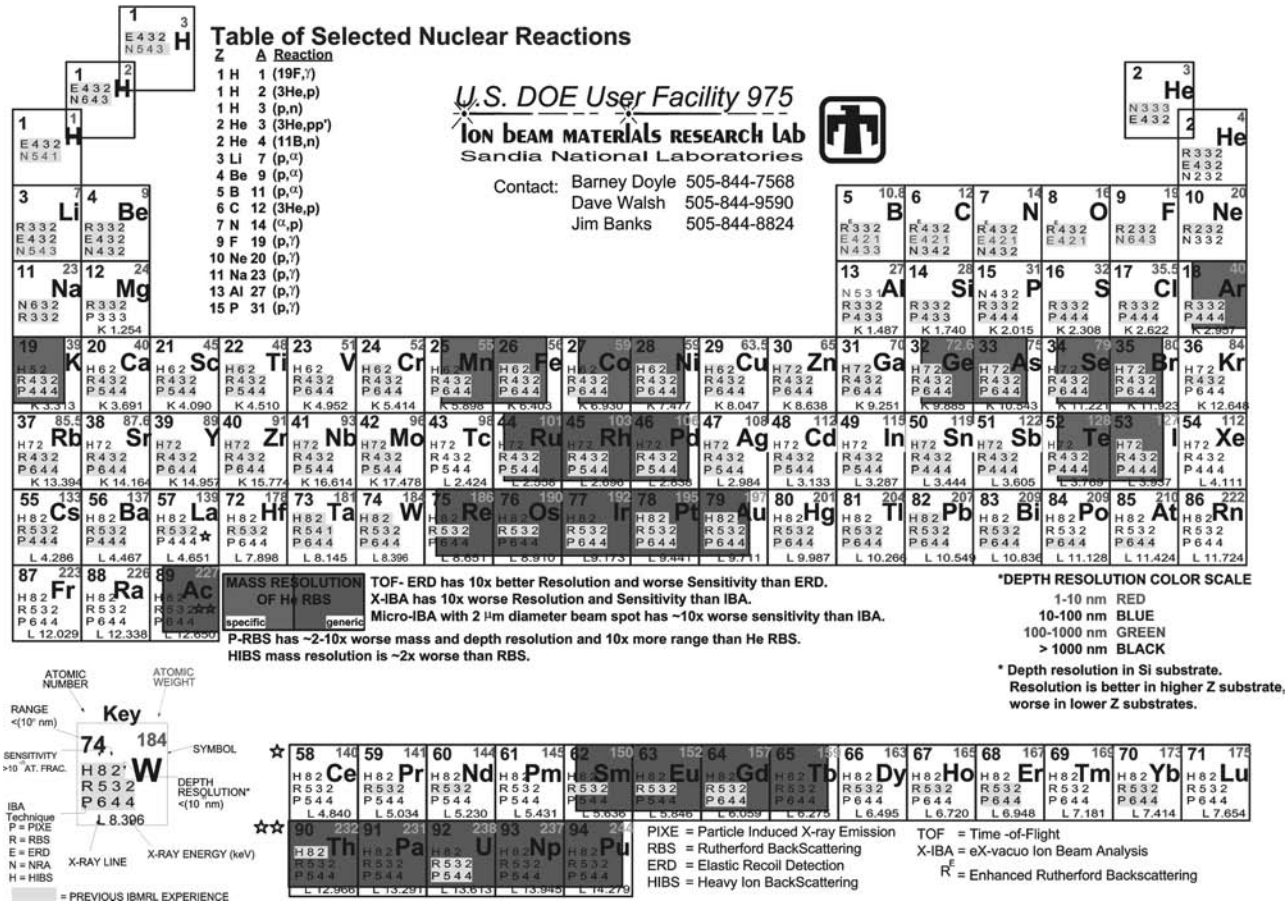


Figure 1. Ion beam analysis table of the elements.

use of magnetic spectrometers to obtain ultrahigh depth resolution is now common, and Dollinger et al. (1995) have obtained single atomic layer resolution by using an Enge Q3D magnetic spectrometer for IBA.

High-Energy IBA is generally thought to apply when the ion beam has energy greater than 0.5 MeV. There are several attributes of high-energy IBA, including (1) the ability to measure virtually any element in the periodic table, at times with isotopic specificity; (2) the ability to determine concentration as a function of depth; (3) that trace element sensitivity is not uncommon; and (4) that the technique is nondestructive with respect to the chemical makeup of the sample.

The periodic table in Figure 1 provides a convenient overview of high-energy IBA capabilities.

ION BEAM ANALYSIS PERIODIC TABLE

The purpose of the periodic table is to quickly provide information on the sensitivity, depth of analysis, and depth resolution of most modern ion beam analysis techniques in a single, easy-to-use format. One can find an interactive version of this table at: *http://www.sandia.gov/1100/1111/Elements/tablefr.htm*.

Key: Technique, Sensitivity, and Depth Resolution

The information is organized in a periodic table to facilitate its use, as most readers may have only little knowledge of IBA but will know what element they want to detect. A key to the information included in this table is found in the lower left for the element W, or tungsten. The first letter indicates the technique. The numbers that follow the technique label are logarithmic scales for sensitivity, analysis range, and depth resolution, respectively. For example (using the key), H (HIBS) can be used to measure W with a sensitivity better than 10^{-8} atomic fraction to a depth range greater than 10^2 nm. Likewise, R (RBS) can measure W with a sensitivity of 10^{-5} atomic fraction to a depth of 10^3 nm with a depth resolution better than 10^2 nm, and P (PIXE) can measure W down to 10^{-6} atomic fraction to a range and depth resolution of 10^4 nm.

This sensitivity and depth information has been calculated based on available ion-atom interaction cross-sections and assumes that (1) 100 counts are measured, (2)

the integrated incident beam current is 10 particle microcoulombs, and (3) the largest practical detector solid angles are used. These calculations ignore backgrounds and other interferences, but IBA scientists routinely use integrated incident beams of greater than 10 μC, and therefore the sensitivities cited are actually rather conservative. In many cases, actual sensitivities are an order of magnitude better. The depth resolutions listed are also very conservative and should be thought of as upper limits.

Mass Resolution

The shaded regions in the table are used to indicate the mass resolution range for He beam RBS. For the first four rows of the table (through Ar), the resolution is adequate to separate all adjacent elements except specific cases of mass interference, which are shaded (e.g., Mn and Fe cannot be separated because their mass differs by only 1 amu). For the remainder of the table, no adjacent elements are separable, and the lightly shaded "generic" mass resolution range is characteristic for an entire row of the periodic table (e.g., the shading of Ru-Rh-Pd indicates that for row 5 of the periodic table the mass resolution is not adequate to separate Ru from Pd but could separate Ru from Ag, and so on for other elements in this row).

High-energy IBA includes elastic scattering, nuclear reactions, and particle-induced X-ray emission (PIXE). We next provide brief descriptions of the techniques that fall under these headings.

ELASTIC SCATTERING (see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS)

Rutherford Backscattering Spectrometry

Rutherford backscattering spectrometry (Chu et al., 1978; Feldman and Mayer, 1986) has matured into one of the most important analytical techniques for the elemental characterization of materials and is used routinely in material science applications in many laboratories around the world. Typically, in conventional RBS, H or He ions, at mega-electron-volt energies, are elastically scattered from a nucleus in a sample and detected in a backward direction in a Si surface barrier detector. For a given detection angle, the kinematics of the elastic scattering is well known, and the energy transferred in the collision can be determined from the Rutherford theory. The scattering kinematics and the energy loss of the ion entering and leaving the sample are well understood. Therefore, the energy of the scattered ion provides information about the mass of the scattering center as well as the depth in the sample. The RBS information may provide the elemental composition and the depth distributions for single- or multi-layered samples at sensitivities as low as a few parts per million. Sensitivities are much better for heavier elements on lighter substrates, although elements with atomic numbers, Z, from 4 to 94 have been detected. Light elements in heavy matrices are much more difficult to analyze. Depth resolutions may vary from 1 to 100 nm. Analysis depth ranges are 0.1 μm to tens of micrometers, and sensitivities vary from 0.1 at. % to ~10 ppm, depending on the atomic number of the trace element.

Heavy-Ion Backscattering

Heavy-ion backscattering spectrometry (HIBS) (Doyle et al., 1989; Weller et al., 1990; Knapp and Banks, 1993) employs heavier incident ions. This increases the elastic scattering cross-section, which is proportional to $(Z_1Z_2/E)^2$, and increases the sensitivity for detection as well as the mass resolution. Here, Z_1 and E are the atomic number and energy, respectively, of the incident ion and Z_2 is the atomic number of the target atom. Because of the heavier mass of the incident ion, the depth analysis range is reduced to 0.01 to 0.1 μm while the sensitivity is increased to 0.1 to 1 ppm.

Elastic Recoil Detection Analysis

Elastic recoil detection analysis (ERDA) (L'Ecuyer et al., 1976; Doyle and Peercy, 1979; Green and Doyle, 1986; Barbour and Doyle, 1995; Dollinger et al., 1995) uses a high-energy, heavy incident ion to recoil light target elements from a sample in the forward direction. Since RBS is not suitable when $M_1 \geq M_2$, where M_1 is the mass of the incident ion and M_2 is the mass of the target element, ERDA is especially useful for light elements such as H through F. Elastic recoil detection analysis depends on the ability to distinguish between incident ions scattered in the forward direction and light recoiling atoms. Since the heavier incident ions have a larger energy loss than the lighter recoil target atoms, a thin foil is sometimes used to block the incident ions from the Si surface barrier detector. This technique may be used with a time-of-flight detector to achieve much higher depth resolution (1 nm) and even permit the separation of different isotopes of light elements. Sensitivities of 10 to 1000 ppm have been obtained. The glancing angle limits the depth analysis range to ~1 μm.

NUCLEAR REACTION ANALYSIS (see NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION)

Nuclear reaction analysis (Bird, 1989; Moore et al., 1975) and resonant nuclear reaction analysis (RNRA) (Vizkelethy, 1995) are very element specific analysis techniques in which nucleus a bombards target A resulting in a light product b and residual nucleus B, $A(a,b)B$. Nonresonant nuclear reactions (NRA) may occur when the energy of the incident nucleus is above a threshold value for the reaction. Nuclear reaction analysis allows bulk measurements of target nuclei distributed throughout a sample. The sensitivity of NRA depends upon the cross-section for the nuclear reaction. Resonant NRA may be used to depth profile the concentration of target nuclei below the surface in a sample. For both NRA and RNRA, the concentration of nuclei is then used to determine elemental concentration from isotopic compositions of the sample. Depending upon the resonant nuclear reaction employed, the technique may be used to profile from the surface to a few micrometers in depth with a resolution of a few to hundreds of nanometers. For example, commonly used

resonant nuclear reactions for H detection are $^{1}H(^{15}N,\alpha\gamma)^{12}C$ (resonance at 6.385 MeV) and $^{1}H(^{19}F,\alpha\gamma)^{16}O$ (resonance at 6.418 MeV). Depending upon the cross-section for the nuclear reaction, sensitivities may vary from 1 to 1000 ppm.

PARTICLE-INDUCED X-RAY EMISSION
(see PARTICLE-INDUCED X-RAY EMISSION)

Particle- (or equivalently proton-) induced X-Ray emission is a multielement, nondestructive analysis technique with sensitivities from major element down to parts-per-million (ppm) trace element, depending upon the elements of interest (Rickards and Ziegler, 1975; Mitchell and Ziegler, 1977; Johansson and Campbell, 1988). With PIXE, a low-atomic-number (Z_1) incident ion beam up to a few mega-electron-volts in energy bombards a sample producing atomic excitations of the elements (Z_2) present in the sample. Here, Z_2 is the atomic number of the target element. Subsequently, characteristic K, L, or M X rays are emitted from the excited elements and are counted to determine the concentrations of the trace elements. Up to 30 elements can be analyzed simultaneously.

Since, the target ionization cross-sections are proportional to Z_1^2/Z_2^4, the number of target element X rays decreases with increasing target atomic number for a given electronic shell. The target ionization cross-section increases with higher incident ion energy E until the velocity of the ion matches the orbital velocity of the electron being ionized and then decreases as $\sim 1/E$. Typically, Si(Li) detectors with Be entrance windows allow detection of elements of atomic number from 12 to 92. Windowless Si(Li) detectors allow elements down to carbon to be analyzed. Sensitivities of 1 to 100 ppm atomic (ppma) have been achieved. Analysis ranges are 1 μm to tens of micrometers.

SUMMARY

The units that follow provide the details required to select the optimum high-energy IBA technique for specific applications. The IBA periodic table in this introduction clearly portrays the interrelationships of these units.

LITERATURE CITED

Barbour, J. C. and Doyle, B. L. 1995. Elastic recoil detection: ERD. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer, M. Nastasi, J. C. Barbour, C. J. Maggiore, and J. W. Mayer, eds.). pp. 83–138. Materials Research Society, Pittsburgh, Pa.

Bird, J. R. 1989. Nuclear reactions. *In* Ion Beams for Materials Analysis (J. R. Williams and J. R. Bird, eds.). pp. 149–207. Academic Press, Marrickville, Australia.

Chu, W-K., Mayer, J. W, and Nicolet, M.-A. 1978. Backscattering Spectrometry. Academic Press, New York.

Dollinger, G., Bergmaier, A., Faestermann, T., and Frey, C. M. 1995. High-resolution depth profile analysis by elastic recoil detection with heavy ions. *Fresenius J. Anal. Chem.* 353:311–315.

Doyle, B. L. and Peercy, P. S. 1979. Technique for profiling H-1 with 2.5-MeV van de Graaff accelerators. *Appl. Phys. Lett.* 34:811–813.

Doyle, B. L., Knapp, J. A., and Buller, D. L. 1989. Heavy-ion backscattering spectrometry (HIBS): An improved technique for trace-element detection. *Nucl. Instrum. Methods* B42:295–297.

Enge, H., Buechner, W. W., and Sperduto, A. 1952. Magnetic analysis of the $Al^{27}(d,p)Al^{28}$ reaction. *Phys. Rev.* 88:963–968.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. North-Holland, New York.

Green, P. F. and Doyle, B. L. 1986. Silicon elastic recoil detection studies of polymer diffusion: Advantages and disadvantages. *Nucl. Instrum. Methods* B18:64–70.

Johansson, S. A. E and Campbell, J. L. 1988. PIXE: A Novel Technique for Elemental Analysis. John Wiley & Sons, Chichester.

Knapp, J. A. and Banks, J. C. 1993. Heavy-ion backscattering spectrometry for high sensitivity. *Nucl. Instrum. Methods* B79:457–459.

L'Ecuyer, J., Brassard, C., Cardinal, C., Chabbal, J., and Deschenes, L. 1976. Elastic recoil detection. *J. Vac. Sci. Technol.* 14:492.

Mitchell, I. V. and Ziegler, J. F. 1977. Ion induced X-rays. *In* Ion Beam Handbook for Material Analysis (J. W. Mayer and E. Rimini, eds.). pp. 311–484. Academic Press, New York.

Moore, J. A., Mitchell, I. V., Hollis, M. J., Davies, J. A., and Howe, L. M. 1975. Detection of low-mass impurities in thin films using MeV heavy-ion elastic scattering and coincidence detection techniques. *J. Appl. Phys.* 46:52–61.

Rickards, J. and Ziegler, J. F. 1975. Trace element analysis of water samples using 700 keV protons. *Appl. Phys. Lett.* 27:707–709.

Rubin, E. 1949. Chemical analysis by proton scattering. *Bull. Am. Phys Soc.* December 1949.

Rubin, E. 1950. Chemical analysis by proton scattering. *Phys. Rev.* 78:83.

Vizkelethy, G. 1995. Nuclear reaction analysis: Particle-particle reactions. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer, M. Nastasi, J. C. Barbour, C. J. Maggiore, and J. W. Mayer, eds.). pp. 139–165. Materials Research Society, Pittsburgh, Pa.

Weller, M. R., Mendenhall, M. H., Haubert, P. C., Döbeli, M., and Tombrello, T. A. 1990. Heavy ion Rutherford backscattering. *In* High Energy and Heavy Ion Beams in Materials Analysis (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). pp. 139–151. Materials Research Society, Pittsburgh, Pa.

BARNEY DOYLE
Sandia National Laboratories
Albuquerque, New Mexico

FLOYD DEL MCDANIEL
University of North Texas
Denton, Texas

ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS

INTRODUCTION

Composition analyses for virtually all of the elements on the periodic table can be performed through a combination of elastic ion-scattering techniques: Rutherford

backscattering spectrometry (RBS), elastic recoil detection (ERD), and enhanced-cross-section backscattering spectrometry (EBS). This unit will review the basic concepts used in elastic ion scattering for composition analysis and give examples of their use. Elastic ion scattering is also used for structure determination utilizing techniques such as ion channeling (see, e.g., Feldman et al., 1982). Composition determination using ion-beam analysis (IBA) falls into two categories: composition-depth profiling and determination of the total integrated composition of a thin region. Unknown species in a sample may be identified using elastic ion-scattering techniques, but in general, analyses involving characteristic x rays or γ rays are more accurate for determining the presence of unknown constituents. For elastic ion scattering, if the ion beam and ion energy are chosen to have sufficiently high energy loss in the sample, then a composition profile is determined as a function of depth by collecting one ion-scattering spectrum. To represent accurately the composition profile as a function of depth in the sample, an accurate knowledge of the sample's volume density is required because IBA techniques are insensitive to density. If the density of the sample is unknown, then the depth in the sample is given in terms of an areal density: depth times volume density. When the incident ion beam is very light or the ion energy very high, then the composition as a function of depth may not be resolved fully, in which case the average composition of the layer of interest is determined by integrating the signal. In this way, the number of atoms per unit area (areal density) is determined from the integrated signal. Table 1 lists typical examples of composition analyses using different elastic ion-scattering techniques with different ion beams and ion-beam energies.

The term elastic refers to the fact that the target nucleus is unchanged after the scattering event (except for kinetic energy), and therefore inelastic conversion between mass and energy or excitation of the nucleus to a higher energy state is absent. Energy and momentum are conserved, and in the center-of-mass frame of reference, the position of the center of mass remains unchanged. The term Rutherford is used to indicate that the scattering cross-section σ is predominantly a result of Coulomb scattering, and therefore σ can be described by the Rutherford formula (Rutherford, 1911). Elastic ion-scattering techniques are often thought of as nondestructive because they do not require activation of the sample, sputtering of the sample, or any other change of the sample to determine the composition of the sample. However, the ion beam does interact with the sample by breaking bonds, exciting electrons, producing phonons, and possibly creating displacements. Therefore, techniques that are sensitive to these properties (e.g., electrical and optical measurements or x-ray diffraction) should be performed prior to IBA.

Now consider a general composition analysis experiment in which a flux of particles is incident upon a sample. These particles can be electrons, ions, or photons that are scattered elastically or absorbed in the sample through energy loss of the particle. If the particle causes an electronic transition in the target atoms, then the energy of emitted radiation is characteristic of that atom and is used to identify the atom. The number of atoms in a sample is determined from knowledge of the type of interaction between the incident particle and the target atom and the probability that an interaction will occur. Each type of interaction has a different cross-section, which is the effective area for an interaction to occur between one of the incident particles and one of the target atoms. The probability for an interaction is then given by the cross-section times the areal density. This is a general description for many types of composition analysis techniques, including elastic ion scattering, and a single equation describes the quantitative composition analysis of a material: $Y = Q\sigma N\tau\Omega\eta$, where Y is the ion-scattering yield, Q is the number of incident particles, σ is the cross-section for a given process, N is the atomic density, τ is the layer thickness, Ω is the detector solid angle, and η is the detector efficiency. The yield is a measured quantity, and thus all of the details needed to determine N, the composition, reduce to a determination of the other variables in the yield equation. The rest of this unit will describe how these other variables are measured or calculated from measured

Table 1. Elastic Ion-Scattering Techniques for Typical Composition Analysis

	Energy (MeV)	θ or ϕ	Elements Analyzed	Sensitivity (at.%)	Analysis Range (μm)	Depth Resolution (μm)	Comments
He RBS	1–3.5	164	Na to U	0.001–0.1	~1.5	0.03	
H RBS[a]	1–2	164	Na to U	0.01–1.0	~10	0.08	
He ERD	2–3	164	H and D	0.1	~1 (polymers)	0.1	Non-Rutherford cross-sections
Si ERD	12–16	30	H to Li	0.05	~0.4	0.02	Range for H
	18–28		H to N	0.1	~0.4	0.03	Range for N
	28–32		H to N and O	0.2	~0.2	0.05	Range for O
He EBS	3.5	164	C	0.01	~0.5	0.03	Depth profile
	3.5		N	0.01	~0.5	0.03	Depth profile
	8.7		O	0.01	~7	0.05	Depth profile
H EBS[a]	2.5	164	O	0.1	~4	0.2	Depth profile
	4.4		S	0.1	~1		Areal density

[a] *CAUTION:* Risks possible radiation exposure to people from prompt nuclear reactions or activation of the samples. Consult your health physicist before doing high-energy proton irradiation.

quantities. In addition, IBA techniques provide depth profiles because the energy loss of the incident and scattered particles are well characterized through analytical formulas (Ziegler, 1980, 1987; Ziegler et al., 1985; Tesmer et al., 1995), and therefore the energy of these particles is described well along the path of the particles. The yield is then calculated as a function of depth in the sample because the energy at each depth in the sample is well known.

Elastic ion scattering methods using both RBS and ERD are simple and fast for quantitative elemental depth profiling in thin films (≤ 1 μm thick) for every element on the periodic table. Elastic recoil detection is often used for profiling elements with atomic number $Z \leq 8$. Sometimes enhanced-cross-section RBS is used to profile elements in the range $5 \leq Z \leq 8$, but RBS is often used for profiling elements with $Z \geq 8$. Once the system parameters are calibrated, this technique is standardless, and the spectra are turned into concentration profiles within minutes. The many attributes of RBS analysis are also shared by ERD analysis, and therefore collection of ERD spectra can be done on the same accelerator and in the same analysis chamber as RBS. The analysis of the data is fully analytical and generally independent of bonding in the sample. This is a tremendous benefit over other techniques that depend upon extensive comparison to standards and also on sputter rates that may vary from sample to sample. Finally, conventional RBS and ERD give adequate depth resolution (tens of nanometers) and sensitivity (~0.1 at.%) for many types of thin-film analyses performed in materials science. If greater depth resolution or greater sensitivity is required, then more exotic detection schemes can be used [e.g., time-of-flight (TOF) detectors, which produce a depth resolution an order-of-magnitude greater than that for conventional surface barrier detectors].

Before considering the principles of the IBA method, it is useful to place some practical limitations on the conventional use of IBA in comparison to other common materials analysis techniques. Composition analysis with approximately mega-electron-volt ions, in a rather standard IBA configuration, is less surface sensitive than techniques such as Auger electron spectroscopy (see AUGER ELECTRON SPECTROSCOPY) or x-ray photoelectron spectroscopy (XPS). Further, XPS can be used to identify changes in bonding even when the composition is not changing. For determination of the material microstructure, transmission electron microscopy (TEM) should be considered (see TRANSMISSION ELECTRON MICROSCOPY and SCANNING TRANSMISSION ELECTRON MICROSCOPY: Z-CONTRAST IMAGING), and if local compositional analysis of regions ~10 nm in size is needed, then TEM combined with energy-dispersive x-ray spectroscopy (EDS) should be considered (see ENERGY-DISPERSIVE SPECTROMETRY). Ion beam analysis is insensitive to film density, and therefore accurate measurement of film thickness is best made by other techniques, such as ellipsometry, x-ray reflectivity, or simple step-height measurements from profilometry. Finally, for trace element analysis (ppm or less), techniques such as secondary ion mass spectroscopy (SIMS) or nuclear reaction analysis (NRA; see NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION) should be considered. In any case, thin-film composition analysis and monitoring reactions and interdiffusion of thin films are performed very effectively using RBS, ERD, and EBS, especially when complimented by the other techniques listed here.

PRINCIPLES OF THE METHOD

This unit will describe the typical IBA experiment using Si surface barrier detectors (SBDs), although several types of IBA techniques have been developed in which the detector is changed to improve the depth resolution or elemental sensitivity. Once the principles are understood for conventional IBA, these principles are extended easily for the case of specialized detectors, and the reader is referred to the literature for those cases (see, e.g., Thomas et al., 1986; Wielunski et al., 1986; Bird and Williams, 1989; Gebauer et al., 1990; Whitlow, 1990; Tesmer et al., 1995; and the Additional References for Special Detectors and Detection Schemes following the Literature Cited section). Rutherford backscattering was first done by Geiger and Marsden (1913), but this technique waited until the 1960s for wider recognition (outside the nuclear physics community) that heavy element analysis on lighter substrates could be performed using RBS (e.g., see Turkevich, 1961). One of the most useful modern IBA techniques for easy depth profiling of light elements is ERD, which was first introduced in 1976 by L'Ecuyer et al. (1976).

The physical concepts governing RBS and ERD are identical, and through the use of a computer, their spectra are easily manipulated to obtain full quantitative analyses. However, even before manipulation of the spectra to obtain a quantitative analysis, it is useful to learn to interpret the spectra in order to understand the data as it is being collected. Therefore, we will begin with a description of the concepts common to RBS and ERD, and then demonstrate how to read the spectra.

Both RBS and ERD depend on only four physical concepts: (1) the kinematic factor describes the energy transfer in an elastic two-body collision; (2) the differential scattering cross-section gives the probability for the scattering event to occur; (3) the stopping powers give the average energy loss of the projectile and scattered atoms as they traverse the sample, thereby establishing a depth scale; and (4) the energy straggling gives the statistical fluctuation in the energy loss. By applying these four physical concepts, the spectra are transformed into a quantitative concentration profile as a function of depth. The main difference in the treatment of ERD data and RBS data is in the calculation of stopping powers: the incident projectile and recoil atoms are different for ERD data, but these incident and detected ion species are the same for RBS data. As a result of the similarity of these two techniques, the analytical expressions describing the backscattering process and the elastic recoil process will be derived in parallel.

A comparison of RBS and ERD is shown in Figure 1 with ~1-MeV/amu projectiles: 1-MeV protons are used to profile Si in RBS analysis and 28-MeV Si ions are used to profile H in ERD analysis. At the top of the figure the experimental setup is shown schematically. This setup is

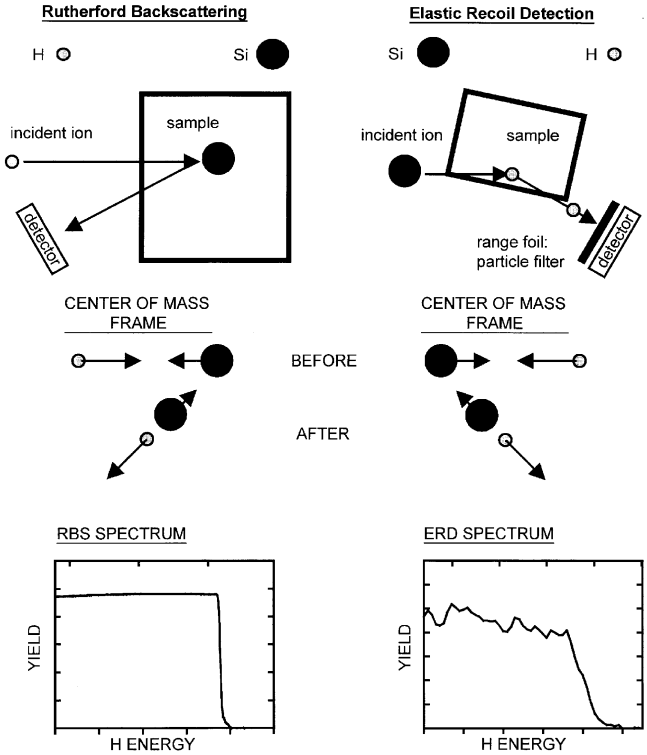


Figure 1. Schematic comparison of RBS and ERD scattering geometries in the laboratory (top) and center-of-mass (middle) frames of reference. The larger filled circle represents the Si atom and the smaller gray circle represents the H atom. In the center-of-mass frame, the scattering events for RBS and ERD are equivalent. The representative spectra for the collection of scattered H atoms for both RBS and ERD are shown at bottom.

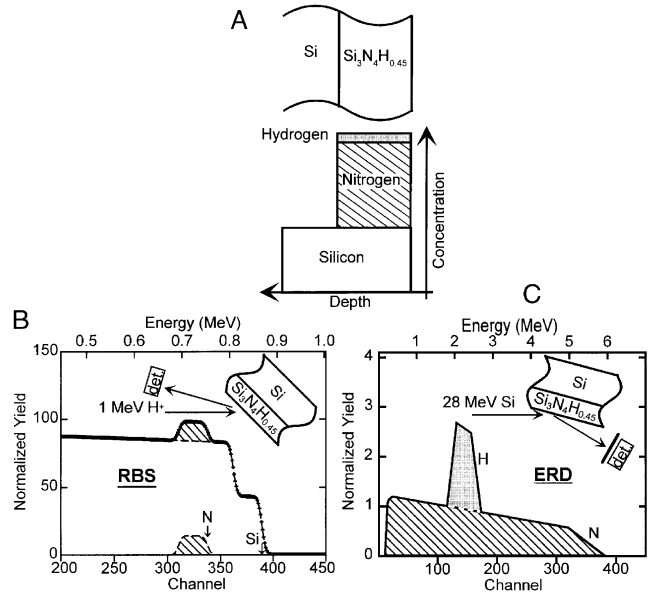


Figure 2. Collection of RBS (**B**) and ERD (**C**) spectra from a 300-nm-thick $Si_3N_4H_{0.45}$ film on Si, shown schematically (**A**), with the concentration profile shown in the middle (depth increasing to the left). The incident projectiles are 1-MeV protons for RBS and 28-MeV Si ions for ERD.

essentially the same for RBS and ERD but with the addition of a range foil for the ERD detector. In fact, RBS and ERD can be set up in the same vacuum chamber by positioning one detector in the backscattering direction and another detector with range foil in the forward-scattering position and by giving the sample holder the rotation capability to allow for grazing incidence of the projectile and exit of the recoiled atoms. The incident projectiles lose energy as they enter the sample and kinematically scatter hydrogen atoms (recoil H atoms for ERD and backscatter H atoms for RBS), and then the H atoms lose energy as they traverse the sample out to a particle detector. In both analyses, the atom entering the detector is H, but the incident ion projectiles differ. The energy loss in the range foil must be calculated when analyzing the ERD data. The differential Rutherford scattering cross-section ($d\sigma/d\Omega$) for the recoil scattering event is ~1.7 b/sr while the differential Rutherford scattering cross-section for the backscattering event is ~0.26 b/sr. The middle portion of Figure 1 schematically shows the scattering events in the center-of-mass reference frame in which RBS and ERD appear as symmetrical scattering events. The yield of H as a function of detected H energy is shown in the bottom portion of the figure for the RBS Si spectrum and the ERD H spectrum.

Both spectra in Figure 1 give the yield of detected H particles as a function of the H energy, but the important point to note is that for RBS the detected particle is the same species as the incident ion, whereas for ERD the detected particle is coming from the sample. In fact, for samples containing several light elements, many of these elements can be detected simultaneously in a single ERD spectrum. In this latter case, the detected particle energy is that for several different recoiled particles causing several overlapping ERD spectra for each detected particle. Schematically, an example is shown in Figure 2 of overlapping spectra for both RBS and ERD data representative of a 300-nm-thick Si_3N_4 layer containing 6 at.% H on a Si substrate. The elemental concentration profiles are given (Fig. 2A), one on top of the other, as stacked bars in which the height of each bar is the concentration in atomic percent and the width of the bar corresponds to depth in the sample. Simulated RBS and ERD spectra from this sample are shown in Figure 2B and C in which the signals from the different elements overlap. The signals are shaded in order to show the contributions coming from the different elements in the sample. Also, the contribution to the RBS spectrum (data points shown as plus symbols) from N alone is shown as a dashed line between channels 300 and 350. Therefore, just as for the bar graph in Figure 2A, the individual contributions to the spectra from the different elements are stacked one on top of the other. These spectra would appear on a multichannel analyzer (MCA) as the yield versus channel, and then the top energy axis is determined from an energy calibration of the channel axis. Note that a signal for H does not appear in the RBS spectrum because the incident protons are not backscattered by the H in the sample, and a signal for Si does not appear in the ERD spectrum because the Si in the sample is not scattered effectively and the incident Si is stopped in a range foil.

The height of each peak in the spectra corresponds to a concentration for that element as function of depth, which is correlated to the energy axis. The high-energy edge (front edge) of a given peak is a result of scattering from the surface of the sample and the depth scale increases to the left (decreasing energy). First, note that the front-edge position is representative of the mass of the target atom in the scattering event. For RBS, the greater the target mass, the greater is the backscattered proton energy, and therefore the increasing energy axis in this spectrum also represents increasing target mass. Similarly, the energy axis in the ERD spectrum is correlated to surface scattering for different target masses, but this correlation is made more complicated by the use of the range foil, as is discussed below, and the surface scattering for a heavy element is not always detected at a higher energy than for a lighter mass element. The surface scattering positions for each detected element are marked in the spectra of Figure 2 with the symbol for that element. The energy axis is often calibrated relative to the measured channels by using the front edge of peaks (surface scattering) for several elements present at the surface of the sample and by calculating the energy for surface scattering. Although the N has the highest concentration in this film, the height of the Si signal in the RBS spectrum and the height of the H signal in the ERD spectrum are both greater than the height of the N as a result of the difference in cross-sections and effective stopping powers relative to those for N. The height of the Si signal in the RBS spectrum below channel 350 is a result of proton scattering from Si in the substrate (100 at.% Si), and the relative decrease in the Si signal between channels 350 and 400 is a result of proton scattering from Si in the silicon nitride film with a decreased Si concentration (40 at.% Si).

The energy scale of the spectra is directly related to depth scale for each element through knowledge of the stopping power for the incident ion and the different scattered (detected) atoms. The width of the N signal in the RBS spectrum corresponds to the width of the 300-nm-thick silicon nitride film, and similarly the width of the Si signal from the $Si_3N_4H_{0.45}$ film corresponds to a depth of 300 nm. In fact, the width of these two signals are equal because the scattered particle in both cases is a proton, and therefore only one type of stopping power is involved in determining the depth scale. In contrast, the width of the N and H signals in the ERD spectrum both correspond to 300 nm of $Si_3N_4H_{0.45}$, but because the scattered particle is a H atom in one case and a N atom in the other case, two different stopping powers are involved in determining the energy-to-depth-scale conversions. Thus, each element profiled in an ERD spectrum has its own depth scale, which complicates the comparison of yields from different species as a function of depth during data collection. To accurately determine the composition-depth profiles from the RBS and ERD data, the spectra need to be separated for each element and the yield converted to atomic density N, as described under Equations for IBA (see Data Analysis and Initial Interpretation).

In contrast to RBS, where light masses produce a signal at low energies with a low yield and heavy masses produce a signal at high energies with high yield, the energy of the

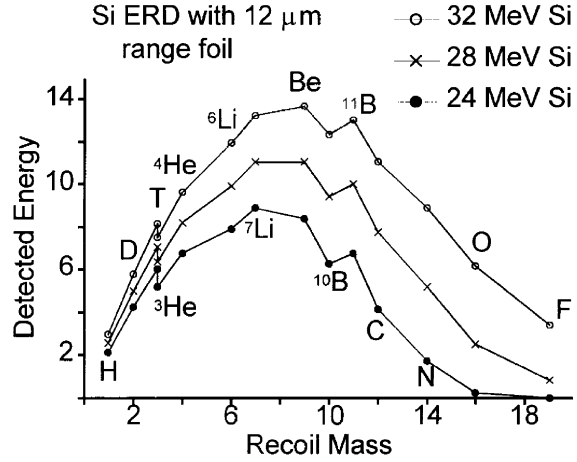


Figure 3. Detected energy calculated for recoiled atoms as a function of recoil mass for 24-, 28-, and 32-MeV Si ERD using a 12-μm Mylar range foil. For these calculations, the Si ions were incident upon the samples at an angle of 75° and the detector was at a scattering angle of 30°.

detected particle, E_d, and the yield of the ERD signals for the different masses depend strongly on the stopping power and thickness of the range foil. A heavier mass may or may not be detected at a higher energy than a light mass depending on the choice of projectile, projectile energy, type of range foil, and thickness of the range foil. Figure 3 shows E_d for recoiled atoms as a function of mass of the recoiled atom for 24-, 28-, and 32-MeV Si ion beams incident at 75° with a scattering angle of 30° and using a 12-μm Mylar range foil. If no range foil were used for these calculations, then the detected recoil energy would steadily increase with recoil mass. These curves show that the detected energy depends strongly on the increased stopping power in the range foil for the heavier recoiled atoms (amu $\geq$ 9). Mass resolution is the ability to separate in the spectra along the detected-energy axis the scattering signal from different target atoms. For surface scattering, E_d in RBS analyses depends solely on the change in kinematic scattering as a function of the target mass, whereas in ERD analyses E_d depends on the energy imparted to a recoiled atom and its stopping in the range foil. The kinematic factor K is the ratio of the scattered particle energy to the projectile energy for an elastic collision. Consequently, the selectivity for different masses in RBS analyses is determined by K, while the mass resolution in ERD analyses is determined by K and the effect of stopping in the range foil, as shown by the curves in Figure 3. For RBS the mass resolution is improved for a given ion beam by increasing the ion energy, and this is often true for ERD, but not always. For example, Figure 3 shows the ability to resolve B from He in a Si ERD spectrum is greatly improved by increasing the Si ion energy from 24 to 32 MeV, causing the difference $|E_d(B) - E_d(He)|$ to increase from 0.25 to 3.1 MeV; however, the same increase in Si ion energy causes $|E_d(He) - E_d(N)|$ to decrease from 5.1 to 0.73 MeV.

Depth resolution is the ability to separate, in the spectrum along the energy axis, the signal coming from

scattering events at different depths in the sample. Therefore, the smallest resolvable detected energy δE_d determines the smallest resolvable depth interval δx. Experimentally, the energy width δE_d is taken from the energies corresponding to 12% and 88% of full signal height for an abrupt change in sample composition (e.g., at the sample surface). An example of the difference in energy width is demonstrated in Figure 2, in which the width of the front edge (measured between the 12% and 88% yields) for the H signal is less than for the N signal in the ERD spectrum, and even though the energy-to-depth conversions for the H and N signals differ significantly, the depth resolution for the H signal is better than that for the N. The depth resolution for both RBS and ERD is improved by using lower energy projectiles or using heavier projectiles. For example, most thin-film RBS analyses are performed using ^{4}He ion beams rather than protons because of the increased depth resolution and increased cross-section. Also, ^{4}He ion beams, like beams of ^{1}H, are relatively easy to produce from simple ion sources.

The description given above for ERD analysis requires the use of a range foil in front of an SBD. However, one rare example of ERD analysis using a Si SBD without using a range foil is when a very heavy projectile is used with moderate-to-light-mass substrates. For this case, if the geometry and beam are chosen such that the recoil scattering angle ϕ is $> \sin^{-1}(M_{\text{substrate}}/M_{\text{incident ion}})$, where M is the mass of the particle, then the incident ion is not scattered into the detector and no range foil is required. This condition is particularly useful for high-depth-resolution ERD analysis with a Au ion.

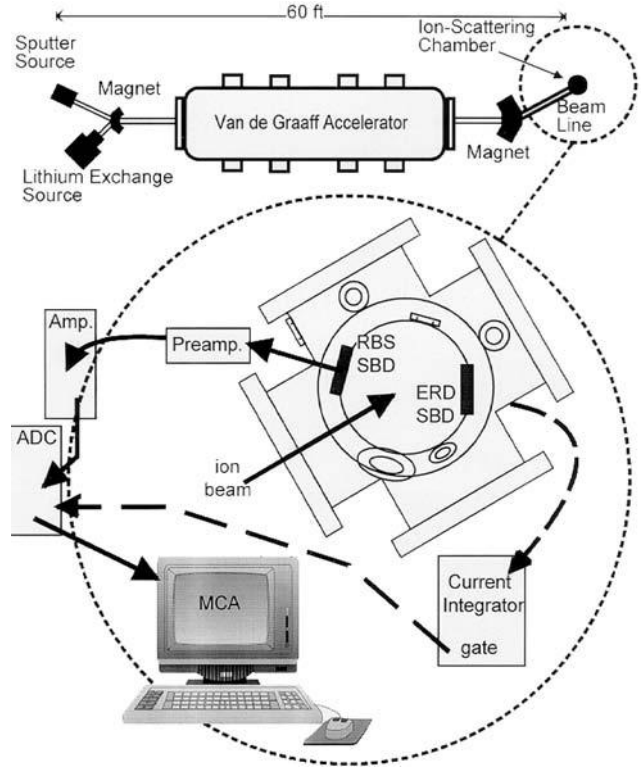


Figure 4. Typical experimental setup for RBS and ERD. Ion-beam analysis facilities are generally <100 ft in length, but this depends on the size of the accelerator being used. The expanded view, at bottom, schematically shows the ion beam scattering chamber and the electronics used to measure the target current and signal detected in an SBD (ADC = analog-to-digital converter).

PRACTICAL ASPECTS OF THE METHOD

Typical Experimental Setup

Now we describe the apparatus used in a conventional RBS or ERD experiment. Rather than give an exhaustive treatment, this discussion is intended to allow a first-time practitioner to become familiar with the purpose of different parts of the apparatus before entering the laboratory. A more detailed treatment is found in Tesmer et al. (1995), Bird and Williams (1989), Chu et al. (1978), and Mayer and Rimini (1977).

First, the ion beam is generated from a plasma or sputtered target (Fig. 4). These ions are extracted at low energy from the ion source and accelerated to millions of electron volts. A magnetic field is used to bend the ions into a specific direction along the analysis beam line (evacuated tube), and by separating the ions according to their mass-to-charge ratio for a given energy, the proper ion beam is selected for the IBA technique of interest. The ion beam is then steered onto a sample that is in an ion-scattering vacuum chamber. A variety of IBA experiments may be performed in an ion-scattering chamber, depending on sample manipulation capabilities and detector geometry. In addition, the chamber may be configured to perform *in situ* materials science experiments such as thermal treatments and gas exposures. The first variable to be determined in the yield equation is the detector solid angle Ω, which is measured with the use of a standard sample if all other parameters are known. However, most ion-scattering chambers are designed with fixed dimensions; therefore the solid angle can be calculated from simple geometry:

$$\Omega = \int_{\text{angle subtended by detector}} \sin\phi \, d\phi \, d\varepsilon \approx \frac{\text{detector area}}{z^2} \quad (1)$$

where ε is the stopping cross-section and z is the sample-to-detector distance.

The next variable to be calculated for the yield equation is Q, the number of incident projectiles, which is determined from the target current and the projectile charge. The target current may be measured either in front of the ion-scattering chamber on a rotating wheel containing slots to allow the ion beam to enter the chamber or directly from the sample of interest. In either case, care is taken to accurately measure all of the incident charge by suppressing the escape of secondary electrons from the target. One solution is to electrically isolate the scattering chamber from ground and then measure the total current on the chamber relative to ground. This technique effectively treats the entire chamber as a Faraday cup for accurate charge collection (see Chapter 12 in Tesmer et al., 1995). To ensure that low-energy secondary electrons are not

lost from the chamber in the direction of the incident beam, a suppression bias of −500 to −1000 V is applied to a wire mesh or metal aperture, which is isolated from the chamber and ground, just in front of the ion beam entrance to the chamber. The ion beam is collimated with movable slits or steered to avoid scattering from the aperture along its path to the sample. A more complete discussion of Faraday cup designs is found in Tesmer et al. (1995). The target current is then sent into a current integrator to measure the total charge incident upon the sample during the data collection, as shown in Figure 4. Generally, the amount of charge is preset on the integrator (typically to a value between 1 and 50 μC for most experiments) such that when the preset value is reached, a signal is sent to gate off the data collection. The length of time or amount of charge collected for an experiment depends on the counting statistics desired. The value needed for Q is the number of particles (not the charge), and therefore the measured charge in microcoulombs is divided by 1.6022×10^{-13} μC per ionic charge and by the ion's charge state to obtain Q (e.g., charge state 2 for He^{2+}, or 5 for Si^{5+}).

The signal measured in RBS or ERD comes from scattered atoms collected in an SBD (Fig. 4). For RBS, the backscattered beam is the same as the incident beam, while for ERD the forward-scattered beam may be composed of many atom species with mass less than that of the incident beam. Also, the incident beam may be forward scattered into the detector for ERD, and therefore a stopping foil (the range foil) is used to stop heavy species such as the incident beam from entering the detector. The number of incident projectiles that scatter into the direction of the detector is generally much greater than the number of recoiled atoms; therefore the foil is needed to prevent damage to the detector and to minimize the background signal from the primary beam.

The SBD is a reverse-biased surface barrier diode (Mayer and Lau, 1990) that contains a depletion region large enough to stop the highest energy particle to be detected. The depletion region for n-type Si is established through the application of a positive bias typically <130 V. As the detected particle traverses the depletion region, the particle loses energy, and that energy lost to electron excitation gives rise to electron-hole pair production. The electrons and holes are swept to opposite terminals of the diode and provide a pulse of current. The detected particle loses its energy quickly enough so that all of the electrons and holes created by this particle are collected in a single pulse of current. Therefore, the greater the energy of the detected particle, the more electron-hole pairs are produced, and the larger the current pulse is generated at the diode terminals. This current pulse is then sent to a charge-sensitive preamplifier to produce a signal proportional to the total charge. This signal is then sent to a spectroscopy amplifier, which outputs an approximately Gaussian signal proportional in amplitude to the size of the input pulse. The unipolar output signal from the amplifier is then sent to an analog-to-digital converter (ADC) to create a digital signal corresponding to the maximum amplitude of each pulse. A useful exercise at the beginning of each experiment is to examine on an oscilloscope the signal coming from the bipolar output of the amplifier. This signal should be examined as the bias voltage is applied to the detector in order to examine the noise reduction obtained for the optimum detector bias voltage. As the data are collected, different amplitude Gaussian peaks are seen on the oscilloscope corresponding to different energies of the detected particles. The gain of the amplifier is adjusted such that all of the most energetic particles will appear at a voltage less than the maximum input voltage for the ADC (typically 8- to 10-V maximum signal).

The digital signal coming from the ADC is recorded as a data point in a channel on an MCA in which the channel number is proportional to the amplitude of the digital signal. This creates a histogram (spectrum) of data in which the channel number represents the height of detected pulses and the yield for a given channel equals the number pulses detected for that pulse height. Each new pulse-height signal that is collected is then added to and stored in the spectrum as one additional count for that channel. This type of data acquisition is known as pulse-height analysis (PHA). Thus, the yield recorded on the MCA as a function of channel number corresponds to the yield of scattered particles with a detected energy E_d in which each channel number corresponds to a value of E_d. As mentioned above, the gate signal from the current integrator (used to measure the incident beam current) is used to turn off the ADC in order to stop the data collection. The remainder of this unit will review how the RBS and ERD spectra are converted from yield to composition and provide concrete examples.

Other Detectors and Detection Geometries

So far this unit has concentrated on conventional ERD analysis geometries, but another possible geometry for ERD is the transmission mode. For transmission ERD, the sample must be thinner than the range of the recoiled atom to be profiled. The projectile beam usually impinges on the sample at or near normal incidence, and the detector is placed at a recoil scattering angle of 0°. Additional range foils are used to stop the high-intensity projectile beam or the sample itself may be sufficient to stop the projectile. The main advantage of transmission ERD is increased sensitivity, by as much as 2 orders of magnitude, in comparison to conventional reflection geometry ERD. This increase in sensitivity stems from several factors: (1) since the analysis range scales with $\cos \Theta_1$, more material is probed with transmission ERD; (2) larger solid angles are used without significant kinematic broadening in transmission ERD since $dK/d\phi = 0$ at $\phi = 0°$ and finally, (3) the background caused by surface H is reduced considerably in the transmission mode due to smaller multiple scattering cross-sections. Wielunski et al. (1986) demonstrated a one- atom ppm sensitivity for detecting H in Ni using 4- to 6-MeV He projectiles in the transmission mode.

The scattering dynamics of ERD and RBS are quite similar, especially when viewed in the center-of-mass frame of reference. Some of these similarities are good, such as the ease of acquiring both types of spectra; but other shared aspects are actually bad, such as the built-in

mass-depth ambiguity that complicates the interpretation of both ERD and RBS energy spectra. This ambiguity results because the energy of the backscattered or recoiled ions depends both on the mass of the target atom and on the depth in the sample where the scattering occurred. Further, in the case of ERD, a mass ambiguity for scattering from the surface of a sample due to the energy loss suffered in the range foil also exists. Avoiding this ambiguity is nearly impossible in RBS because the detected ion is always the same as the incoming ion. However, in ERD, the recoil ion energy and mass are measured independently either by employing an $E - \Delta E$ particle telescope (Gebauer et al., 1990) or by combining a measurement of the ion's flight time (i.e., velocity) to that of total energy (Thomas et al., 1986; Whitlow, 1990). This latter technique is referred to as time of flight. An additional benefit of TOF techniques is that the depth resolution is improved considerably, and the use of a TOF detector is beneficial to increase depth resolution for both RBS and ERD analyses. This improvement in depth resolution results because the timing resolution involved with TOF is generally much better than the energy resolution involved with SBDs. Time-of-flight detection involves the simultaneous measurement of both the velocity and total energy of the detected atoms. The detected atom velocity is determined by measuring the time t required to pass along a preset flight path of length L, and then the detected energy is given by $E_d = ML^2/2t^2$.

DATA ANALYSIS AND INITIAL INTERPRETATION

Equations for IBA

The analytical expressions that relate the observed RBS and ERD energy spectra (the yield equation) to target atom type and concentration with depth are quite similar and will be developed here in brief. A more exhaustive treatment in the development of these equations is found in Chapters 4 and 5 of Tesmer et al. (1995), Chapter 3 of Chu et al. (1978), Brice (1973), Foti et al. (1977), and Doyle and Brice (1988). The geometries for RBS and ERD kinematic scattering in the laboratory reference frame are given in Figure 5, in which the incident beam and the directions of detection are coplanar. In this figure, for simplicity, the RBS and ERD geometries are combined, and the atom configurations before a scattering event are shown with cross-hatched circles and the atom configurations after a scattering event are shown schematically with an open circle for RBS and solid circles for ERD. However, the backscattering angle θ for RBS and forward-scattering angle ϕ for ERD are not necessarily coplanar, and the following development is made independent of that representation in the figure. The scattering angles are both measured relative to the incident beam direction. An ion beam with energy E_0, mass M_1, and atomic number Z_1 is incident onto a target atom with mass M_2 and atomic number Z_2. For RBS, the projectile is typically protons or He ions and $M_1 < M_2$, while the projectile for ERD varies widely from He ions to Au ions and $M_1 > M_2$. Through conservation of energy and momentum, the energy of the scattered atom, E_2, is related to E_0 by the kinematic factor K such that

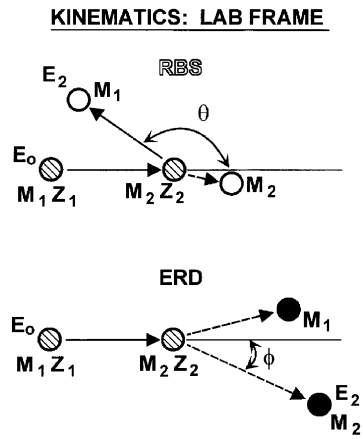


Figure 5. Schematic representation of two different elastic collision events occurring between a projectile of mass M_1, atomic number Z_1, and energy E_0 and a target of mass M_2 (atomic number Z_2) that is at rest before the collision. For RBS (open circle), the incident ion is backscattered with an energy E_2 at an angle θ measured relative to the incident beam direction. For ERD (filled circles), the target mass is recoiled forward at an angle ϕ with an energy E_2.

$$E_2 = KE_0 \tag{2}$$

and

$$K = \left(\frac{\sqrt{M_2^2 - M_1^2 \sin^2 \theta} + M_1 \cos \theta}{M_1 + M_2} \right)^2 \tag{3}$$

for RBS or

$$K = \frac{4M_1 M_2 \cos^2 \phi}{(M_1 + M_2)^2} \tag{4}$$

for ERD.

Equations 2 to 4 relate the energy before and after the scattering event, independent of depth in the sample.

The following development is based on a slab analysis in which the variables needed for the yield equation are evaluated at energies corresponding to uniform increments of depth in the sample. Figure 6 gives the geometry for the ion-scattering processes occurring at a depth x in the sample measured along the sample normal. This figure describes either the forward-scattering process using a range foil with thickness $X^{(0)}$ or the backscattering process without a range foil [$X^{(0)} = 0$]. The projectile ion beam is incident upon the sample at an angle Θ_1 measured from the surface normal to the beam direction, and the detected beam is collected in an SBD at angle Θ_2 measured from the surface normal. In an experiment, the scattering angles are fixed by the apparatus while Θ_1 and Θ_2 vary depending on the sample tilt. In Figure 6, Θ_1 and Θ_2 are coplanar and related to the scattering angle by $\theta = \pi - (\Theta_1 + \Theta_2)$ for

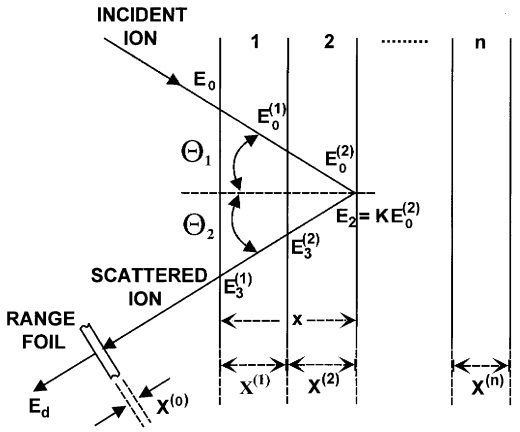


Figure 6. Schematic configuration for an elastic ion-scattering experiment used to describe either RBS or ERD analysis in which the sample is divided into n layers of thickness $X^{(n)}$) for calculations. The superscripts in parentheses refer to layer number and the subscripts on E refer to the type of atom: 0 = incident ion, 2, 3 = scattered ion. The parameters are incident ion energy E_0, incident and exit angles Θ_1 and Θ_2 (measured relative to the sample normal), the energy (E_2) of the scattered atom at depth x, the energy of the scattered atom (E_3) as it traverses the sample, and the energy detected in an SBD (E_d).

RBS or $\phi = \pi - (\Theta_1 + \Theta_2)$ for ERD [note that $\Theta_1 + \Theta_2 > \pi/2$ for ERD]; however, the formalism developed below does not require that Θ_1 be coplanar with Θ_2, and in a general experimental apparatus they may not be coplanar. The incident projectile enters the sample with energy E_0 and loses energy (dE/dx) along its path in the sample, and at a depth x the projectile has an energy $E_0^{(2)}$ just prior to a scattering event. Thus, the initial energy of the detected atom after the scattering event is $E_2 = KE_0^{(2)}$, and after traversing the sample this atom exits with energy $E_3^{(1)}$. For ERD, the recoiled atom is then incident normal to a range foil, which is assumed to be parallel to the detector surface, and this atom then loses more energy in the range foil. For either RBS or ERD, the atom is then detected with energy E_d in an SBD ($E_d = E_3^{(1)}$ for RBS).

The yield Y of detected atoms in energy channel E_d with a channel width E_d (in kilo-electron-volts perchannel) is given by

$$Y(E_d) = \frac{QN(x)\sigma(E_0^{(n)})\Omega\ \delta E_d}{\cos\ \Theta_1 dE_d/dx} \tag{5}$$

where Q is the number of incident projectiles, $N(x)$ is the atomic density of the target atom at depth x, $\sigma(E_0^{(n)})$ is the average scattering cross-section in laboratory coordinates at an energy $E_0^{(n)}$, Ω is the detector solid angle, and dx is the increment of depth at x corresponding to an increment of energy dE_d. The quantities Q, Ω, and Θ_1 are fixed by the experimental setup. A defining aperture is usually placed in front of the detector in ERD to maximize the depth resolution, which is degraded by the trajectory of the particles entering the detector at different scattering angles. Rigorously, Ω is the integral of the differential surface area element in spherical coordinates, but it is often calibrated by the yield from a standard sample (generally a H profile in ERD or a Si substrate profile in RBS). Care should be taken to avoid ion beam channeling in a crystalline sample when using the spectrum's surface edge height to calibrate a detector solid angle. Typical solid angles are $\sim$5 msr for ERD analyses and 1 msr $\leq \Omega \leq$ 10 msr for RBS analyses. To transform the yield to a concentration profile, analytical expressions for $\sigma(E_0^{(n)})$ and dE_d/dx must be determined.

First, the cross-sections for many instances are governed by Coulomb scattering, as modeled by Rutherford (1911), and expressed in the laboratory reference frame as (Marion and Young, 1968)

$$\frac{d\sigma_R(E_0^{(n)},\theta)}{d\Omega} = \left(\frac{Z_1 Z_2 e^2}{4E_0^{(n)}}\right)^2 \left(\frac{4}{\sin^4\ \theta}\right) \times \frac{\sqrt{1 - [(M_1/M_2)\sin\theta]^2} + \cos\theta)^2}{\sqrt{1 - [(M_1/M_2)\sin\theta]^2}} \tag{6}$$

for RBS and

$$\frac{d\sigma_R(E_0^{(n)},\phi)}{d\Omega} = \left(\frac{Z_1 Z_2 e^2}{4E_0^{(n)}}\right)^2 \left(\frac{4}{\cos^3\phi}\right) \frac{(M_1 + M_2)^2}{(M_2)^2} \tag{7}$$

for ERD, where $(e^2/E)^2 = 0.020731$ b for $E = 1$ MeV ($1\ b = 1 \times 10^{-24}$ cm^2). In practice, the differential cross-section is averaged over a finite volume Ω and given simply as σ_R. For those cases in which the cross-section is non-Rutherford (see Problems), the simplest solution for the thin-film approximation is to multiply the σ_R in Equations 6 and 7 by a scaling factor f in order to model the true cross-section, i.e., $\sigma = f\sigma_R$. It is important to note the functional dependencies of σ_R in Equations 6 and 7.

1. Both cross-sections are inversely proportional to the square of the projectile energy, and therefore $Y(E_d)$ increases as $(1/E_0^{(n)})^2$. Thus, the projectile energy should be decreased in order to increase the yield or sensitivity for a given element.
2. Since atomic mass is proportional to atomic number, the RBS σ_R is proportional to $(Z_1)^2$, and the ERD σ_R is $\sim (Z_1)^2$ for $M_1 \approx M_2$ and $\sim (Z_1)^4$ for $M_1 \gg M_2$. Therefore the yield for a given element is increased by using heavier projectiles.
3. The RBS σ_R is proportional to $(Z_2)^2$, and the ERD σ_R is $\sim (Z_2)^2$ for $M_1 \approx M_2$; therefore the scattering probability is greater for heavy targets than for light targets. The yield in RBS is greater for heavy targets than for light targets, but the yield in ERD also depends upon the stopping powers for the recoiled atom. For $M_1 \gg M_2$, the ERD σ_R is less dependent upon Z_2.
4. Both cross-sections are axially symmetric with respect to the incident beam direction and independent of Θ_1 and Θ_2. Thus the sample can be tilted and rotated to change the ion beam path lengths without changing the angular dependence in σ_R. For RBS,

the $(1/\sin\theta)^4$ dependence increases the yield as the backscattering angle is increased while the $\cos\Theta_1$ term in the yield equation tends to cancel the $1/\cos^3\phi$ increase in the ERD cross-section.

As described above (see Introduction), the areal density (AD) is the integrated concentration N over the layer thickness τ. From the yield Equation 5,

$$\mathrm{AD} = \int_0^\tau N(x)\,dx = \frac{\cos\Theta_1}{Q\Omega}\int_0^\tau \frac{S^{\mathrm{eff}}}{\sigma\,\delta E_\mathrm{d}} Y(E_\mathrm{d})dx = \frac{\cos\Theta_1}{Q\Omega}\int_{E_\mathrm{d}(0)}^{E_\mathrm{d}(x)} \frac{Y(E_\mathrm{d})}{\sigma}\frac{dE_\mathrm{d}}{\delta E_\mathrm{d}} \tag{8}$$

where S^{eff} is the effective stopping power $(= dE_\mathrm{d}|dx|_x)$. Equation 8 shows that the effective stopping power term is eliminated by integration when determining the areal density. In practice, the final integral of Equation 8 is evaluated as a summation; and for a surface peak, the cross-section σ_0 is evaluated at the specific energy E_0:

$$\mathrm{AD} = \frac{\cos\Theta_1}{Q\Omega\sigma_0}\sum_{\mathrm{channel}[E_\mathrm{d}(0)]}^{\mathrm{channel}[E_\mathrm{d}(x)]} Y(\mathrm{channel}) \tag{9}$$

Expressions for $\sigma(E_0^{(n)})$ and dE_d/dx are needed to convert the yield to a concentration profile, and these expressions are derived in terms of the energies $E_0^{(n)}$ and stopping powers (dE/dx) evaluated at $E_0^{(n)}$ corresponding to uniform increments of depth in the sample. A simple method to calculate the apparent energy loss in the detected beam as a function of depth is to divide the sample into a series of slabs of equal thickness, and then create a table of the corresponding energies and stopping powers for each layer or slab of the material, e.g., as shown schematically in Figure 6. The energy and depth relationships are expressed by the equations given below using the following convention for superscripts and subscripts. The subscripts refer to the atom type, except for the notation given above for the angles Θ_1 and Θ_2, and the superscripts refer to the layer of the target (1, 2, ...) or to the range foil (superscript 0). For example, the projectile energy and stopping powers are $E_0^{(n)}$ and $S_\mathrm{p}^{(n)}$, respectively; the energy for the detected atom is either $E_2^{(n)}$, or $E_3^{(n)}$ and the stopping power is $S_\mathrm{r}^{(n)}$ for recoiled atoms and $S_\mathrm{b}^{(n)}$ for backscattered atoms. The electronic energy lost by a particle as it traverses a slab containing more than one type of atom is approximated using the principle of additivity of stopping cross-sections first given by Bragg and Kleeman (1905). The stopping cross-section $\varepsilon \equiv (1/N)(dE/dx)$ is the proportionality factor between the amount of energy lost in a thin slab and the areal density of that slab. The energy loss in a material composed of several atom species is the sum of the losses for each species weighted according to the composition. This is known as Bragg's rule, and for a material with formula unit $\mathrm{A}_\alpha\mathrm{B}_\beta$,

$$\varepsilon^{\mathrm{A}_\alpha\mathrm{B}_\beta} = \alpha\varepsilon^\mathrm{A} + \beta\varepsilon^\mathrm{B} \quad \text{and} \quad S = N^{\mathrm{A}_\alpha\mathrm{B}_\beta}\varepsilon^{\mathrm{A}_\alpha\mathrm{B}_\beta} \tag{10}$$

The length of path traveled by the projectile crossing the nth slab with thickness $X^{(n)}$ is $X^{(n)}/\cos\Theta_1$, and the energy lost by the projectile is $E_0^{(n)} - E_0^{(n-1)}$. For $X^{(n)}$ sufficiently small, $S_\mathrm{p}^{(n)}$ changes little across the slab and is approximated by the value at energy $E_0^{(n-1)}$ such that:

$$E_0^{(n)} = E_0^{(n-1)} - \frac{X^{(n)}}{\cos\Theta_1}S_\mathrm{p}^{(n)}(E_0^{(n-1)} \tag{11}$$

Similarly, the energy $E_0^{(n)}$ for the detected particle as a function of depth in the sample is written in terms of a recursion relation for the energy from the previous slab boundary, or in terms of the kinematic relationship Equation 2 just after a scattering event. For an RBS scattering event occurring in slab n,

$$E_3^{(n)} = KE_0^{(n)} - \frac{X^{(n)}}{\cos\Theta_2}S_\mathrm{b}^{(n)}(KE_0^{(n)}) \tag{12}$$

and for an RBS event occurring in slab $j > n$,

$$E_3^{(n)} = E_3^{(n+1)} - \frac{X^{(n)}}{\cos\Theta_2}S_\mathrm{b}^{(n)}(E_3^{(n+1)}) \tag{13}$$

Equations 12 and 13 for ERD have a similar form and are given below (see Appendix B). In practice, the depth in the sample is related to the detected energy measured in a spectrum, and therefore the following equations relate the detected energy for the nth slab, $E_\mathrm{d}^{(n)}$, to the thickness $X^{(n)}$:

$$E_\mathrm{d}^{(n)} = K\left[E_0 - \frac{X^{(n)}}{\cos\Theta_1}\sum_{j=1}^{n} S_\mathrm{p}^{(j)}(E_0^{(j-1)})\right] - \frac{X^{(n)}}{\cos\Theta_2}\left[S_\mathrm{b}^{(n)}(KE_0^{(n)} + \sum_{j=n}^{2} S_\mathrm{b}^{(j-1)}(E_3^{(j)})\right] \tag{14}$$

for RBS and

$$E_\mathrm{d}^{(n)} = K\left[E_0 - \frac{X^{(n)}}{\cos\Theta_1}\sum_{j=1}^{n} S_\mathrm{p}^{(j)}(E_0^{(j-1)})\right] - \frac{X^{(n)}}{\cos\Theta_2}\left[S_\mathrm{r}^{(n)}(KE_0^{(n)} + \sum_{j=n}^{2} S_\mathrm{r}^{(j-1)}(E_3^{(j)})\right] - \Delta E_\mathrm{foil} \tag{15}$$

for ERD. Equations 14 and 15 are very similar with the change of notation to indicate the species of the detected atom [backscattered (b) vs. recoil (r)] and with the inclusion of stopping by the recoiled atom in the range foil for ERD calculations. The energy lost by the recoil atom in traversing the foil (ΔE_foil) should be calculated from a slab analysis of the foil containing NF slabs of thickness $\Delta x^{(0)}$ (i.e., $\mathrm{NF} = X^{(0)}/\Delta x^{(0)}$) as

$$\Delta E_\mathrm{foil} = \Delta x^{(0)}\sum_{j=1}^{\mathrm{NF}} S_\mathrm{r}^{(0)}(E_\mathrm{f}^{(j-1)}) \tag{16}$$

where the recoil atom stopping power is evaluated at the energy $E_\mathrm{f}^{(j-1)}$ before traversing the jth slab (e.g., $E_\mathrm{f}^{(0)} =$

$E_3^{(1)}$). Finally, the effective stopping power for the deleted particle at a given projectile energy, dE_d/dx, is given as

$$\left.\frac{dE_d}{dx}\right|_n = [S]_{p,b}^{(n)} \frac{S_b(E_d)}{S_b(KE_0^{(n)})} \tag{17}$$

for RBS and

$$\left.\frac{dE_d}{dx}\right|_n = [S]_{p,r}^{(n)} \prod_n \tag{18}$$

for ERD, where the energy loss factors $[S]$ corresponding to a backscattering or recoil event at energy $E_0^{(n)}$ are

$$[S]_{p,b}^{(n)} = \frac{KS_p^{(n)}(E_0^{(n)})}{\cos\Theta_1} + \frac{S_b^{(n)}(KE_0^{(n)})}{\cos\Theta_2} \tag{19}$$

for RBS and

$$[S]_{p,r}^{(n)} = \frac{KS_p^{(n)}(E_0^{(n)})}{\cos\Theta_1} + \frac{S_r^{(n)}(KE_0^{(n)})}{\cos\Theta_2} \tag{20}$$

for ERD. The product term $\prod_n$ in Equation 18 for ERD analysis is given by

$$\prod_n = \frac{S_r^{(n)}(E_3^{(n)})}{S_r^{(n)}(E_2)} \frac{S_r^{(n-1)}(E_3^{(n-q)})}{S_r^{(n-1)}(E_3^{(n)})} \cdots \frac{S_r^{(0)}(E_d)}{S_r^{(0)}(E_3^{(1)})} \tag{21}$$

which is similar to the product term $S_b(E_d)/S_b(KE_0^{(n)})$ for RBS analysis. Definitions for variables and the equations used for IBA are summarized below (see Appendices A and B).

Several methods can be followed to determine composition profiles from RBS and ERD spectra using the equations given above. In general, an iterative process is used in which the stopping powers, energies, and cross-sections for each slab layer are calculated from an estimated composition. A spectrum is simulated and compared to the data or the data are directly converted into a concentration-depth profile. Based on these results, the composition or slab thicknesses are changed and the energies, S, and σ are recalculated for another simulation or conversion of the data to a profile until a convergence between the estimated composition and results are obtained. This type of approach is used in RBS analysis programs such as RUMP (Doolittle, 1986, 1994). The method of directly converting the data to a concentration profile by scaling the data with cross-sections and dE_d/dx as a function of depth is known as spectral scaling. In this approach, a table is made containing the sample depth, $E_0^{(n)}$, $\sigma(E_0^{(n)})$, E_d, and dE_d/dx as a function of slab layer. The data are then scaled one channel at a time to produce a concentration-depth profile by interpolating between values in the table, and the energy scale is transformed to depth while the yield scale converts to concentration. The yield equation 5 is thereby converted to appear as though the cross-section and effective stopping power are constant through the sample. Analytically, the scaling is written as

$$\begin{aligned} N^{(n)}(x) &= \frac{\cos\Theta_1}{Q\Omega\,\delta E_d} \frac{dE_d/dx|_n}{\sigma_R(E_0^{(n)})} Y^{(n)}(E_d) \\ &= \frac{\cos\Theta_1}{Q\Omega\,\delta E_d} \frac{S_0^{eff}}{\sigma_0} \left[\frac{\sigma_R(E_0)}{\sigma_R(E_0^{(n)})} \frac{dE_d/dx|_n}{dE_d/dx|_0} Y^{(n)}(E_d) \right] \end{aligned} \tag{22}$$

where S_0^{eff} and σ_0 are evaluated at the incident energy E_0. The cross-section scaling factor for Rutherford cross-sections is $\sigma_R(E_0)/\sigma_R(E_0^{(n)} = [(E_0^{(n)})/E_0]^2$, and the stopping power scaling factor, $(dE_d/dx\,|_n)/(dE_d/dx\,|_0)$, is determined by interpolating from tables. Examples of spectral scaling and slab analysis of RBS and ERD spectra will be given below.

PROBLEMS

Non-Rutherford Cross-Sections

Deviations from the Rutherfordscattering cross-sections given in Equations 6 and 7 may result from Coulomb barrier penetration, which occurs most often for low-Z projectiles at high energies, or from excess screening effects, which occur most often for high-Z projectiles at low energies. At high energies, resonances occur in the backscattering cross-sections, and those resonances for which σ is enhanced relative to σ_R are useful for enhanced sensitivity of light elements with ion backscattering (EBS). These resonances occur because the incident particle has sufficient energy to penetrate the Coulomb potential barrier and probe the nuclear potential, which produces an attractive, strong, short-range force. The Coulomb potential produces a relatively weak, long-ranged repulsive force. At high energies, the projectile may penetrate into the nucleus of the target atom such that the two nuclei combine for a finite time as a single excited-state "compound nucleus." The width of a given resonance depends on the lifetime of the excited state in the compound nucleus, and for large resonance widths or energies with overlapping resonances, σ is relatively constant with energy in order to permit EBS depth profiling. Bozoian et al. (1990), Bozoian (1991a,b), and Hubbard et al. (1991) modeled the case of Coulomb barrier penetration and determined the energies E_{NR}, where the cross-sections may become non-Rutherford by 4%. In the lab reference frame their formulation of this energy is given as (Bozoian et al., 1990)

$$E_{NR}(\text{Me V}) \approx (0.12 \pm 0.01)Z_2 - (0.5 \pm 0.1) \tag{23}$$

for proton backscattering with $160^\circ < \theta(\text{lab}) < 180^\circ$,

$$E_{NR}(\text{Me V}) \approx (0.25 \pm 0.01)Z_2 + (0.4 \pm 0.2) \tag{24}$$

for ^{4}He backscattering with $160^\circ < \theta(\text{lab}) < 180^\circ$, and

$$E_{NR}(\text{Me V}) \approx \frac{-1.1934(Z_1Z_2)(M_1+M_2)}{M_1^{4/3}M_2 \ln(0.001846Z_1M_2/Z_2)} \tag{25}$$

for ERD with $\phi = 30°$. Furthermore, Andersen et al. (1980) modeled the case where screening effects cause significant deviations from the Rutherford formula, and in the laboratory reference frame for 1% screening effects (i.e., $\sigma/\sigma_R = 0.99$), the non-Rutherford energy boundary is given by

$$E_{NR}(\text{MeV/amu}) = 99 V_{LJ} \frac{M_1 + M_2}{M_1 M_2} \tag{26}$$

where

$$V_{LJ} = 48.73 Z_1 Z_2 (Z_1^{2/3} + Z_2^{2/3})^{1/2} \tag{27}$$

is the Lenz-Jensen potential in electron volts.

The use of a He ion beam for ERD analysis is nearly always in an energy range where penetration of the Coulomb barrier occurs, and the use of a Au ion beam for ERD is nearly always in an energy range where significant screening effects occur (for current IBA accelerator technology). A quick method to determine the shape of α particle and proton backscattering cross-sections at high energies is to examine the nuclear data sheet of the compound nucleus for the given nuclear reaction (see Internet Resources). As an example, the nuclear reaction $^{16}O(\alpha,\alpha)^{16}O$ for high-energy α particles backscattering from the ^{16}O nucleus creates a compound nucleus of ^{20}Ne. Figure 7 shows a portion of the data sheet (Tilley et al., 1998) for ^{20}Ne at the top and a magnified section of the $^{16}O(\alpha,\alpha)^{16}O$ cross-section at the bottom. This figure shows that sharp resonances occur at laboratory energies (in mega-electron-volts) of 2.52, 3.04, 3.08, 3.37, 3.89, 4.9, ..., and a broader resonance occurs from ~8.35 to 8.85 MeV, which is particularly useful for oxygen depth profiling. Even though the shape of the cross-section is shown here, the actual value of an enhanced cross-section should be determined from a well-characterized standard for the energy and scattering angle used in a given experiment.

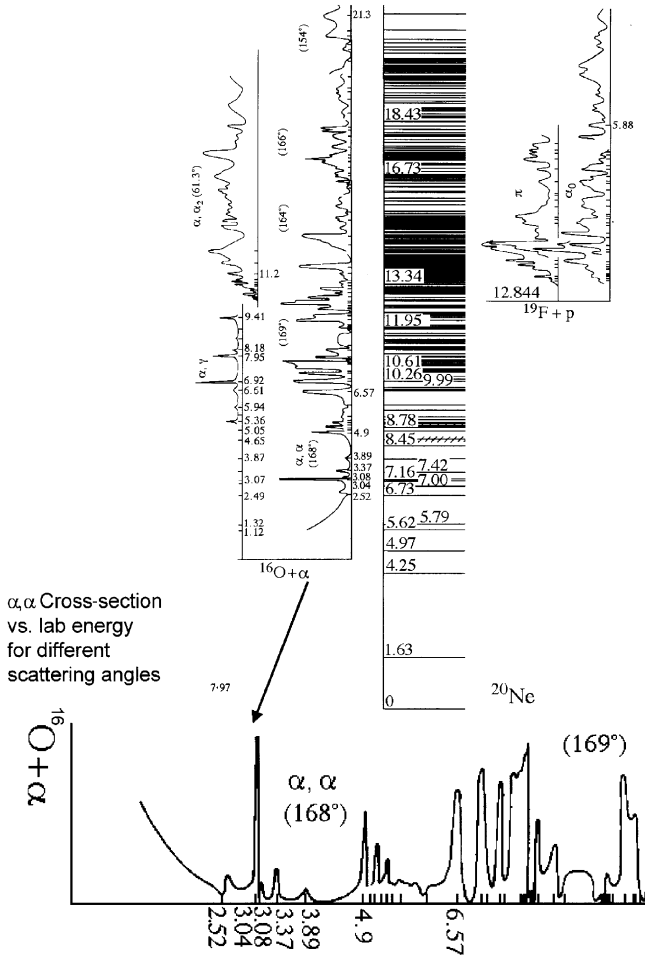


Figure 7. A portion of the nuclear data sheet for ^{20}Ne is shown at top, and a magnified view of the $^{16}O(\alpha,\alpha)^{16}O$ cross-section from this data sheet is shown at bottom.

Radiation Hazards

When the incident beam energy exceeds the Coulomb barrier of the target, nuclear reactions other than elastic scattering are possible. High-energy proton scattering can produce unwanted γ or neutron radiation exposures to workers, even along sections of the beam line away from the sample chamber. This concern is particularly significant for irradiation of light elements, but it is also important for elements such as Cu. Neutrons and γ rays are the main prompt radiation hazard while activation of the sample produces a longer lasting radiation hazard (of particular importance for high-energy proton beams). Activated samples can be easily controlled to avoid accidental radiation exposure of workers as well as satisfy regulatory agencies, but a knowledge of possible activating nuclear reactions is necessary for this purpose. Novices should seek help from a qualified radiation-health physicist to assess the hazards when using a high-energy proton beam.

Enhanced-Cross-Section Backscattering Spectrometry

Rutherford backscattering spectrometry was shown above to be highly useful for thin-film composition analysis. However, one of the disadvantages of RBS is its relative insensitivity to light elements, particularly lighter elements on heavy element substrates. Enhanced-cross-section backscattering with He ions offers sensitivity for light elements such as B, C, N, and O, and EBS with protons offers increased sensitivity for these light elements as well as even heavier elements such as S. The disadvantages of EBS as opposed to RBS is that the depth resolution for the higher energy EBS technique is generally less than that for RBS and the cross-section does not obey the Rutherford equations 6 and 7. Therefore σ_{EBS} must be calibrated from a standard. The decreased sensitivity to light elements in RBS results directly from the fact that σ_R is proportional to $(Z_1 Z_2/E)^2$. Furthermore, RBS analysis is complicated by the fact that both H and He projectiles show deviations from σ_R for light elements at relatively low energy, and in the low-energy regime where low-Z element cross-sections are still Rutherford, the depth of analysis is

rather small (<0.5 μm) and the background from heavier substrates is high.

The analytical formulas for EBS analysis are identical to those for RBS analysis with a change in the value of the cross-section. The elastic scattering cross-sections for high-energy α particles and protons were some of the first cross-sections measured with the emergence of modern particle accelerators and nuclear physics in the mid-20th century. The elements with the most useful α high-energy cross-sections are ^{11}B, ^{12}C, ^{14}N, and ^{16}O, while elements out to ^{32}S have been analyzed using high-energy proton cross-sections but with much less depth resolution. One difficulty found for EBS analysis is when the matrix contains moderate- to low-Z elements such as Si or Al, which have sharp resonances in their cross-sections, producing wild fluctuations in the backscattering signal from the matrix. These fluctuations make a determination of the light atom concentration nearly impossible, as is the case for oxygen profiling from SiO_2 or Al_2O_3 using 8.7-MeV ^4He ion EBS analysis.

EXAMPLES

Several examples are now given to demonstrate mass resolution, energy resolution, and depth of analysis for backscattering and ERD as a function of ion energy and ion species. Also, examples of how to convert spectra to composition-depth profiles for both RBS and ERD analyses are demonstrated. Figure 8 shows an RBS spectrum (+) for 2.2-MeV He^+ ions incident upon a film of $La_xSr_yCo_zO_3$ grown on a $LaAlO_3$ substrate. The spectrum is plotted from channel 15 to channel 615 similar to how it would appear when collected as data on an MCA. An MCA gives yield in counts, but for convenience the yield in Figure 8 is normalized by the number of incident particles ($Q = 20$ pμC), the solid angle ($\Omega = 6.95$ msr), and the channel energy width (3.314 keV). [Note that 20 pμC (particle micro- coulombs) is equal to $(20 \times 10^{-6}/1.602 \times 10^{-19})$ particles.] The sample was tilted 10° in order to decrease channeling in the substrate, and the detector was mounted at a scattering angle $\theta = 164°$. The top (energy) axis was determined using the surface scattering energies for O, Co, and La. These energies are given in Table 2 along with the detected energies for scattering from O, Al, and La at the film-substrate interface. Also, the surface scattering positions are marked with arrows in the figure. The Al surface energy is shown in parentheses in the table because Al is only present in the substrate, and therefore the Al edge energy is decreased from 1.224 to 1.040 MeV. The O substrate edge energy is marked in parentheses because the signal from oxygen in the substrate is small and difficult to differentiate from the background. Therefore, the O substrate energy given in Table 2 is based on the relative position of the O surface energy and the energy width of the layer as determined from the La signal.

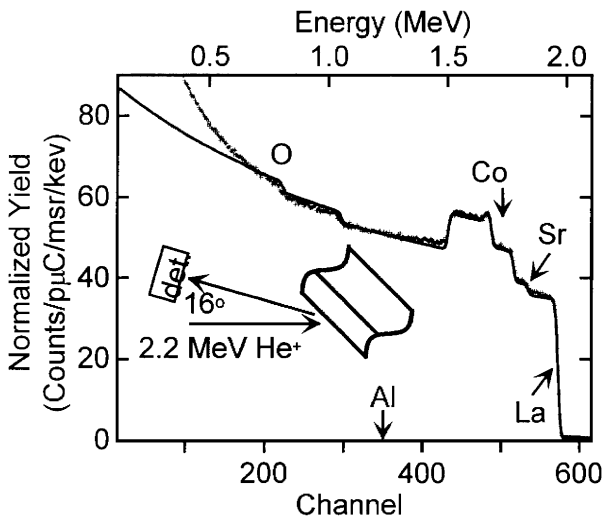


Figure 8. Backscattering spectrum (+) and simulation (solid line) obtained from 2.2-MeV He^+ ions scattered off $La_xSr_yCo_zO_3$/$LaAlO_3$ at 164° relative to the incident beam direction, with $Q = 20$ pμC, $\Omega = 6.95$ msr, $\Theta_1 = 10\circ$, and channel energy width 3.314 keV/channel.

Table 2. Energy Scale for RBS Spectrum in Figure 8

Element	Surface Scattering Energy (MeV)	Substrate Scattering Energy (MeV)
O	0.8069	(0.620)
Al	(1.224)	1.040
Co	1.685	
Sr	1.839	
La	1.965	1.780

Figure 8 shows a spectrum from a relatively simple sample, a thin film with a uniform composition (≈190 nm thick). Yet, the spectrum contains many overlapping peaks, which makes composition-depth profiling of a single element difficult because of the variation in the background signal due to the other elements in the film and the substrate. In these circumstances, a simulation of the spectrum often suffices to determine the composition of the film. A simulation of Rutherford backscattering from a $La_{0.85}Sr_{0.15}Co_{0.95}O_3$ film with an areal density of 1.6×10^{18} atoms/cm^2 is shown in this figure as a solid line using the analysis program RUMP (Doolittle, 1994). The oxygen content was determined from the reduced height of the individual metallic constituents in the film and later confirmed with greater accuracy using 8.7-MeV He^{2+} EBS. The height of the simulated substrate signal confirmed the values previously measured for Q and Ω. Deviations from the measured values occur for large dead times on the ADC or if the incident ion channels into the substrate; however, the agreement between the values given above and those determined from the simulation is better than 3%. At energies below ~0.6 MeV, the data and the simulation begin to deviate substantially as a result of multiple scattering in the substrate, which was not taken into account in the simulation. The layer thickness is given as areal density in the simulation because the exact density of the film is unknown, but if a density of 0.85×10^{23} atoms/cm^3 is assumed, then the film thickness is 188 nm. For more simple spectra in which

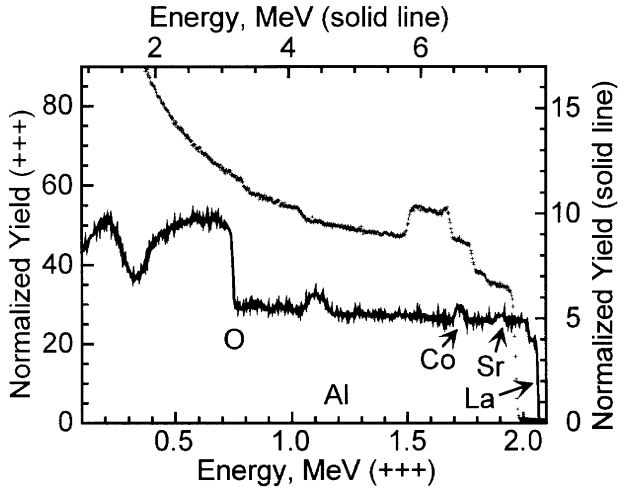


Figure 9. Spectra for backscattering off the same sample as in Figure 8. The 2.2-MeV He^+ spectra is reproduced here (+) in addition to a spectrum (solid line) obtained from 8.7-MeV He^{2+} ion scattering at 164°, with $Q = 18$ pμC, $\Omega = 6.95$ msr, $\Theta_1 = 45°$, and channel energy width 7.82 keV/channel. The left and bottom axes are for the low-energy RBS spectrum, and the right and top axes are for the high-energy EBS spectrum.

peaks of the individual elements are easily separated from a background signal, each peak is integrated to determine the areal density for that element and thus provide the relative areal densities for the composition of the layer. Also, if the ion energy were increased for analysis of this sample, then the mass resolution would increase and the individual peaks would separate in the spectrum.

A He ion energy of 8.7 MeV was used to analyze the same sample as given in Figure 8, and the spectrum is shown as a solid line in Figure 9, with the normalized yield plotted on the right axis and the energy plotted along the top axis. The parameters used for this spectrum were $Q = 18$ pμC, $\Omega = 6.95$ msr, channel energy width 7.82 keV/channel, $\Theta_1 = 45°$, and $\theta = 164°$. The sample was tilted 45° to make the apparent thickness increase for increased depth resolution. For comparison, the 2.2-MeV He^+ RBS spectrum (+) is plotted versus energy along the bottom axis and normalized yield on the left axis. To put these spectra on the same figure in a readable manner, two separate sets of axes were used because the energy and yield scales differ considerably between the spectra. In fact, it is this difference in the energy scale that gives rise to the separation of signals from different masses. The surface energies for 8.7-MeV He EBS are given in Table 3 along with the detected energies for scattering from O, Al, and La at the film-substrate interface, and the surface scattering positions for the EBS spectrum are marked with arrows in Figure 9. A comparison of the surface energies in Tables 2 and 3 show that the mass resolution, which is given by the difference in energy between surface edges, is greater for 8.7-MeV He scattering than for 2.2-MeV He scattering by a factor of 4. At the same time, the depth resolution is decreased by using a higher energy ion beam. Depth resolution δx is given by the resolution measured along the detected energy axis divided by the effective stopping power for the detected atom: $\delta x = E_d/(dE_d/dx)$. The front edges for the peaks shows that $E_d \sim 2$ times smaller for the low-energy spectrum than for the high-energy spectrum, and the effective stopping power is greater for the low-energy spectrum than for the high-energy spectrum. Thus, the depth resolution is considerably better at low energies than at high energies.

Table 3. Energy Scale for EBS Spectrum in Figure 9 (Solid Curve)

Element	Surface Scattering Energy (MeV)	Substrate Scattering Energy (MeV)
O	3.192	3.03
Al	(4.841)	4.68
Co	6.663	
Sr	7.272	
La	7.770	7.61

The high-energy ion beam with the lower stopping cross-section does afford a greater depth of analysis that is useful for thick samples. The depth of analysis can be estimated from stopping and range computer codes such as TRIM (Ziegler, 1987), and as a practical limit the maximum depth is between one-tenth and one-third the projected range. Moreover, the greatest advantage to using higher energy ion beams in an EBS analysis is demonstrated by the increased yield from O relative to La. The yield from the low-energy backscattering is considerably greater than the yield from the high-energy backscattering because of a $1/E^2$ dependence for the Rutherford cross-section. At 8.7 MeV, the He scattering from both Sr and La obeys the Rutherford formula, and therefore the yield is less than at 2.2 MeV, but the O cross-section is $22\sigma_R$ at 8.7 MeV ($\theta = 164°$, $\Omega = 6.95$ msr), and therefore the signal-to-background noise for the O signal is improved at the higher energy. The Co and Al also have non-Rutherford σ at 8.7 MeV, but $\sigma(\text{Co}) < \sigma_R(\text{Co})$ and the σ(Al) varies considerably. The rapid variation in the Al cross-section for 8.7-MeV He EBS adds uncertainty to the background below the O signal. The addition of sharp resonance background peaks below an O signal becomes much worse for Si, Al, or Al_2O_3 substrates.

A sample more complicated than the example just given can still yield a composition-depth profile as shown in the next analysis. Figure 10 is an RBS spectrum (+) collected from a thick film of Al deposited on a Si substrate in such a manner that the oxygen content in the film varies greatly with depth. A 2.8-MeV He^+ ion beam was used for this analysis to avoid sharp resonance peaks from Al or Si and to give a compromise between depth of analysis and depth resolution. Also, σ(O) was calibrated previously at this energy and for this detector geometry such that $\sigma(\text{O})/\sigma_R(\text{O}) = 1.25$. The energy scale for the spectrum in Figure 10 was calibrated from another sample such that the channel energy width is 3.11 keV/channel, $\Theta_1 = 5°$, $Q = 20$ pμC, $\Omega = 6.95$ msr, and $\theta = 164°$. The surface energies for O, Al, and Si are marked in the figure at energies 1.027, 1.558, and 1.595 MeV, respectively. To simulate this spectrum fairly well, the film was divided into six separate layers in which the thickness and composition were allowed to vary, as shown in Table 4 with the layer number

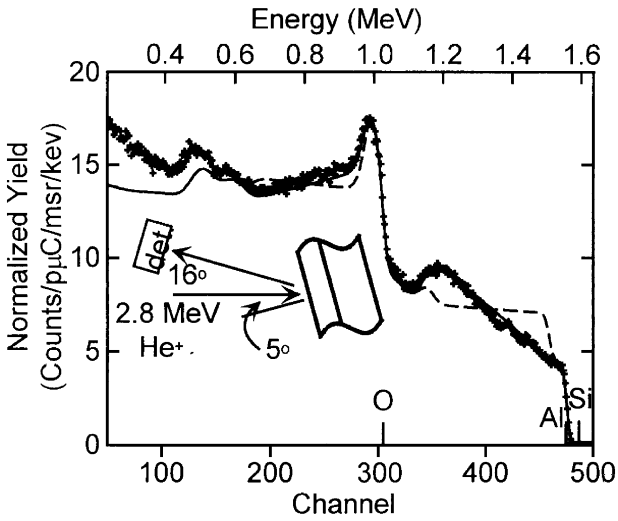


Figure 10. An RBS spectrum (+) and simulations (solid line and dashed line) for 2.8-MeV He^+ backscattering from an AlO_x/Si sample with a greatly varying oxygen content. The scattering angle is 164°, $\Theta_1 = 5°$, $Q = 20$ pμC, and $\Omega = 6.95$ msr.

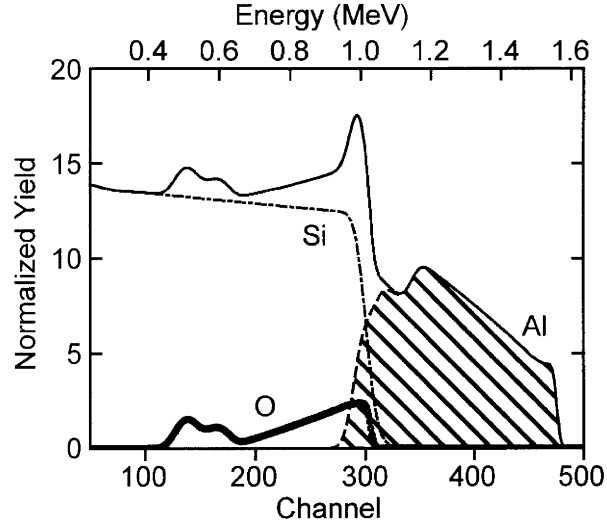


Figure 11. Simulated RBS spectrum (solid line) and the contributions to this simulation from the individual elements: Si (dash-dotted line), O (open circles), and Al (dashed line delineating the cross-hatched region). The simulation is for 2.8-MeV He^+ backscattering as described in Figure 10.

increasing in depth. Again, the thickness is given as a real density because the volume density of each layer is not known for such a complicated sample. The thicker layers were divided into two or more equal-thickness sublayers to increase the accuracy of the energy and stopping power calculations as a function of depth. Layer 2 was simulated as ten equal-thickness sublayers such that the composition across layer 2 graded linearly, as shown This simulation is presented in Figure 10 as a solid line in addition to a simulation (dashed line) assuming that layer 2 has a constant average composition $Al_{0.675}O_{0.325}$. The dashed-line simulation fits the data poorly. For such a complicated sample, the simulated sample structure in Table 4 may not be unique, but it will serve well to calculate the stopping powers and energies as a function of depth, and then the spectrum is scaled to give the composition-depth profile from the data, as demonstrated below. Further, the figure shows that the number of sample subdivisions used in the calculation is quite important, as determined from the comparison of simulations for layer 2.

The linearly graded sample simulation is given again for clarity as a solid line in Figure 11, and the individual contributions to the simulation are presented as a dash-dotted line for Si, open circles for O, and a dashed line with a cross-hatched region for Al. The contributions to the RBS spectrum from all three elements overlap in the spectral region from ~0.9 to 1.1 MeV, and this overlap produces a sharp peak at 1 MeV. The RBS signal from each element is separated from the background in the spectrum by subtracting the individual contributions to the simulation. Figures 12 and 13 show the RBS signals from the O (open circles) and the Al (+) after the simulated Si and Al (or O) signals were subtracted from the data. The solid lines are the contributions to the simulation for that element. The yields in Figures 12 and 13 were then scaled with cross-sections and stopping powers as a function of energy to give the concentration as a function of depth (in atoms per square centimeter). These scaled data are

Table 4. Sample Structure Used for Simulation in Figure 10

Layer	Thickness (10^{15} atoms/cm^2)	Composition
1	800 (with 2 sublayers)	$Al_{0.4}O_{0.6}$
2	4900 (with 10 sublayers)	Linearly graded from $Al_{0.4}O_{0.6}$ to $Al_{0.95}O_{0.05}$
3	200	$Al_{0.95}O_{0.05}$
4	1000 (with 2 sublayers)	$Al_{0.7}O_{0.3}$
5	200	$Al_{0.95}O_{0.05}$
6	1000 (with 2 sublayers)	$Al_{0.6}O_{0.4}$

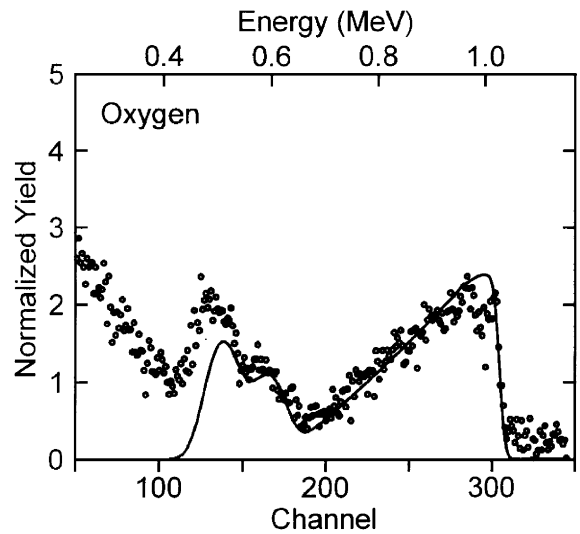


Figure 12. Oxygen signal (open circles) from the RBS spectrum of Figure 10 after the background Si and Al signals were subtracted. The simulated O signal is shown as a solid line.

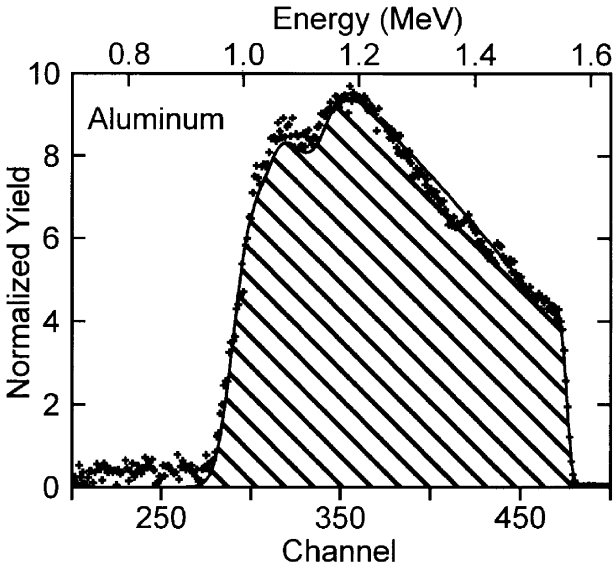


Figure 13. Aluminum signal (+) from the RBS spectrum of Figure 10, after the background Si and O signals were subtracted. The simulated Al signal is shown as a solid line delineating a cross-hatched region.

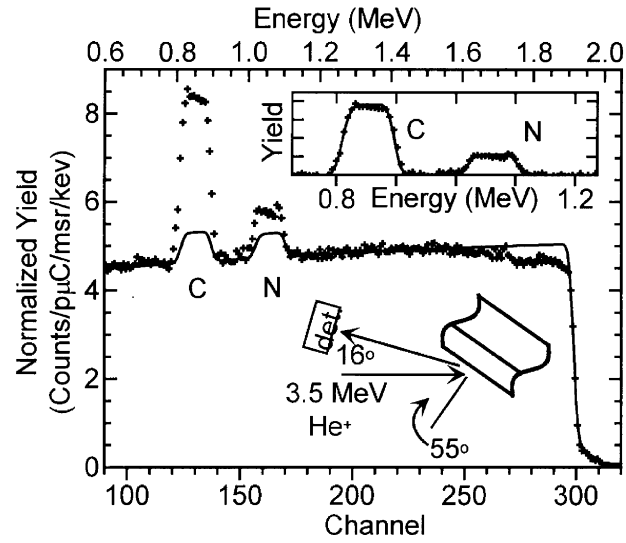


Figure 15. An RBS spectrum (+) and simulation based on Rutherford scattering (solid line) for 3.5-MeV He^+ backscattering from a CN_x/Si sample. The scattering angle is 164°, $\Theta_1 = 55°$, $Q = 20$ pμC, and $\Omega = 6.95$ msr. The inset shows the C and N signals above background overlayed by a simulation (solid line) based on enhanced scattering cross-sections ($5.6\sigma_R$ for C and $2\sigma_R$ for N).

plotted in Figure 14 with the depth (in micrometers) calculated from an assumed average volume density of 0.831×10^{23} atoms/cm^3 (about half way between Al_2O_3 and Al). Note that the depth scale is plotted with depth increasing to the right.

The previous analysis was given as an RBS example because the Al and Si cross-sections obeyed the Rutherford formula and the O cross-section was very nearly Rutherford. Still, it shows that there is no clear boundary between naming a technique as RBS or EBS, and convention allows that RBS is used as the acronym for backscattering analysis independent of the cross-section formulation. Also, σ was non-Rutherford for O at a relatively low He^+ ion energy, showing that a check is always required for the σ calibration involving He scattering from elements with $Z \leq 9$.

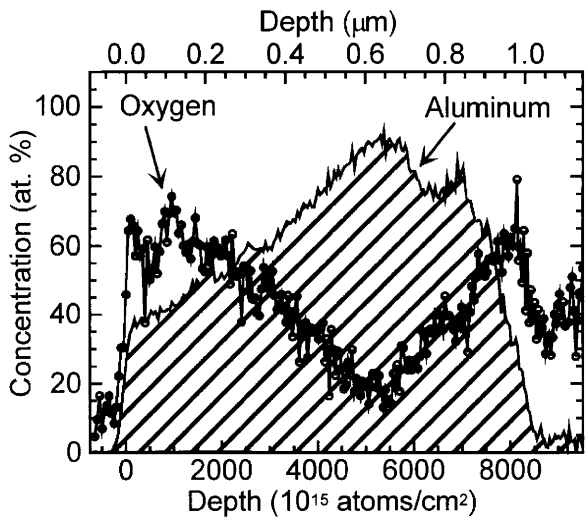


Figure 14. Oxygen (open circles) and aluminum (solid line delineating a cross-hatched region) concentration-vs.-depth profiles determined from the data given in Figures 12 and 13.

Another example of useful enhanced-scattering cross-sections is given by analysis of a CN_x film in which both the C and N cross-sections are non-Rutherford. An example of a high-energy backscattering spectrum (+) taken from a 44-nm-thick $C_{0.65}N_{0.35}$ sample deposited on a Si substrate is shown in Figure 15. The analysis beam was 3.5-MeV He^+ incident at 55° from the sample normal and $\theta = 164°$. This angle of incidence was chosen to increase the depth resolution. The surface energy for C and N are ~0.9 and 1.1 MeV, respectively. (The ERD analysis showed that the sample contains less than 2% H.)

A simulated spectrum for a film with composition $C_{0.65}N_{0.35}$ and using Rutherford cross-sections is shown in Figure 15 as a solid line. The Si cross-section is Rutherford at this energy, and thus a small amount of channeling in the Si substrate is evident in the data from channel 250 to channel 300, but channeling does not affect the height of the C and N signals because this $C_{0.65}N_{0.35}$ film is amorphous. The carbon cross-section was calibrated from a pure C film on Si, and the N cross-section was calibrated from an amorphous Si_3N_4 film on Si. The following enhancements to σ were determined for this energy and scattering geometry: carbon $\sigma = 5.6\sigma_R$, and nitrogen $\sigma = 2\sigma_R$. The C and N signals above background (+) are shown inset at the top, with the simulated signal (solid line) determined using these factors for the enhanced cross-sections. The depth resolution in these signals is 9 to 10 nm.

Great care should be taken when choosing a carbon standard. A sample that contains a considerable amount of H has a significantly different height for the C signal

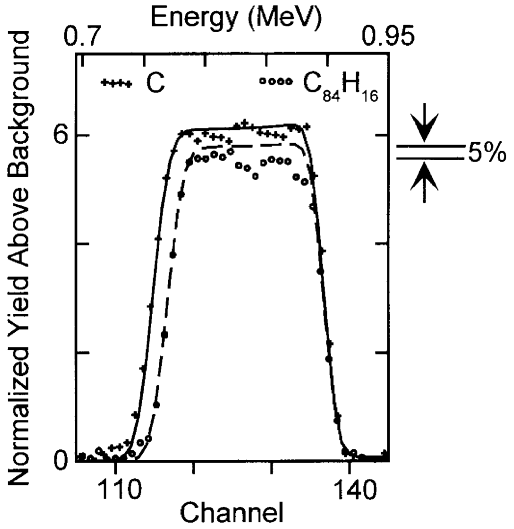


Figure 16. Carbon signal above background for RBS spectra from a pure carbon sample (+) and a $C_{84}H_{16}$ sample (open circles). The spectra were obtained using 3.5-MeV He^+ with $\theta = 164°$; $\Theta_1 = 55°$ and $\Theta = 45°$, respectively; $Q = 20$ pμC; and $\Omega = 6.95$ msr. Each spectrum (solid line and dashed line, respectively) was simulated based on the enhanced scattering cross-section given in Figure 15.

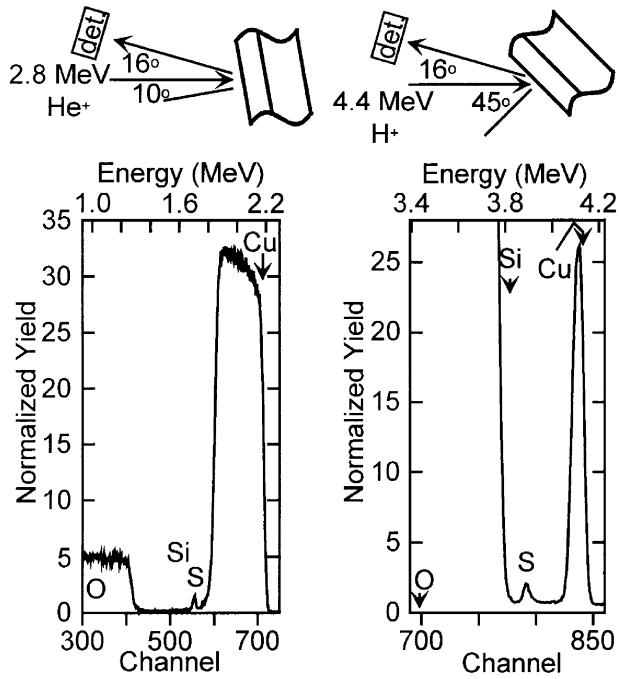


Figure 17. Helium (left) and H (right) backscattering spectra from a sulfidized Cu layer on a Si substrate. The 2.8-MeV He^+ spectrum was collected using $\theta = 164°$, $\Theta_1 = 10°$, $Q = 10$ pμC, and $\Omega = 6.95$ msr. The 4.4-MeV H^+ spectrum was collected using $\theta = 164°$, $\Theta_1 = 45°$, $Q = 10$ pμC, and $\Omega = 6.95$ msr.

than a sample that is pure C. Figure 16 shows the C signal above background for a C standard (+) containing <0.05 at.% H and for a $C_{0.84}H_{0.16}$ sample (open circle). If the $C_{0.84}H_{0.16}$ sample were used for determining σ without considering the H content, then a value of only $5\sigma_R$ would be obtained, but the low-H-content standard has a C signal that is $5.6\sigma_R$, as demonstrated by the solid line. Thus the knowledge of the H content of the standard is important to avoid errors (~10%) when calibrating the cross-section. Also, the stopping powers of C and C-H samples are often corrected for bonding effects that influence the composition analysis and thickness determination. A correction of ~5% must be applied to the simulation (dashed line) in order to obtain the proper height on the C signal from the $C_{0.84}H_{0.16}$ sample. This correction then causes an ~5% difference in the thickness of the film determined using RBS. The bonding effects on the stopping power add an increased uncertainty to the composition analysis of a CN_x film containing H because the nature of the bonding is often unknown and its effects on the stopping power are also unknown, and therefore the inherent limit to the accuracy of the composition analysis is ~5%.

An example of enhanced RBS comparing H^+ and He^+ backscattering is demonstrated by the composition analysis of sulfur on a heavy element substrate. Analysis of sulfur on Si or Al substrates is relatively easy with He^5 RBS because S is heavier than Si and Al, but analysis of a thin layer containing S on bulk Cu is more difficult. The spectra shown in Figure 17 are for 2.8-MeV He^+ and 4.4-MeV H^+ backscattering from a thin CuS_x film on a Cu layer on a Si substrate. The Θ_1 for each analysis are as indicated in the figure. For this case of He backscattering, the small S peak is just separated from the low-energy side of the Cu signal because the Cu layer was only 325 nm thick. However, a thicker Cu layer would cause the S signal to be lost on top of a large Cu background signal. Aldridge et al. (1968) examined the high-energy backscattering cross-sections for α particles incident on ^{32}S up to an energy of 17.5 MeV. The absence of enhanced cross-sections for $(alpha, \alpha)$ scattering necessitates the use of proton scattering for thick Cu substrates. Proton scattering cross-sections for the $^{32}S(p,p)^{32}S$ reaction at moderate energies are larger than σ_R (Abbandanno et al., 1973), and similar to the example given above for the $^{16}O(\alpha,\alpha)^{16}O$ reaction, a flat-top cross-section exists for the $^{32}S(p,p)^{32}S$ reaction in the energy range ≈ 4.3 to ≈ 4.5 MeV. A CdS standard was used to calibrate the sulfur cross-section for 4.4-MeV incident protons scattering at an angle of 164° such that $\sigma = 6.55\sigma_R$.

Analysis of the proton scattering S and Cu peaks in Figure 17 give the following composition in areal densities: 0.45×10^{17} S/cm^2 and 2.79×10^{18} Cu/cm^2. Although the sample was tilted to 45° relative to the incident beam in order to increase depth resolution, the sulfidized layer remained unresolved; therefore only areal densities are determined rather than a composition-depth profile. In fact, the loss of depth resolution is one of the difficulties encountered when using proton RBS rather than α RBS, as shown in this figure. Nevertheless, the enhanced S cross-section gives an enhanced yield relative to the Cu signal and good counting statistics for the areal density, yielding an uncertainty of ±10%.

Finally, focus is placed on one sample analyzed with different ion beam species and energies in order to demonstrate

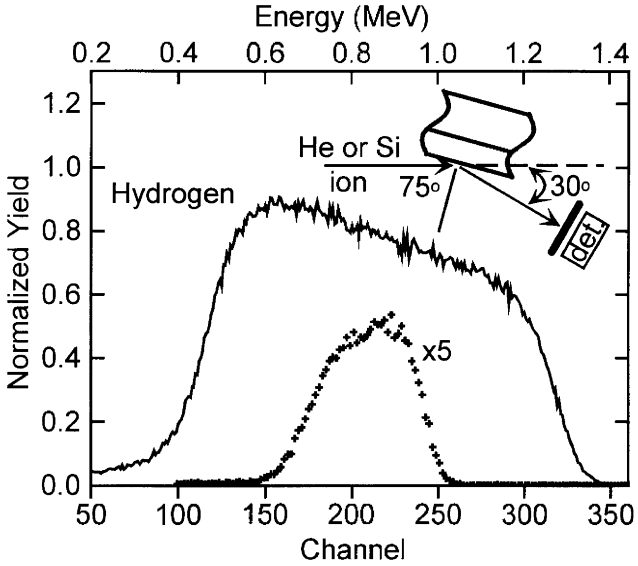


Figure 18. The ERD spectra collected from a ≈300-nm-thick Si_3N_4 layer on Si. The only detected recoil atom for these spectra is H. The 16-MeV Si ERD spectrum is shown as a solid line, and the 2.8-MeV He ERD spectrum is magnified 5 times and shown as plus symbols. For both spectra, $\phi = 30°$, $\Theta_1 = 75°$, $Q = 4$ pμC, and $\Omega = 5.26$ msr.

ERD analysis. Figure 18 shows two ERD spectra collected from the same Si_3N_4 film on a Si substrate using either 2.8-MeV He^+ ions (+) or 16-MeV Si^{3+} ions (solid line) incident at 75° relative to the sample normal. For greater visibility, the yield for the He ERD spectrum is magnified 5 times in this figure, even though the recoil cross-section for ^{4}He recoiling ^{1}H is ~2.5 times greater than Rutherford. The recoil cross-section for ^{28}Si recoiling ^{1}H is Rutherford for 16- to 30-MeV Si. Both spectra were collected at a scattering angle $\varphi = 30°$ and using a 12-μm Mylar range foil in front of the SBD. The only signal visible in these spectra arises from recoiled H atoms because all other species are either not recoiled or stopped in the range foil. This figure demonstrates that the yield and depth resolution are greatly increased by using a heavier incident projectile. Nevertheless, He ERD is often done (Green and Doyle, 1986) because either a He ion beam causes less radiation damage in the sample (e.g., polymers) or the heavy ion sources and high energies required for the heavy ion ERD are unavailable on a given accelerator.

A comparison of ERD spectra collected from this same silicon nitride sample for different incident Si ion energies is shown in Figure 19. Each spectrum was collected for $\Theta_1 = 75°$ and $\varphi = 30°$ using a 12-μm Mylar range foil. This figure shows that increasing the incident Si ion energy above 16 MeV decreases the yield for H and introduces a signal from the recoiled N that can now get through the Mylar foil. The front edges of the N and H nearly overlap for the 24-MeV Si ERD spectrum but are well separated for the 30-MeV Si ERD spectrum. As a result of recoil atom stopping in the range foil, this variation in mass resolution does not follow a simple analytical formula and is best determined by experiment or using a computer code such as SERDAP (Barbour, 1994) to determine the expected surface energies. Further, the N signal from the entire thickness of the film was observed using 30-MeV Si ERD, whereas only a portion of the N signal was observed using 24-MeV Si ERD. Thus the depth of analysis is increased by using a higher energy ion beam. The H signals are easily separated from the background signals using a linear fit of the background. These signals were then scaled by the cross-sections and stopping powers as a function of depth for the 16- and 30-MeV Si ERD spectra. The H profiles obtained from this spectral scaling, shown in Figure 20, demonstrate the increased depth

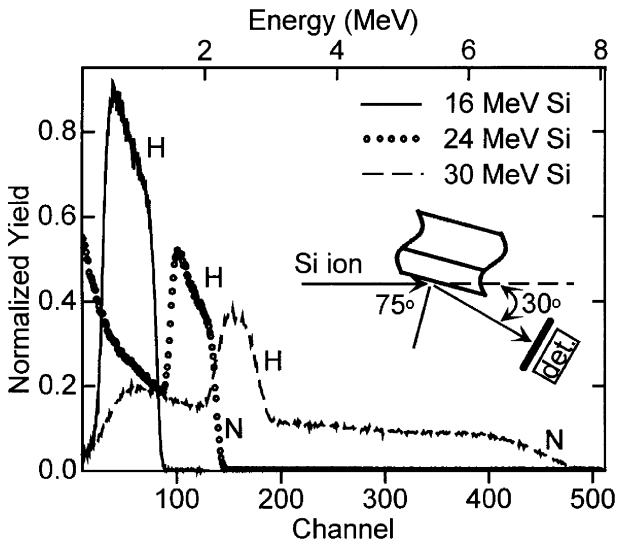


Figure 19. The ERD spectra from the same sample as in Figure 18 collected using three different ion energies: 16-MeV Si (solid line), 24-MeV Si (open circles), and 30-MeV Si (dashed line). The higher energy ERD spectra have the signal from recoiled H atoms sitting on the signal from recoiled N atoms. For all three spectra, $\phi = 30°$, $\Theta_1 = 75°$, and $\Omega = 5.26$ msr. For the 16- and 24-MeV Si spectra, $Q = 4$ pμC, whereas $Q = \Omega$ pμC for the 30-MeV spectrum.

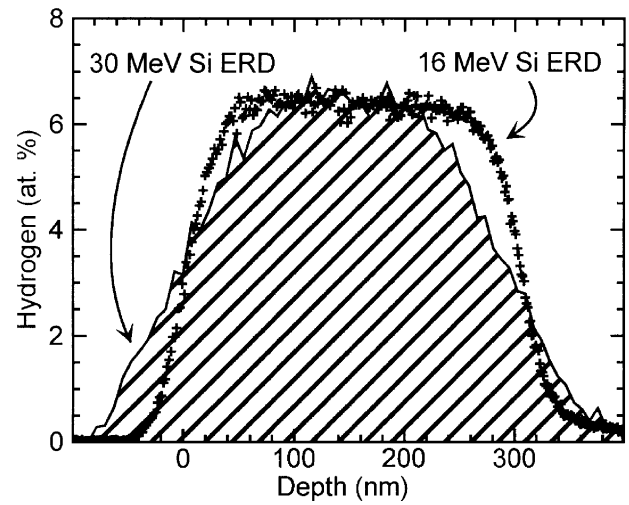


Figure 20. The H concentration-vs.-depth profiles determined from the 16-MeV Si (+) and 30-MeV Si (solid line) ERD data given in Figure 19. The N background was subtracted with a linear fit of the N signal under the H signal for the 30-MeV Si ERD data.

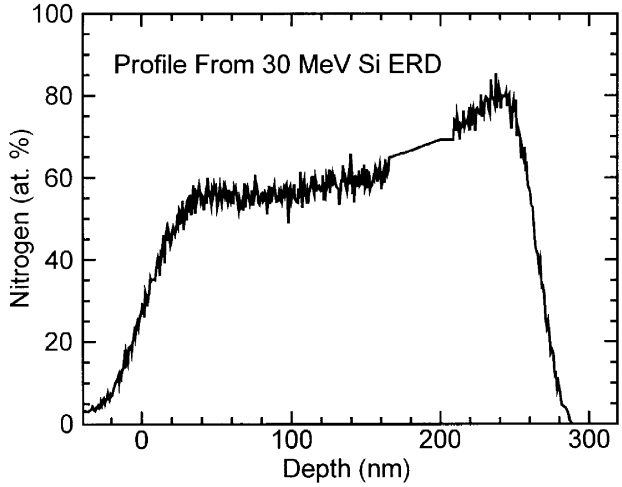


Figure 21. The N concentration-vs.-depth profile determined from the N signal for the 30-MeV Si ERD data given in Figure 19. The straight-line profile from ~160 to 210 nm has no visible statistical noise because of the linear fit used to eliminate the H signal sitting on top of the N signal.

resolution obtained with the lower energy ion beam. The depth and concentration scales were determined using the volume density of 0.903×10^{23} atoms/cm^3 for Si_3N_4.

The N profile was obtained only for the higher energy IBA and is given in Figure 21 using the same assumed volume density as in Figure 20. Comparing Figures 20 and 21 shows that the depth resolution for N is better at a given Si energy than for H as a result of the greater stopping by N than by H in the range foil. Comparison of these two figures also gives a check on the accuracy of the nitrogen stopping power in the Mylar range foil. Independently, the H and N depth profiles yield approximately the same Si_3N_4 film thickness, and therefore the stopping power used for N is acceptable. The uncertainty in the N stopping power may, however, cause inaccurate scaling of the N concentration at depth. The N profile in Figure 21 should remain constant with depth but appears to increase beyond ~160 nm. This increase may result from inaccurately calculating the N stopping power, from multiple scattering in the sample, or from a pulse-height defect effect called the nuclear deficit (i.e., some ion energy is deposited in the detector as atomic displacements rather than through the process of producing electron-hole pairs). Further work is needed to accurately determine the N stopping power in Mylar for the N energy range from 1 to 8 MeV.

LITERATURE CITED

Abbondanno, U., Lagonegro, M., Pauli, G., Poliani, G., and Ricci, R. A. 1973. Isospin-forbidden analogue resonances in ^{33}Cl, II: Levels of ^{33}Cl and higher $T = 3/2$ resonances. *Nuovo Cimento* 13A:321.

Aldridge, J. P., Crawford, G. E., and Davis, R. H., 1968. Optical-model analysis of the $^{32}S(\alpha,\alpha)^{32}S$ elastic scattering from 10.0 MeV to 17.5 MeV. *Phys. Rev.* 167:1053.

Andersen, H. H., Besenbacher, F., Loftager, P., and Moller, W., 1980. Large-angle scattering of light ions in the weakly screened Rutherford region. *Phys. Rev. A* 21:1891.

Barbour, J. C. 1994. SERDAP computer program. Sandia National Laboratories, Dept. 1111, Albuquerque, New Mex.

Bird, J. R. and Williams, J. S. (eds.). 1989. Ion Beams for Materials Analysis. Academic Press, San Diego.

Bozoian, M. 1991a. Threshold of non-Rutherford nuclear cross-sections for ion beam analysis. *Nucl. Instrum. Methods B* 56/57:740.

Bozoian, M. 1991b. Deviations from Rutherford backscattering for $Z = 1$, 2 projectiles. *Nucl. Instrum. Methods B* 58:127.

Bozoian, M., Hubbard, K. M., and Nastasi, M. 1990. Deviations from Rutherford-Scattering cross-sections. *Nucl. Instrum. Methods B* 51:311.

Bragg, W. H. and Kleeman, R. 1905. The α particles of radium and their loss of range in passing through various atoms and molecules. *Philos. Mag.* 10:318.

Brice, D. K. 1973. Theoretical analysis of the energy spectra of back-scattered ions. *Thin Solid Films* 19:121.

Chu, W.-K., Mayer, J. W., and Nicolet, M.-A. 1978. Backscattering Spectrometry. Academic Press, New York.

Doolittle, L. R. 1986. Algorithms for the rapid simulation of Rutherford backscattering spectra. *Nucl. Instrum. Methods B* 9:344.

Doolittle, L. R. 1994. Computer code "RUMP." Computer Graphics Services, Ithaca, N.Y.

Doyle, B. L. and Brice, D. K. 1988. The analysis of elastic recoil detection data. *Nucl. Instrum. Methods B* 35:301.

Feldman, L. C., Mayer, J. W., and Picrauz, S. T. 1982. Materials Analysis by Ion Channeling. Academic Press, New York.

Foti, G., Mayer, J. W., and Rimini, E. 1977. Backscattering spectrometry. *In* Ion Beam Handbook for Material Analysis (J. W. Mayer and E. Rimini, eds.). p. 22. Academic Press, New York.

Gebauer, B., Fink, D., Goppelt, P., Wilpert, M., and Wilpert, Th. 1990. A multidimensional ERDA spectrometer at the VICKSI heavy ion accelerator. *In* High Energy and Heavy Ion Beams in Materials Analysis (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). p. 257. Materials Research Society, Pittsburgh.

Geiger, H. and Marsden, E. 1913. The laws of deflexion of α particles through large angles. *Philos. Mag.* 25:604.

Green, P. F. and Doyle, B. L. 1986. Silicon elastic recoil detection studies of polymer diffusion: Advantages and disadvantages. *Nucl. Instrum. Methods B* 18:64.

Hubbard, K. M., Tesmer, J. R., Nastasi, M., and Bozoian, M. 1991. Measured deviations from Rutherford backscattering cross sections using Li-ion beams. *Nucl. Instrum. Methods B* 58:121.

L'Ecuyer, J., Brassard, C., Cardinal, C., Chabbal, J., Deschenes, L., Labrie, J. P., Terrault, B., Martel, J. G., and St.-Jacques, R. 1976. An accurate and sensitive method for the determination of the depth distribution of light elements in heavy materials. *J. Appl. Phys.* 47:381.

Marion, J. B. and Young, F. C. 1968. Nuclear Reaction Analysis. North-Holland Publishing, New York.

Mayer, J. W. and Lau, S. S. 1990. Electronic Materials Science: For Integrated Circuits in Si and GaAs. Macmillan, New York.

Mayer, J. W. and Rimini, E. (eds.). 1977. Ion Beam Handbook for Materials Analysis. Academic Press, New York.

Rutherford, E. 1911. The scattering of α and β particles by matter and the structure of the atom. *Philos. Mag.* 21:669.

Tesmer, J. R., Nastasi, M., Barbour, J. C., Maggiore, C. J., and Mayer, J. W. (eds.). 1995. Handbook of Modern Ion Beam Materials Analysis. Materials Research Society, Pittsburgh.

Thomas, J. P., Fallavier, M., and Ziani, A., 1986. Light elements depth-profiling using time-of-flight and energy detection of recoils. *Nucl. Instrum. Methods B* 15:443.

Tilley, D. R., Cheves, C. M., Kelley, J. H., Raman S., and Weller, H. R. 1998. Energy levels of light nuclei $A = 20$. *Nucl. Phys. A* 636(3):249–364.

Turkevich, A. L., 1961. Chemical analysis of surfaces by use of large-angle scattering of heavy charged particles. *Science* 134:672.

Whitlow, H. J. 1990. Mass and energy dispersive recoil spectrometry: A new quantitative depth profiling technique for microelectronic technology. *In* High Energy and Heavy Ion Beams in Materials Analysis (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). p. 73. Materials Research Society, Pittsburgh.

Wielunski, L., Benenson, R., Horn, K., and Lanford, W. A. 1986. High sensitivity hydrogen analysis using elastic recoil detection. *Nucl. Instrum. Methods B* 15:469.

Ziegler, J. F. 1980. Handbook of Stopping Cross-Sections for Energetic Ions in All Elements. Pergamon Press, New York.

Ziegler, J. F., 1987. RSTOP computer code from TRIM-91. IBM Research Center, Yorktown Heights, N.Y.

Ziegler, J. F., Biersack, J. P., and Littmark, U., 1985. The Stopping and Range of Ions in Solids. Pergamon Press, Elmsford, N.Y.

ADDITIONAL REFERENCES FOR SPECIAL DETECTORS AND DETECTION SCHEMES

Bowman, J. D. and Heffner, R. H. 1978. A novel zero time detector for heavy ion spectroscopy. *Nucl. Instrum. Methods* 148:503.

Chu, W.-K. and Wu, D.T., 1988. Scattering recoil coincidence spectroscopy. *Nucl. Instrum. Methods B* 35:518.

Gossett, C. R. 1986. Use of a magnetic sector spectrometer to profile light elements by elastic recoil detection. *Nucl. Instrum. Methods B* 15:481.

Hofsass, H. C., Parich, N. R., Swanson, M. L., and Chu, W.-K. 1990. Depth profiling of light elements using elastic recoil coincidence spectroscopy (ERCS). *Nucl. Instrum. Methods B* 45:151.

Kraus, R. H., Vieira, D. J., Wollnik, H., and Wouters, J. M. 1988. Large-area fast-timing detectors developed for the TOFI spectrometer. *Nucl. Instrum. Methods A* 264:327.

Nagai, H., Hayashi, S., Aratani, M., Nozaki, T., Yanokura, M., Kohno, I., Kuboi, O., and Yatsurugi, Y., 1987. Reliability, detection limit and depth resolution of the elastic recoil measurement of hydrogen. *Nucl. Instrum. Methods B* 28:59.

Whitlow, H. J. 1990. Time of flight spectroscopy methods for analysis of materials with heavy ions: A tutorial. *In* High Energy and Heavy Ion Beams in Materials Analysis (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). p. 243. Materials Research Society, Pittsburgh.

Zebelman, A. M., Meyer, W. G., Halbach, K., Poskanzer, A. M., Sextro, R. G., Gabor, G., and Landis, D. A., 1977. A zero-time detector utilizing isochronous transport of secondary electrons. *Nucl. Instrum. Methods* 141:439.

KEY REFERENCES

Bird and Williams, 1989. See above.

Good general overview of ion beam analysis techniques.

Chu et al., 1978. See above.

Tutorial-style book for learning RBS.

Tesmer et al., 1995. See above.

Tutorial handbook for learning ion-beam analysis.

INTERNET RESOURCES

http://www.sandia.gov/1100/ibanal.html

http://corpbusdev.sandia.gov/Facilities/Descriptions/beam-materials. htm

The home page of the Ion Beam Materials Research Laboratory at Sandia National Laboratories

http://www.tunl.duke.edu/nucldata/groupinfo.html

A web site for nuclear data sheets maintained by Nuclear Data Group, Triangle Univesities Nuclear Laboratory (TUNL)

http://fai.idealibrary.com:80/cgi-bin/fai.idealibrary.com_8011/iplogin/toc/ds

A web site for nuclear data sheets maintained by Academic Press.

APPENDIX A: GLOSSARY OF TERMS AND SYMBOLS

AD	Areal density (number of atoms per unit area)
dE/dx	Energy loss of a particle (generally as a result of electronic stopping)
dE_d/dx	Effective stopping power for the detected particle at a given projectile energy (an incremental change in E_d corresponds to an incremental change in depth, dx, at x)
dx	Increment of depth at x corresponding to an increment of energy dE_d
E_0	Energy of the incident projectile particle
E_0'	Projectile energy at a depth x'
$E_0^{(n)}$	Projectile energy incident upon slab $n+1$
$E_3^{(n)}$	Energy of the backscattered atom (or recoiled atom for ERD) as it emerges from slab n
E_d	Energy of detected particle
$E_d^{(n)}$	Detected energy for the nth slab
$E_f^{(j-1)}$	Energy in foil before recoil atom traverses the jth slab
E_{NR}	Energy where σ becomes non-Rutherford
K	Kinematic factor (ratio of the scattered particle energy to the projectile energy for an elastic collision)
L	Length of light tube
M_1	Mass of the incident projectile particle

M_2	Mass of a target atom involved in a scattering event
N	Volume density of atoms in the target
NF	Number of slabs in foil
Q	Number of incident projectiles (determined from a target current and the projectile charge)
S^{eff}	Effective stopping power
S_{p}	Projectile particle stopping power (energy loss), $= dE/dx$ projectile
S_{r} (S_{b})	Recoil (backscattered) particle stopping power (energy loss)
$[S]_{\text{p,r}}$ or $[S]_{\text{p,b}}$	Energy loss factor corresponding to a recoil (backscattering) event at energy E
t	Layer thickness; also time
V_{LJ}	Lenz-Jensen potential
x	Distance (depth) in the sample, measured along the sample normal
$X^{(0)}$	Thickness of range foil
$X^{(n)}$	Thickness of slab (sublayer) n within the sample
$Y(E_{\text{d}})$	Yield (height) of detected particles at a channel corresponding to energy E_{d}
z	Sample-to-detector distance
Z_1	Atomic number of the incident projectile particle
Z_2	Atomic number of a target atom involved in a scattering event
δE_{d}	Energy width of a channel in the multi- channel analyzer
δx	Smallest resolvable depth interval
ΔE_{foil}	Energy lost by a recoil atom in traversing the range foil
$\Delta x^{(0)}$	Incremental slab thickness for range foil containing $X^{(0)}/\Delta x^{(0)}$ slabs
ε	Stopping cross-section, $\equiv (1/N)(dE/dx)$
η	Detector efficiency
θ	Backscattering angle measured in the laboratory frame of reference
Θ_1	Incident angle, measured between the sample normal and the incident pro jectile direction
Θ_2	Detection angle, measured between the sample normal and the scattered particle direction
σ	Average differential scattering cross- section (also known as the scattering cross-section)
σ_0	Scattering cross-section of surface peak
σ_{R}	Rutherford scattering cross-section
τ	Layer thickness
ϕ	Forward (recoil) scattering angle measured in the laboratory frame of reference
Ω	Detector solid angle

APPENDIX B: RBS AND ERD EQUATIONS

Figure 22 shows equations applicable to both RBS and ERD, whereas Figure 23 shows equations applicable to either RBS or ERD. Although these analytical expressions appear complicated, their implementation through the use of a computer is simple and fast.

$$\Omega = \int_{\substack{\text{area subtended}\\ \text{by detector}}} \sin\phi \, d\phi \, d\varepsilon \approx \frac{\text{detector area}}{\text{z}^2} (\text{z} = \text{sample-to-detector distance}) \tag{1}$$

$$Y(E_{\text{d}}) = \frac{QN(x)\sigma(E_0^{(n)})\Omega\delta E_{\text{d}}}{\cos\Theta_1 \, dE_{\text{d}}/dx} \tag{5}$$

$$\text{AD} = \int_0^{\tau} N(x)(dx) = \frac{\cos\Theta_1}{Q\Omega} \int_0^{\tau} \frac{S^{\text{eff}}}{\sigma\delta E_{\text{d}}} Y(E_{\text{d}}) dx = \frac{\cos\Theta_1}{Q\Omega} \int_{E_{\text{d}}(0)}^{E_{\text{d}}(x)} \frac{Y(E_{\text{d}})}{\sigma} \frac{dE_{\text{d}}}{\delta E_{\text{d}}} \tag{8}$$

$$\text{AD} = \frac{\cos\Theta_1}{Q\Omega\sigma_0} \sum_{\text{channel}[E_{\text{d}}(0)]}^{\text{channel}[E_{\text{d}}(x)]} Y(\text{channel}) \tag{9}$$

$$\text{Bragg's rule:} \quad \varepsilon^{\text{A}}\alpha^{\text{B}}\beta = \alpha\varepsilon^{\text{A}} + \beta\varepsilon^{\text{B}} \quad \text{and} \quad S = N^{\text{A}}\alpha^{\text{B}}\beta\varepsilon^{\text{A}}\alpha^{\text{B}}\beta \tag{10}$$

$$E_0^{(n)} = E_0^{(n-1)} - \frac{X^{(n)}}{\cos\Theta_1} S_{\text{p}}^{(n)}(E_0^{(n-1)}) \tag{11}$$

$$\text{Spectral scaling:} \quad N^{(n)}(x) = \frac{\cos\Theta_1}{Q\Omega\,\delta E_{\text{d}}} \frac{dE_d/dx|_n}{\sigma_{\text{R}}(E_0^{(n)})} Y^{(n)}(E_{\text{d}}) = \frac{\cos\Theta_1}{Q\Omega\delta E_{\text{d}}} \frac{S_0^{\text{eff}}}{\sigma_0} \left[\frac{\sigma_{\text{R}}(E_0)}{\sigma_{\text{R}}(E_0^{(n)})} \frac{dE_{\text{d}}/dx|_n}{dE_{\text{d}}/dx|_0} Y^{(n)}(E_{\text{d}}) \right] \tag{22}$$

Figure 22. Equations applicable to both RBS and ERD.

RBS	Equation(s)	ERD
$K = \left(\dfrac{\sqrt{M_2^2 - M_1^2 \sin^2\theta} + M_1\cos\theta}{M_1 + M_2}\right)^2$	3,4	$K = \dfrac{4M_1M_2\cos^2\phi}{(M_1+M_2)^2}$
$Y^{(n)}(E_\mathrm{d}) = \dfrac{Q\Omega N^{(n)}(x)\sigma(E_0^{(n)},\theta)\delta E_\mathrm{d}}{\cos\Theta_1\, dE_\mathrm{d}/dx\|_n}$	5	$Y_\mathrm{r}^{(n)}(E_\mathrm{d}) = \dfrac{Q\Omega N_\mathrm{r}^{(n)}(x)\sigma_\mathrm{r}(E_0^{(n)},\phi)\delta E_\mathrm{d}}{\cos\Theta_1\, dE_\mathrm{d}/dx\|_n}$
$\dfrac{d\sigma_\mathrm{R}(E_0^{(n)},\theta)}{d\Omega} = \left(\dfrac{Z_1Z_2e^2}{4E_0^{(n)}}\right)^2 \left(\dfrac{4}{\sin^4\theta}\right) \times \dfrac{(\sqrt{1-[(M_1/M_2)\sin\theta]^2} + \cos\theta)^2}{\sqrt{1-[(M_1/M_2)\sin\theta]^2}}$	6,7	$\dfrac{d\sigma_\mathrm{R}(E_0^{(n)},\phi)}{d\Omega} = \left(\dfrac{Z_1Z_2e^2}{4E_0^{(n)}}\right)^2 \left(\dfrac{4}{\cos^3\phi}\right) \dfrac{(M_1+M_2)^2}{(M_2)^2}$
For a scattering event occurring in slab n:		For a scattering event occurring in slab n:
$E_3^{(n)} = KE_0^{(n)} - \dfrac{X^{(n)}}{\cos\Theta_2} S_\mathrm{b}^{(n)}(KE_0^{(n)})$	12	$E_3^{(n)} = KE_0^{(n)} - \dfrac{X^{(n)}}{\cos\Theta_2} S_\mathrm{r}^{(n)}(KE_0^{(n)})$
For a scattering event occurring in slab $j > n$:		For a scattering event occurring in slab $j > n$:
$E_3^{(n)} = E_3^{(n+1)} - \dfrac{X^{(n)}}{\cos\Theta_2} S_\mathrm{b}^{(n)}(E_3^{(n+1)})$	13	$E_3^{(n)} = E_3^{(n+1)} - \dfrac{X^{(n)}}{\cos\Theta_2} S_\mathrm{r}^{(n)}(E_3^{(n+1)})$
$E_\mathrm{d}^{(n)} = K\left[E_0 - \dfrac{X^{(n)}}{\cos\Theta_1}\sum_{j=1}^{n} S_\mathrm{p}^{(j)}(E_0^{(j-1)})\right] - \dfrac{X^{(n)}}{\cos\Theta_2}\left[S_\mathrm{b}^{(n)}(KE_0^{(n)} + \sum_{j=n}^{2} S_\mathrm{b}^{(j-1)}(E_3^{(j)})\right]$	14,15	$E_\mathrm{d}^{(n)} = K\left[E_0 - \dfrac{X^{(n)}}{\cos\Theta_1}\sum_{j=1}^{n} S_\mathrm{p}^{(j)}(E_0^{(j-1)})\right] - \dfrac{X^{(n)}}{\cos\Theta_2}\left[S_\mathrm{r}^{(n)}(KE_0^{(n)}) + \sum_{j=n}^{2} S_\mathrm{r}^{(j-1)}(E_3^{(j)})\right] - \Delta E_\mathrm{foil}$
	16	$\Delta E_\mathrm{foil} = \Delta x^{(0)} \sum_{j=1}^{\mathrm{NF}} S_\mathrm{r}^{(0)}(E_\mathrm{f}^{(j-1)})$
$\left.\dfrac{dE_\mathrm{d}}{dx}\right\|_n = [S]_\mathrm{p,b}^{(n)} \dfrac{S_\mathrm{b}(E_\mathrm{d})}{S_\mathrm{b}(KE_0^{(n)})}$	17,18	$\left.\dfrac{dE_\mathrm{d}}{dx}\right\|_n = [S]_\mathrm{p,r}^{(n)} \prod_n$
$[S]_\mathrm{p,b}^{(n)} = \dfrac{KS_\mathrm{p}^{(n)}(E_0^{(n)})}{\cos\Theta_1} + \dfrac{S_\mathrm{b}^{(n)}(KE_0^{(n)})}{\cos\Theta_2}$	19,20	$[S]_\mathrm{p,r}^{(n)} = \dfrac{KS_\mathrm{p}^{(n)}(E_0^{(n)})}{\cos\Theta_1} + \dfrac{S_\mathrm{r}^{(n)}(KE_0^{(n)})}{\cos\Theta_2}$
	21	$\prod_n = \dfrac{S_\mathrm{r}^{(n)}(E_3^{(n)})}{S_\mathrm{r}^{(n)}(E_2)} \dfrac{S_\mathrm{r}^{(n-1)}(E_3^{(n-1)})}{S_\mathrm{r}^{(n-1)}(E_3^{(n)})} \cdots \dfrac{S_\mathrm{r}^{(0)}(E_\mathrm{d})}{S_\mathrm{r}^{(0)}(E_3^{(1)})}$

Figure 23. Equations specific to either RBS or ERD.

J. C. BARBOUR
Sandia National Laboratories
Albuquerque, New Mexico

NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION

INTRODUCTION

Nuclear reaction analysis (NRA) and proton- (particle-) induced gamma ray emission (PIGE) are based on the interaction of energetic (from a few hundred kilo-electron-volt to several mega-electron-volt) ions with light nuclei. Every nuclear analytical technique that uses nuclear reactions has a unique feature—isotope sensitivity. Therefore, it is insensitive to matrix effects, and there is much less interference than in methods where signals from different elements overlap. In NRA the nuclear reaction produces charged particles, while in PIGE the excited nucleus emits gamma rays. Sometimes the charged particle emission and the gamma ray emission occur simultaneously, such as in the $^{19}\mathrm{F}(\mathrm{p},\alpha\gamma)^{16}\mathrm{O}$ reaction. Both NRA and PIGE measure the concentration and depth distribution of elements in the surface layer (few micrometers) of the sample. Both techniques are limited by the available nuclear reactions. Generally, they can be used for only light elements, up to calcium. This is the basis for an important property of these methods: the nuclear reaction technique is one of the few analytical techniques that can quantitatively measure hydrogen profiles in solids. Since these techniques are sensitive to the nuclei in the sample, they are unable to provide information

about the chemical states and bonds of the elements in the sample. For the same reason they cannot provide information about the microscopic structure of the sample. Combined with channeling, NRA or PIGE can provide information about the location of the measured element in a crystalline lattice.

The sensitivity and depth resolution of these methods depend on the specific nuclear reaction. The sensitivity typically varies between 10 and 100 ppm while the depth resolution can be as good as a few nanometers or as large as hundreds of nanometers. The lateral resolution of the method depends on the size of the bombarding ion beam. Good nuclear microprobes are currently in the few-micrometer range. Also, using nuclear microprobes, an elemental/isotopic image of the sample can be recorded. Both NRA and PIGE are nondestructive, although some materials might be lattice damaged after long, high-current bombardment.

Although NRA and PIGE are quantitative, in most cases standards have to be used. Depending on the shape of the particular nuclear cross-section, nonresonant or resonant depth profiling can be used. The resonant profiling (in most cases PIGE uses resonances but there are a few charged particle reactions that have usable resonance) typically can give very good resolution, but the measurement takes much longer than in nonresonant profiling; therefore, the probability of inducing changes in the sample by the ion beam is higher.

These techniques require a particle accelerator capable of accelerating ions up to several mega-electron-volts. This limits their availability to laboratories dedicated to these methods or that have engaged in low-energy nuclear physics in the past (i.e., they have a particle accelerator that is no longer adequate for modern nuclear physics experiments because of its low energy but is quite suitable for nuclear analytical techniques).

This unit will concentrate on the specific aspects of these two nuclear analytical techniques and will not cover the details of the ion-solid interaction or the general aspects of ion beam techniques (e.g., stopping power, detection of ions). These are described in detail elsewhere in this part (see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS and MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY).

Competitive and Complementary Techniques

Practically every analytical technique that measures the concentration and depth profile of elements in the top few micrometers of a solid competes with NRA and PIGE. However, either most of these techniques are not isotope sensitive or their isotope resolution is not very good. The competing techniques can be divided into two groups: ion beam techniques and other techniques. The ion beam techniques that compete with NRA and PIGE are Rutherford backscattering spectrometry (RBS; see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS), elastic recoil detection (ERD or ERDA; see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS), proton-induced x-ray emission (PIXE; see PARTICLE-INDUCED X-RAY EMISSION), secondary ion mass spectroscopy (SIMS), medium-energy ion scattering (MEIS; see MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY), and ion scattering spectroscopy (ISS; see HEAVY-ION BACKSCATTERING SPECTROMETRY).

Rutherford backscattering spectrometry and ERD can be considered special cases of NRA in which the nuclear reaction is just an elastic scattering. The main advantage of RBS and ERD is that they are able to see almost all elements in the periodic table (with the obvious exception in RBS when the element to be detected is lighter than the bombarding ions). This can be a disadvantage when a light element has to be measured in a heavy matrix (e.g., the measurement of oxygen in tantalum by RBS). This is the typical case when NRA can solve the problem but RBS cannot. In ERD, since the forward recoiled atoms have to be mass analyzed, the measurement becomes more complicated than in NRA, requiring more sophisticated (and more expensive) instruments. The resolution and sensitivity achieved when using RBS, ERD, and NRA are on the same order, although when using very heavy ions, the sensitivity and depth resolution can be an order of magnitude better in ERD than in NRA. Using heavy ions presents other problems, but these are beyond the scope of this unit (see HEAVY-ION BACKSCATTERING SPECTROMETRY). For more discussion see Green and Doyle (1986) and Davies et al. (1995).

Proton-induced x-ray emission (see PARTICLE-INDUCED X-RAY EMISSION) can detect most elements, except H and He. Also, the detection of x rays from low-Z elements (below Na) requires a special, windowless detector. Another drawback of PIXE is that it generally cannot provide depth information.

Secondary ion mass spectroscopy can be used for most of the analyses performed in NRA, including hydrogen profiling. Its sensitivity and depth resolution are superior to those of NRA. The main advantage of NRA over SIMS is that while NRA is a nondestructive technique, SIMS depth profiling destroys the sample. Also, the depth scale with SIMS depends on accurate knowledge of sputtering rates for a specific sample.

Although MEIS and ISS can be considered competing techniques, their probing depth is much smaller (which also means much better depth resolution) than that of NRA.

Among the non–ion beam techniques, Auger electron spectroscopy (AES, see AUGER ELECTRON SPECTROSCOPY), x-ray photoelectron spectroscopy (XPS), x-ray fluorescence (XRF, see X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS), and neutron activation analysis (NAA) should be mentioned.

Auger electron spectroscopy and XPS not only provide concentration information but also are sensitive to the chemical states of the elements; therefore, they give information about chemical bonds. Since both methods get the information from the first few nanometers of the sample, to measure a depth profile, layers of the sample have to be removed (usually by sputtering); therefore, the sample is destroyed. X-ray fluorescence can provide concentration information about elements heavier than Na (again, it depends on the x-ray detector) but it cannot measure the depth profile. The only technique listed that has the same isotope sensitivity as NRA is NAA. However, NAA is a bulk technique and does not provide any depth information.

Most of the above-mentioned techniques are complementary techniques to NRA, and in many cases are, used concurrently with NRA. [It is quite common in ion beam analysis (IBA) laboratories for an ion-scattering chamber to have an AES or XPS spectrometer mounted on it.] The most frequent combination is RBS and NRA, since they are closely related and use the same equipment. Whereas, NRA can measure light elements in a heavy matrix but cannot see the matrix itself (at least not directly), RBS is very sensitive and has a good mass resolution for heavy elements but cannot see a small amount of some light element in a heavy matrix.

PRINCIPLES OF THE METHOD

Although NRA and PIGE are similar to other high-energy ion beam techniques (see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS and PARTICLE-INDUCED X-RAY EMISSION), here we will discuss only the principles specific to NRA and PIGE.

Both NRA and PIGE measure the prompt reaction products from nuclear reactions. The yield (number of detected particles, γ rays) provides information about the concentration of elements, and the energy of the detected charged particles provides information about the depth distribution (depth profile) of the elements. Depending on whether the reaction cross-section has a sharp resonance or not, the methods are distinguished as either resonant or nonresonant (e.g., PIGE uses only the resonant method).

When a high-energy ion beam is incident on a target, the ions slow down in the material and either undergo elastic scattering (RBS, ERDA) or induce a nuclear reaction at various depths. In a nuclear reaction, a compound nucleus is formed and almost immediately a charged particle (p, d, ^{3}He, and ^{4}He) and/or a γ photon is emitted. Many (p, α) reactions have an associated γ photon. The emitted charged particle/photon then leaves the target and is detected by an appropriate detector. The energy of the charged particle depends on the angle of the incident ion and the emitted particle, the kinetic energy of the incident particle, and the Q value of the reaction, where Q is the energy difference between the initial and final nuclear states. The energy of the emitted γ photon is determined by the energy structure of the compound nucleus (for the formulas, see Appendix). The detected number of emitted particles is proportional to the number of incident particles, the solid angle of the detector, the concentration of the atoms participating in the nuclear reaction, and the cross-section of the reaction. The charged reaction products will lose energy on their way out from the target; therefore, their energy will carry information about the depth where the nuclear reaction occurred. Since the energy of the γ photons does not change while they travel through the target, they do not provide depth information.

Nonresonant Methods

Overall Near-Surface Content. If the cross-section changes slowly in the vicinity of the E_0 bombarding energy, the absolute value of nuclei per square centimeters

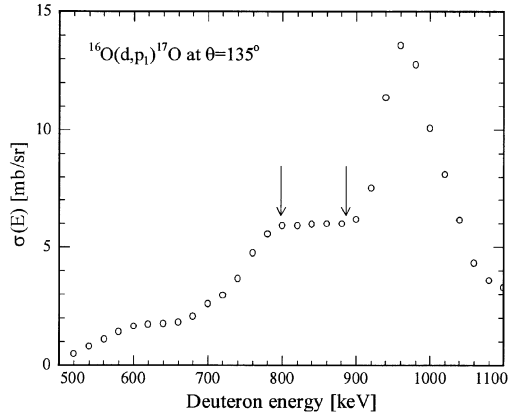


Figure 1. Cross-section of the ^{16}O(d, p$_1$)^{17}O reaction; θ is the scattering angle (Jarjis, 1979).

can be determined in thin layers independent of the concentration profile and the other components of the target. Assuming that the incident ions lose ΔE energy in the thin layer and the reaction cross-section $\sigma(E) \approx \sigma(E_0)$ for $E_0 > E > E - \Delta E$, the number of particles detected is

$$Y = \frac{Q_C \Omega \sigma(E_0) Nt}{\cos \alpha_{in}} \qquad (1)$$

where Q_C is the collected incident ion charge, is the solid angle of the detector, α_{in} is the angle between the direction of the incident beam and the surface normal of the target, and Nt is the number of nuclei per square centimeter. The spectrum of emitted particle would contain a peak with an area Y. An example of such a cross-section is shown in Figure 1. The ^{16}O(d, p$_1$)^{17}O reaction has a plateau between 800 and 900 keV, as indicated by arrows in the figure. This reaction is frequently used to determine oxygen content of thin layers or thickness of surface oxide layers up to several hundreds of nanometers.

Depth Profiling. When the thickness of the sample becomes larger and the cross-section cannot be considered constant, the spectrum becomes the convolution of the concentration profile, the energy resolution of the incident and the detected beam, and the cross-section. A typical scattering geometry is shown in Figure 2. Although the figure shows only backward geometry, forward geometry is used in certain cases as well. The depth scale can be calculated by determining the energy loss of the incident ions

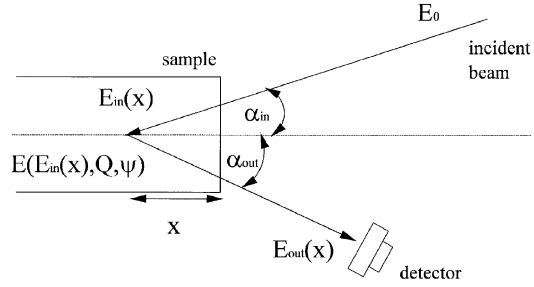


Figure 2. Typical scattering geometry used in NRA experiments.

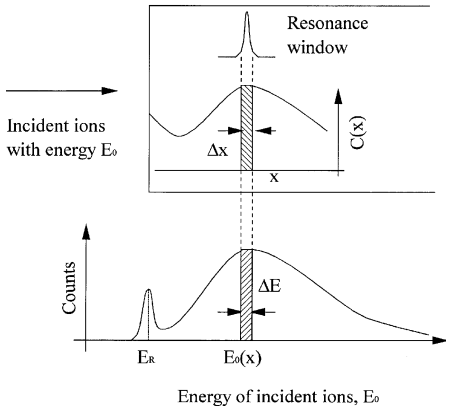


Figure 3. Principle of resonant depth profiling.

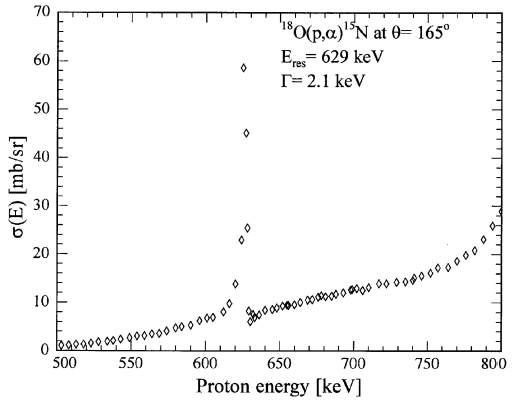


Figure 4. Cross-section of the ^{18}O(p, α)^{15}N reaction around the 629-keV resonance; θ is the scattering angle (Amsel and Samuel, 1967).

before they induce the nuclear reaction and the energy loss of the reaction product particles:

$$E_{\text{out}}(x) = E(E_{\text{in}}(x),\ Q,\ \psi) - \int_0^{x/\cos\alpha_{\text{out}}} S_{\text{out}}(E)\,dx \qquad (2)$$

$$E_{\text{in}}(x) = E_0 - \int_0^{x/\cos\alpha_{\text{in}}} S_{\text{in}}(E)\,dx \qquad (3)$$

where x is the depth at which the nuclear reaction occurs, E_0 is the energy of the incident ions, Q is the Q value of the reaction, ψ is the angle between the incident and detected beams, α_{in} and α_{out} are angles of the incident and detected beams to the surface normal of the sample, and S_{in} and S_{out} are the stopping powers of the incident ions and the reaction products. The term $E(E_{\text{in}}(x), Q, \psi)$ is the energy of the reaction product calculated from Equation 7. (Also see Fig. 7.)

In most cases, to get quantitative results, a nonlinear fit of the spectrum to a simulated spectrum is necessary (Simpson and Earwaker, 1984; Vizkelethy, 1990; Johnston, 1993). For a discussion of energy broadening, see the Appendix.

Resonant Depth Profiling

In case of sharp resonances in the cross-section, most particles/γ photons come from a very narrow region in the target that is equal to the energy-loss-related width of the resonances. As the energy of the incident beam increases, the incident ions slow to the resonance energy at deeper and deeper depth; therefore, the thin layer from which the reaction products are coming lies deeper and deeper. In this way a larger depth can be probed and a depth profile can be measured. The principle of the method is illustrated in Figure 3.

The depth x and the energy E_0 of the incident ions are related through the equation

$$E_0(x) = E_{\text{R}} + \int_0^{x/\cos\alpha_{\text{in}}} S_{\text{in}}(E)\,dx \qquad (4)$$

where E_{R} is the resonance energy. In the figure $C(x)$ is the concentration of the element to be detected. Figure 4 shows the cross-section of the ^{18}O(p, α)^{15}N reaction. The resonance at 629 keV is widely used in ^{18}O tracing experiments where oxygen movement is studied. This is an excellent example of how useful NRA and PIGE are, since these methods are sensitive to only ^{18}O and not to the other oxygen isotopes. The measured spectrum ("excitation curve") is the convolution of the concentration profile and the energy spread of the detection system. To extract the depth profile from the excitation function, we should deconvolve it with the depth-dependent energy spread or fit the excitation function to a simulation of it. The theory and simulation of narrow resonances are discussed in detail in Maurel (1980), Maurel et al. (1982), and Vickridge (1990).

PRACTICAL ASPECTS OF THE METHOD

Equipment

The primary equipment used in NRA and PIGE is an electrostatic particle accelerator, which is the source of the high-energy ion beam. The accelerators used for IBA are either Cockroft-Walton or van de Graaff type (or some slight variation of them, such as pelletron accelerators) with terminal voltage in the range of a few million volts. When higher energies are needed or heavy ions have to be accelerated, tandem accelerators are used. In general, the requirements for the accelerators used in NRA and PIGE are similar to those of any other IBA technique. In addition, to perform resonance profiling, the accelerators must have better energy resolution and stability than are necessary for RBS or ERDA and be capable of changing the beam energy easily. Usually, accelerators with slit feedback energy stabilization systems are satisfactory. Another NRA-specific requirement is needed because of the deuteron-induced nuclear reactions. When deuterons are accelerated, neutrons are generated mainly from the D(d, n) ^{3}He reaction. Therefore, if deuteron-induced nuclear reactions are used, additional shielding is necessary to protect personnel against neutrons. Many laboratories using ion beam techniques do not have this shielding.

Most NRA and PIGE measurements take place in vacuum, with the exception of a few extracted beam experiments. The required vacuum, i.e., $>10^{-3}$ Pa,

depends on the particular application; it usually ranges from 10^{-3} to 10^{-7} Pa. The main problem caused by bad vacuum is hydrocarbon layer formation on the sample under irradiation, which can interfere with the measurement. Having a load-lock system attached to the scattering chamber makes the sample changing convenient and fast.

In most cases, standard surface barrier or ion-implanted silicon particle detectors are used to detect charged particles. These detectors usually have an energy resolution around 10 to 12 keV, which is adequate for NRA. When better energy resolution is required, electrostatic or magnetic spectrometers or time-of-flight (TOF) techniques can be used.

In PIGE, NaI, BGO (bismuth germanate), Ge(Li), or HPGE (high-purity germanium) detectors are used to detect γ rays. The NaI and BGO detectors are used when high efficiency is needed (measuring low concentrations, or the cross-section is too small) using well-isolated resonances. The NaI detectors were used mainly before BGO detectors became available, but currently BGO detectors are preferred since their efficiency is higher for a given size (Kuhn et al., 1990) and have a better signal-to-background ratio. The Ge(Li) or HPGE detectors are used when there are interfering peaks in the spectrum that cannot be resolved using other detectors.

To process the signals from the detectors, standard NIM (nuclear instrument module) or CAMAC (computer-automated measurement and control) electronics modules, such as amplifiers, single-channel analyzers (SCA), and analog-to-digital converters (ADCs), are used. In most cases, the ADC is on a card in a computer or connected to a computer through some interface (e.g., GPIB, Ethernet). Therefore, the spectrum can be examined while it is being collected and the data can be evaluated immediately after the measurement.

Filtering of Unwanted Particles (NRA)

From Figure 2 it is obvious that not only the reaction products but also all the backscattered particles will reach the detector. Since the Q value of the most frequently used reactions is positive, the backscattered particle will have less energy than the reaction products; therefore, the signal from the backscattered particles will not overlap in the spectrum. [The exceptions are the (p, α) reactions with low Q value. Due to the higher stopping power of the α particles than that of the protons, the energy of an α particle coming from a deeper layer can be the same as the energy of the proton coming from another layer.] However, the cross-section of the Rutherford scattering is usually much larger than the cross-section of the nuclear reaction we want to use. Therefore, much less reaction product would be detected in a unit time than backscattered particles. Since every detection system has finite dead time, the number of backscattered particles would limit the maximum incident ion current, so the NRA measurement would take a very long time. To make the measurement time reasonable, the backscattered particles should be filtered out.

The simplest method to get rid of the backscattered particles is to use an absorber foil, since the energy of the reaction products is higher than the energy of the backscattered particles. The energy of the reaction products after the absorber foil is

$$E_{\text{abs}}(x) = E_{\text{out}}(x) - \int_0^{x_{\text{abs}}} S_{\text{abs}}(E)\,dx \qquad (5)$$

where x_{abs} and S_{abs} are the thickness and stopping power of the absorber foil and $E_{\text{out}}(x)$ is from Equation 2. Choosing an absorber foil thickness such that E_{abs} is larger than the energy range of the backscattered particles will eliminate the unwanted particles. Usually Mylar or aluminum foils are used as absorbers. The main disadvantage of the method is the poor depth resolution due the large energy spread (straggling) in the absorber foil.

There are more sophisticated methods to filter the backscattered particles. Here we briefly list them without detailed discussion:

1. Electrostatic or magnetic deflection (or the use of electrostatic or magnetic spectrometers) gives much better resolution than the absorber method but is complicated, time consuming, and expensive (spectrometers). For applications see Möller (1978), Möller et al. (1977), and Chaturvedi et al. (1990).
2. The TOF technique, a standard technique used in nuclear physics to distinguish particles, can be used to select the reaction products only. This technique requires sophisticated electronics and a two-dimensional multichannel analyzer.
3. The thin-detector technique is used when protons and α peaks overlap in a spectrum. Since the stopping power of protons is much smaller than the stopping power of α particles, the protons would lose only the fraction of their energies in a thin detector while the α particles stop completely in it. Thus the α particles will be separated from the mix. This technique uses either "dE/dx detectors," which are quite expensive, or low-resistivity detectors, with low bias voltage. Recently the use of low-resistivity detectors was studied in detail by Amsel et al. (1992).

The TOF and thin-detector techniques have the disadvantage that although they select the reaction product, the large flux of backscattered particles still reach and can damage the detector.

Energy Scanning (for Resonance Depth Profiling)

To do resonance depth profiling, the energy of the incident ion beam has to be changed in small steps. After acceleration, the ion beam goes through an analyzing magnet that selects the appropriate energy. Generally, the energy is changed by adjusting the terminal voltage and the magnetic field. Automatic energy-scanning systems have been developed using electrostatic deflection plates after the analyzing magnet by Amsel et al. (1983) and Meier and Richter (1990). Varying the voltage on these plates, the magnetic field is kept constant and the terminal voltage is changed by the slit feedback system. This

system allows energy scanning up to several hundred kilo-electron-volts.

Background, Interference, and Sensitivity

Since NRA and PIGE are both sensitive to specific isotopes, the background, interference from the sample matrix, and sensitivity, all depend on which nuclear reaction is used. When the nonresonant technique is used, the spectrum is generally background free. Interference from elements in the matrix is possible, especially in the deuteron-induced reaction case. An element-by-element discussion of the possible interferences can be found in Vizkelethy (1995) for NRA and in Hirvonen (1995) for PIGE.

In resonant depth profiling, there are two sources of background. One is natural background radiation (present only for PIGE). The effect of this background can be taken into account by careful background measurements or can be minimized by using appropriate active and passive shielding of the detector (Damjantschitsch et al., 1983; Kuhn et al., 1990; Horn and Lanford, 1990). The second background source is also present if the resonance is not an isolated resonance that sits on top of a non-resonant cross-section. In this case, the background is not constant and depends on the concentration profile of the element we want to measure. To extract a reliable depth profile from the excitation curve, a non-linear fit to a simulation is necessary.

As mentioned above, the sensitivity depends on the cross-section used and the composition of the sample. Generally the sensitivities of NRA and PIGE are on the order of 10 to 100 ppm. Lists of sensitivities for special applications can be found in Bird et al. (1974) and Hirvonen (1995).

Cross-Sections and Q values

The most important data in NRA and PIGE are the cross-sections, the Q values of the reactions, and the energy level diagrams of nuclei. The most recent compilation of Q values and energy level diagrams can be found for $A = 3$ in Tilley et al. (1987); $A = 5, \ldots, 10$ in Ajzenberg-Selove (1988); $A = 11, 12$ in Ajzenberg-Selove (1990); $A = 13, \ldots, 15$ in Ajzenberg-Selove (1991); $A = 16, 17$ in Tilley et al. (1993); f$A = 18, 19$ in Tilley et al. (1995); $A = 18, \ldots, 20$ in Ajzenberg-Selove (1987); and $A = 21, \ldots, 44$ in Endt and van der Leun (1978). Data of frequently used cross-sections can be found in Foster et al. (1995), and an extensive list of (p, γ) resonances can be found in Hirvonen and Lappalainen (1995).

The data are also available on the Internet (see Internet Resources). The energy levels are available from Lund Nuclear Data Service (LNDS) and from Triangle Universities Nuclear Laboratories (TUNL). An extensive database of nuclear cross-sections is available from the National Nuclear Data Center, NNDC) and from the T-2 Nuclear Information Services, although these serve more the needs of nuclear physicists than people working in the field of IBA. Nuclear reaction cross-sections used in IBA are available from the Sigmabase together with many (p, p) and (α,α) cross-sections.

Usage of Standards

There are two basic reasons to use standards with NRA and PIGE. First, the energy of the bombarding beam is determined by reading the generating voltmeter (GVM) of the accelerator or by reading the magnetic field of the analyzing magnet. Usually these readings are not absolute and might shift with time. Therefore, in depth profiling it is necessary to use standards to determine the energy of the resonance (with respect to the accelerator readings) and the energy spread of the beam.

Second, since the yield is proportional to the collected charge, the cross-section, and the solid angle of the detector, to make an absolute measurement, the absolute values of these three quantities must be known precisely. In most cases, these values are not available. Ion bombardment always causes secondary electron emission. Imperfect suppression of the secondary electrons falsifies the current (and therefore the collected charge) measurement. It is not easy to design a measurement setup that can measure current with high absolute precision, but high relative precision and reproducibility are easy to achieve. Also, the solid angle of the detector and the absolute cross-section are usually not known precisely. Using reference targets that compare the reference yield to the yield from the unknown sample can eliminate most of these uncertainties.

A good reference target must satisfy the following requirements: (1) have high lateral uniformity; (2) be thick enough to provide sufficient yield in reasonable time but thin enough not to cause significant change in the cross-section; (3) have a standard that is amorphous (to avoid accidental channeling); (4) be stable in air, in vacuum, and under ion bombardment; and (5) be highly reproducible.

A detailed discussion of reference targets used in nuclear microanalysis can be found in Amsel and Davies (1983).

METHOD AUTOMATION

Automation of acquisition of NRA spectra is the same as in RBS. The only task that can be automated is changing the sample. Since most of the measurements are done in high vacuum, there is a need for some mechanism to change samples without breaking the vacuum. If the scattering chamber is equipped with a load-lock mechanism, the sample change is easy and convenient and does not require automation. Without the load lock, usually large sample holders that can hold several samples are used. In this case stepping motors connected to a computer can be used to move from one sample to another.

In PIGE or resonant depth profiling, an automatic energy scanning system synchronized with data acquisition electronics is desired. Sophisticated energy scanning systems have been developed by Amsel et al. (1983) and Meier and Richter (1990). An alternative to these methods is the complete computer control of the accelerator and the analyzing magnet.

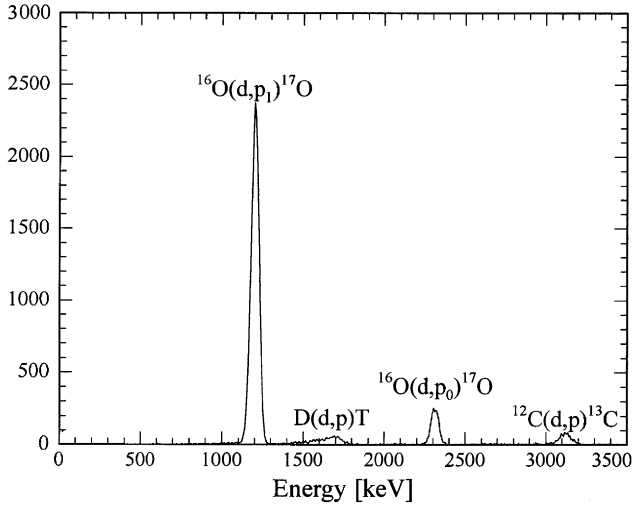


Figure 5. Measured spectrum from an 834-keV deuteron beam on 100 nm SiO_2 on Si.

DATA ANALYSIS AND INITIAL INTERPRETATION

Overall Surface Content (Film Thickness)

As mentioned above, when the cross-section can be considered flat across the thickness of the film, the overall surface content can be determined easily. Figure 5 shows a spectrum from an 834-keV deuteron beam incident on 100 nm SiO_2 on bulk Si. The ^{16}O(d, p)^{17}O reaction has two proton groups that show up as two separate peaks in the spectrum. Since the (d, p_1) cross-section has a plateau around 850 keV, we need to consider only the p_1 peak.

Using Equation 1, we can easily calculate the number of oxygen atoms per square centimeter. In most cases reference standards are used (thin Ta_2O_5 layers on Ta backing or thin SiO_2 layer on Si); therefore, the exact knowledge of the detector solid angle, cross-section, and absolute value of the collected charge is not necessary. If the reference sample contains Nt_{ref} oxygen atoms per square centimeter, the number of oxygen atoms per square centimeter in the unknown sample is

$$Nt = Nt_{\mathrm{ref}} \frac{Y}{Y_{\mathrm{ref}}} \frac{Q_{\mathrm{C}}^{\mathrm{ref}}}{Q_{\mathrm{C}}} \qquad (6)$$

where $Q_{\mathrm{C}}^{\mathrm{ref}}$ and Q_{C} are collected charges and Y_{ref} and Y are the peak areas of the p_1 protons for the reference standard and for the unknown sample, respectively. There are two peaks in the spectrum that need further explanation. The peak at higher energies is the result of the ^{12}C(d, p)^{13}C reaction from the hydrocarbon layer deposited during ion bombardment. The broad peak between the p_0 and p_1 peaks is due to the D(d, p)T reaction. Since this particular sample was used as a reference standard, during the measurements a considerable amount of deuterium had been implanted into it and is now a detectable component of the target.

Nonresonant Depth Profiling

When the cross-section cannot be considered constant, the spectrum becomes the convolution of the concentration profile with the energy broadening of the incident ions and the detected reaction particles. The interpretation is not that simple now. The energy-depth conversion can be calculated using Equations 2 and 3. If the cross-section changes only slowly, the spectrum shows the main features of the concentration profile, but this cannot be considered as quantitative analysis. To extract the concentration profile from the spectrum, the spectrum has to be simulated and the assumed concentration profile should be changed until an acceptable agreement between the simulation and the measurement is reached. Several computer programs are available that can do the simulations and fits, e.g., ANALRA (Johnston, 1993) and SENRAS (Vizkelethy, 1990). These programs are available from the Sigmabase (see Internet Resources).

Resonant Depth Profiling

In resonant depth profiling, the spectrum is usually not recorded, but the particles or γ photons are counted in an energy window. The data that carry the depth profile information is the excitation function, i.e., the number of counts vs. the incident energy. The excitation function is the convolution of the resonance cross-section, the energy profile of the incident beam as a function of depth, and the concentration profile. Qualitative features of the concentration profile can be deduced from the excitation function. Quantitative concentration profiles can be obtained by simulating the excitation function and using least-squares fitting to match it to the measured excitation functions. [More detailed discussion can be found in Hirvonen (1995).] Also, several computer programs have been developed for profiling purposes, most recently in Smulders (1986); Lappalainen (1986), and Rickards (1991). The SPACES program (Vickridge and Amsel, 1990) was developed especially for high-resolution surface studies using very narrow resonance. This program is also available from the Sigmabase.

A simple example is shown in Figure 6. In this experiment an ^{18}O profile was measured in YBaCuO using the already mentioned ^{18}O(p, α)^{15}N reaction. From the spectrum

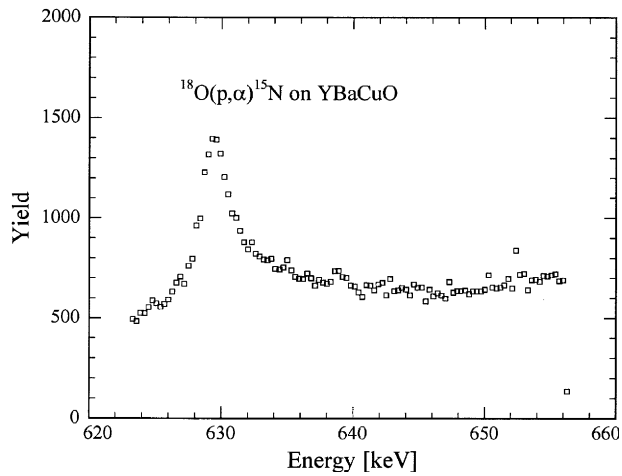


Figure 6. Excitation function of the ^{18}O(p, α)^{15}N reaction measured on an ^{18}O-enriched YBaCuO sample (Cheang-Wong, 1992).

we can deduce qualitatively that there is an ^{18}O enrichment on the surface. Using Equation 6, we can estimate the thickness of the enriched region (width of the peak divided by the stopping power of proton around the resonance energy), which is about 20 nm. Using simulation of the excitation curve and a nonlinear fit to the measured excitation function gives a 15-nm-thick layer enriched to 30% ^{18}O with 3.5% constant volume enrichment.

A recent review of the computer simulation methods in IBA can be found in Vizkelethy (1994).

SPECIMEN MODIFICATION

Although NRA and PIGE are nondestructive techniques, certain modifications to the samples can occur due to the so-called beam effect. Since a high-energy ion beam is bombarding the sample surface during the measurement, considerable heat transfer occurs. A 1-MeV ion beam with 1-$\mu A/cm^2$ current (not unusual in NRA and PIGE) means 1-W/cm^2 heat is transferred to the sample. This can be a serious problem in biological samples, and the measurement can lead to the destruction of the samples. In good heat-conducting samples, the heat does not cause a big problem, but diffusion and slight structural changes can occur. Poor heat-conducting samples will suffer local overheating that can induce considerable changes in the composition of the sample.

Apart from the heat delivered to the sample, the large measuring doses can cause considerable damage to single-crystal samples. It is especially significant in heavy-ion beams, such as ^{15}N, which is used to profile hydrogen using the $^1H(^{15}N, \alpha\gamma)^{12}C$ reaction. In this case significant hydrogen loss can occur during the analysis. Heavy ions also can cause ion beam mixing and sputtering.

To avoid these problems, the beam should be spread out over a large area on the sample if high current must be used or the current should be kept at the minimum acceptable level.

PROBLEMS

Several factors can lead to false results. Most of them can be minimized with careful design of the measurement.

A frequently encountered problem is the inability to precisely measure the current. The secondary electrons emitted from the sample can falsify the current measurement and thus measurement of the collected charge. To minimize the effect of secondary electron emission, several techniques can be used:

1. A suppression cage around the sample holder will minimize the number of escaping electrons, but it is not possible to build a perfect suppression cage. A much simpler technique is to apply positive bias to the sample, but this alternative cannot be used for every sample.
2. Isolating the scattering chamber from the rest of the beamline and allowing its use as a Faraday cup is also a good solution, but in many cases the chamber is connected electrically to other equipment that cannot be isolated.
3. There are different transmission Faraday cup designs where the current is measured not on the sample but in a separate Faraday cup that intercepts the beam in front of the sample several times per second.
4. As described above (see Practical Aspects of the Method), standards can be used to accurately measure current.

Another problem is hydrocarbon layer deposition during measurement. This can cause serious problems when narrow resonances are used for depth profiling. The incident ions will lose energy in the hydrocarbon layer that has an unknown thickness. This can lead to a false depth scale if the energy loss is significant in the hydrocarbon layer. Since the hydrocarbon layer formation is slower in higher vacuum, the best way to minimize its effect is to maintain good vacuum (see GENERAL VACCUM TECHNIQUES). Cold traps in the beamline and in the scattering chamber can significantly reduce hydrocarbon layer formation. Another way to minimize the effect of the carbon layer in case of resonance profiling (when many measurements have to be made on the same sample) is to frequently move the beam spot to a fresh area.

Interference from the elements of the matrix can be another problem. This is a potential danger in deuteron-induced reactions. Many light elements have (d, p) or (d, α) reactions, and the protons and α particles can overlap in the spectrum. Careful study of the kinematics and the choice of appropriate absorber foil thickness can solve this problem, although in certain cases the interference cannot be avoided. In PIGE, when there are interfering reactions, the use of a Ge(Li) or HPGE detector can help. The better resolution of germanium detectors allows the separation of the γ peaks from the different reactions, but at the cost of lower efficiency.

A special problem arises in single-crystal samples—accidental channeling. When the direction of the incident ions coincides with one of the main axes of the crystal, the ions are steered to these channels and the reaction yield becomes lower than it would otherwise be. (Also there is a difference between the energy loss of ions in a channel and in a random direction, which can affect the energy-to-depth conversion.) The solution is to tilt the samples to 5° to 7° with respect to the direction of the beam. (There are cases when channeling is preferred. Using NRA or PIGE combined with channeling, the lattice location of atoms can be determined.)

Electrically nonconducting samples will become charged under ion bombardment up to voltages of several kilovolts, and this can modify the beam energy. This can be especially detrimental when narrow (few hundred electron volts wide) resonances are used for depth profiling. There are several solutions to avoid surface charging: (1) A thin conducting coating (usually carbon) on the surface of the sample can reduce the charging significantly. (2) Supplying low-energy electrons from a hot filament will neutralize the accumulated positive charge on the sample surface.

A problem can arise from the careless use of the fitting programs. As in any deconvolution problem, there is usually no unique solution determined by the measured yield curve only. (One example is the fact that these deconvolutions tend to give an oscillating depth profile, which is in most cases obviously incorrect.) Extra boundary conditions and assumptions for the profile are needed to make the solution acceptable.

LITERATURE CITED

Ajzenberg-Selove, F. 1987. Energy levels of light nuclei $A = 18-20$. *Nucl. Phys.* A475:1–198.

Ajzenberg-Selove, F. 1988. Energy levels of light nuclei $A = 5-10$. *Nucl. Phys.* A490:1.

Ajzenberg-Selove, F. 1990. Energy levels of light nuclei $A = 11-12$. *Nucl. Phys.* A506:1–158.

Ajzenberg-Selove, F. 1991. Energy levels of light nuclei $A = 13-15$. *Nucl. Phys.* A523:1–196.

Amsel, G., d'Artemare, E., and Girard., E. 1983. A simple, digitally controlled, automatic, hysteresis free, high precision energy scanning system for van de Graaff type accelerators. *Nucl. Instrum. Methods.* 205:5–26.

Amsel, G. and Davies, J. A. 1983. Precision standard reference targets for microanalysis with nuclear reactions. *Nucl. Instrum. Methods* 218:177–182.

Amsel, G., Pászti, F., Szilágyi, E., and Gyulai, J. 1992. p, d, and α particle discrimination in NRA: Thin, adjustable sensitive zone semiconductor detectors revisited. *Nucl. Instrum. Methods* B63:421–433.

Amsel, G. and Samuel, D. 1967. Microanalysis of the stable isotopes of oxygen by means of nuclear reactions. *Anal. Chem.* 39:1689–1698.

Bird, J. R., Campbell, B. L., and Price, P. B. 1974. Prompt nuclear analysis. *Atomic Energy Rev.* 12:275–342.

Blatt, J. M. and Weisskopf, V. 1994. Theoretical Nuclear Physics. Dover Publications, New York.

Chaturvedi, U. K., Steiner, U., Zak, O., Krausch, G., Shatz, G., and Klein, L. 1990. Structure at polymer interfaces determined by high-resolution nuclear reaction analysis. *Appl. Phys. Lett.* 56:1228–1230.

Cheang-Wong, J. C., Ortega, C., Siejka, J., Trimaille, I., Sacuto, A., Balkanski, M., and Vizkelethy, G. 1992. RBS analysis of thin amorphous YBaCuO films: Comparison with direct determination of oxygen contents by NRA. *Nucl. Instrum. Methods* B64:169–173.

Damjantschitsch, H., Weiser, G., Heusser, G., Kalbitzer, S., and Mannsperger, H. 1983. An in-beam-line low-level system for nuclear reaction γ-rays. *Nucl. Instrum. Methods* 218:129–140.

Davies, J. A., Lennard, W. N., and Mitchell, I. V. 1995. Pitfalls in ion beam analysis. *In* Handbook of Modern Ion Beam Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 343–363. Materials Research Society, Pittsburgh, Pa.

Endt, P. M. and van der Leun, C. 1978. Energy levels of $A = 21-44$ nuclei. *Nucl. Phys.* A310:1–752.

Foster, L., Vizkelethy, G., Lee, M., Tesmer, J. R., and Nastasi, M. 1995. Particle-particle nuclear reaction cross sections. *In* Handbook of Modern Ion Beam Materials Analysis (J.R. Tesmer and M. Nastasi, eds.). pp. 549–572. Materials Research Society, Pittsburgh, Pa.

Green, P. F. and Doyle, B. L. 1986. Silicon elastic recoil detection studies of polymer diffusion: Advantages and disadvantages. *Nucl. Instrum. Methods* B18:64–70.

Hirvonen, J.-P. 1995. Nuclear reaction analysis: Particle-gamma reactions. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.) pp. 167–192. Materials Research Society, Pittsburgh, Pa.

Hirvonen, J.-P. and Lappalainen, R. 1995. Particle-gamma data. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 573–613. Materials Research Society, Pittsburgh, Pa.

Horn, K. M. and Lanford, W. A. 1990. Suppression of background radiation in BGO and NaI detectors used in nuclear reaction analysis. *Nucl. Instrum. Methods* B45:256–259.

Jarjis, R. A. 1979. Internal Report, University of Manchaster, U.K.

Johnston, P. N. 1993. ANALRA—charged particle nuclear analysis software for the IBM PC. *Nucl. Instrum. Methods* B79:506–508.

Kuhn, D., Rauch, F., and Baumann, H. 1990. A low-background detection system using a BGO detector for sensitive hydrogen analysis with the $^{1}H(^{15}N,\alpha\gamma)^{12}C$ reaction. *Nucl. Instrum. Methods* B45:252–255.

Lappalainen, R. 1986. Application of the NRB method in range, diffusion, and lifetime measurements, Ph.D. Thesis, University of Helsinki, Finland.

Maurel, B. 1980. Stochastic Theory of Fast Charged Particle Energy Loss. Application to Resonance Yield Curves and Depth Profiling. Ph.D. Thesis, University of Paris, France.

Maurel, B., Amsel, G., and Nadai, J. P. 1982. Depth profiling with narrow resonances of nuclear reactions: Theory and experimental use. *Nucl. Instrum. Methods* 197:1–14.

Meier, J. H. and Richter, F. W. 1990. A useful device for scanning the beam energy of a van de Graaff accelerator. *Nucl. Instrum. Methods* B47:303–306.

Möller, W. 1978. Background reduction in $D(^{3}He, \alpha)H$ depth profiling experiments using a simple electrostatic deflector. *Nucl. Instrum. Methods* 157:223–227.

Möller, W., Hufschmidt, M., and Kamke, D. 1977. Large depth profile measurements of D, ^{3}He, and ^{6}Li by deuterium induced nuclear reactions. *Nucl. Instrum. Methods* 140:157–165.

Rickards, J. 1991. Fluorine studies with a small accelerator. *Nucl. Instrum. Methods.* B56/57:812–815.

Simpson, J. C. B. and Earwaker, L. G. 1984. A computer simulation of nuclear reaction spectra with applications in analysis and depth profiling of light elements. *Vacuum* 34:899–902.

Smulders, P. J. M. 1986. A deconvolution technique with smooth, non-negative results. *Nucl. Instrum. Methods* B14:234–239.

Tilley, D. R., Weller, H. R., and Hassian, H. H. 1987. Energy levels of light nuclei $A = 3$. *Nucl. Phys.* A474:1–60.

Tilley, D. R., Weller, H. R., and Cheves, C. M. 1993. Energy levels of light nuclei $A = 16-17$. *Nucl. Phys.* A564:1–183.

Tilley, D. R., Weller, H. R., Cheves, C. M., and Chesteler, R. M. 1995. Energy levels of light nuclei $A = 18-19$. *Nucl. Phys.* A595:1–170.

Vickridge, I. C. 1990. Stochastic Theory of Fast Ion Energy Loss and Its Application to Depth Profiling Using Narrow Nuclear Resonances. Applications in Stable Isotope Tracing Experiment for Materials Science. PhD Thesis, University of Paris, France.

Vickridge, I. C. and Amsel, G. 1990. SPACES: A PC implementation of the stochastic theory of energy loss for narrow resonance profiling. *Nucl. Instrum. Methods* B45:6–11.

Vizkelethy, G. 1990. Simulation and evaluation of nuclear reaction spectra. *Nucl. Instrum. Methods* B45:1–5.

Vizkelethy, G. 1994. Computer simulation of ion beam methods in analysis of thin films. *Nucl. Instrum. Methods* B89:122–130.

Vizkelethy, G. 1995. Nuclear reaction analysis: Particle-particle reactions. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 139–165. Materials Research Society, Pittsburgh, Pa.

KEY REFERENCES

Amsel, G. and Lanford, W. A. 1984. Nuclear reaction technique in materials analysis. *Ann. Rev. Nucl. Part. Sci.* 34:435–460.

Excellent review paper of RNA.

Deconnick, G. 1978 Introduction to Radioanalytical Chemistry. Elsevier, Amsterdam.

Excellent reviews of the methods.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. North-Holland Publishing, New York. See chapter on "Nuclear techniques: Activation analysis and prompt radiation analysis," pp. 283–310.

Good discussion of the fundamental physics involved in NRA.

Hirvonen, 1995. See above.

Detailed discussion of PIGE with several worked-out examples.

Lanford, W. A. 1995. Nuclear reactions for hydrogen analysis. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 193–204. Materials Research Society, Pittsburgh, Pa.

Detailed discussion of hydrogen detection with NRA.

Peaisach, M. 1992. Nuclear reaction analysis. *In* Elemental Analysis by Particle Accelerators (Z. B. Alfassi and M. Peisach, eds.). pp. 351–383. CRC Press, Boca Raton, Fla.

General discussion of NRA with valuable references.

Vizkelethy, 1995. See above.

Detailed discussion of NRA with worked-out examples and discussion of useful reactions.

INTERNET RESOURCES

Lund Nuclear Data Service, *http://nucleardata.nuclear.lu.se/nucleardata*

National Nuclear Data Center, *http://www.nndc.bnl.gov/*

Sigmabase, *http://ibaserver.physics.isu.edu/sigmabase*

T-2 Nuclear Information Service, *http://t2.lanl.gov/*

Triangle Universities Nuclear Laboratory, *http://www.tunl.duke.edu/NuclData*

APPENDIX

Energy Relations in Nuclear Reactions

In a charged particle reaction, the energies of the emitted particles in the laboratory coordinate system are (see Fig. 7):

$$E_{\text{light}} = B\left(\cos\psi \pm \sqrt{\frac{D}{B} - \sin^2\psi}\right)^2 (E_0 + Q) \qquad (7)$$

$$E_{\text{heavy}} = A\left(\cos\varphi \pm \sqrt{\frac{C}{A} - \sin^2\varphi}\right)^2 (E_0 + Q) \qquad (8)$$

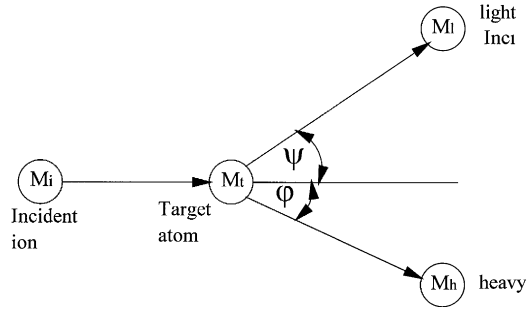


Figure 7. Reaction kinematics.

where E_0 is the energy of the incident ions; Q is the Q value of the reaction; and A, B, C, and D are given as

$$\begin{aligned} A &= \frac{M_1 M_4}{(M_1+M_2)(M_3+M_4)} \frac{E_0}{E_0+Q} \\ B &= \frac{M_1 M_3}{(M_1+M_2)(M_3+M_4)} \frac{E_0}{E_0+Q} \\ C &= \frac{M_2 M_3}{(M_1+M_2)(M_3+M_4)} \left(1 + \frac{M_1 Q}{M_2(E_0+Q)}\right) \\ D &= \frac{M_2 M_4}{(M_1+M_2)(M_3+M_4)} \left(1 + \frac{M_1 Q}{M_2(E_0+Q)}\right) \end{aligned} \qquad (9)$$

where M_1, M_2, M_3, and M_4 are the masses of the incident ion, the target nucleus, the lighter reaction product, and the heavier product, respectively.

In Equation 7 only plus signs are used unless $B > D$, in which case $\psi_{\max} = \sin^{-1}\sqrt{D/B}$, and in Equation 8 only plus signs are used unless $A > C$, in which case $\varphi_{\max} = \sin^{-1}\sqrt{C/A}$. The emitted particle then loses energy along its path until it leaves the target. The energy loss of ions inward and outward is described by Equations 8 and 9.

Energy Spread (Depth Resolution)

Equations 2 and 3 are true only on the average. Since the energy loss is a stochastic process, the ions, after passing a certain thickness, will have an energy distribution rather than a sharp energy. This phenomenon is called straggling. There are several models to describe the energy straggling; in most cases, the simplest one, Bohr straggling, is satisfactory. [One exception is the case of very narrow resonances, which is described in details in Maurel (1980) and Vickridge (1990).]

There are other factors that cause broadening of the energy in the detected spectrum. The following contribute to the energy spread of detected particles in NRA: (1) initial energy spread of the incident beam, (2) straggling of the incident ions, (3) multiple scattering, (4) geometric spread due to the finite acceptance angle of the detector and to the finite beam size, (5) straggling of the outgoing particles, (6) energy resolution of the detector, and (6) energy straggling in the absorber foil.

A detailed treatment of most of these factors can be found in Vizkelethy (1994). In PIGE and resonant depth profiling with NRA, only a few contribute to the depth resolution. Since the energy of the detected particles or γ

rays is not used to extract concentration profile information, only the initial energy spread, the straggling of the incident beam, and the resonance width determine the depth resolution. (The contribution from the multiple scattering is negligible since then normal incidence is used.)

Calculation of the Excitation Function for Resonance Depth Profiling

The excitation function at an E_0 incident energy is

$$Y(E_0) \propto \sigma(E_0) * h(E_0) * \int_0^\infty c(x) g(E_0; x)\, dx \qquad (10)$$

where $\sigma(E_0)$ is the reaction cross-section, $h(E_0)$ is the initial energy distribution of the ion beam, $c(x)$ is the concentration profile of the atoms to be detected, $g(E; x)$ is the probability that an ion loses E energy when penetrating a thickness x, and the asterisk denotes convolution. [This convolution is a very complicated calculation; for details see Maurel (1980) and Vickridge (1990).] When the resonance is not extremely narrow, the initial energy spread of the ions is not too small, and the profile is not measured very close to the surface (the energy loss is large enough that the straggling can be considered Gaussian); the function $g(E; x)$ can then be approximated with a Gaussian:

$$g(E; x) = \frac{1}{\sqrt{2\pi s^2(x)}} e^{-[E-E(x)]^2/2s^2(x)} \qquad (11)$$

where $E(x)$ is the same as $E_{\text{in}}(x)$ in Equation 3 and $s(x)$ is the root-mean-square energy straggling at x. If the resonance cross-section is given by a Breit-Wigner resonance (Blatt and Weiskopf, 1994),

$$\sigma(E_0) = \sigma_0 \frac{\Gamma^2/4}{\Gamma^2/4 + (E - E_{\text{R}})^2} \qquad (12)$$

where σ_0 is the strength, Γ is the width, and E_{R} is the energy of the resonance, then the straggling, the Doppler-broadening, and the initial energy spread can be combined into a single Gaussian with depth-dependent width. Evaluating Equation 10, the excitation function is

$$Y(E_0) \propto \int_0^\infty c(x)\sigma_0 \frac{\Gamma}{\sqrt{8}S(x)} \text{Re}\, w\left(\frac{E_0 - E_R}{\sqrt{2}S(x)} - i\frac{\Gamma}{\sqrt{8}S(x)}\right) dx \qquad (13)$$

where $\text{Re}\, w(z)$ is the real part of the complex error function and $S(x)$ the width of the combined Gaussian.

GYÖRGY VIZKELETHY
Idaho State University
Pocatello, Idaho

PARTICLE-INDUCED X-RAY EMISSION

INTRODUCTION

Particle-induced x-ray emission (PIXE) is an elemental analysis technique that employs a mega-electron-volt energy beam of charged particles from a small electrostatic accelerator to induce characteristic x-ray emission from the inner shells of atoms in the specimen. The accelerator can be single ended or tandem. Most PIXE work is done with 2- to 4-MeV proton beams containing currents of a few nanoamperes within a beam whose diameter is typically a few millimeters; a small amount of work with deuteron and helium beams has been reported, and use of heavier ion beams has been explored. The emitted x rays are nearly always detected in the energy-dispersive mode using a Si(Li) spectrometer. Although wavelength-dispersive x-ray spectrometry would provide superior energy resolution, its low geometric efficiency is a significant disadvantage for PIXE, in which potential specimen damage by heating limits the beam current and therefore the emitted x-ray intensity. In the proton microprobe, magnetic or electrostatic quadrupole lenses are employed to focus the beam to micrometer spot size; this makes it possible to perform micro-PIXE analysis of very small features and, also, by sweeping this microbeam along a preselected line or over an area of the specimen, to determine element distributions in one or two dimensions in a fully quantitative manner. Most proton microprobes realize a beam spot size smaller than a few micrometers at a beam current of at least 0.1 nA. In trace element analysis or imaging by micro-PIXE, the high efficiency of a Si(Li) detector, placed as close to the specimen as possible, is again mandatory for x-ray collection. To date, PIXE and micro-PIXE have most often been used to conduct trace element analysis in specimens whose major element composition is already known or has been measured by other means, such as electron microprobe analysis. However, both variants are capable of providing major element analysis for elements of atomic number down to $Z = 11$.

Competing Methods

In the absence of beam focusing, conventional PIXE is an alternative to x-ray fluorescence analysis (XRFA; see X-RAY MICROPROBE FOR FLUORESCENCE AND DIFFRACTION ANALYSIS) but requires more complex equipment. Micro-PIXE is a powerful and seamless complement to electron microprobe analysis (EPMA), a longer established microbeam x-ray emission method based on excitation by kilo-electron-volt electron beams; detection limits of a few hundred parts per million in EPMA compare with limits of a few parts per million in micro-PIXE. While the electron microprobe is ubiquitous, there are only about 50 proton microprobes around the world, but many of these can provide beam time on short notice. Conducted with the highly polarized x radiation from a synchroton storage ring equipped with insertion devices such as wigglers and undulators, XRFA now provides a strong competitor to micro-PIXE in terms of both spatial resolution (a few micrometers) and detection limits (sub−parts per million). The x-ray spectrum from an undulator is highly collimated, extremely intense, and nearly monochromatic; the exciting x-ray energy can be tuned to lie just above the absorption edge energy of the element of greatest interest, where the photoabsorption cross-section is highest, thereby optimizing the detection limit. However, there are only a handful of synchrotron facilities, and beam time is at a premium. A

much cheaper method of obtaining an x-ray microbeam is through the use of a conventional fine-focus x-ray tube and a nonimaging optical system based upon focusing capillaries (Rindby, 1993). Total-reflection XRFA (Klockenkamper et al., 1992) offers similar detection limits, but this method is restricted to very thin films deposited on a totally reflecting substrate and lacks the versatility of PIXE and XRFA as regards different specimen types. One of PIXEs merits relative to XRFA is the possibility of deploying other ion beam analysis techniques simultaneously or sequentially; these include Rutherford backscattering spectroscopy (see MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY and HEAVY-ION BACKSCATTERING SPECTROMETRY), nuclear reaction analysis (NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION), ionoluminescence, and scanning transmission ion microscopy (STIM).

Optical spectroscopy methods based on emission (AES—atomic emission spectroscopy) and absorption (AAS—atomic absorption spectroscopy) are the workhorses of trace element analysis in materials that can be reduced to liquid form or dissolved in a liquid and then atomized. The inductively coupled plasma (ICP) has become the method of choice for inducing light emission, and ICP-AES attains detection limits of 1 to 30 ng/mL of liquid; the limits referred to the original material depend on the dilution factor and typically reach down to 0.1 ppm. Interference effects among optical lines are more complex than is the case in PIXE with x-ray lines. Use of mass spectrometry (ICP-MS) reduces interferences, provides a more complete elemental coverage than ICP-AES, and has much less element-to-element variation in sensitivity than does ICP-AES; its matrix effects are less straightforward to handle than those of PIXE. Overall, PIXE is more versatile as regards specimen type and preparation than ICP-AES and ICP-MS and has the advantages of straightforward matrix correction, smoothly varying sensitivity with atomic number, and high accuracy. But for conventional bulk analysis, the highly developed optical and mass spectrometry methods will frequently be more accessible and entirely satisfactory and provide excellent detection limit without the need for an accelerator. For particular specimen types such as aerosol particulates and awkwardly shaped specimens (e.g., in archaeology) requiring nondestructive analysis, PIXE has unique advantages. The ability to handle awkwardly shaped specimens by extracting the proton beam through a thin window into the laboratory milieu is particularly valuable in archaeometric applications.

The nondestructive nature, high spatial resolution, parts per million detection limits, and high accuracy of micro-PIXE together with the simultaneous deployment of STIM and backscattering make it very competitive with other microprobe methods. It has been applied intensively in the analysis of individual mineral grains and in zoning phenomena within these, individual fly-ash particles, single biological cells, and thin tissue slices containing features such as Alzheimer's plaques (Johansson et al., 1995). Secondary ion mass spectrometry (SIMS) often provides better detection limits but is destructive. The combination of laser ablation microprobe with inductively coupled plasma (ICP) excitation and mass spectrometry offers resolution of some tens of micrometers and detection limits as low as 0.5 ppm; however, the ablation basis of this method renders matrix effects and standardization more complex than in micro-PIXE.

Detailed accounts of the fundamentals and of many applications can be found in two recent books on PIXE (Johansson and Campbell, 1988; Johansson et al., 1995), the first of which provides considerable historical perspective. The present offering is an overview, including some typical applications. The proceedings of the triennial international conferences on PIXE provide both an excellent historical picture and an account of recent developments and new applications.

PRINCIPLES OF THE METHOD

The most general relationship between element concentrations in a specimen and the intensity of detected x rays of each element is derived for the simple case of K x rays (see Appendix A). The L and M x rays may be dealt with in similar manner, with the added complexity that in these cases ionization occurs in three and five subshells, respectively, and there is the probability of vacancies being transferred by the Coster-Kronig effect to higher subshells prior to the x-ray emission occurring. In principle, Equation 6 could be used to perform standardless analysis; this would involve a complete reliance on the underlying theory and the database and on the assumed properties of the x-ray detector. At the other extreme, by employing standards that very closely mimic the major element (matrix) composition of a specimen, the PIXE analyst working on trace elements could completely avoid dependence upon theory and rely only upon calibration curves relating x-ray yield to concentration. In practice, there is much practical merit in adopting an intermediate stance. We recommend a fundamental parameter approach using the well-understood theories of the interaction of charged particle beams and photon beams with matter and of characteristic x-ray emission from inner shell vacancies but also relying on a small set of standards that need bear only limited resemblance to the specimens at hand.

In this approach the equation used to derive element concentrations from measured x-ray intensities is

$$Y(Z) = Y_1(Z) H \varepsilon_Z t_Z C_Z Q \qquad (1)$$

where $Y(Z)$ is the measured intensity of x rays in the principal line of element Z (concentration C_Z), $Y_1(Z)$ is the theoretically computed intensity per unit beam charge per unit detector solid angle per unit concentration, H is an instrumental constant that subsumes the detector solid angle and any calibration factor required to convert an indirect measurement of beam charge to units of microcoulombs, Q is the direct or indirect measure of beam charge, t_Z is the x-ray transmission fraction through any absorbers (see Practical Aspects of the Method) that are deliberately interposed between specimen and detector, and ε_Z is the detector's intrinsic efficiency. It is also straightforward to

calculate the contribution of secondary x rays fluoresced when proton-induced x rays are absorbed in the specimen and to correct for these. The various atomic physics quantities required are obtained from databases that have been developed by fitting appropriately parameterized expressions (in part based on theory and in part semiempirical) to either compilations of experimental data or theoretically generated values. The H value is determined via Equation 1 by use of appropriate standards or standard reference materials. It is implicit in the direct use of Equation 1 that the major elements have already been identified and that their concentrations are known *a priori*. This permits computation of the integrals that involve matrix effects (slowing down of the protons and attenuation of the x rays); Equation 1 then provides the trace element concentrations. However, if the x-ray lines of the major elements are observed in the spectrum, Equation 1 may be solved in an iterative manner to provide both major and trace element concentrations.

The method is fully quantitative, and its accuracy is determined in part by the accuracy of the database. The theoretically computed x-ray intensity uses ionization cross-sections and stopping powers for protons, x-ray mass attenuation coefficients, x-ray fluorescence yields and Coster-Kronig probabilities for the various subshells, and relative intensities of the lines within each x-ray series (e.g., K, L_1, ...) emitted by a given element. For ionization cross-sections, the choice is among the hydrogenic model calculations of Cohen and Harrigan (1985, 1989) or of Liu and Cipolla (1996) based on the so-called ECPSSR model of Brandt and Lapicki (1981), the equivalent but more sophisticated ECPSSR treatment of Chen and Crasemann (1985, 1989) based on self-consistent field wave functions, or experimental cross-section compilations such as those of Paul and Sacher (1989) for the K shell and Orlic et al. (1994) for the L subshells. Proton stopping powers (i.e., rate of energy loss as a function of distance traveled in an element) are usually taken from one of the various compilations of Ziegler and his colleagues, which are summarized in a report by the International Commission on Radiation Units and Measurements (1993). A variety of mass attenuation coefficient schemes ranging from theoretical to semiempirical have been used, and the XCOM database provided by the National Institute of Science and Technology (NIST; Berger and Hubbell, 1987) appears to us to be admirably suited. The relative x-ray intensities within the K and L series are invariably taken from the calculations of Scofield (1974a,b, 1975) and from fits to these by Campbell and Wang (1989), although corrections must be made for the K x rays of the elements $21 < Z < 30$ where configuration mixing effects arising from the open $3d$ subshell cause divergences; relative intensities for the M series are given by Chen and Crasemann (1984). Atomic fluorescence and Coster-Kronig probabilities may be taken from the theoretical calculations of Chen et al. (1980a,b, 1981, 1983) or from critical assessments of experimental data such as those of Krause (1979) and Bambynek (see Hubbell, 1989). Campbell and Cookson (1983) have assessed how the various uncertainties in all these quantities are transmitted into accuracy estimates for PIXE, but the best route for assessing overall accuracy remains the analysis of known reference materials.

The intrinsic efficiency of the Si(Li) x-ray detector enters Equation 1 and so must be known. This efficiency is essentially unity in the x-ray energy region between 5 and 15 keV. At higher energies ε_Z falls off because of penetration through the silicon crystal (typically 3 to 5 mm thick). At lower energies, ε_Z falls off due to attenuation in the beryllium or polymer vacuum window, the metal contact, and internal detector effects such as the escape of the primary photoelectrons and Auger electrons created in x-ray interactions with silicon atoms and loss of the secondary ionization electrons due to trapping and diffusion. The related issues of detector efficiency and resolution function (lineshape) are dealt with below (see Appendix B).

A basic assumption of the method is that the element distribution in the volume analyzed is homogeneous. The presence of subsurface inclusions, for example, would negate this assumption. Of course, micro-PIXE can be used to probe optically visible inhomogeneities such as zoning and grain boundary phenomena in minerals and to search for optically invisible inhomogeneities in, for example, air particulate deposits from cascade impactors. It can then be used to select homogeneous regions for analysis. Given (1) homogeneity and (2) knowledge of the major element concentrations, PIXE or micro-PIXE analysis is fully quantitative, with the accuracy and the detection limits determined by the type of specimen.

A great deal of PIXE work is conducted on specimens that are so thin that the proton energy is scarcely altered in traversing the specimen and there is negligible attenuation of x rays. Films of fine particulates collected from the ambient air by sampling devices are an example. In such cases, a similar approach may be taken with concentration C_Z replaced in Equation 1 by the areal density of elements in the specimen.

The main engineering advance responsible for PIXE was the advent of the Si(Li) x-ray detector in the late 1960s. Micro-PIXE was made possible by the development of high-precision magnetic quadrupole focusing lenses in the 1970s. The acceptance of PIXE and other accelerator-based ion beam techniques stimulated advances in accelerator design that have resulted in a new generation of compact, highly stable electrostatic accelerators provided primarily for ion beam analysis.

PRACTICAL ASPECTS OF THE METHOD

Most PIXE work is done with nanoampere currents of 2- to 4-MeV protons transmitted in a vacuum beam line to the specimen. However, some laboratories extract the beam into the laboratory through a thin Kapton window to deal with unwieldy or easily damaged specimens such as manuscripts or objects of art (Doyle et al., 1991).

As with EPMA, most of the technical work for the user lies in specimen preparation (see Sample Preparation). Conduct of PIXE analysis requires an expert accelerator operator to steer and focus the beam onto the specimen.

As indicated above (see Principles of the Method), the intrinsic efficiency of the Si(Li) detector as a function of

x-ray energy is an important practical aspect. This function may be determined from, e.g., the manufacturer's data on crystal thickness and the contact and window thickness. More accurate determinations of these quantities may be effected via methods outlined below (see Appendix B). The lineshape or resolution function of the detector is required (see Data Analysis and Initial Interpretation) for the interpretation of complex PIXE spectra with their many overlapping lines. Guidance in this direction is also provided below (see Appendix B).

Specimens are generally classified as thin or thick. Thin specimens are those that cause negligible reduction in the energy of the protons, and they are usually deposited on a substrate of trace-element-free polymer film; examples include microtome slices of tissue, residue from dried drops of fluid containing suspended or dissolved solids, or films of atmospheric particulate material collected by a sampling device. Thick specimens are defined as those having sufficient areal density to stop the beam within the specimen, and they may be self-supporting or may reside on a substrate. As indicated below (see Appendix A), their analysis requires full attention to the effect of the matrix in slowing down the protons and in attenuating the excited characteristic x rays. The x-ray production decreases rapidly with depth but is significant for some 10 to 30 μm. Examples of thick specimens include mineral grains, geological thin sections, metallurgical specimens, and archaeological artefacts such as jewelry, bronzes, and pottery. Specimens between these limiting cases are referred to as having intermediate thickness, and their thickness must be known if matrix effects are to be accurately handled. In early PIXE work, many intermediate specimens were treated as if they were thin; i.e., matrix effects were neglected. However, it is now customary to correct for these, using, for example, an auxiliary proton scattering or transmission measurement to determine specimen thickness and major element composition.

With specimens that stop the beam, current and charge measurements are usually accomplished very simply by having the specimen electrically insulated from the chamber and by connecting it to a charge integrator. If thick specimens are not conducting, they must be coated with a very thin layer of carbon to prevent charging; neglecting this results in periodic spark discharge, which in turn causes intense bremsstrahlung background in the x-ray spectrum. Secondary electrons are emitted from the specimen, potentially causing an error of up to 50% in beam charge determination. These electrons must be returned to the specimen by placing a suitable electrode or grid close by, at a negative potential of typically 100 V. In thin specimens, the beam is transmitted into a graphite-lined Faraday cup, also with an electron suppressor, and the charge is integrated.

Electronic dead time effects must be accounted for. After detection of an event, there is a finite processing time in the electronic system, during which any subsequent events will be lost. A dead time signal from the pulse processor may be used to gate off the charge integrator and effect the necessary correction. Alternatively, an on-demand beam deflection system may be used. Here, the proton beam passes between two plates situated about 1 m upstream of the specimen and carrying equal voltages; detection of an x ray triggers a unit that grounds one plate as rapidly as possible, thereby deflecting the beam onto a tantalum collimator. The beam is restored when electronic processing is complete, so no corrections for dead time are required. This approach has two further advantages. The first is that specimen heating effects are reduced by removing the beam when its presence is not required. The second is that pile-up of closely spaced x-ray events is substantially decreased, thereby removing undesired artefacts from the spectrum.

The finite counting rate capacity of a Si(Li) detector and its associated pulse processor and multichannel pulse height analyzer demand that unnecessary contributions in the x-ray spectrum be minimized and as much of the capacity as possible be used for the trace element x rays of interest. Most thin-specimen analyses in the atmospheric or biological science context are performed with a Mylar absorber of typically 100 to 200 μm thick employed to reduce the intense bremsstrahlung background that lies below 5 keV in the spectrum and whose intensity increases rapidly toward lower energy. With thick metallurgical, geological, or archaeological specimens, x-ray lines of major elements dominate the spectrum, making it necessary to reduce their intensities with an aluminum absorber of thickness typically 100 to 1000 μm. For example, in silicate minerals, the *K* x rays of the major elements Na, Mg, Al, and Si occupy most of the counting rate capability; an aluminum foil of thickness 100 μm reduces the intensity of these x rays to a level such that they are still present at useful intensity in the spectrum, but the x rays from trace elements of higher atomic number can now be seen, with detection limits approaching the 1-ppm level. The thickness of such filters must be accurately determined.

As a thin-specimen example, Figure 1 shows the PIXE spectrum of an air particulate standard reference material. Proton-induced x-ray emission has proven extremely powerful in aerosol analysis, and special samplers have been devised to match its capabilities (Cahill, 1995). The low detection limits facilitate urban studies with short sampling intervals in order to determine pollution patterns over a day. At the other extreme, in remote locations, robust samplers of the IMPROVE type (Eldred et al., 1990) run twice weekly for 24-h periods at sites around the world in a project aimed at assembling continental and global data on visibility and particulate composition (Eldred and Cahill, 1994). The low detection limits match PIXE perfectly with cascade impactors, where the sample is divided according to particle diameter. Figure 2 (Maenhaut et al., 1996) presents measured detection limits for two cascade impactors that are well suited to PIXE in that they create on each impaction stage a deposit of small diameter that may be totally enveloped by a proton beam of a few millimeters in diameter. Now, PIXE appears in about half of all analyses of atmospheric aerosols in the standard aerosol journals.

Figure 3 shows an example spectrum from an "almost thin" specimen where an auxiliary technique is used to determine thickness and major element composition so that matrix corrections may be applied in the data reduction. The specimen is a slice of plant tissue, and the

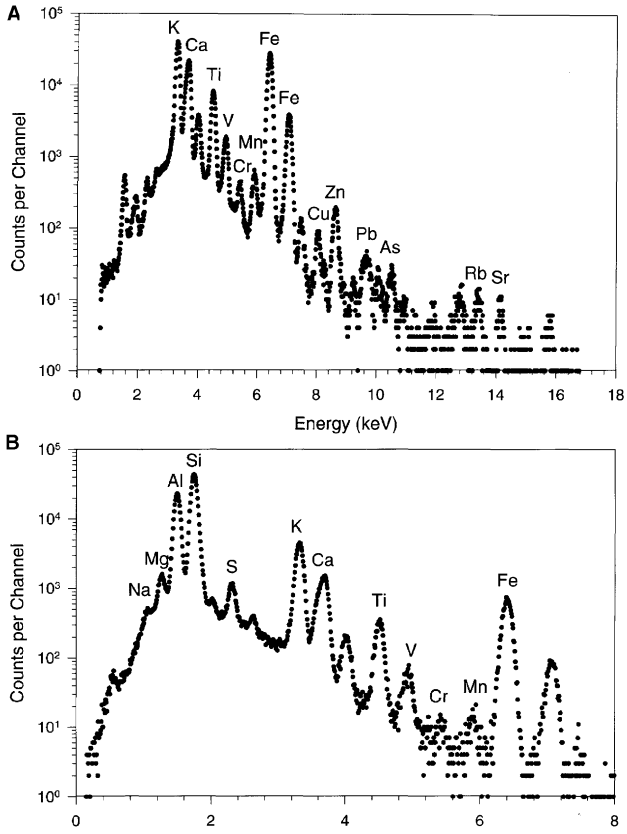


Figure 1. PIXE spectra of the BCR-128 fly ash reference standard, measured at Guelph with two x-ray detectors: **(A)** 8-μm window, no absorber; **(B)** 25-μm window, 125-μm Mylar absorber. Proton energy was 2.5 MeV.

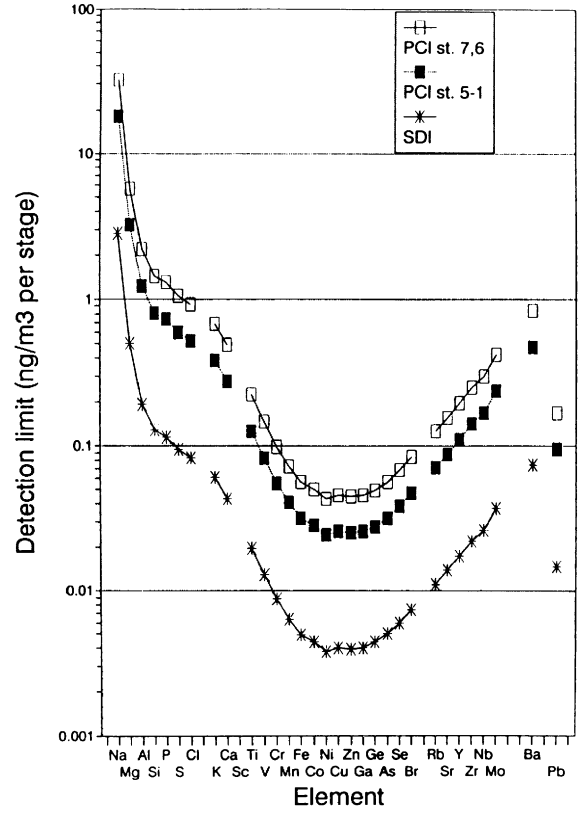


Figure 2. Detection limits for the PIXE International cascade impactor (PCI) and a small-deposit low-pressure impactor (SDI). Note that the detection limit depends upon the impactor stage. From Maenhaut et al. (1996) with permission of Elsevier Science.

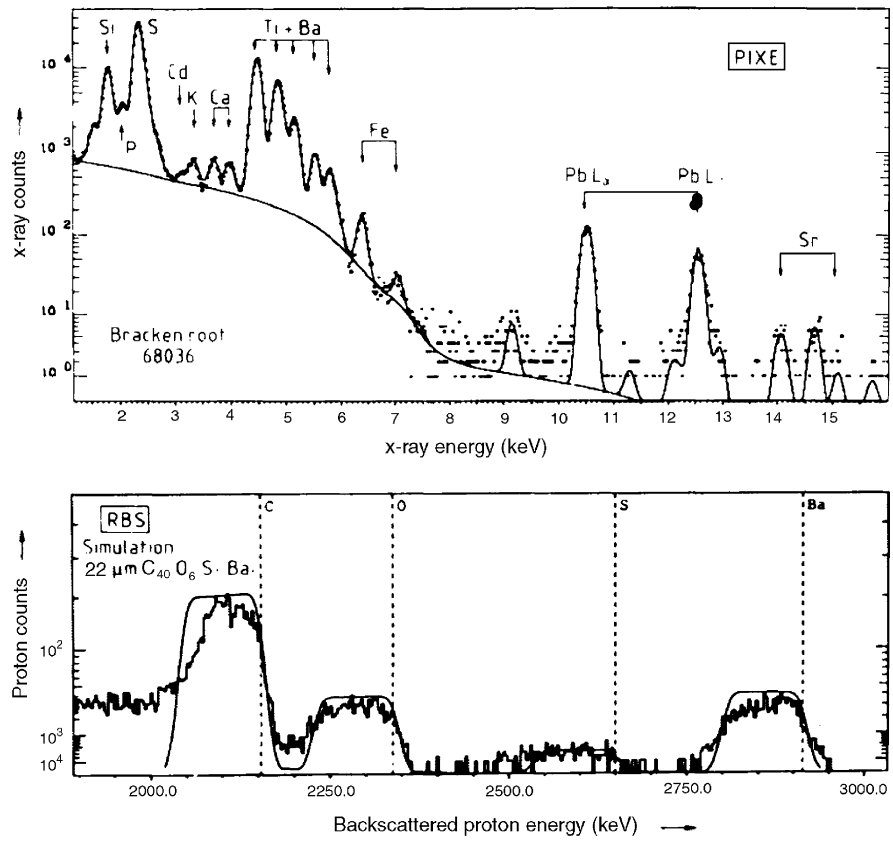


Figure 3. PIXE and proton backscattering energy spectra from a focal deposit of heavy metals in a plant root (Watt et al., 1991). Proton energy was 3 MeV. The continuous curves represent a fit (PIXE case) and a simulation (BS case). Reproduced by permission of Elsevier Science.

auxiliary technique is proton backscattering, whose results, acquired simultaneously, are in the lower panel of the figure. Many PIXE systems employ an annular silicon surface barrier detector upstream from the specimen and situated on the beam axis so that the beam passes through the annulus. This affords a detection angle close to 180° for scattered particles, which results in optimum energy resolution. This combination has proved to be a powerful means to determine major element concentrations in a large range of biological applications. Another powerful technique ancillary to micro-PIXE in the analysis of thin tissue slices is STIM. Here, a charged particle detector situated downstream from the specimen at about 15° to 25° to the beam axis records the spectrum of forward-scattered particles that reflects energy losses from both nuclear scattering and ion-electron collisions. This technique is effective in identifying structures such as individual cultured cells on a substrate, individual aerosol particles, and lesions in medical tissue. It has been used to identify neuritic plaques in brain tissue from Alzheimer's disease patients (Landsberg et al., 1992), thereby eliminating the need for immunohistochemical staining and so removing a sample preparation step that has the potential for contamination. Simultaneous recording of STIM, backscatter, and PIXE spectra and two-dimensional images has become a standard approach with biological specimens whose structure remains unaffected by preparation.

Micro-PIXE elegantly complements electron probe microanalysis in the trace element analysis of mineral grains using the same specimens and extending detection limits down to the parts per million level (Campbell and Czamanske, 1998). One major niche is the study of precious and other metals in sulfide ores. Figure 4 shows spectra of three iron-nickel sulfides (pentlandites) from a Siberian ore deposit; the large dynamic range between major elements such as iron (over 30% concentration in this case) and trace elements at tens to hundreds of parts per million concentration is typical of the earth science application area, and it necessitates use of absorbing filters (several hundred micrometers of aluminum) to depress the intensities of the intense x rays of the lighter major elements. Another important area is the study of trace elements in silicate minerals, where PIXE has aided in the development of geothermometers and fingerprinting approaches that are proving immensely powerful in, for example, assessing kimberlites for diamond potential (Griffin and Ryan, 1995).

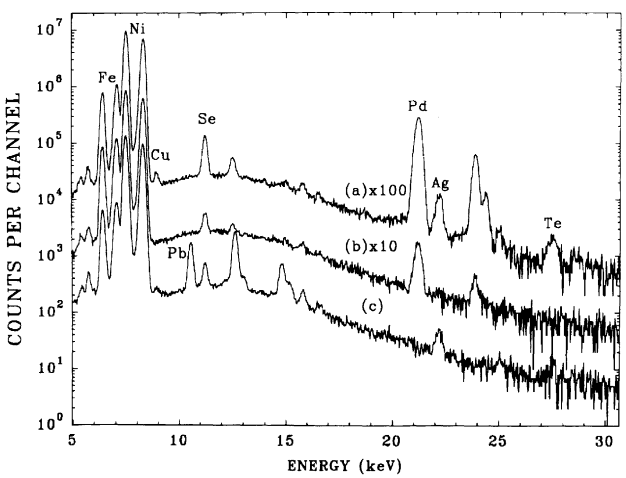


Figure 4. PIXE spectra of three Siberian pentlandites (Czamanske et al., 1992). The spectra were recorded using 3-MeV protons and an aluminum filter of 352 mm thickness. Element concentrations (in ppm) are (*a*) Se 261, Pd 2540, Ag 112, Te 54; (*b*) Se 80, Pd 132; (*c*) Se 116, Pd< 5, Ag 34, Te 36, Pb 1416.

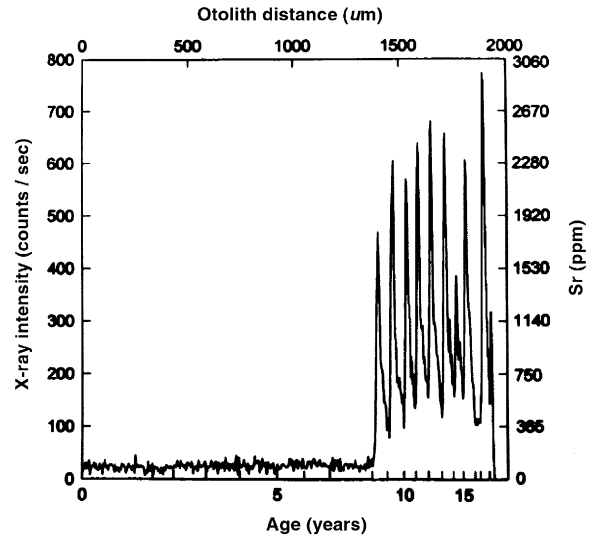


Figure 5. Strontium concentration profile in the growth zones of an otolith from an Arctic char collected from the Jayco River in the Northwest Territories of Canada (Babaluk et al., 1997). A 30-min micro-PIXE linescan was conducted using a 10 × 10-mm beam of 3-MeV protons. Reproduced by permission of the Arctic Institute of Canada.

Micro-PIXE is also widely used in trace element studies of zoned minerals, where the zoning contains information on the history of the mineral. Figure 5 shows a recently developed zoning application in calcium carbonate of biological origin—the otolith of an Arctic fish (Babaluk et al., 1997). The oscillatory behavior of Sr content in annual growth rings reflects annual migration of the fish from a freshwater lake to the open ocean, starting at age 9 years in the example shown. Micro-PIXE is now being applied in population studies for stock management.

Many more examples, including exotic ones in the fields of art and archaeometry, may be found in the book by Johansson et al. (1995). A unique example of nondestructive analysis is the use of micro-PIXE to identify use of different inks in individual letters in manuscripts. An example of a partly destructive analysis is the extraction of a very narrow core from a painting and the subsequent scanning of a microbeam along its length to identify and characterize successive paint layers.

In the absence of interfering peaks, the detection limit for an element is determined by the intensity of the continuous background on which its principal x-ray peak is superimposed in the measured spectrum. This background is due mainly to secondary electron bremsstrahlung,

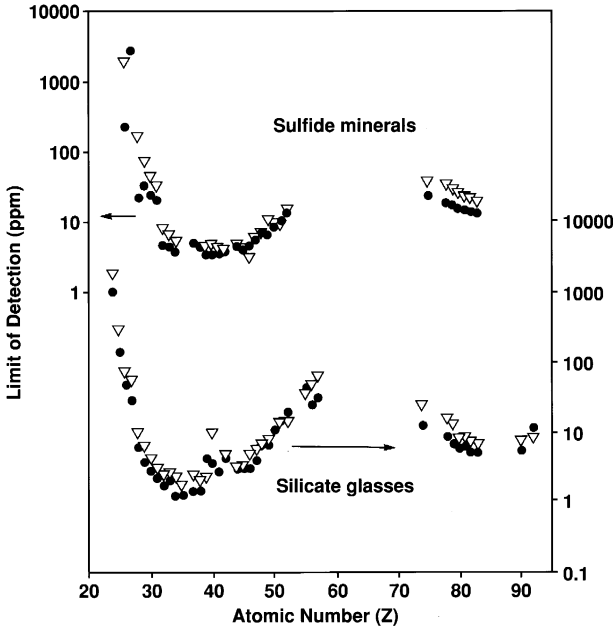


Figure 6. Micro-PIXE detection limits measured at Guelph for sulfide minerals and silicate glasses. Upper plot (left scale): triangles, pentlandites (350-μm Al filter); dots, pyrrhotite (250-mm Al filter); 10 mC of 3-MeV protons in a 5 × 10-mm spot. Lower plot (right scale): triangles, BHVO-1 basalt fused to glass; dots, rhyolitic glass RLS-158; 2.5 μC of 3-MeV protons in 5 × 10-μm spot; Al filter 250 mm.

whose intensity diminishes rapidly with increasing photon energy, but it can be augmented by gamma rays from nuclear reactions in light elements such as Na and F; these produce an essentially flat contribution to the background, visible at higher x-ray energies beyond the lower energy region where the bremsstrahlung dominates. The usual criterion is that a peak is detectable if its integrated intensity (I) exceeds a three-standard-deviation ($3\sqrt{B}$) fluctuation of the underlying background intensity (B). The desirability of a thin substrate in the case of thin specimens is obvious, since the substrate contributes only background to the spectrum. The ratio $I/3\sqrt{B}$ increases as the square root of the collected charge, suggesting that detection limits will be minimized by maximizing beam current and measuring time. Beam current is often limited by specimen damage, and measuring time may be limited by cost. Detection limits should be quoted for a given beam charge and detector geometry. The Z dependence of detection limits is U shaped, as shown in the examples of Figure 6, and this reflects the theoretical x-ray intensity expression that is included in Equation 1. For many thick specimen types, major elements contribute intense peaks and both pile-up continua and pile-up peaks to the spectrum, and these can locally worsen detection limits considerably; electronic pile-up rejection or on-demand beam deflection, which minimize such artefacts, are therefore necessary parts of a PIXE system.

The hazards of the technique are those involved with operation of any small accelerator operating with low beam currents. While radiation fields just outside the specimen chamber are very small, steps are necessary to shield radiation from beam-defining slits and from the accelerator itself. Special precautions (Doyle et al., 1991) are necessary if the beam is extracted through a window into the laboratory to analyze large or fragile specimens.

METHOD AUTOMATION

Specimen chambers usually accommodate many specimens, and computer-controlled stepping motors, either inside or outside the vacuum, are employed to expose these sequentially to the beam and to control all other aspects of data acquisition, e.g., charge measurement and selection of absorbing filter. In the case of micro-PIXE, the coordinates for each point analysis, linescan, or area map may be recorded by optical examination at the outset, permitting subsequent unsupervised data acquisition. Human intervention is only needed to break vacuum and insert new specimen loads into the analysis chamber, reevacuate, and input the parameters of the measurement (e.g., integrated charge per specimen). However, the accelerator itself is rarely automated, and a technical expert is required to steer and focus the beam prior to the analysis.

DATA ANALYSIS AND INITIAL INTERPRETATION

On the assumption that the volume sampled by the beam is homogeneous, the method is fully quantitative. Prior to analysis of measured spectra, checks are necessary to ensure that the expected detector energy resolution was maintained and that counting rate was in the proper domain.

The main step in data analysis is a nonlinear least-squares fit of a model energy-dispersive x-ray spectrum to the measured spectrum. The software requires the database mentioned earlier to describe the x-ray production process, a representation (Gaussian with corrections, as outlined below; see Appendix B) of the peaks in the energy-dispersed spectrum, and a means of dealing with continuous background; the latter may be fitted by an appropriate expression or removed by applying an appropriate algorithm prior to the least-squares fitting. Accurate knowledge of the properties of the detector and of any x-ray absorbers introduced between specimen and detector is necessary.

The quantities determined by the fit are the heights, and therefore the intensities, of the principal x-ray line of each element present in the spectrum. The intensities of all the remaining lines of each element are then set by that of the element's main line together with the relative x-ray intensity ratios provided by the database; these ratios are adjusted to reflect the effects of both absorber transmission and detector efficiency. For thick specimens, the database intensity ratios are also adjusted to reflect the effects of the matrix. This last adjustment is straightforward in a trace element determination in a known matrix; the matrix corrections to the ratios remain fixed through the iterations of the fit. The more complex case of major element measurement will be discussed below. Given the

very large number of x-ray lines present, the number of pile-up combinations can run into hundreds, and these double and triple peaks must also be modeled. A sorting process is used to eliminate pile-up combinations that are too weak to be of importance.

The final intensities provided by the fit for the principal lines of each element present are corrected for secondary fluorescence contributions, for pulse pile-up, and dead time effects (if necessary), prior to conversion to concentrations using Equation 1.

There is a linear relationship between the x-ray energy and the corresponding channel number of the peak centroid in the spectrum, and there is a linear relationship between the variance of the Gaussian peaks and the corresponding x-ray energy. The four system calibration parameters inherent in these relationships may be fixed or variable in the fitting process. The latter option has merit in that it accounts for small changes due to electronic drifts and does not require that the system be preset in any standard way.

If the major element concentrations are known *a priori*, then the second step, the conversion of these peak intensities to concentrations, is accomplished directly from Equation 1 or a variant thereof using the instrumental constant H, which is determined either using standards or from known major elements in the specimen itself. If the major elements have to be determined by the PIXE analysis, then Equation 1 has to be solved by an iterative process, starting with some estimate of concentrations and repeating the least-squares fit until consistency is achieved between the concentrations utilized and generated by Equation 1.

Various software packages are available for the above tasks (Johansson et al., 1995). They differ mainly in the databases that they adopt and in their approaches to handling the continuous background component of the spectra. The first option for the background is to add an appropriate expression to the peak model, thus describing the whole spectrum in mathematical terms, and then determine the parameter values in the background expression via the fitting process; the most common choice to describe the background has been an exponential polynomial of order up to 6 for the electron bremsstrahlung plus a linear function for the gamma ray component. The first of these two expressions must be modified to describe filter transmission and detector efficiency effects; the second needs no modification, because high-energy gamma rays are not affected by the absorbers and windows. While one particular form of expression may cope well, in an empirical sense, with a given type of specimen, there is little basis for assuming that the expressions described will be universally satisfactory. This has led to approaches that involve removing the background continuum by mathematically justifiable means. The simplest such approach, developed in the electron microprobe analogue, involves convoluting the spectrum with the "top-hat" filter shown in Figure 7. Such a convolution reduces a linear background to zero and will therefore be effective if the continuum is essentially linear within the dimensions of the filter. These dimensions are prescribed by Schamber (1977) as UW = 1 FWHM and LW = 0.5 FWHM, where FWHM is the full width at half-maximum of the resolution function at the

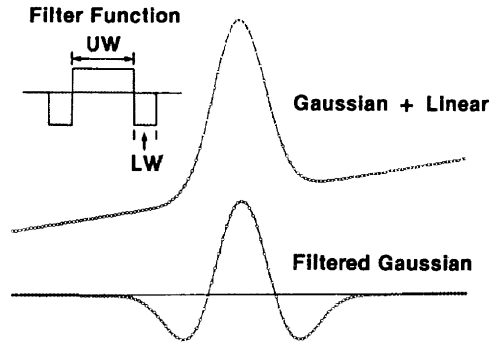


Figure 7. Top-hat filter and its effect on a Gaussian peak with linear background: UW and LW are the widths of the central and outer lobes.

x-ray energy specified. A more sophisticated multistep approach to continuum removal uses a peak-clipping algorithm to remove all peaks from the measured spectrum, and Ryan et al. (1988) have developed a version of this that is configured to cope with the very large dynamic range in peak heights that are encountered in PIXE spectra. First, the measured spectrum is smoothed by a "low-statistics digital filter" of variable width; at each channel of the spectrum this width is determined by the intensity in that channel and its vicinity. As the filter moves from a valley toward a peak, the smoothing interval is reduced to prevent the peak intensity from influencing the smoothing of the valley region. In a second step, a double-logarithmic transformation of channel content is applied, i.e.,

$$z = \ln[\ln(y + 1)] \tag{2}$$

in order to compress the dynamic range. The third step is a multipass peak clipping, in which each channel content $z(c)$ is replaced by the lesser of $z(c)$ and

$$\bar{z} = \frac{1}{2}[z(c + w) + z(c - w)] \tag{3}$$

where the scan width w is twice the FWHM corresponding to channel c. After no more than 24 passes, during which w is reduced by a factor of 2, a smooth background remains. The final step is to transform this background back through the inverse of Equation 2. The accuracy of this so-called SNIP (statistics-insensitive nonlinear peak-clipping algorithm) approach has been demonstrated in many PIXE analyses of geochemical reference materials (Ryan et al., 1990).

The VAX software package GEOPIXE (Ryan et al., 1990) is designed for thick specimens encountered in geochemistry and mineralogy and can also provide accurate elemental analysis of subsurface fluid inclusions (Ryan et al., 1995). It is installed in various micro-PIXE laboratories concerned with geochemical applications. It takes the SNIP approach to dealing with the background continuum. The widely used PC package GUPIX (Maxwell et al., 1989, 1995) deals with thin-, thick-, intermediate-, and multiple-layer specimens, employing the simple top-hat filter method to strip background. GUPIX offers the

option, useful in mineralogy, of analyzing in terms of oxides rather than elements; this enables the user to compare the sum of generated oxide concentrations to 100% and draw appropriate conclusions. GUPIX also allows the user to include one "invisible" element in the list of elements to be included in the calculation, requiring the sum of concentrations of all "visible" elements and the invisible element to sum to 100%; this can be used to determine oxygen content in oxide minerals or sulfide content in sulfide minerals when the oxygen and sulfur K x rays are not observable in the spectrum (as is usually the case). The GUPIX code can provide analysis and thickness determination of multilayer film structures, provided that the elements whose concentrations in a given layer are to be determined are not also present at unknown concentrations in any other layer. GUPIX is presently being extended to cope with analysis using deuteron and helium beams (Maxwell et al., 1997).

The above has been concerned with PIXE analysis for element concentrations in a fixed area on the specimen. In imaging applications, the beam is scanned over the specimen, and a record is built of the coordinates and channel number of each recorded x-ray event. From this record, a one- or two-dimensional image may be reconstructed using the spectrum intensity in preset "windows" to represent amounts of each element in a semiquantitative fashion. Obviously, such an approach can encounter errors due to peak overlap, but these may be dealt with by appropriate nonlinear least-squares spectrum fitting upon conclusion of the analysis. It is, however, desirable to transform observed window intensities into element concentrations in real time, updating as the data accumulate. Such a dynamic analysis method has been developed by Ryan and Jamieson (1993), and it enables on-line accumulation of PIXE maps that are inherently overlap resolved and background subtracted.

Returning now to direct analysis, standard reference materials of a similar nature to each specimen type of interest should be analyzed to demonstrate, e.g., accuracy and detection limits, which are determined in part by the measurement and in part by the fitting procedure. There are many examples of the accurate analysis of standard reference materials (SRMs). Table 1 shows the results of PIXE analysis by Maenhaut (1987) of the European Community fly-ash reference material used to simulate exposed filters from air particulate samplers. Agreement between measured and certified values is excellent, except for the elements of very low atomic number; these light elements tend to occur in larger soil-derived particles, and the data treatment used did not account for particle size effects. Table 2 summarizes a PIXE study of eight biological standard reference materials prepared for PIXE analysis by two different methods; one method (A) involved direct deposition of freeze-dried powder on the polymer substrate, and the other (D) involved Teflon bomb digestion of the powder followed by deposition and drying of aliquots. Table 3 shows micro-PIXE analyses of geochemical reference standards from three laboratories that specialize in earth science applications of micro-PIXE; these are the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia, the University of Guelph in Canada, and the National Accelerator Center at Faure in South Africa; details of these analyses are given by Campbell and Czamanske (1998). In such studies, the accuracy of the PIXE technique is usually assessed by comparing the mean element concentrations over several replicate samples (or spots on the same sample in the micro-PIXE case). In these and other such studies, the accuracy is typically a few percent when the concentration is well above the detection limit. The matter of reproducibility or precision is discussed below (see Problems).

Table 1. Analysis by PIXE of Thin Films (BCR 128) Containing Certified Fly Ash (BCR 38)[a]

Element	Reference Value[a,b]	PIXE Value[c]	PIXE/ Reference[d]
Mg[e]	9.45 ± 0.34[f]	8.3 ± 1.1	0.88 ± 0.12
Al[e]	127.5 ± 4.9[f]	104.0 ± 1.6	0.82 ± 0.01
Si[e]	227 ± 12[f]	176.0 ± 2.1	0.78 ± 0.01
P[e]		1.60 ± 0.15	
S[e]	3.89 ± 0.13[f]	4.17 ± 0.03	1.07 ± 0.01
K[e]	34.1 ± 1.6[f]	30.6 ± 0.7	0.90 ± 0.02
Ca[e]	13.81 ± 0.63[f]	13.35 ± 0.16	0.97 ± 0.01
Ti[e]		5.07 ± 0.12	
V	334 ± 23	340 ± 2	1.02 ± 0.07
Cr	178 ± 12	161 ± 16	0.90 ± 0.09
Mn	479 ± 16	454 ± 10	0.95 ± 0.02
Fe	33.8 ± 0.7	32.9 ± 0.5	0.97 ± 0.01
Ni	194 ± 26	192 ± 8	0.99 ± 0.04
Cu	176 ± 9	169 ± 8	0.96 ± 0.05
Zn	581 ± 29	572 ± 18	0.98 ± 0.03
Ga		56.2 ± 4.5[g]	
Ge		16.3 ± 4.7[g]	
As	48.0 ± 2.3	49 ± 11[g]	1.02 ± 0.23
Br		29 ± 7[g]	
Rb		183 ± 18[g]	
Sr		187 ± 18[g]	
Zr		169 ± 23[g]	
Pb	262 ± 11	225 ± 26[g]	0.86 ± 0.10

[a] Concentrations are given in micrograms per gram unless indicated otherwise.
[b] Reference values and associated errors are for the certified fly ash (BCR 38) unless indicated otherwise.
[c] PIXE data are averages and standard deviations, based on the analysis of five thin-film samples, unless indicated otherwise.
[d] Values were obtained by dividing both the PIXE result and its associated error by the reference concentration.
[e] Concentration in millligrams per gram.
[f] Concentration and standard deviation obtained by averaging round-robin values for the fly ash.
[g] Result derived from the sum spectrum, obtained by summing the PIXE spectra of the five films analyzed; the associated error is the error from counting statistics.

SAMPLE PREPARATION

In thin-specimen work, it is straightforward to deposit thin films of powder or fluid (which is then dried) or microtome slices of biological tissue onto a polymer substrate. In aerosol specimens, the particulate material is deposited on an appropriate substrate, usually Nuclepore or Teflon filters, by the air-sampling device. Reduction of original bulk

Table 2. Comparison between Certified Values and PIXE Data[a] for 16 Elements in 8 Reference Materials

Element	Number of Certified Values	Average Absolute % Difference Relative to Certified Value[b]	
		Method A	Method D
K	7	2.1 (7)	5.7 (5)
Ca	7	6.6 (7)	51 (7)
Cr	5	[17 (4)]	[19 (4)]
Mn	7	3.3 (6)	3.1 (6)
Fe	8	5.9 (8)	3.7 (8)
Ni	2	[19 (2)]	[25 (2)]
Cu	8	5.5 (8)	3.0 (8)
Zn	6	2.3 (6)	4.0 (6)
As	6	7.8 (2)	6.2 (2)
Se	3	[0.6 (1)]	[15 (1)]
Br	1	0.7 (1)	
Rb	7	2.0 (7)	5.6 (7)
Sr	6	2.6 (5)	3.1 (5)
Mo	3	[0.6 (1)]	[1.4 (1)]
Ba	1		
Pb	7	4.4 (4)	3.0 (4)

[a] With standard deviation from counting statistics <15%.
[b] The numbers in parentheses designate how many values the averages are based upon; brackets are used to indicate that all PIXE data for the certified element have standard deviation values in the range 5% to 15%.

Source: From Maenhaut et al., 1987. Reproduced by permission of Springer-Verlag.

material to a fluid or powder draws on the standard repertoire of grinding, ashing, and digestion techniques. Powdered material may be suspended in a liquid that is pipetted onto the substrate. Similarly, an aliquot of liquid containing dissolved material may be allowed to dry on the substrate. However, nonuniform deposits left from drying liquid drops may result in poorer than optimal reproducibility. The beam diameter is kept less than the specimen diameter to avoid problems of nonuniform specimen at the edges. If specimens are thin enough, there are no charging problems, but on occasion graphite coating has been necessary.

With thick specimens, a polished surface must be presented to the beam. If the specimen is an insulator, it must be carbon coated to prevent charge buildup. Geological specimens are prepared for micro-PIXE precisely as they are for electron probe microanalysis. The options are thin sections (30 to 50 μm) or multiple grains "potted" in epoxy resin.

SPECIMEN MODIFICATION

Overly large energy deposition per unit area may cause heating damage to specimens. This is observed as loss of matter from biological specimens or as crazing of mineral grains, and its avoidance is largely a matter of experience. Obviously the risk of damage increases with finer beam diameters. Volatile elements can be lost from biological materials, e.g., some compounds of sulfur, chlorine, and bromine. A guideline for tolerable beam current densities on thin biological specimens is 5 nA/mm^2 *in vacuo*

Table 3. Micro-PIXE Analyses of Fused Rock Standards at Three Laboratories[a]

	W-2 (CSIRO)[b]		BHVO-1 (Guelph)[c]		BCR-1 (Faure)[d]	
	Measured	Nominal	Meassured	Nominal	Measured	Nominal
K	6130 ± 500	5200	nm	—	nm	—
Ca	8.1 ± 0.3%	78%	nm	—	nm	—
Ti	6450 ± 100	6350	nm	—	nm	—
V	289 ± 40	262	nm	—	nm	—
Cr	nd	93	nm	—	nm	—
Mn	1270 ± 40	1260	1319 ± 25	1300	1358 ± 219	1400
Fe	7.5%	7.5%	8.62%	8.64%	9.5%	9.5%
Ni	64 ± 8	70	119 ± 6	121	17 ± 5	13
Cu	45 ± 4	103	160 ± 3	136	14 ± 2	18
Zn	74 ± 5	77	113 ± 3	105	137 ± 8	125
Ga	20 ± 1.4	20	22 ± 2	21	21 ± 1	22
As	1.4 ± 0.6	(1.2)	nd	—	nd	—
Rb	20 ± 0.8	20	9 ± 1	11	43 ± 2	47
Sr	195 ± 6	194	398 ± 3	403	330 ± 8	330
Y	19 ± 1	24	23.5 ± 0.5	27.6	32 ± 4	39
Zr	100 ± 3	94	170 ± 3	179	183 ± 5	190
Nb	5.7 ± 0.7	(8)	18 ± 1	19	13 ± 2	16
Ba	179 ± 15	182	119 ± 17	139	652 ± 21	680
Pb	nd	(9)	nd	—	16 ± 3	15
Th	2.1 ± 1.2	2.2	nd	—	nd	—

[a] Measured concentration in ppm (except Fe in wt.%); nominal concentration in ppm (except Fe in wt.%); brackets signify that the nominal value is only an estimate, as opposed to a formal recommendation. nm, not measured; nd, not detected.
[b] Concentrations from a sum spectrum from 10 spots each having 3-μC charge.
[c] Mean concentrations from spectra of 10 spots, each having 2.5-μC charge.
[d] Concentrations from a sum spectrum from 6 spots, each having 1-μC charge.

(Maenhaut, 1990), but it is advisable to conduct tests on a particular specimen type. In geological materials, e.g., volcanic glass, mobility of sodium results in significant analytical error. Careful design is needed when analyzing certain works of art and archaeologic specimens; glass, ceramics, and paper may be discolored (Malmqvist, 1995), and the intrinsic merit of in-air PIXE analysis, which provides continuous cooling of the specimen, is obvious.

PROBLEMS

Lack of homogeneity in the volume sampled by the ion beam is a potential problem, especially in mineralogical work, where a subsurface grain may be present but invisible just under the region being analyzed or microinclusions may be present in an optically homogeneous crystallite.

Spectrum artefacts may be interpreted as element peaks by the spectrum-fitting software, leading to erroneous detection of elements that are not present. Peak tailing effects on the predominantly Gaussian response function, double and triple pile-up peaks, and escape peaks (which arise when the x-ray interaction causes creation of a silicon K x ray that escapes from the detector) must be dealt with rigorously in the code, and this requires precharacterization of the lineshape as a function of x-ray energy. Overlaps between major x-ray lines of neighboring elements can cause error, and this places demands upon the accuracy of the x-ray physics portion of the database. Inadequate description of the continuous background underlying peaks of low intensity is another source of error. As a general rule, a fitting code should be directed to fit only those elements that visual inspection suggests are contributing to the spectrum; given an extensive list of elements that is not limited by inspection, fitting codes will happily generate small concentrations that represent nothing more than minor spectral artifacts such as imperfect description of tailing.

Precision for a given specimen type is generally determined by replicate analyses that provide a standard deviation to accompany each mean concentration, as in the example of Table 3. Analysis of the same region will test the reproducibility of all aspects except the specimen itself; analysis of different spots will include any specimen inhomogeneity in the variance that is determined. Usually, if the reduced chi-squared of the fit is good, the standard deviation for a suite of replicates will be close to that expected from the counting statistics of a single measurement; if it is in excess, this constitutes evidence for sample inhomogeneity. Comparison of mean concentrations determined in this manner for standard reference materials with formally accepted or recommended concentration values provides estimates of systematic errors; these tend to derive from database errors (e.g., in x-ray attenuation coefficients or relative x-ray intensities) that affect the accuracy of the fit.

LITERATURE CITED

Babaluk, J. A., Halden, N. M., Reist, J. D., Kristofferson, A. H., Campbell, J. L., and Teesdale, W. J. 1997. Evidence for nonanadromous behavior of Arctic char from Lake Hazen, Ellesmere Island, Northwest territories, Canada, based on scanning proton microprobe analysis of otolith strontium distribution. *Arctic* 50:224–233.

Berger, M. J. and Hubbell, J. H. 1987. Photon cross-sections on a personal computer. National Bureau of Standards Report NBSIR 87–3597.

Brandt, W. and Lapicki, G. 1981. Energy-loss effect in inner-shell Coulomb ionization by heavy charged particles. *Phys. Rev. A* 23:1717–1729.

Cahill, T. A. 1995. Compositional analysis of atmospheric aerosols. *In* Particle-Induced X-ray Emission Spectrometry (S. A. E. Johansson, J. L. Campbell, and K. G. Malmqvist, eds.). pp. 237–312. John Wiley & Sons, New York.

Campbell, J. L. 1996. Si(Li) detector response and PIXE spectrum fitting. *Nucl. Instrum. Methods* B109/110:71–78.

Campbell, J. L. and Cookson, J. A. 1983. PIXE analysis of thick targets. *Nucl. Instrum. Methods* B3:185–197.

Campbell, J. L. and Czamanske, G. C. 1998. Micro-PIXE in earth science. *Rev. Econ. Geol.* 7:169–185.

Campbell, J. L. and Wang, J.-X. 1989. Interpolated Dirac-Fock values of L-subshell x-ray emission rates including overlap and exchange effects. *At. Data Nucl. Data Tables* 43:281–291.

Chen, M. H. and Crasemann, B. 1984. M x-ray emission rates in Dirac-Fock approximation. *Phys. Rev. A* 30:170–176.

Chen, M. H. and Crasemann, B. 1985. Relativistic cross-sections for atomic K and L ionization by protons, calculated from a Dirac-Hartree-Slater model. *At. Data Nucl. Data Tables* 33:217–233.

Chen, M. H. and Crasemann, B. 1989. Atomic K-, L- and M-shell cross-sections for ionization by protons: A relativistic Hartree-Slater calculation. *At. Data Nucl. Data Tables* 41:257–285.

Chen, M. H., Crasemann, B., and Mark, H. 1980a. Relativistic K-shell Auger rates, level widths and fluorescence yields. *Phys. Rev. A* 21:436–441.

Chen, M. H., Crasemann, B., and Mark, H. 1980b. Relativistic M-shell radiationless transitions. *Phys. Rev. A* 21:449–453.

Chen, M. H., Crasemann, B., and Mark, H. 1981. Widths and fluorescence yields of atomic L vacancy states. *Phys. Rev. A* 24:177–182.

Chen, M. H., Crasemann, B., and Mark, H. 1983. Radiationless transitions to atomic $M_{1,2,3}$ shells: Results of relativistic theory. *Phys. Rev. A* 27:2989–2993.

Cohen, D. D. and Harrigan, M. 1985. K and L shell ionization cross-sections for protons and helium ions calculated in the ECPSSR theory. *At. Data Nucl. Data Tables* 33:256–342.

Cohen, D. D. and Harrigan, M. 1989. K-shell and L-shell ionization cross-sections for deuterons calculated in the ECPSSR theory. *At. Data Nucl. Data Tables* 41:287–338.

Czamanske, G. C., Kunilov, V. E., Zientek, M. L., Cabri, L. J., Likhachev, A. P., Calk, L. C., and Oscarson, R. L. 1992. A proton microprobe analysis of magmatic sulfide ores from the Noril'sk-Talnakh district, Siberia. *Can. Mineralogist* 30:249–287.

Doyle, B. L., Walsh, D. S., and Lee, S. R. 1991. External micro-ion-beam analysis. *Nucl. Instrum. Methods* B54:244–257.

Eldred, R. A. and Cahill, T. A. 1994. Trends in elemental concentrations of fine particles at remote sites in the U.S.A. *Atmos. Environ.* 5:1009–1019.

Eldred, R. A., Cahill, T. A., Wilkinson, L. K., Feeney, P. J., Chow, J. C., and Malm, W. C. 1990. Measurement of fine particles and their chemical components in the IMPROVE/NPS networks. *In*

Visibility and Fine Particles (C. V. Mathai, ed.). pp. 187–196. Air and Waste Management Association, Pittsburgh, Pa.

Griffin, W. L. and Ryan, C. G. 1995. Trace elements in indicator minerals: Area selection and target evaluation in diamond exploration. *J. Geochem. Exploration* 53:311–337.

Hubbell, J. H. 1989. Bibliography and current status of *K*, *L* and higher shell fluorescence yields for computations of photon energy-absorption coefficients. National Institute of Standards and Technology Report NISTIR 89–4144, U.S. Department of Commerce.

International Commission on Radiation Units and Measurements (ICRUM). 1993. Stopping Powers and Ranges for Protons and Alpha Particles. ICRUM, Bethesda, Md.

Johansson, G. I. 1982. Modifications of the HEX program for fast automatic resolution of PIXE spectra. *X-ray Spectrom*. 11:194–200.

Johansson, S. A. E. and Campbell, J. L. 1988. PIXE: A Novel Technique for Elemental Analysis. John Wiley & Sons, Chichester.

Johansson, S. A. E., Campbell, J. L. and Malmqvist, K. G. 1995. Particle-Induced X-ray Emission Spectrometry. John Wiley & Sons, New York.

Klockenkamper, R., Knoth, J., Prange, A., and Schwenke, H. 1992. Total reflection x-ray fluorescence spectroscopy. *Anal. Chem*. 64:1115A–1121A.

Krause, M. O. 1979. Atomic radiative and radiationless yields for *K* and *L* shells. *J. Phys. Chem. Ref. Data* 8:307–327.

Landsberg, J. P., McDonald, B., and Watt, F. 1992. Absence of aluminium in neuritic plaque cores in Alzheimer's disease. *Nature (London)* 360:65–68.

Larsson, N. P.-O., Tapper, U. A. S., and Martinsson, B. G. 1989. Characterization of the response function of a Si(Li) detector using an absorber technique. *Nucl. Instrum. Methods* B43:574–580.

Liu, Z. and Cipolla, S. J. 1996. A program for calculating *K*, *L* and *M* cross-sections from ECPSSR theory using a personal computer. *Comp. Phys. Commun*. 97:315–330.

Maenhaut, W. 1987. Particle-induced x-ray emission spectrometry: An accurate technique in the analysis of biological, environmental and geological samples. *Anal. Chim. Acta* 195: 125–140.

Maenhaut, W. 1990. Multi-element analysis of biological materials by particle-induced x-ray emission. *Scan. Microscy.* 4: 43–62.

Maenhaut, W., Hillamo, R., Makela, T., Jaffrezo, J.-L., Bergin, M. H., and Davidson, C. I. 1996. A new cascade impactor for aerosol sampling with subsequent PIXE analysis. *Nucl. Instrum. Methods* B109/110:482–487.

Maenhaut, W., Vandenhaute, J., and Duflour, H. 1987. Applicability of PIXE to the analysis of biological reference materials. *Fresenius Z. Anal. Chem*. 326:736–738.

Malmqvist, K. G. 1995. Application in art and archaeology. *In* Particle-Induced X-Ray Emission Spectrometry (S. A. E. Johansson, J. L. Campbell, and K. G. Malmqvist, eds.). pp. 367–417. John Wiley & Sons, New York.

Maxwell, J. A., Teesdale, W. J., and Campbell, J. L. 1989. The Guelph PIXE software package. *Nucl. Instrum. Methods* B43:218–230.

Maxwell, J. A., Teesdale, W. J., and Campbell, J. L. 1995. The Guelph PIXE software package II. *Nucl. Instrum. Methods* B95:407–421.

Maxwell, J. A., Teesdale, W. J., and Campbell, J. L. 1997. Private communication.

Orlic, I., Sow, C. H., and Tang, S. M. 1994. Semiempirical formulas for calculation of *L*-subshell ionization cross sections. *Int. J. PIXE* 4:217–230.

Paul, H. and Sacher, J. 1989. Fitted empirical reference cross-sections for *K*-shell ionization by protons. *At. Data Nucl. Data Tables* 42:105–156.

Rindby, A. 1993. Progress in X-ray microbeam spectroscopy. *X-Ray Spectrom*. 22:187–191.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., and Cousens, D. R. 1988. SNIP—a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrum. Methods* B34:396–402.

Ryan, C. G., Cousens, D. R., Sie, S. H., Griffin, W. L., and Suter, G. F. 1990. Quantitative PIXE microanalysis of geological material using the CSIRO proton microprobe. *Nucl. Instrum. Methods* B47:55–71.

Ryan, C. G., Heinrich, C. A., van Achterberg, E., Ballhaus, C., and Mernagh, T.P. 1995. Microanalysis of ore-forming fluids using the scanning proton microprobe. *Nucl. Instrum. Methods* B104:182–190.

Ryan, C. G. and Jamieson, D. N. 1993. Dynamic analysis: On-line quantitative PIXE microanalysis and its use in overlap-resolved elemental mapping. *Nucl. Instrum. Methods* B77: 203–214.

Schamber, F. H. 1977. A modification of the non-linear least-squares fitting method which provides continuum suppression. *In* X-Ray Fluorescence Analysis of Environmental Samples (T. G. Dzubay, ed.). pp. 241–248. Ann Arbor Science, Ann Arbor, Mich.

Scofield, J. H. 1974a. Exchange corrections of *K* x-ray emission rates. *Phys. Rev. A* 9:1041–1049.

Scofield, J. H. 1974b. Hartree-Fock values of *L* x-ray emission rates. *Phys. Rev. A* 10:1507–1510.

Scofield, J. H. 1975. Erratum. *Phys. Rev. A* 12:345.

Watt, F., Grime, G., Brook, A. J., Gadd, G. M., Perry, C. C., Pearce, R. B., Turnau, K., and Watkinson, S. C. 1991. Nuclear microscopy of biological specimens. *Nucl. Instrum. Methods* B54:123–143.

KEY REFERENCES

Johansson and Campbell, 1988. See above.

The first book on PIXE, providing extensive practical detail on all aspects and covering the early history.

Johansson et al., 1995. See above.

With the technique at maturity, this book provides the basics and presents a very wide range of examples of PIXE and micro-PIXE analysis in biology, medicine, earth science, atmospheric science, and archaeometry.

Nuclear Instruments and Methods in Physics Research, vols. 142 (1977), 181 (1981), B3 (1984), B22 (1987), B49 (1990), B75 (1993), B109/110 (1996), B150 (1999).

Proceedings of the seven international conferences on PIXE, they reflect its development and increasing sophistication.

APPENDIX A: RELATIONSHIP BETWEEN X-RAY INTENSITIES AND CONCENTRATIONS

Consider the general case of a proton beam incident with energy E_0 at angle α to the specimen normal and a detector at angle Θ_{T_0} to the specimen surface; the specimen thickness is t. To an excellent approximation, the proton travels

in a straight line, its path length being $t/\cos\alpha$, and emerges from the specimen with energy E_f. The proton's energy profile (E) along its path (direction x) is given by the stopping power of the matrix S_M:

$$S_M(E) = \rho^{-1}\frac{dE}{dx} \tag{4}$$

which is the concentration-weighted sum of the stopping powers of the major (matrix) elements comprising the specimen, whose density is ρ.

We consider only K x-ray production. The treatment for L and M x rays is similar, but the existence of subshells (L_1 to L_3, M_1 to M_5) renders matters more complex. The ionization cross-section $\sigma_Z(E)$ for a constituent element Z (concentration C_Z, atomic mass A_Z) decreases along the proton path as the energy decreases.

When an ionization occurs, the probability of K x-ray emission is the fluorescence yield ω_Z, and the fraction of the K x rays appearing in the principal line ($K\alpha_1$) is b_Z. The x rays from each successive element of the path have a transmission probability $T_Z(E)$ for reaching the detector through the overlying matrix,

$$T_Z(E) = \exp\left[-\left(\frac{\mu}{\rho}\right)_{Z,M}\frac{\cos\alpha}{\sin\Theta_{To}}\int_{E_0}^{E}\frac{dE}{S_M(E)}\right] \tag{5}$$

in which the matrix mass attenuation coefficient $(\mu/\rho)_{Z,M}$ is the concentration-weighted sum of the mass attenuation coefficients $(\mu/\rho)_{Z,M}$ of the major (or matrix) elements.

Integration along the proton track gives the total x-ray intensity due to a total incident beam charge Q as

$$Y(Z) = \frac{(N_{av}/e)\omega_Z b_Z t_Z \varepsilon_Z(\Omega/4\pi)}{A_Z} Q C_Z \int_{E_0}^{E_f}\frac{\sigma_Z(E)T_Z(E)}{S_M(E)}dE \tag{6}$$

in which N_{av} is Avogadro's number, e is the electronic charge, t_Z represents the transmission of the x rays through any absorbing filter used, $\Omega/4\pi$ is the solid angle fraction subtended by the detector, and ε_Z is the detector's intrinsic efficiency.

For a thick target ($E_f = 0$), it follows that

$$Y(Z) = Y_1(Z)H\varepsilon_Z t_Z C_Z Q \tag{7}$$

where $Y_1(Z)$ is the yield computed from the PIXE database per steradian per microcoulomb per unit concentration (ppm) and H is an instrumental constant that subsumes both detector solid angle and any beam change calibration factor.

For a thin target ($E_f \sim E_0$), $T_Z(E) \sim 1$, $\sigma_Z(E)$ becomes $\sigma_Z(E_0)$, and the integral in Equation 7 reduces to $t/\cos$ a. We then have

$$Y(Z) = Y_1(Z)H\varepsilon_Z t_Z m_a(Z)Q \tag{8}$$

where $Y_1(Z)$ is the computed x-ray yield per microcoulomb per unit areal density of element Z per steradian, H is the same instrumental constant, and $m_a(Z)$ is the areal density of element Z.

Specimens of intermediate thickness and multiple-layer cases are dealt with by straightforward extensions of this formalism.

APPENDIX B: RESPONSE FUNCTION AND EFFICIENCY OF SI(LI) DETECTOR

A hypothetically perfect Si(Li) detector would provide a delta function response in the spectrum to a monoenergetic x ray. In reality, x rays have an intrinsic Lorentzian energy distribution whose width is small compared to the Gaussian broadening that arises from the statistics of charge formation and from the electronic noise. The main feature of the lineshape is therefore a Gaussian. Partial escape from the front surface of photoelectrons and Auger electrons created in the silicon causes a flat shelf extending down to low energy. Some detectors, but not all, exhibit a truncated shelf whose origin is likely due to diffusion of thermalized ionization electrons out of the front surface. Most detectors show a quasi-exponential tail that may be due in part to this out-diffusion and includes the effects of Lorentzian broadening but may also arise in part from other origins. The mathematical model for a given peak may be assembled from these features, which are shown in Figure 8. Various authors have determined the values of the parameters of these components (e.g., height, width, and truncation energy) as a function of x-ray energy by using monochromatized x rays (Campbell, 1996). PIXE analysts have recourse to an adequate, albeit less accurate, characterization by fitting PIXE spectra of pure single elements. The KMM, KLL, and KLM radiative Auger satellites present in K x-ray spectra also contribute to apparent quasi-exponential tailing features and therefore have to be allowed for in this process. An elegant method for dealing with this complication through use of multiple absorbers is given by Larsson et al. (1989).

The detectors's intrinsic efficiency, i.e., the counts in the Gaussian peak relative to the number of x-ray photons incident on the detector window, is given by

$$\varepsilon_i = \frac{(\exp[-\sum_i^4 \mu_i t_i])f_{CCE}\, f_E[1-\exp(-\mu_{Si}D)]}{(1+z/d)^2} \tag{9}$$

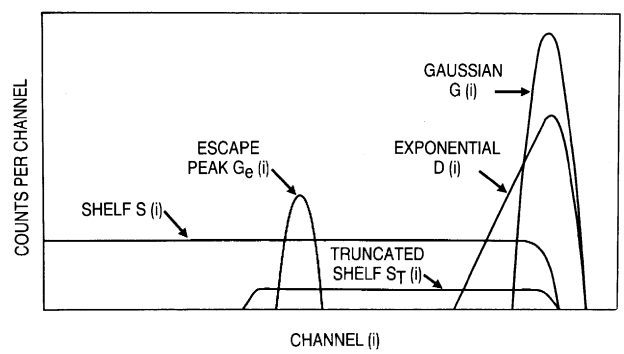


Figure 8. Typical components of the line shape of a Si(Li) detector.

where

$$z = \frac{1 - [\exp(-\mu_{Si}D)(1 + \mu_{Si}D)]}{\mu_{Si}[1 - \exp(-\mu_{Si}D)]} \quad (10)$$

and D is detector thickness, d the distance from specimen to detector surface, and μ_{Si} the linear attenuation coefficient of incident x rays in silicon. The first factor in the numerator of Equation 9 reflects x-ray attenuation in, successively, the vacuum window (beryllium or plastic), any ice layer on the cooled detector surface, the metal contact (most often gold), and any silicon dead layer that acts as a simple absorber. The second factor describes x-ray interactions that happen near the front surface of the active silicon and from which charge carriers escape via the processes described above and any others; the resulting signals fall in the shelves or the exponential features and not in the Gaussian. The factor f_E describes loss to the "escape peak," which occurs when a silicon K x ray, resulting from deexcitation after a photoelectric interaction, escapes from the crystal. The fourth factor describes penetration through the silicon. The reader will recall that the detector's solid angle was incorporated into Equation 1. That is the solid angle defined by the front surface of the silicon. Such a definition of solid angle is correct for only low-energy x rays that interact at the front of the crystal. As x-ray energy increases, the effective interaction depth is greater and the effective solid angle changes. The denominator of Equation 9 incorporates this nonnegligible effect in the intrinsic efficiency.

The detector manufacturer usually provides the thicknesses of the window, the contact, and the crystal. The presence of an ice layer is difficult to determine, but a steady decrease in efficiency at low energies betrays its presence; a detector that incorporates a heater for ice removal is advantageous. Detailed spectra measurements on pure elements are necessary to elucidate the correction for charge loss, which varies greatly among detectors. The term f_E has been determined accurately by various authors, e.g., Johansson (1982).

To determine the distance from specimen to detector crystal, one requires the distance of the crystal inside the window, which the manufacturer provides only approximately. This may be determined by recording the K x-ray spectrum from a ^{55}Fe radionuclide point source placed at accurately determined positions along the axis of the detector and invoking the inverse square law to describe its distance dependence.

The crystal thickness can be determined by constructing an efficiency curve using radionuclide point source standards and by least-squares fitting the model of Equation 9 to this curve with D as a variable. Suitable nuclides emitting x rays in the energy region 5 to 60 keV and enabling efficiency determination at accuracies of 1% to 3% are supplied by the Physikalisches-Technisches Institut in Braunschweig, Germany, and by other national laboratories charged with responsibility for standards.

JOHN L. CAMPBELL
University of Guelph
Guelph, Ontario, Canada

RADIATION EFFECTS MICROSCOPY

INTRODUCTION

Radiation effects microscopy is a general name that includes both ion-beam-induced charge (IBIC) microscopy and single-event upset (SEU) imaging. Ion-beam-induced charge microscopy was initially developed as a means of imaging the distribution of buried *pn* junctions in microelectronic devices (Breese et al., 1992) and dislocation networks in semiconductors (Breese et al., 1993a). More recently it has found a wider range of applications, e.g., in the study of solar cells (Beckman et al., 1997), segmented silicon detectors (R.A. Bardos, pers. comm.), and chemical vapor deposition (CVD) diamond films (Manfredotti et al., 1995, 1996). Many other applications are reviewed in the literature (Breese et al., 1996), and Doyle et al. (1997) is a good source of up-to-date applications.

Ion-beam-induced charge microscopy uses a beam of mega-electron-volt light ions, mainly protons or helium, to create electron-hole pairs in semiconductor material. This energy transfer from the incident ions to the material causes them to lose energy and eventually stop in the material at a well-defined distance called the ion range. The number of these electron or hole charge carriers that are collected at a junction within the sample is measured and displayed as a function of position of the focused beam within the scanned area. This process is analogous to electron-beam-induced current (EBIC) microscopy (Leamy, 1982), which uses a focused kilo-electron-volt electron beam, and optical-beam-induced current (OBIC) microscopy (Acciarri et al., 1996), which uses a focused laser beam to study similar problems in semiconductor technology. They are also used to image dislocations and inversion layers in semiconducting materials and depletion regions in devices and to give quantitative measurements of the minority carrier diffusion length (Wu and Wittry, 1978; Chi and Gatos, 1979). A description of different EBIC imaging modes as well as more detailed accounts of the theory and technical aspects of image generation is given elsewhere (Holt et al., 1974; Leamy, 1982; Piqueras et al., 1996).

As microelectronic device features continually shrink, the amount of charge that defines the different logic levels becomes smaller. This means that the charge generated by the passage of ionizing radiation through the device is more likely to cause a change in its logic state, called a soft upset, or a SEU. This phenomenon is particularly important for satellite-based devices, which are exposed to a high flux of ionizing cosmic radiation. It is a serious problem in the design of high-density semiconductor memories (May and Woods, 1979). Much work has been done using unfocused high-energy heavy-ion beams from large tandem accelerators (Knudson and Campbell, 1983), pulsed lasers (Buchner et al., 1988), and pulsed electron beams as sources of ionizing radiation to study upset mechanisms. However, the lack of spatially resolved information led to difficulties in determining which part of the device caused the upsets. Single-event upset analyses with collimated heavy-ion beams revealed the benefits of spatially resolving the results (Geppert et al., 1991;

Nashiyama et al., 1993). This prompted the development of SEU microscopy at Sandia (Doyle et al., 1992; Horn et al., 1992; Sexton et al., 1993a). This has many similarities to IBIC, which is usually performed in conjunction with SEU microscopy to image the device components present within the area analyzed for upsets. The two techniques use almost all the same nuclear microprobe hardware and electronics. Whereas IBIC mainly uses mega-electron-volt light-ion beams for analysis, SEU analysis uses heavier ions, such as carbon or silicon, to create a plasma density of electron-hole charge carriers high enough to cause upsets (McLean and Oldham, 1982; Knudson and Campbell, 1983).

There is no displacement damage in silicon using kilo-electron-volt electrons for EBIC, and a low-power laser beam such as Ar or He-Ne used for OBIC also produces no material damage. In comparison, mega-electron-volt ions generate defects in semiconductors. This is a major drawback in IBIC and SEU imaging as it imposes a limitation on the maximum number of ions that can be used to form an image, and care must be taken that the material property that is being measured is not altered beyond what can be accounted for during data collection. The effects of ion-induced damage on the measured charge pulse height spectrum is outlined in Specimen Modification for IBIC and are more fully described by Breese et al. (1996) and for SEU analysis by Sexton et al. (1995).

A major advantage of IBIC and SEU microscopy over competing electron beam techniques is their ability to give high-spatial-resolution analysis in thick layers. This is due to the different shapes of the charge carrier generation volume of mega-electron-volt ions compared with kilo-electron-volt electrons, which are compared in Figure 1. The measured carrier generation volume is shown in the form of concentration contours. The axes are normalized to the particle range. For kilo-electron-volt electrons the generation volume is approximately spherical. The lateral extent of the generation volume can be reduced by lowering the electron beam energy, but at the expense of further decreasing the analytical depth. Because of the large lateral extent of the generation volume of the electron beam within the silicon, the high spatial resolution of the electron beam on the sample surface is degraded, so deeply buried areas cannot be imaged with high spatial resolution. For 3-MeV protons there is little lateral scattering of the proton beam in the top few micrometers, and most occurs close to the end of range, giving a teardrop-shaped carrier generation volume. If the ion range is much greater than both the depletion depth and the diffusion length of the sample (which is usually the case with mega-electron-volt protons), then carriers generated deep within the sample, where there is significant lateral scattering, will not be measured.

The main advantages of IBIC microscopy over electron and optical methods are therefore a large pentration depth and a high spatial resolution for analysis of buried features within microelectronic devices (Breese et al., 1994; Kolachina et al., 1996) or semiconductor wafers. The main advantage of SEU microscopy is its ability to give spatially resolved information on which components within microelectronic devices are responsible for changes

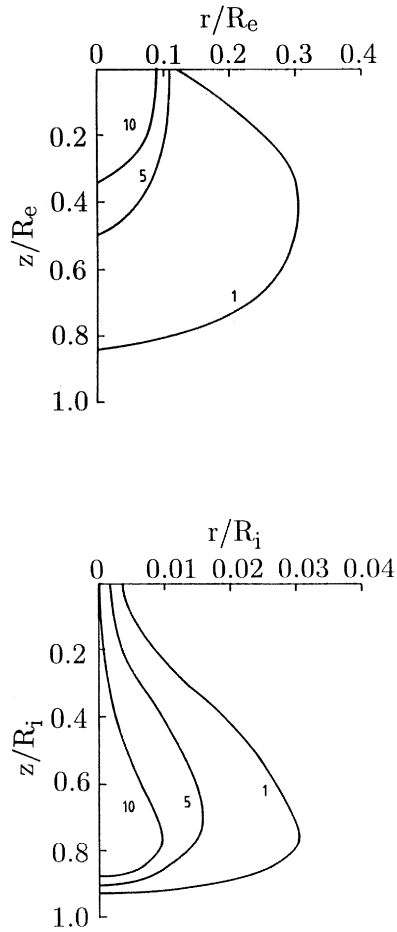


Figure 1. (**A**) Carrier generation contours for 10-keV electrons and-3 MeV protons in silicon. The lateral distance r and the depth z are plotted as fractions of (A) the electron range $R_e$and (**B**) the ion range R_i. The numbers 10, 5, and 1 refer to the intensity of the individual generation contour. Courtesy of Breese et al. (1992). Copyright 1992 American Institute of Physics.

in the device logic states due to the passage of high-energy ionizing radiation. A practical disadvantage of both forms of radiation effects microscopy is that they are only available in conjunction with nuclear microprobes located on megavolt particle accelerators (Grime and Watt, 1984; Watt and Grime, 1987; Breese et al., 1996). This makes them expensive compared with competing electron and laser beam techniques, and they are only available in laboratories that specialize in ion beam analysis, of which there are however a considerable number.

PRINCIPLES OF THE METHOD

The basic semiconductor theory relevant for IBIC and SEU imaging is reviewed here. Other sources (Sze, 1981; Kittel, 1986) are recommended for in-depth discussions of these and other theoretical aspects.

Semiconductor Theory

Ionizing radiation such as mega-electron-volt ions, kilo-electron-volt electrons, or high-frequency photons can create mobile charge carriers in semiconducting materials by transferring enough energy to the atoms to move valence electrons to the conduction band, leaving behind a positively charged hole. The average energy (E_{eh}) needed to create this electron-hole pair does not depend on the type of ionizing radiation and is constant for a given material (Klein, 1968). Charge carriers produced by ionizing radiation diffuse randomly through the semiconductor lattice, and if it contains no electric field, they become trapped and recombine until all excess carriers are removed.

If the semiconductor contains an electric field, the charge carriers are separated in the field region, and this charge flow can be measured in an external circuit. Because of the large number of charge carriers produced by individual mega-electron-volt ions, a charge pulse from each incident ion can be detected above the thermal noise level. The use of mega-electron-volt ions differs in this respect from electron and optical methods such as EBIC and OBIC, in which variations about a steady-state injection level produced by a large incident flux of ionizing radiation is detected rather than individual charge pulses, since these would not be resolved from the thermal noise of the sample.

Typical sample geometries used for IBIC microscopy are shown in Figure 2. The electric field is usually provided by *pn* junctions in microelectronic devices or by a Schottky barrier for the analysis of semiconductor wafers. Microelectronic devices typically consist of a semiconductor substrate with a patterned array of *pn* junctions at the substrate surface, which comprise the different transistors and other electronic components present (Sze, 1988). Above the semiconductor surface there are usually thick, patterned layers of insulating and metal tracks that make up the interconnecting device layers, with a thick passivation layer over the device surface. The total thickness of all the surface layers present can be up to 10 μm. As mentioned above, this is difficult to penetrate with good spatial resolution using an electron beam, and a laser beam suffers high attenuation in passing through any metallization layers, so analysis of small underlying junctions is very limited.

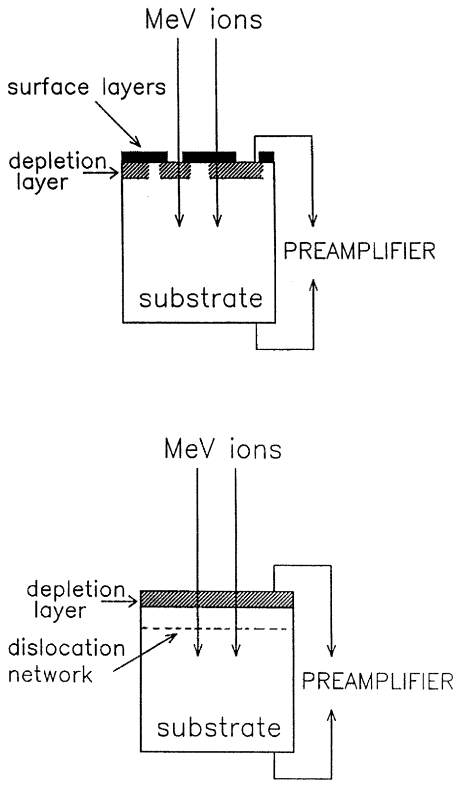


Figure 2. Schematic of geometry used for IBIC analysis of (**A**) microcircuits and (**B**) semiconductor and insulator wafers.

For both IBIC and SEU microscopy a focused beam current of ~1 fA of mega-electron-volt ions is incident on the front face of the sample, and each ion generates charge carriers along its trajectory. The number of charge carriers generated by each incident ion is measured for IBIC microscopy using a contact from the relevant depletion layer at the front face and a contact to the rear surface, each connected to a charge-sensitive preamplifier. This gives an output voltage that is amplified and measured by the data acquisition computer of the nuclear microprobe.

Quantitative Interpretation of the IBIC Pulse Height

A simple one-dimensional theory used to calculate the IBIC pulse height as a function of ion type and energy, minority carrier diffusion length, and surface and depletion layer thicknesses is described in the literature (Breese, 1993), and only a brief summary is given here. The ion-induced charge pulse height is given here in units of kilo-electron-volts. For example, a measured charge pulse height of 1 MeV from a 2-MeV ion means that half of the beam energy has generated charge carriers that contribute to the measured charge pulse height, and the other 1 MeV has been lost in passing through the surface layers or through carrier recombination within the substrate.

The total measured charge pulse height of Π kilo-electron-volts can be calculated by integrating the rate of ion electronic energy loss dE/dz between the semiconductor surface ($z = 0$) and the ion range, i.e., mean penetration depth ($z = R_i$):

$$\Pi = \int_0^{z_d} \frac{dE}{dz} dz + \int_{z_d}^{R_i} \frac{dE}{dz} \exp\left[-\frac{z - z_d}{L}\right] dz \qquad (1)$$

where L is the minority carrier diffusion length in the sample. The first term is the contribution from the charge carriers generated within the depletion region of the collecting junction, which is z_d micrometers thick. The second term is the contribution from the charge carriers diffusing to the collecting junction from the substrate. Figure 3 shows the calculated charge pulse height for protons and ^{4}He ions for different values of L, based on Equation 1. For a short diffusion length, the charge pulse height reaches a maximum value at a certain proton energy and then decreases as the proton energy is raised further. This is because at a low beam energy there is a high rate of

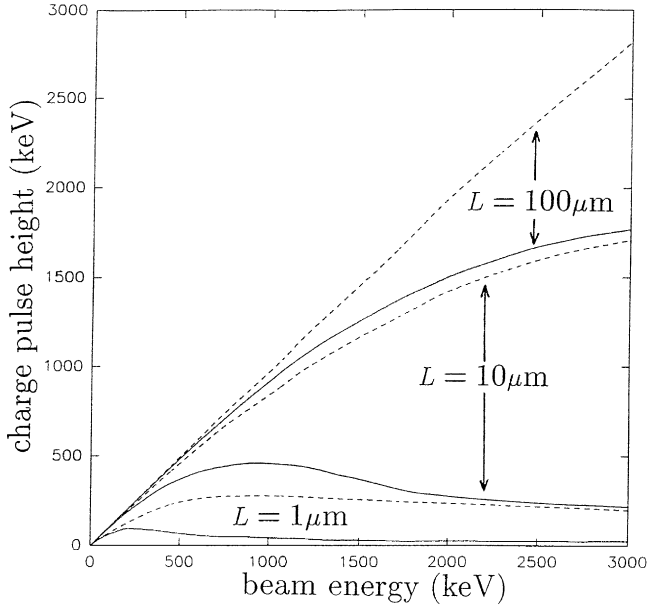


Figure 3. Variation of the charge pulse height with protons (solid lines) and ^{4}He ions (dashed lines) versus energy for diffusion lengths $L = 1$, 10, 100 μm. Courtesy of Breese et al. (1993). Copyright 1993 American Institute of Physics.

energy loss close to the surface where the generated charge carriers can be measured. As the beam energy is raised further, there is a gradual reduction in the rate of energy loss at the surface, and most carriers are generated too deep to be measured. Similar behavior is also found for kilo-electron-volt electrons (Wu and Wittry, 1978).

IBIC Topographical Contrast

It is important to know how the IBIC image contrast changes as a function of changes in surface layer thickness in order to correctly interpret the distribution of charge pulses (Breese et al., 1995). Figure 4A shows the rate of electronic energy loss with distance traveled for 2-MeV protons and 3-MeV ^{4}He ions. Since they have the same range of ~12 μm as 3-MeV ^{4}He ions, 38-keV electrons have been included, so that the charge pulse height from these different charged particles with the same range can be compared.

Figure 4B shows the charge pulse height resulting from a sample consisting of a depletion layer thickness of 1 μm and a substrate diffusion length $L = 6$ μm, with increasing surface layer thickness for the same incident charged particles. The maximum slope of the charge pulse height variation with surface layer thickness occurs close to the end of range, i.e., between 10 and 12 μm, for 3-MeV ^{4}He ions, whereas the charge pulse height for 2-MeV protons for similar thickness surface layers changes at a rate that is a hundred times slower. Thus, mega-electron-volt ^{4}He ions can produce IBIC images that are highly sensitive to device topography. In comparison, similar energy protons produce IBIC images that are almost independent of device topography, in which the contrast just depends on the distribution of the underlying *pn* junctions, as will be demonstrated in Figure 10.

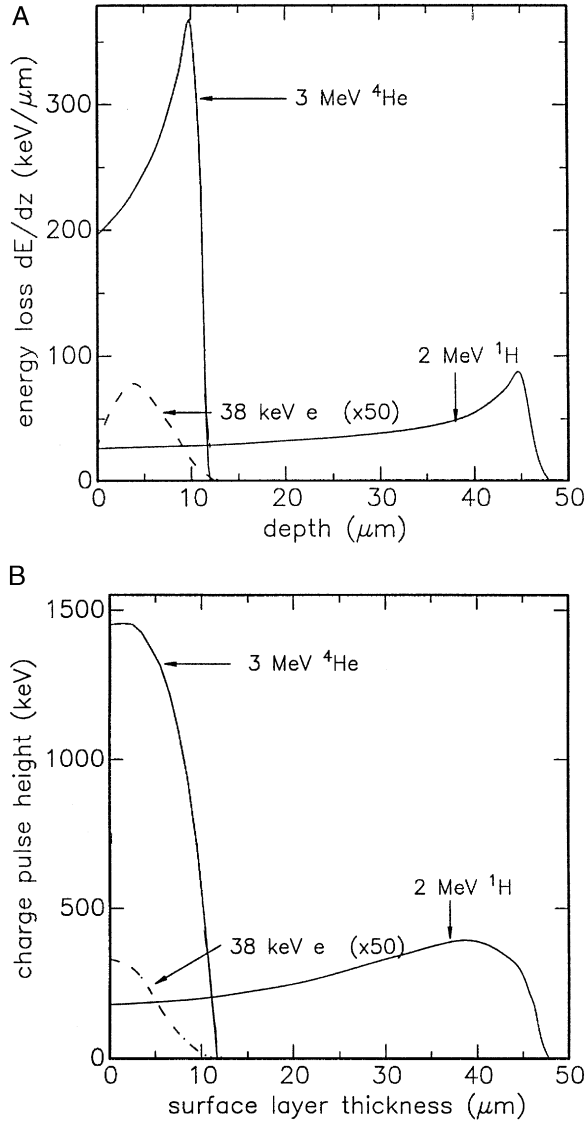


Figure 4. (**A**) Electronic energy loss with depth for 2-MeV protons, 3-MeV He ions, and 38-keV electrons (dashed line with the energy loss is multiplied by 50). (**B**) Charge pulse height resulting from a depletion layer thickness of 1 μm and a diffusion length of 6 μm as a function of increasing surface layer thickness for the same charged particles. Courtesy of Breese et al. (1995). Copyright 1995 American Institute of Physics.

The charge pulse height from 38-keV electrons is shown on the same vertical scale as for mega-electron-volt ions to compare the sensitivity of EBIC to changes in surface layer thickness (in practice, EBIC uses a steady-state large incident beam current). The maximum slope of the resultant charge pulse height variation is typically between that measured with protons and with mega-electron-volt ^{4}He ions, showing that IBIC can be made either sensitive or insensitive to topographical contrast whereas EBIC does not have this flexibility. Optical methods also lack this flexibility owing to strong attenuation by metallization layers.

This section has briefly described a basic model of ion-induced charge collection, which provides some insight into how the measured charge pulses vary with beam energy and surface layer thickness. There are two routes to obtaining better quantitative understanding for more detailed analysis and interpretation. First, better analytical models of ion-induced charge diffusion and collection are being developed (Donolato et al., 1996; Nipoti et al., 1998). These take into account two-dimensional lateral charge diffusion effects, which should result in better quantitative understanding of how the proximity of *pn* junctions, grain boundaries, and dislocations affect IBIC image contrast. The second route to acquiring a better insight into charge collection mechanisms is with detailed computer simulations. These seem to be the best path to a quantitative understanding of three-dimensional charge collection for use with IBIC and SEU microscopy, as described above.

SEU Microscopy

The passage of a high-energy, heavy ion through a semiconductor can produce a region with a very high density of excess charge carriers along the ion path because of the ion's high rate of electronic energy loss. The dense plasma created around heavy ions is a reasonable experimental analogy for the passage of strongly ionizing cosmic radiation passing through device memories in satellite-based microelectronic devices, which is why this has become an important research area. If the carrier density reaches a critical level, then enough charge might be collected at a sensitive device junction to change the device logic state, causing a corruption of the information stored, i.e., a SEU has occurred. Figure 5 gives a schematic of the SEU imaging process within a nuclear microprobe. A focused heavy-ion beam is scanned over the chip, and the information stored within the device is sampled at each beam position within the scanned area, so that an image can be constructed showing the upset probability at each location.

If a heavy ion passes through a depletion region of a device memory (Fig. 6), the resultant dense plasma formation can distort its shape, giving rise to the phenomenon of charge funneling (McLean and Oldham, 1982). The carrier density can reach up to 10^{20} cm^{-3} along the path of a heavy ion, which is much greater than the typical device substrate doping density. As a consequence, the junction depletion layer in the vicinity of the ion path is quickly neutralized, and the high electric field in the junction is screened by electrons being drawn off at the electrode. The electric field associated with the junction is reduced in size and becomes elongated in the direction of the ion path, and Figure 6C shows the distorted equipotential line associated with this effect. The effect of this charge funnel is to give a much larger measured charge pulse than would otherwise occur by charge drift and diffusion alone. The charge is also much more localized around the

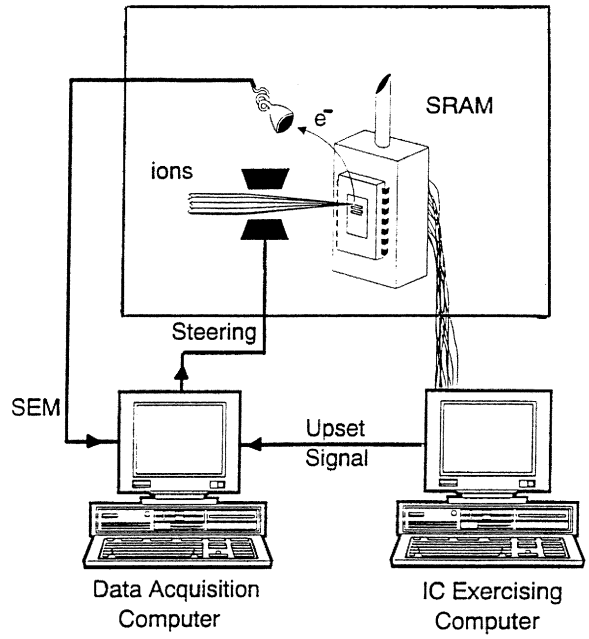


Figure 5. Schematic of apparatus used for SEU analysis. Courtesy of Doyle et al. (1992).

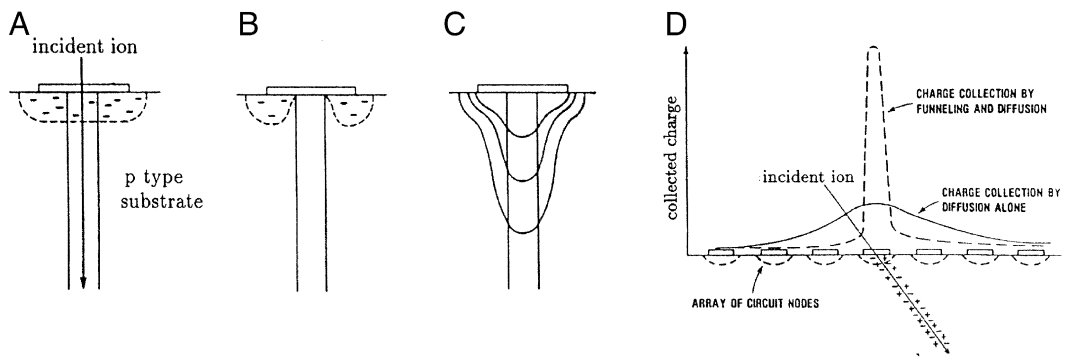


Figure 6. Schematic of charge funneling showing (**A**) a heavy ion hitting a junction, (**B**) the depletion region being neutralized by the resultant plasma column, and (**C**) equipotential lines from the junction being extended down. (**D**) Resultant charge collection profile over a circuit array, with enhanced charge collection due to charge funneling. Adapted from McLean et al. (1982). Reprinted in modified form. Copyright 1982 IEEE.

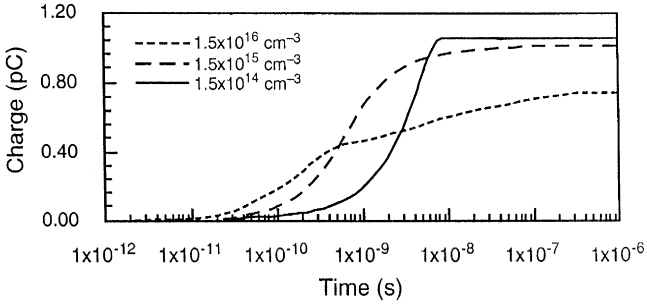


Figure 7. Simulated charge transients generated by a 100-MeV Fe ion in three n^+/p silicon diodes with substrate doping densities shown in the top left. Courtesy of Dodd et al. (1994).

struck junction, as less charge diffuses away to be absorbed by other junctions, resulting in a much increased susceptibility of the device to SEUs.

Charge-funneling effects are also important to IBIC as well as for SEU analysis because they should be taken into account when calculating the charge pulse height due to heavy ions if the ionization carrier density is much greater than the substrate density. Sophisticated two- and three-dimensional computer codes (Dodd et al., 1994; Dodd, 1996) can be used to model the effects of SEU and charge-funneling effects in order to gain a detailed understanding of the basic mechanisms that can upset devices and to distinguish funneling from the effects of charge drift and diffusion.

An example is a heavily doped (5×10^{20} cm^{-3}) n^+ diffusion in a p-type silicon substrate with different doping concentrations (Dodd et al., 1994). The passage of a 100-MeV Fe ion through this system was modeled to investigate under what conditions it could upset the device. Figure 7 shows the simulated charge collection transients from one quadrant of the symmetrical device area for three different substrate doping densities. The total measured amount of charge decreases with increasing substrate doping density, because it is less likely that significant additional amounts of charge will be collected via the funneling effect. For the lightly doped substrates, nearly all the charge carriers are collected by a charge-funneling process, so there is little extra charge to be collected by diffusion after the funnel collapses ~10 ns after the passage of the ion through the device. For the more heavily doped substrate of 1.5×10^{16}cm^{-3}, the funnel collapses 400 ps after ion impact, leaving a considerable fraction of charge carriers free to diffuse through the device. Some carriers diffuse to the collecting junction and others recombine, so less charge is measured in total.

Further Advances

An interesting development that should further enhance the capabilities of IBIC and SEU analysis is the production of time-resolved, or transient, IBIC images. Presently being developed by the Sandia microprobe group (Schone, 1998), this enables those charge components from the depletion region and from the substrate to be measured directly, being distinguished by their different charge collection time scales. The ion-induced transient response is measured with a timing resolution of tens of picoseconds, as has already been done with unfocused ion beams (Wagner et al., 1988). The experimental procedure is to store the digitized ion-induced current transient along with the (x,y) position coordinates of the beam. From the stored data, the shape of the current transient can be studied as a function of position as well as time, enabling direct measurements of the critical charge needed to cause SEUs. These also help to test numerical three-dimensional simulations of charge diffusion and charge funneling.

A similar system has been built by the Melbourne microprobe group (Laird et al., 1999). The installation of a cold stage on this system capable of cooling samples to liquid helium temperature should enable transient IBIC pulses to be studied as a function of temperature, in a similar manner to deep level transient spectroscopy; see DEEP-LEVEL TRANSIENT SPECTROSCOPY (Lang, 1974).

PRACTICAL ASPECTS OF THE METHOD

The basic layout used for IBIC analysis of microcircuits and Schottky barrier samples is shown in Figure 2. The IBIC pulses are measured using standard charged particle detection electronics, similar to those used for other ion beam analysis techniques. Each pulse is fed to a low-noise charge-sensitive preamplifier whose output voltage is proportional to the measured number of carriers generated by individual ions.

Charge-sensitive preamplifiers are ideal for use in IBIC experiments because they integrate the induced charge on a feedback capacitor. A charge-sensitive preamplifier typically has an open-loop gain of $\sim 10^4$ so that it appears as a large capacitance to the sample, rendering the gain insensitive to changes in the sample capacitance. Details of preamplifier design, pulse shaping, and methods of noise reduction are described in many texts (e.g., Bertolini and Coche, 1968; England, 1974). The preamplifier output voltage V_o, is given as

$$V_o = \frac{1000\Pi e}{E_{eh}C} \tag{2}$$

where Π is the measured charge pulse height in kiloelectron-volts, given by Equation 1. With a feedback capacitance $C = 1$ pF, a typical preamplifier output pulse size is 44 mV per mega-electron-volts of energy of the incident ion. The small output pulses from the preamplifier are fed into an amplifier, which gives an output voltage of ~1 V. This is then fed into the data acquisition computer and an image is generated showing either variations in the average measured charge pulse height or the intensity of counts from different "windows" of the spectrum at each pixel within the scanned area.

A beam current of ~1 fA (i.e., ~6000 ions/s) is typically used for both IBIC and SEU microscopy since the maximum data acquisition rate available with most nuclear microprobes is less than 10 kHz. To produce a focused spot containing such a small current, a much larger beam current of 100 pA is first focused in a conventional

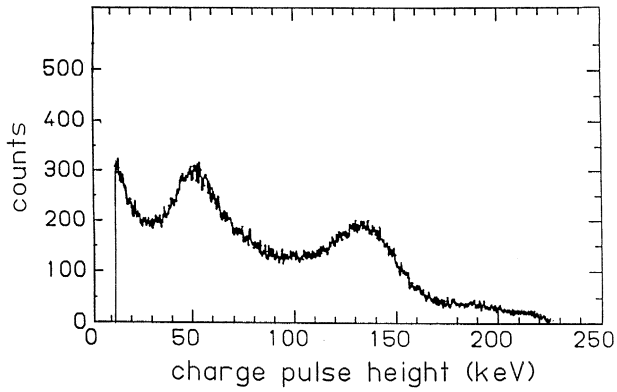


Figure 8. IBIC pulse height spectrum from the 300 × 300 μm^2 device area shown in Figure 9, produced with 3-MeV protons. The preamplifier was connected between the data pin and ground.

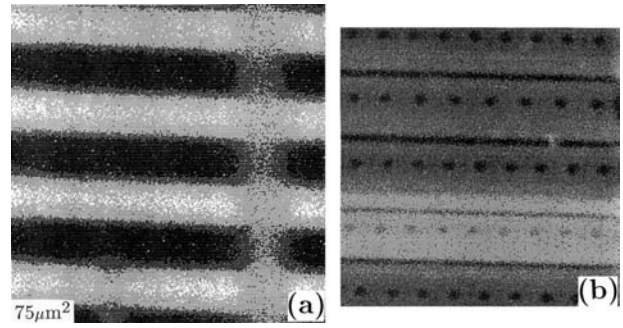


Figure 10. (**A**) 75 × 75 μm^2 average pulse height IBIC image of the central region in Figure 9 using 3-MeV protons. (**B**) 2.3-MeV ^{4}He IBIC images of the same area, under the same conditions, with a scan size of 40 × 40 μm^2. Reprinted from (Breese et al. 1994) with permission. Copyright 1994 American Institute of Physics.

manner (Breese et al., 1996). The object and divergence slits are then closed until the remaining current is a few thousand ions per second, as measured by a semiconductor detector placed in the path of the ion beam. This procedure ensures that the sample is not irradiated with a high ion dose prior to analysis. The sample is then moved onto the beam axis, and a large-area IBIC image is used to identify and position the region of interest. With SEU analysis, damage effects are even more critical since heavier ions are used. Therefore, initial sample positioning with SEU microscopy is also best carried out using IBIC with light ions, since they will cause much less damage prior to SEU analysis.

IBIC Examples

A typical IBIC charge pulse height spectrum measured from a 300 × 300 μm^2 area of an extended programmable read-only-memory (EPROM) chip using a focused 3-MeV proton beam is shown in Figure 8. The charge pulses in this case are up to five times larger than the thermal noise level, which occupies the lower end of the spectrum. The lower input threshold of the amplifier should be raised to a level such that very few noise pulses are measured or else the high noise level will saturate the data acquisition system and the ion-induced charge pulses will not be measured. However, if the threshold is raised too far, then the smallest charge pulses will not be measured, demonstrating the problem of a poor signal-to-noise ratio, which must always be borne in mind in IBIC experiments. This may also influence the choice of incident ion used for the IBIC analysis, since heavier ions usually give larger charge pulses (see Fig. 4).

In this example the device is an *n*-type metal-oxide-semiconductor (NMOS) memory chip. The area analyzed contains two output driver field effect transistors and also two transistors comprising the input buffer. The metallization consists of a 1-μm-thick layer of Al(1% Si). Over the surface of the device is an ∼1.5-μm-thick SiO_2 passivation layer, with a total surface layer thickness up to 4 μm. Figure 9 shows three IBIC images of this 300 × 300 μm^2 area, but with different preamplifier connections and device pin voltages in each case. In each case, spectra similar to that shown in Figure 8 are measured, and IBIC images showing the average charge pulse height at each pixel are produced. The differing contrasts are explained in detail elsewhere (Breese et al., 1992), but they demonstrate here that IBIC can generate images showing the different distributions of active components buried several micrometers beneath the surface of the functioning device.

Figure 10 shows IBIC images of the central region of Figure 9c with the device under the same operating conditions. Figure 10 shows a 75 × 75 μm^2 image generated using 3-MeV protons and a 40 × 40 μm^2 image generated with 2.3-MeV ^{4}He ions. Figure 10b shows strong topographical contrast where the circles along the lengths of the horizontally running metallized source and drain regions can be seen. These are not visible in Figure 10a, which only shows IBIC contrast due to the layout of the underlying

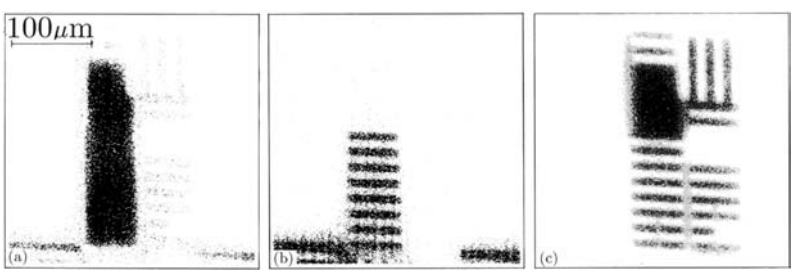


Figure 9. Three 300 × 300 μm^2 IBIC images of the same device area. The preamplifier connections are (**A**) between the supply voltage pin, with $V_{cc} = +5$ V and ground. The output driver voltage $V_{o=+5}$ V, and the data pin is not connected here. (**B**) Same as (A) except the data pin is now also connected to a different preamplifier. (**C**) Measured between the data pin and ground but with other transistors on.

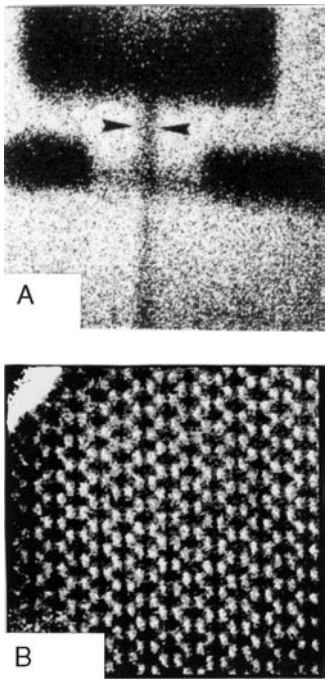


Figure 11. (**A**) $20 \times 20\ \mu m^2$ IBIC image of a GaAs device, between the *p*- and *n*-type contacts. A 0.8-μm-wide depletion region is indicated. (**B**) $65 \times 65\ \mu m^2$ IBIC image of a DRAM. Reprinted from Breese et al. (1993b) with permission from Elsevier Science B.V., Amsterdam, The Netherlands.

pn junctions. This demonstrates the insensitivity of IBIC to device topography using mega-electron-volt protons, as calculated in Figure 4B. The maximum signal-to-noise level in Figure 10b is ~50, whereas in the proton IBIC images of this same area it was ~5, so in this case the use of heavier ions has greatly increased the pulse sizes.

Other IBIC Examples

An IBIC image of a GaAs high-electron-mobility transistor is shown in Figure 11a. The surface of the device was patterned using a focused Be^+ ion beam, which formed *p*-type material and resulted in the formation of a field effect transistor with a very narrow gate depletion region that was confined both laterally and in depth. The IBIC images were generated using contacts between the *p*- and *n*-type material, with a 10-μm-thick Al foil placed between the focused 3-MeV protons and the device surface. On the IBIC image, a 0.8-μm-wide depletion region is indicated by arrows; this high spatial resolution resulted from the low lateral straggling of mega-electron-volt protons through thick surface layers.

Figure 11b shows an IBIC image of a $65 \times 65\ \mu m^2$ area of the memory array in a 4-Mbit dynamic random-access memory (DRAM) device. Smaller charge pulses were measured from the light-colored honeycomb structure of 1-μm-wide trench cells that comprise parts of the individual memory cells. To prepare a sample suitable for IBIC microscopy, all the device surface layers were removed using hydrofluoric acid, leaving just the *p*-type substrate and the *n*-doped trench walls. Electrical contacts were then produced by depositing a thin gold layer onto the back surface to form an ohmic contact and a thin aluminum layer onto the front surface to form a Schottky contact in the geometry shown in Figure 2B. This device also formed part of an IBIC study of methods of compensating for the effects of beam-induced damage in IBIC images, as described by Breese et al. (1993b).

DATA ANALYSIS AND INTERPRETATION

SEU analysis of SRAMs

An example of the SEU analysis of radiation-hardened static random-access memory (SRAM) devices is given here. An SEU image generated using a low-intensity beam of 24-MeV Si^{6+} ions scanning over a $40 \times 40\ \mu m^2$ area of the device is shown in Figure 12a, and the schematic of the device layout within this area is shown in Figure 12b. The SEU image was produced by interrogating the device logic state as a function of irradiation with the focused ion beam at each pixel within the scanned area. The SEU image shows variations in the upset probability within this area as darker shading at the center and lighter shading toward the edges of the upset-generating region.

An IBIC image of the same area generated using 2.4-MeV ^4He ions is shown in Figure 12c. The ion impact locations that give rise to upsets within this scanned area can now be directly identified by comparing the SEU and IBIC images with the device layout. In this case, the upset

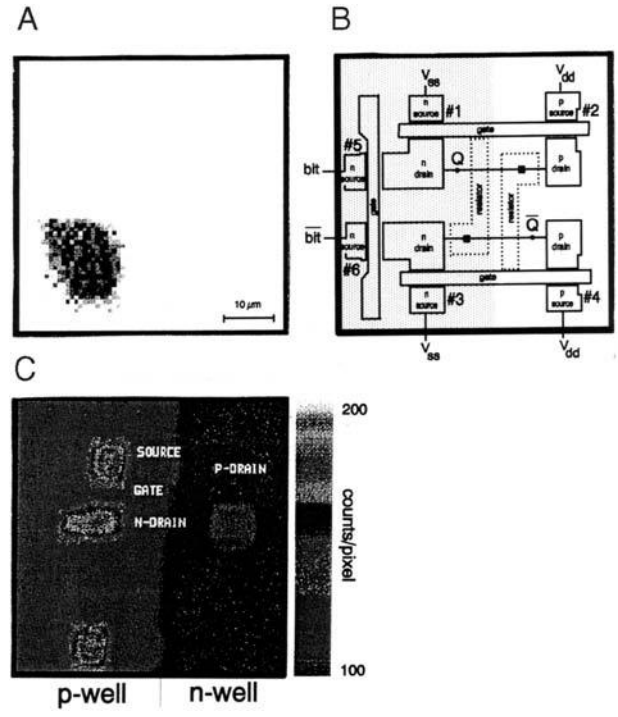


Figure 12. (**A**) $40 \times 40\ \mu m^2$ SEU image of the SRAM device. (**B**) Schematic of the device layout in this area and (**C**) IBIC image from the same area. Courtesy of Horn et al. (1993). Reprinted from Sexton et al. (1993b) with permission from Elsevier Science B.V., Amsterdam, The Netherlands.

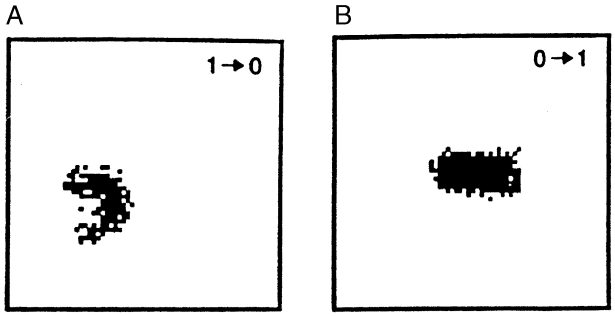


Figure 13. SEU images of an SRAM memory cell showing the logic state dependence: (**A**) with logic 1 initially stored in the memory cell and (**B**) with logic state 0 initially stored. Courtesy of Horn et al. (1993).

area was localized around an *n*-channel transistor drain. In the IBIC image, a large charge pulse was measured from the *n* drains within the *p* well of the device. Within the *p* well, a large charge pulse was also measured from the *n* source, but from the SEU image this is shown to be insensitive to upsets. There is a weaker region of charge collection within the *n*-well region on the right of the IBIC image, corresponding to a *p* drain. A detailed explanation of the observed IBIC image contrast in terms of the device layout is given in the literature (Horn et al., 1993; Sexton et al., 1993b), where the behavior observed in the SEU image was correlated with the observed IBIC contrast. It was postulated that regions of larger than expected charge pulses could arise from a "shunt" effect (Kreskovsky and Grubin, 1986) caused by the high ion-induced carrier density along the ion path.

Single-event upset microscopy also gives the opportunity to measure logic-state-dependent upsets. Figure 13 shows an SEU image from the same SRAM device structure for each of two different initial logic states. As can be seen, the upsets occur in different positions, and the sensitive volume is of a different shape and size, depending on the initial value stored. This observation is typical of the additional information to be gained by the ability to spatially resolve the locations of SEUs.

SPECIMEN MODIFICATION

The main damage mechanism of mega-elextron-volt ions in semiconductors is the creation of vacancy/interstitial pairs, called Frenkel defects. This is caused by direct collisions between the incident ions and the lattice nuclei, in which a sufficient amount of energy is imparted to the nucleus to displace it from its lattice site. Their primary effect for IBIC and SEU microscopy is a reduction in the minority carrier diffusion length, because the defects act as trapping and recombination centers. Since many charge pulses should ideally be measured at each pixel to reduce the statistical noise in IBIC and SEU images, ion-induced damage is a major drawback as it limits the beam dose, which can be used to produce an image before the image contrast greatly changes.

Prior to the development of IBIC, it was demonstrated that a nuclear microprobe could be used to spatially resolve ion beam damaged areas in a *pn* junction by measuring the charge pulses from the focused beam (Angell et al., 1989). Since the subsequent development of IBIC as an analytical technique, the effects of ion-induced damage and methods of compensating for its effects have been studied extensively (Breese et al., 1993b, 1995).

A study into the observed damage occurring during SEU analysis (Sexton et al., 1995) investigated whether the observed damage effects with heavy ions were due to the ionizing energy loss component, i.e., to possible charge buildup in insulating layers, or the nonionizing component, i.e., to displacement effects. Devices were irradiated with X-rays to model the effects of ionizing radiation, and it was found that no significant changes were observed in subsequent SEU behavior. In comparison, identical devices irradiated with large doses of 30-MeV copper ions underwent significant changes in SEU behavior after a certain dose, with more components becoming sensitive to upsets with cumulative beam dose. It was postulated that this was due to a reduction of the critical charge threshold for upset produced by beam damage effects of the nonionizing component of the energy loss.

Using the same simple theory to characterize charge collection as described in Principles of the Method, a rough guide to the behavior of IBIC pulses with cumulative beam dose can be derived. The calculated rate of charge pulse height reduction for different energy protons and α-particles is shown in Figure 14. In brief, the charge pulse height reduces with cumulative beam dose because of a reduction in the minority carrier diffusion length. The reduction is faster for heavier ions because they introduce more defects and thus decrease the diffusion length more rapidly.

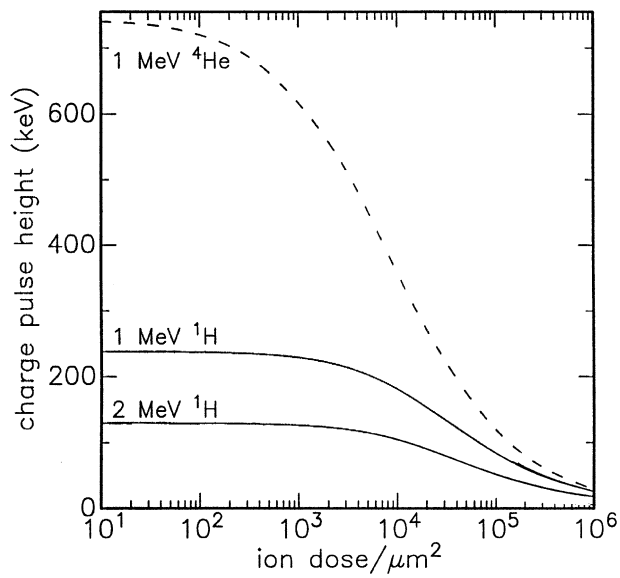


Figure 14. Calculated charge pulse height with cumulative ion irradiation with 1- and 2-MeV protons (solid lines) and 1-MeV ^{4}He ions (dashed lines). An initial diffusion length of 5 μm is assumed. Except where indicated, reprinted from Breese (1993) with permission. Copyright 1993 American Institute of Physics.

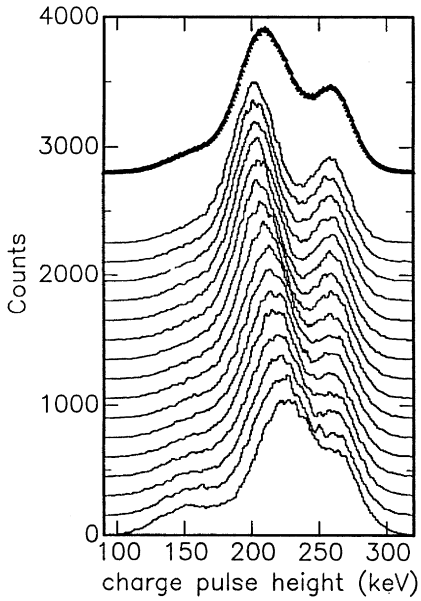


Figure 15. Charge pulse height spectra from a 80×80 μm^2 area produced using 2-MeV ^4He ions. The measured charge pulse height spectra with cumulative sequential dose increments of 9 ions/μm^2 are offset vertically by a fixed arbitrary amount for clarity. The uppermost spectrum is the cumulative spectrum from which the sequential spectra were extracted. Reprinted from Breese et al. (1995) with permission. Copyright 1995 American Institute of Physics.

Therefore, although heavier ions give a larger measured IBIC pulse than similar energy light ions, this advantage might be outweighed by the more rapid rate of damage. This is another consideration when choosing the optimum type of ion for use in IBIC or SEU analysis.

The best type of data acquisition system to use for IBIC and SEU microscopy is an event-by-event system (O'Brien et al., 1993), which stores the measured charge pulses along with their position coordinates for subsequent processing and analysis. The specific advantage of this is that the measured charge pulse height data set can be sliced into sequential dose increments so that the evolution of the charge pulse height spectrum and image contrast from different parts of the scanned area can be examined with the cumulative ion dose.

Figure 15 shows the IBIC pulse height spectra from a 80×80 μm^2 area of a different microelectronic device. Here 2-MeV ^4He ions are used, and the resultant spectrum is shown at the start of data collection and also after fixed sequential dose increments. It can be seen that different parts of the charge pulse height spectrum behave in different ways; the left-hand peak decreases in size with increasing beam dose, whereas the right-hand peak remains the same size. The explanation for this type of behavior is given below. Here it suffices to say that even though the charge pulse height in adjacent pixels may vary differently with cumulative ion dose, i.e., the contrast may change, the charge pulses can still be used to generate an IBIC image when an event-by-event acquisition system is used.

Minimizing the Effects of Ion-Induced Damage

Ion-induced damage of the semiconductor substrate decreases the measured amount of charge because the increased recombination of the slowly diffusing charge carriers causes a reduction in the carrier diffusion length, as described below. The mechanism by which the amount of charge measured from the depletion region reduces with ion-induced damage is more complex. Charge carriers generated in the depletion region have a much lower recombination probability than those generated in the substrate because of the associated electric field. This accelerates the carriers in opposite directions and reduces the probability of them becoming trapped and recombining at any defects present.

Figure 16 shows a schematic of the final locations of ions with a fixed energy penetrating through an increasingly thick surface layer into a narrow depletion region and then into the substrate. For a thin surface layer, the ions penetrate deep into the substrate, and most of the measured charge is due to carriers diffusing from the substrate to the collecting junction. Since the distribution of the nonionizing energy loss along the ion trajectory follows a similar trend as the electronic energy loss, most of the defects are created in the substrate, where they have a large effect on the measured charge pulse height. For a thicker surface layer, all the ions may be stopped in the depletion layer at the semiconductor surface. The maximum defect generation rate also occurs here, where they have a much smaller effect on the measured charge pulse height since charge carriers are much less likely to be trapped here. For a still thicker surface layer (*c*), most of the ions are stopped in the surface layers and only a few penetrate into the depletion region. If the ion does not

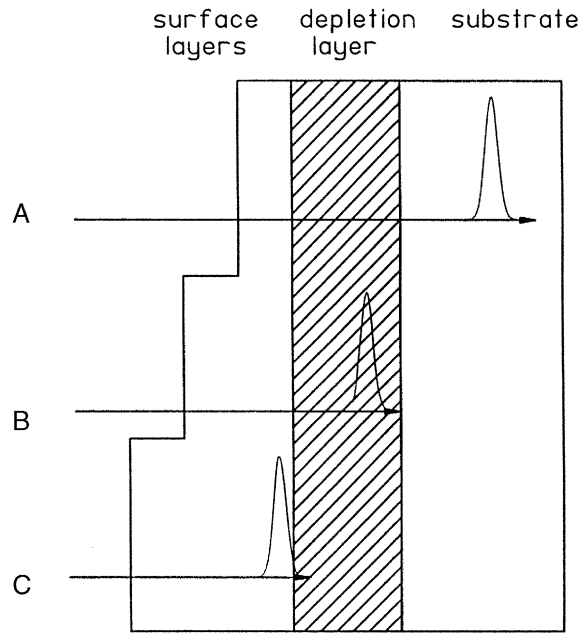


Figure 16. Schematic showing the locations of the end-of-range defects (approximated as Gaussian distributions) for increasing surface layer thicknesses from (**A**) to (**C**).

generate enough charge carriers to be measured above the noise level, then it will not be detected.

The different behavior of the two charge peaks in Figure 15 is due to the effect shown in Figure 16. The left-hand peak is due to ions that penetrate into the substrate and so reduces in height rapidly, whereas the right-hand peak is due to ions that are stopped in depletion regions present.

PROBLEMS

Two experimental problems commonly encountered during IBIC experiments are difficulties in producing measurable charge pulses and the associated large noise level present from the microelectronic device or semiconductor wafer.

Absence of Observed Charge Pulses

It is important to test the electrical contacts to the IBIC or SEU sample before irradiating them in the microprobe. Otherwise there is a high probability of unintentionally damaging the sample by mistaking the absence of measured IBIC pulses for the absence of a beam irradiating it. First the sample's sensitivity to light should be checked by measuring the current flowing through it. Light generates charge carriers in a similar manner to ions, so on most types of samples there should be a large increase in the measured current with light shining on them. Next the sample in the shielded microprobe chamber should be connected and a determination made as to whether the noise level measured with an oscilloscope (with the chamber lights off) is small enough to resolve charge pulses. Excessively large noise level might be due to poor barrier preparation or wrong connections to the microelectronic device. A further wise step is to check that charge pulses generated by α-particles from a test source can be resolved from the noise level. If not, then the use of heavier ions to produce larger charge pulses should be considered. Also, the amplifier polarity should be set so as to give positive pulses into the data acquisition system—a commonly made mistake.

Methods of Noise Reduction

If the noise level in an SEU or IBIC sample is too high, then it is difficult to sufficiently resolve the pulses. Great care must therefore be taken to reduce the noise level as much as possible before the start of data collection.

The noise level typically encountered with IBIC and SEU samples is $\sim$1 mV at the preamplifier output for microcircuit devices. Noise in charge-sensitive preamplifiers has three main sources: the input field effect transistor, the input capacitance, and the preamplifier resistance. The noise contribution from the input capacitance increases at a typical rate of 15 to 20 eV/pF, so any excess capacitance should be removed. Leads between the preamplifier and the sample should be as short as possible since the capacitance increases with lead length, and ideally the preamplifier should be mounted inside the target chamber to be as close as possible to the sample. For IBIC analysis of semiconductor wafers, the area of the Schottky barrier or the number of pins connected to the preamplifier should be as small as possible to minimize the total active area, as this reduces the capacitance. The leads should be well screened, and the sample should be isolated from earth loops, circulating currents, and radio frequency pickup from other components of the microprobe electronics. For SEU analysis, which typically requires a larger number of feedthroughs and connections to the device in the microprobe chamber in order to check its logic state after every incident ion, it is even more important that no external noise is introduced that might affect the device.

The noise level in the measured charge pulse height spectrum also depends on the amplifier time constant. A typical value of 1 μs is frequently used for the best signal-to-noise ratio. The optimum value depends on both the current flowing in the sample and its capacitance. Decreasing the time constant too much can decrease the measured charge pulse height because not all the trapped carriers may have detrapped, so it may be necessary to make a compromise between the best signal-to-noise ratio and quantifying the diffusion length from the measured charge pulse height. Once all excess input capacitance to the preamplifier has been eliminated, the remaining measured noise level is dominated by thermal generation of charge carriers. Cooling the material reduces the thermally generated noise level. The charge pulse height spectra measured from Schottky barriers for IBIC microscopy have been very noisy to date, because large-area barriers have been used in order to manually attach a thin wire to the front surface using silver paint. Since both the capacitance and the thermally generated noise contributions increase with the barrier area, this should ideally be as small as possible.

LITERATURE CITED

Acciarri, M., Binetti, S., Garavaglia, M., and Pizzini, S. 1996. Detection of junction failures and other defects in silicon and III–V devices using the LBIC technique in lateral configuration. *Mat. Sci. Eng.* B42:208–212.

Angell, D., Marsh B. B., Cue, N., and Miao, J. W. 1989. Charge collection microscopy: Imaging defects in semiconductors with a positive ion microbeam. *Nucl. Instrum. Methods* B44:172–178.

Beckman, D. R., Saint, A., Gonon, P., Jamieson, D. N., Prawer, S., and Kalish, R. 1997. Spatially resolved imaging of charge collection efficiency in polycrystalline CVD diamond by the use of ion beam induced current. *Nucl. Instrum. Methods* B130:518–523.

Bertolini, G. and Coche, A. 1968. Semiconductor Detectors. North-Holland, Amsterdam.

Breese, M. B. H. 1993. A theory of ion beam induced charge collection. *J. Appl. Phys.* 74(6):3789–3799.

Breese, M. B. H., Grime, G. W., and Watt, F. 1992. Microcircuit imaging using ion beam induced charge. *J. Appl. Phys.* 72(6):2097–2105.

Breese, M. B. H, King, P. J. C., Grime, G. W., and Wilshaw, P. R. 1993a. Dislocation imaging using ion beam induced charge. *Appl. Phys. Lett.* 62:3309–3311.

Breese, M. B. H., Grime, G. W., and Dellith, M. 1993b. The effects of ion induced damage on IBIC images. *Nucl. Instrum. Methods* B77:332–338.

Breese, M. B. H., Laird, J. S., Moloney, G. R., Saint, A., and Jamieson, D. N. 1994. High signal to noise ion beam induced charge images. *Appl. Phys. Lett.* 64(15):1962–1964.

Breese, M. B. H., Saint, A., Sexton, F. W., Schone, H., Doyle, B. L., Laird, J. S., and Legge, G. J. F. 1995. Optimisation of ion beam induced charge microscopy for the analysis of integrated circuits. *J. Appl. Phys.* 77(8):3734–3741.

Breese, M. B. H., Jamieson, D. N., and King, P. J. C. 1996. Materials Analysis Using a Nuclear Microprobe. John Wiley & Sons, New York.

Buchner, S., Knudson, A. R., Kang, K., Campbell, A. B. 1988. Charge collection from focused pico-second laser pulses. *IEEE Trans. Nucl. Sci.* NS-35:1517–1522.

Chi, J. C. and Gatos, H. C. 1979. Determination of dopant concentration diffusion lengths and lifetime variations in silicon by scanning electron microscopy. *J. Appl. Phys.* 50(5):3433–3440.

Dodd, P. E. 1996. Device simulation of charge collection and single-event upset. *IEEE Trans. Nucl. Sci.* NS-43(2):561–575.

Dodd, P. E., Sexton, F. W., and Winokur, P. S. 1994. Three-dimensional simulation of charge collection and multiple-bit upset in Si devices. *IEEE Trans. Nucl. Sci.* NS-41(6):2005–2017.

Donolato, C., Nipoti, R., Govoni, D., Egeni, G. P., Rudello, V., and Rossi, P. 1996. Images of grain boundaries in polycrystalline silicon solar cells by electron and ion beam induced charge collection. *Mat. Sci. Eng.* B42:306–310.

Doyle, B. L., Horn, K. M., Walsh, D. S., and Sexton, F. W. 1992. Single event upset imaging with a nuclear microprobe. *Nucl. Instrum. Methods* B64:313–320.

Doyle, B. L., Maggiore, C. J., and Bench. G. 1997. *In* Proceedings of the Fifth International Conference on Nuclear Microprobe Technology and Applications (H.H. Andersen and L. Rehn, eds.). pp. 1–750. Elsevier/North-Holland, Amsterdam.

England, J. B. A. 1974. Techniques in Nuclear Structure Physics. MacMillan Press, London.

Geppert, L. M., Bapst, U., Heidel, D. F., and Jenkins, K. A. 1991. Ion Microbeam Probing of Sense Amplifiers to Analyse Single Event Upsets in a CMOS DRAM. *IEEE Trans. Solid-State Circuits* 26(2):132–134.

Grime, G. W. and Watt, F. 1984. Beam Optics of Quadrupole Probe-forming Systems. Adam Hilger, Bristol.

Holt, D. B., Muir, M. D., Grant, P. R., and Boswarva, I. M. 1974. Quantitative Scanning Electron Microscopy. Academic Press, London.

Horn, K. M., Doyle, B. L., Walsh, D. S., and Sexton, F. W. 1992. Nuclear microprobe imaging of single event upsets. *IEEE Trans. Nucl. Sci.* 39(1):7–12.

Horn, K. M., Doyle, B. L., Sexton, F. W., Laird, J. S., Saint, A., Cholewa, M., and Legge, G. J. F. 1993. Ion beam induced charge collection (IBICC) microscopy of ICs: Relation to single event upsets (SEUs). *Nucl. Instrum. Methods* B77:355–361.

Kittel, C. 1986. Introduction to Solid State Physics. John Wiley & Sons, Singapore.

Klein, C. A. 1968. Bandgap dependence and related features of radiation ionisation energies in semiconductors. *J. Appl. Phys.* 39:2029–2038.

Knudson, A. R. and Campbell, A. B. 1983. Investigation of soft upsets in integrated circuit memories and charge collection and charge collection in semiconductor structures by the use of an ion microbeam. *Nucl. Instrum. Methods* 218:625–631.

Kolachina, S., Ong, V. K. S., Chan, S. H., Phang, J. C. H., Osipowicz, T., and Watt, F. 1996. Unconnected junction contrast in ion beam induced charge microscopy. *Appl. Phys. Lett.* 68(4):532–534.

Kreskovsky, J. P. and Grubin, H. L. 1986. Numerical simulation of charge collection in two- and three-dimensional silicon diodes—a comparison. *Solid-State Electron.* 29(5):505–518.

Laird, J. S., Bardos, R. A., Jagadish, C., Jamieson, D. N., and Legge, G. J. F. 1999. Scanning ion deep level transient spectro-scopy. *Nucl. Instrum. Method*s B158:464–469.

Lang, D. V. 1974. Deep level transient spectroscopy: A new method to characterize traps in semiconductors. *J. Appl. Phys.* 45(7):3023–3032.

Leamy, H. J. 1982. Charge collection scanning electron microscopy. *J. Appl. Phys.* 53:R51-R80.

Manfredotti, C., Fizzotti, F., Vittone, E., Boero, M., Polesello, P., Galassini, S., Jaksic, M., Fazinic, S., and Bogdanovic, I. 1995. IBIC Investigations on CVD diamond. *Nucl. Instrum. Methods* B100:133–140.

Manfredotti, C., Fizzotti, F., Vittone, E., Boero, M., Polesello, P., Galassini, S., Jaksic, M., Fazinic, S., and Bogdanovic, I. 1996. Study of physical and chemical inhomogeneities in semiconducting and insulating materials by a combined use of micro-PIXE and micro-IBIC. *Nucl. Instrum. Methods* B109/110:555–562.

May, T. C. and Woods, M. H. 1979. Alpha-particle-induced soft errors in dynamic memories. *IEEE Trans. Electron Devices* ED-26:2–9.

McLean, F. B. and Oldham, T. R. 1982. Charge funneling in n- and p-type Si substrates. *IEEE Trans. Nucl. Sci.* NS-29:2018–2023.

Nashiyama, I., Hirao, T., Kamiya, T., Yutoh, H., Nishijima, T., and Sekiguti, H., 1993. Single-event current transients induced by high energy ion microbeams. *IEEE Trans. Nucl. Sci.* NS-40(6):1935–1940.

Nipoti, R., Donolato, C., Govoni, D., Rossi, P., Egeni, E., and Rudello, V. 1998. A study of He^+ ion induced damage in silicon by quantitative analysis of collection efficiency data. *Nucl. Instrum. Methods*. B138:1340–1344.

O'Brien, P. M., Moloney, G., O'Connor, A., and Legge, G. J. F. 1993. A versatile system for the rapid collection, handling, and graphics of multidimensional data. *Nucl. Instrum. Methods* B77:52–55.

Piqueras, J., Fernandez, P., and Mendez, B. 1996. *In* Proceedings of the Fourth Workshop on Beam Injection Assessment of Defects in Semiconductors (M. Balkanski, H. Kamimura, and S. Mahajan, eds.). pp. 1–310. Elsevier/North-Holland, Amsterdam.

Schone, H., Walsh, D. S., Sexton, F. W., Doyle, B. L., Dodd, P. E., Aurand, J. F., Flores, R. S., and Wing, N. 1998. Time-resolved ion beam induced charge collection in microelectronics. *IEEE Trans. Nuc. Sci.* 45(6):2544–2549.

Sexton, F. W., Horn, K. M., Doyle, B. L., Laird, J. S., Cholewa, M., Saint, A., and Legge, G. J. F. 1993a. Relationship between IBICC imaging and SEU in CMOS ICs. *IEEE Trans. Nucl. Sci.* NS-40:1787–1794.

Sexton, F. W., Horn, K. M., Doyle, B. L., Laird, J. S., Saint, A., Cholewa, M., and Legge, G. J. F. 1993b. Ion beam induced charge collection imaging of CMOS ICs. *Nucl. Instrum. Methods* B79:436–442.

Sexton, F. W., Horn, K. M., Doyle, B. L., Shaneyfelt, M. R., and Meisenheimer, T. L. 1995. Effects of ion damage on IBICC and SEU imaging. *IEEE Trans. Nucl. Sci.* 42(6):1940–1947.

Sze, S. M. 1981. Physics of Semiconductor Devices. John Wiley & Sons, New York.

Sze, S. M. 1988. VLSI Technology. McGraw-Hill, Singapore.

Wagner, R. S., Bordes, N., Bradley, J. M., Maggiore, C., Knudson, A. R., and Campbell, A. B. 1988. Alpha-, boron-, silicon- and

iron-ion induced current transients. *IEEE Trans. Nucl. Sci.* NS-35(4):1578–1584.

Watt, F. and Grime, G. W. 1987. Principles and Applications of High Energy Ion Microbeams. Adam Hilger, Bristol.

Wu, C. J. and Wittry, D. B. 1978. Investigation of minority carrier diffusion lengths by electron bombardment of Schottky barriers. *J. Appl. Phys.* 49(5):2827–2836.

KEY REFERENCES

Breese et al., 1996. See above.

A comprehensive description of IBIC analysis with many examples. Explains in detail the one-dimensional theory of IBIC. Describes other SEU work.

Doyle et al., 1997. See above.

Supplies the most up-to-date papers on both technological aspects and applications of IBIC and SEU microscopy.

MARK B. H. BREESE
National University of Singapore
Kent Ridge, Singapore

TRACE ELEMENT ACCELERATOR MASS SPECTROMETRY

INTRODUCTION

Accelerator mass spectrometry (AMS; Jull et al., 1997) is a relatively new analytical technique that is being used in over 40 laboratories worldwide for the measurements of cosmogenic long-lived radioisotopes, as a tracer in biomedical applications, and for trace element analysis of stable isotopes. The use of particle accelerators, along with mass spectrometric methods, has allowed low-concentration measurements of small-volume samples. The use of AMS to directly count ions, rather than measure radiation from slow radioactive decay processes in larger samples, has resulted in sensitivities of one part in 10^{15} (Elmore and Phillips, 1987; Jull et al., 1997). The ions are accelerated to mega-electron-volt energies and can be detected with 100% efficiency in particle detectors. The improved sensitivities and smaller sample sizes have resulted in a wide variety of applications in anthropology, archaeology, astrophysics, biomedical sciences, climatology, ecology, geology, glaciology, hydrology, materials science, nuclear physics, oceanography, sedimentology, terrestrial studies, and volcanology, among others (Jull et al., 1997). Accelerator mass spectrometry has opened new areas of research in the characterization of trace elements in materials.

This unit will focus on the technique of trace element accelerator mass spectrometry (TEAMS) and its applications in the analysis of stable isotopes in materials. This technique has been labeled by different groups as TEAMS (McDaniel et al., 1992; Datar et al., 1996, 1997a,b; Grabowski et al., 1997; McDaniel et al., 1998a,b), accelerator secondary-ion mass spectrometry (SIMS; Purser et al., 1979; Anthony et al., 1994a,b; Ender et al., 1997a,b,c,d), secondary-ion accelerator mass spectrometry (SIAMS; Ender et al., 1997a,b; Massonet et al., 1997), super-SIMS (Anthony et al., 1986), atomic mass spectrometry (Anthony et al., 1990b; McDaniel et al., 1990a), AMS (Anthony et al., 1985; Wilson et al., 1997b), and microbeam AMS (Sie and Suter, 1994; Sie et al., 1997a,c; Niklaus et al., 1997; Sie et al., 1999).

Trace element AMS for stable isotopes is very similar to SIMS, which is one of the most sensitive techniques for determination of elemental composition in materials. As with SIMS, TEAMS is a destructive technique that may be used in a static mode for bulk measurements of elemental composition or may be used in a dynamic mode to profile elemental composition as a function of depth in the material. As with SIMS, TEAMS employs a positively charged, primary-ion beam at energies of 3 to 30 keV to sputter secondary ions, atoms, and molecules from the surface or near surface of a material. Of course, most of the particles sputtered from the surface are electrically neutral and cannot be analyzed at all unless one uses a laser or some other means to ionize the neutrals, e.g., sputter-initiated resonance ionization spectrometry (SIRIS). With SIMS, either positive or negative secondary ions may be extracted from the surface through a few kilovolts and then magnetically analyzed to determine composition. This is an advantage because some atoms more readily form positive ions while others more readily form negative ions (Wilson et al., 1989). A disadvantage of SIMS is that both atomic and molecular ions are produced and, in many cases, the molecular mass is nearly the same as the atomic mass of interest. For example, ^{30}SiH has the same nominal mass as ^{31}P. Operating SIMS in a high-mass-resolution mode can reduce the molecular interferences by reducing the slit openings into the magnetic spectrometer, but this is at the cost of reduced signal strength and therefore reduced sensitivity (Wilson et al., 1989).

Trace element AMS employs a tandem electrostatic accelerator with the center terminal at a positive potential. The secondary ions from the sample, which are injected into the tandem accelerator, must be negatively charged to be accelerated to the terminal. At the terminal, the atomic and molecular secondary ions from the sample are stripped of a number of electrons in a gas or foil stripper. The electron stripping causes molecular break-up through a Coulomb explosion into elemental ions, which are then accelerated to higher energy and exit the accelerator. If the diatomic or triatomic molecular ion is stripped to a 3^+ or greater charge state, almost all molecules break up into elemental constituents (Weathers et al., 1991). With TEAMS, the resulting atomic ions, which are in a number of different charge states with correspondingly different energies, are extracted from the high-energy end of the tandem accelerator. The atomic ions are then magnetically analyzed for momentum/charge (mv/q) and electrostatically analyzed for energy/charge (E/q), which together give mass/charge (m/q) discrimination. Here, m is the mass, ν is the velocity, E is the kinetic energy, and q is the charge of the ion. To eliminate atomic interferences from possible break-up fragments that have the same m/q (e.g., $^{28}Si^{2+}$ from Si_2 break-up vs. $^{56}Fe^{4+}$), the total energy of the ion is measured in a surface barrier detector or ionization chamber.

Sensitivities for TEAMS measurements vary for different elements. These sensitivities depend on (1) the ability to form a negative atomic ion or a negatively charged molecule that can be accelerated to the terminal and then broken apart; (2) the charge state used for the measurements since the charge state ($\geq 3^+$) must be chosen to remove molecular interferences; and (3) transmission of the ion from the sample through the accelerator and beamlines to the detector. By calibrating the secondary-ion yield against reference standards and by measuring the depth of the sputtered crater, a quantified depth profile can be obtained. The sensitivity or detection limit depends on the amount or volume of material available to analyze. The sensitivity can be much better for bulk measurements, where the element of interest is uniformly dispersed throughout the material. Then, the sample can be run almost indefinitely until good statistics have been obtained. Typically, sensitivities of 0.1 part per billion (ppb) can be obtained in the depth-profiling mode (McDaniel et al., 1998a,b; Datar et al., 2000) and 5 parts per trillion (ppt) in the bulk analysis mode. Theoretical sensitivities of 1 ppt are expected under ideal conditions. Because sensitivities are so high, it is sometimes difficult to obtain suitable reference standards.

Complementary, Competitive, and Alternative Methods

There are a number of complementary trace analysis techniques of lower sensitivity and lower cost than TEAMS. The logical choice for TEAMS measurements is as a complement to other measurements provided by photon, electron, proton, or heavier ions. Depending upon the concentrations of the trace elements of interest in the sample, one may want to use one of the lower cost methods which can have sensitivities from major-element down to sub-ppb. Since TEAMS can be used for both bulk and depth-profiling measurements, some of the techniques discussed below are primarily for bulk analysis [inductively coupled plasma mass spectrometry and neutron activation analysis (ICPMS, NAA)], and others may be used for depth profile analysis (SIMS, SIRIS). Other competitive techniques that involve high-energy ion beam analysis (IBA) are described in the section introduction (HIGH-ENERGY ION BEAM ANALYSIS).

Secondary-Ion Mass Spectrometry. A complementary, alternative method to TEAMS, and one of the most powerful techniques for the characterization of materials, is SIMS (Benninghoven et al., 1987; Wilson et al., 1989; Lu, 1997; Gillen et al., 1998). In fact, the front end of TEAMS is like a SIMS system. In the next few paragraphs, the operating principles and analytical capabilities of SIMS are reviewed both as an introduction to TEAMS and as a competitive and complementary technique. An excellent SIMS tutorial is provided at the web site of Charles Evans and Associates: http://www.cea.com/tutorial.htm.

With SIMS, a positive primary-ion beam, which can be O_2^+, O^+, Cs^+, Ar^+, Xe^+, Ga^+, or any number of other species, typically 3 to 30 keV in energy, sputters secondary ions from the surface of a material. The bombarding primary ion produces atomic and molecular secondary ions and neutrals of the sample material, resputtered primary ions, and electrons and photons. Typically, O_2^+ is used if the secondary ions are more likely to be electropositive, and Cs^+ is typically used if the secondary ions are more likely to be electronegative. The choice of sputter ion can cause a change in secondary-ion yield of several orders of magnitude. Some elements do not form negative ions (e.g., N, Mg, Mn, Zn, Cd, Hg, and the noble gases), and many elements form negative ions very weakly. Of course, most of the particles sputtered from the surface are electrically neutral and cannot be analyzed at all unless one uses a laser or some other means to ionize the neutrals. Sputter-initiated resonance ionization spectrometry is such a method. The small fraction that are ionized depend strongly upon the matrix from which they are sputtered. The sputtering process is not just with the surface layer but consists of the implantation of the primary species into the sample at depths of 1 to 10 nm and the removal of the atoms or molecules from the surface by the energy loss of the primary ion through cascade collisions. Typical sputter rates in SIMS vary between 0.5 and 5 nm/s and depend upon beam intensity, sample material, and crystal orientation. The secondary ions, which are extracted through a potential of a few kilovolts and mass analyzed, are representative of the composition of the sample being analyzed. The sputter yield is the ratio of the number of atoms sputtered to the number of primary ions impinging on the sample and can range from 5 to 15. An advantage of SIMS is that secondary-ion yields can be measured over a dynamic range from matrix atoms to ppb.

The SIMS detection limits for most trace element are 1×10^{12} to 1×10^{16} atoms/cm^3. Factors that affect sensitivity and detection limits are ionization efficiencies or relative sensitivity factors (RSFs; Wilson et al., 1989); dark currents from electron multiplier detectors, which limits the minimum count rate; vacuum system elements and elements from SIMS components, which produce background counts; and mass interferences. Because both atomic and molecular ions are sputtered from the surface and analyzed with SIMS, the detection sensitivity for some elements is limited by molecular interferences. For example, in Si, ^{56}Fe is masked by $(^{28}Si)_2$; ^{31}P is masked by ^{30}SiH, $^{29}SiH_2$, and $^{28}SiH_3$; and ^{32}S is masked by $(^{16}O)_2$, among others. However, the molecular species do allow chemical information to be extracted from a sample in some cases.

The SIMS instruments may use magnetic sector, radio frequency (RF) quadrupole, or time-of-flight (TOF) mass analyzers. The secondary ions must also be energy analyzed. The magnetic sector mass analyzer separates the secondary ions by their momentum/charge (mv/q). Because of the high extraction voltage used for secondary ions, magnetic sector analyzers have high mass transmission. Magnetic sector analyzers may be used in a high-mass-resolution mode to separate molecular masses that interfere with an atomic mass of interest. This is usually accomplished at the expense of signal strength and therefore reduced sensitivity. RF quadrupole mass analyzers are smaller, more compact, and easier to use. A mass spectrum is produced by applying RF and direct current electric fields to a quadrupole rod assembly. Quadrupole SIMS instruments have loss of transmission and mass resolution

with increasing mass and also cannot operate in a high-mass-resolution mode. The TOF mass analyzers, which are used with pulsed secondary-ion beams, have high transmission and mass resolution. TOF-SIMS provides excellent surface sensitivity, and many commercial instruments couple this with submicrometer primary optics for analysis.

The SIMS data may be taken in either a bulk analysis or a depth-profiling mode. Bulk analysis allows scanning of the mass spectrometer while a sample is being sputtered away. Bulk analysis may also provide greater sensitivity since a larger volume of material is available for analysis. Bulk measurements are mainly useful for uniformly doped impurities or for surface analysis. Depth profiling allows the concentration of impurities to be determined as a function of depth beneath the surface. One or more masses are monitored as the mass spectrometer is sequentially switched. For the mass of interest, the detector signal comes from increasingly greater depths in the sample. Accurate depth profiles require uniform bombardment of the sample analysis area (i.e., a flat bottom crater) with no contribution from the crater walls, adjacent sample surfaces, or nearby instrument surfaces. Usually the primary-ion beam, which may range in size from 1 μm to hundreds of micrometers, is rastered over a region of the sample. Only data from the center part of the scan are retained for analysis to reduce contributions from crater walls.

Secondary-ion yields for elements can vary by many orders of magnitude for a given material or matrix and even from matrix to matrix. It is therefore difficult to quantify SIMS results without using calibration standards. Wilson et al. (1989) have provided tables of RSFs for many elements for both O_2^+ and Cs^+ bombardment of different substrate materials, which help quantify SIMS data.

Sputter-Initiated Resonance Ionization Spectrometry. Sputter-initiated resonance ionization spectrometry is a relatively young and powerful method for selectively ionizing and analyzing materials (Winograd et al., 1982; Pellin et al., 1987, 1990; Bigelow et al., 1999, 2001). As with TEAMS and SIMS, SIRIS also removes atoms and molecules from a sample by sputtering with ions at kiloelectron-volt energies. Again, the majority of the sputtered atoms and molecules are neutral and only a small fraction is ionized. This fraction depends strongly upon the matrix of the material. TEAMS and SIMS mass analyze this ionized fraction. SIRIS uses lasers to ionize the much larger sputtered neutral fraction for a particular species with essentially 100% efficiency and then detects the post-ionized atoms. With SIRIS, two or more lasers are used to first resonantly excite an atom and then ionize it from the excited state. By suitably tuning the lasers, only one atomic species is ionized at a time. Therefore, even without subsequent mass analysis, SIRIS is highly selective and efficient. The yield ratio of atoms detected to atoms sputtered can approach 0.1, which may be several orders of magnitude better than SIMS or TEAMS. The SIRIS technique has been demonstrated to have bulk detection limits as low as 100 ppt (Pellin et al., 1990). Because ≅80% of the sputtered atoms originate from the surface monolayer of the sample and ≅95% from the top two surface monolayers (Hubbard et al., 1989), SIRIS has excellent surface sensitivity. Because of the high useful yield, measurements may be made while sputtering sub-monolayer target thicknesses, making SIRIS essentially nondestructive. However, SIRIS may also be used to measure atoms below the surface if the primary-ion beam is allowed to sputter the sample for a longer time. One limitation of SIRIS is that it is very element specific, and multielement work may require several laser schemes.

Neutron Activation Analysis. Other mature methods that are readily available and provide rapid and economical, multielement, ppb-level detection of trace elements include NAA (Kruger, 1971). While NAA has excellent sensitivity in some cases (e.g., ^{197}Au in Si), the sensitivity of NAA is limited by the neutron capture cross sections, which vary from element to element. In addition, NAA may be primarily useful for Si since irradiation of other compound semiconductor materials like GaAs and HgCdTe will produce higher background radiation from the additional substrate elements.

Inductively Coupled Plasma Mass Spectrometry. Inductively coupled plasma mass spectrometry is a sensitive bulk materials analysis technique (10^{14} atoms/cm^3 or 10 ppb; Houk, 1986; Baumann and Pavel, 1989). Essentially the entire sample is vaporized, and the resulting plasma is analyzed by mass spectrometry. However, there may be a concern about incomplete elemental dissolution and contamination during sample preparation.

Neutron Activation–Accelerator Mass Spectrometry. Neutron activation (NA) followed by AMS has been used to study light impurities in Si (Elmore et al., 1989; Gove et al., 1990). Neutron activation is used to produce long-lived radioisotopes, which generally have lower concentrations in the environment. After activation, AMS can provide more sensitive background-free measurements of the neutron activation products.

For example, neutron activation to convert stable ^{35}Cl to the long-lived radioisotope ^{36}Cl followed by conventional AMS has been used to study both the diffusion of Cl in Si (Datar et al., 1995) and the impurity Cl concentration in Si. This gets around the problem of having to ensure source cleanliness since the ambient ^{36}Cl has an abundance ratio of 10^{-15} vis-à-vis stable Cl. This was especially useful in the Rochester measurements (Datar et al., 1995) since the ion source used for the measurements was heavily contaminated with stable Cl. This technique can be used in principle for any isotope that produces a suitable, relatively long lived isotope by thermal neutron activation. The sensitivity of the measurement can be increased by increasing the total thermal neutron flux incident on the samples.

Choosing the Proper Technique. From the above discussion, the basic criteria for choosing TEAMS over other techniques should be based upon the need to determine elemental concentrations in a sample for those elements

that need high sensitivity or have a molecular interference. Because of the cost and difficulty of a TEAMS measurement, TEAMS should not be used for all elements if the element can be analyzed by an alternative and less expensive method such as Rutherford backscattering spectrometry (RBS; see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS), particle-induced X-ray emission (PIXE; see PARTICLE-INDUCED X-RAY EMISSION), or nuclear reaction analysis (NRA; see NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION). In some cases, chemical information may be needed for a sample, and SIMS would be more useful. For some elements, molecular interferences are not a problem, and SIMS would be easier and less expensive. In addition, for low-mass-resolution applications, SIMS may give lower detection limits due to higher ionization rates with O_2^+ primary-ion bombardment.

PRINCIPLES OF THE METHOD

The Trace Element Accelerator Mass Spectrometry Facility

The TEAMS facility consists of three main components: (1) the ultraclean ion source for the generation of negative secondary ions from a material by primary-ion sputtering, much like in SIMS; (2) the tandem accelerator, which accelerates the secondary ions to mega-electron-volt energies and strips electrons to remove molecular interferences; and (3) the detector system, which includes the high-energy beam transport line with magnetic and electrostatic analysis, and the particle total energy detection system. Different laboratories around the world are employing a number of different approaches to TEAMS, and attempts will be made to contrast these different approaches in this unit. However, the following description of TEAMS will be primarily the University of North Texas (UNT) facility.

Ultraclean Ion Source for Negatively Charged Secondary-Ion Generation

The ion source is designed to produce secondary ions from the sample while minimizing contaminant ions produced from the ion source hardware and from previous samples. The ion source is discussed in detail under Practical Aspects of the Method.

Secondary-Ion Acceleration and Electron-Stripping System

With the sample biased at -13 kV, the negatively charged secondary ions are extracted from the chamber at the energy of 13 keV and passed through a 45° electrostatic analyzer (ESA). The ESA discriminates against energy/charge (E/q) with $E/\Delta E = 75$ for a 2-mm slit width. The ESA passes all masses originating from the sample surface that have the correct extraction energy. Secondary ions originating from other surfaces in the ion source chamber are prevented from being passed down the extraction beamline and into the tandem accelerator. Since the negative ion yield is at most 1% to 2%, a typical secondary-ion beam after the ESA due to 900 nA of 23-keV Cs^+ (10 keV from Cs oven plus 13 keV from negative sample bias) is

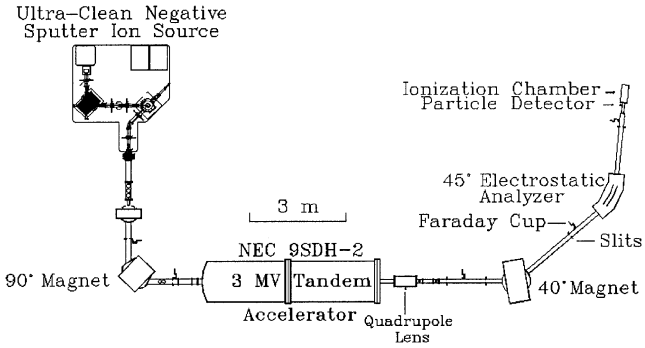


Figure 1. The TEAMS facility in the Ion Beam Modification and Analysis Laboratory at the University of North Texas.

≅50 to 60 nA of ^{28}Si through a 2-mm slit opening. After the ESA, the negative ion beam is passed through a 90° dipole magnet and momentum/charge analyzed (mv/q) with $m/\Delta m = 500$ for a 2-mm slit width. A mass scan after the 90° magnet would be equivalent to a SIMS analysis and would include molecular as well as atomic negative ions. After the 90° magnet, the ^{28}Si secondary-ion beam is ≅12 to 15 nA. Figure 1 shows the TEAMS system at UNT.

The 13-keV negative ions are injected into the tandem accelerator and accelerated to the terminal, which typically is operated at 1.7 MV for TEAMS. For an ion with charge 1^-, injection voltage $U = 13$ kV, and terminal voltage $V = 1.7$ MV, the ion energy is given by

$$E = e(U + V) \tag{1}$$

The higher energy negative ions (1.713 MeV) are then efficiently stripped of one or more electrons in the terminal by an N_2 gas or carbon foil stripper. The gas or foil stripper is sufficiently thick to produce atomic and molecular ions in a distribution of positive charge states q (Betz, 1972; Wiebert et al., 1994). All of the charge states produced will be accelerated again from the terminal down the high-energy side of the tandem accelerator, producing a number of different energy ions. The 1^+ ions will gain another 1.7 MeV of energy; the 2^+ ions will gain another 3.4 MeV of energy, and so on for each subsequent charge state q. The mass m of the ion exiting the accelerator may be less than the mass m_0 of the ion injected (if dissociation occurs in the terminal) or equal to the injected mass if no dissociation occurs. The total energy of the ion of mass m after the accelerator is given by

$$E = e(m/m_0)(U + V) + eqV \tag{2}$$

where the first term is the residual energy from the dissociated ion in the terminal and the second term is the post-stripping energy. This energy neglects the energy loss of the ion in the stripper gas, which can vary from 0.1 keV for H to 7.0 keV for Si and must be taken into account in subsequent analysis (Arrale et al., 1991). The N_2 gas pressure may be used to optimize the production of the charge state of interest. For 1.7 MV terminal voltage, the most populated charge state is 3^+ for masses in the range

from 20 to 60 amu. The charge state chosen after acceleration and for further analysis depends upon possible interferences with other m/q ions, mainly from matrix elements in the sample and upon the charge state needed for molecular break-up. Molecules in a 1^+ or 2^+ charge state will still exist and will be accelerated to the higher energies. There are several molecules that are known to be at least metastable in a 3^+ charge state. While most of these molecular ions are larger clusters or organic molecules (Burdick et al., 1986; Saunders, 1989), a few are relatively small trimers CS_2 and CSe_2 and the dimer S_2 (Morvay and Cornides, 1984). The small dimers B_2 and AlO have been observed to remain in a 3^+ metastable state for many microseconds, enough time to travel all the way from the terminal to the detector (Weathers et al., 1991; Kim et al., 1995). Normally, ions in a 3^+ charge state are chosen to ensure that there are no molecular interferences. Occasionally, for some elements, a charge state of 4^+ or 5^+ might be more desirable. The rule of thumb is that m/q should be a noninteger.

For elements that do not form negative ions or do so very poorly, the sensitivity for atomic detection would be very low or zero. However, for these cases, a negative molecular ion may be injected into the tandem accelerator of the form $(MX)^{1-}$, where M is the matrix atom or some other atom present in the sample (e.g., Si) and X is the impurity atom of interest. After accelerating the molecule to the terminal and breaking it apart by electron stripping, the atomic fragment of interest, X^{q+}, may be analyzed, where q is the charge state of the ion. Therefore, essentially, any element in the periodic table may be analyzed by the TEAMS technique. Indeed, the sensitivity for detection of each element depends upon a number of things, such as negative ion formation probability, matrix composition, sample contamination, surface oxidation or covering, charge state chosen, and transmission through the accelerator.

High-Energy Beam Transport, Analysis, and Detection System

After acceleration to mega-electron-volt energies and break-up of molecular interferences, the total ion beam in all charge states is 20 to 30 nA of ^{28}Si. The atomic ions in the chosen charge state are selected by magnetic analysis at 40° and by electrostatic analysis at 45°, as seen in Figure 1. After the magnetic and electrostatic analysis, the $^{28}Si^{3+}$ beam is ≅6 nA and ≅5 nA, respectively. The magnetic rigidity is given by $Br_b = mv/q$ with $m/\Delta m \cong 300$ for a 2-mm slit width, and the electrostatic rigidity is given by $\mathcal{E}r_e = 2E/q$ with $E/\Delta E \cong 250$. Here, B and $\mathcal{E}$ are the magnetic and electric field strengths, respectively, and r_b and r_e are the radii of curvature in the magnetic and electric fields, respectively. Combining these two analyses gives m/q. To eliminate atomic interferences from possible break-up fragments that have the same m/q (e.g., $^{28}Si^{2+}$ from the break-up of Si_2 and $^{56}Fe^{4+}$), the total energy of the ion is measured in a surface barrier detector or ionization chamber.

For bulk measurements of uniformly distributed impurities or dopants, the beam spot may be spread out to ~2 mm in diameter to increase count rate and the sample may be counted for longer time periods. For depth-profiling measurements, only data collected from the center ≅12% of the rastered area (plus the width of the sputter beam) are gated into the pulse height analyzer. Data from the remaining ≅88% of the rastered area, which includes the secondary ions sputtered from the crater wall, are discarded. It is still possible and even likely that some of the material comprising the ≅88% could be resputtered back onto the sample center ≅12% region, producing a contamination of the sample and reduced sensitivity.

The TEAMS system is automated and under computer control. The software to operate and control the TEAMS system is a combination of in-house and LabView programs at UNT. The software allows magnetic and electrostatic analyzers to scan mv/q and E/q, respectively, to a precision of 1:10,000 and to make either bulk or depth-profiling measurements. The magnetic and electrostatic analyzers are calibrated by analyzing known masses, usually from different matrix elements or molecules. Both atomic and molecular ions and atomic fragments are typically used for calibrating a wide range of magnetic (mv/q) and electric (E/q) rigidities. The software also operates Faraday cups, viewers, detectors, quadrupoles, steerers, and beam attentuators. Transmission through the entire TEAMS system is ≅1% to 8%.

Future Possible Advances in TEAMS

Future possible advances in TEAMS include the development of a tandem accelerator specially designed to directly couple to a SIMS source (with suitably modified injection optics) and the possible development of an efficient charge exchange canal to enable conversion of positive ions to negative ions. The Naval Research Laboratory (NRL) group (Grabowski et al., 1997) is constructing a TEAMS system by attaching a SIMS instrument to a tandem accelerator. Future improvements would include a submicrometer primary-ion beam and a means for secondary electron imaging of features on a sample. The Australia group (Sie et al., 1998) and the Zurich group (Ender et al., 1997c) have already developed microbeam TEAMS systems. Also on the horizon are other techniques to increase ionization probabilities such as laser postionization of neutrals sputtered from a sample (e.g., SIRIS; Winograd et al., 1982; Pellin et al., 1987, 1990; Bigelow et al., 1999, 2001). For some applications, the truly atomic aspects of TEAMS may be used to advantage, for example, in profiles through layered films.

PRACTICAL ASPECTS OF THE METHOD

The uniqueness of TEAMS is that it combines the two techniques of AMS and SIMS and can thereby be used to measure very low levels of stable isotopes of almost any element in a variety of different matrices corresponding to different materials (Rucklidge et al., 1981, 1982; Wilson et al., 1991, 1997a,b; Ender et al., 1997a-c; Massonet et al., 1997; McDaniel et al., 1998a,b; Datar et al., 2000). TEAMS has the capability to achieve ppt detection limits for some

elements in Si, which is considerably better than SIMS. TEAMS analysis is similar to SIMS. Perhaps, the largest benefit of TEAMS results from the passage of the secondary ions through a tandem accelerator where electrons are stripped from the molecular ion, resulting in a Coulomb explosion of the molecule into its elemental components. With TEAMS, after molecular break-up, magnetic (mv/q), electrostatic (E/q), and total energy (E) analyses are performed that uniquely identify the elemental ion. The mega-electron-volt energies of the secondary ions allow nuclear physics particle detectors to be used with essentially 100% detection efficiency because of the reduction in detector noise and scattered particles. The sensitivity of TEAMS is reduced because only one charge state ion is analyzed from the distribution of charge states produced during the electron-stripping process in the tandem accelerator. Since the most dominant charge state is usually chosen for analysis, the reduction in sensitivity is usually by only a factor of 3. The current depth-profiling detection limits at UNT of $\cong$0.1 ppb atomic are primarily due to ion source contamination and suboptimal secondary-ion transmission through the accelerator and analysis system (McDaniel et al., 1998a,b). Removal of source contamination and improvements in system transmission should allow ppt-level detection of stable elements in the depth-profiling mode. Bulk detection limits are lower because more material is available to analyze just by counting longer. Other TEAMS facilities around the world have similar detection limits (see section on existing TEAMS facilities in operation or under construction below).

The TEAMS technique can be used to easily discriminate against molecular interferences. However, it cannot easily discriminate against isobaric interferences, since at relatively low energies the ionization chamber detector cannot provide significant dE/dx information. Thus, in the case of more than one isotope with possible isobaric interferences, the isotope without any interferences should be chosen for analysis.

Ultraclean Ion Source Design Details

The ion source is designed to reduce sample contamination from elements in the ion source hardware and from chemically similar elements in the Cs and cross-contamination from other or previously analyzed samples (commonly called source memory effects). To reduce sample contamination, ion source components exposed to the primary-ion beam may be coated with ultraclean Si or some other pure material. The ion optics should be designed to reduce scattering onto the sample (Kirchhoff et al., 1994). The primary-ion beam is Cs^+, which readily provides electrons for the production of negatively charged atomic and molecular ions. The Cs^+ may be produced by a surface ionization ion source with an energy spread of approximately $kT = 0.1$ eV, where $k = 1.38 \times 10^{-23}$ JK^{-1} is the Boltzmann constant, and T is the temperature in K. For nominal energies of 10 keV incident upon the sample, the fractional energy spread is approximately $\Delta E/E = 10^{-5}$. The sample is held at a negative bias of 11 to 30 kV to allow the secondary ion produced to have enough energy to be injected into the accelerator and transmitted through the entire

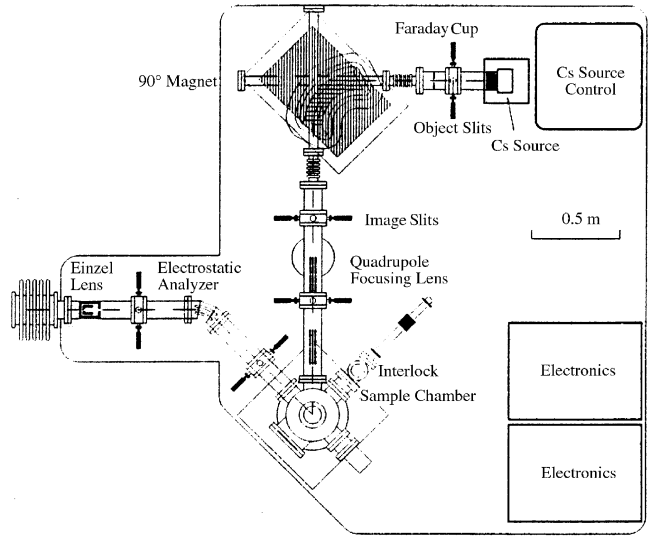


Figure 2. The TEAMS ultraclean negative sputter ion source. Shown are the Cs oven, the 90° magnet for preanalysis of the Cs sputter beam, sample chamber with sample load interlock, and 45° electrostatic analyzer in the secondary-ion beamline. Not shown are the eight Einzel focusing lenses, the six octupole steerers, and Cs beam scanner, which are electrostatic and therefore mass independent. Also, not shown is the aperture plate with laser drilled holes from 2 mm down to 0.2 mm, which allows the beam object size to be varied for microbeam production. Because of the many components for focusing and steering the primary Cs beam and the secondary ions from the sample, the entire source is under computer control at the source and in the control room.

TEAMS system. The lower sample bias allows better depth resolution in the sample. For the sample biased at -13 kV, the total primary-ion impact energy is 23 keV. While the Cs^+ ion beam may be very monoenergetic, it is not very pure. The primary Cs^+ ion beam should also be magnetically preanalyzed (mv/q) before striking the sample to reduce contamination of the sample (Kirchhoff et al., 1994). The ion source is normally capable of depth profiling with a microbeam primary ion as well as bulk impurity measurements in materials. For depth profiling, the last optical element before the sample is usually used to raster the primary-ion microbeam over a selected region of the sample. Secondary-ion data are then accumulated only from the center (12% to 20%) region of the crater sputtered by the primary ion to reduce wall contributions. It is necessary to have a microscope in the sample chamber to observe and monitor the beam spot on the sample.

Figure 2 shows the main components of the ultraclean ion source developed at the Ion Beam Modification and Analysis Laboratory (IBMAL) at the UNT (Kirchhoff et al., 1994). The intensity of the ion source is determined by the brightness of the primary Cs source, which is a UNT-modified, model 133 Cs source available from HVEC-Europa. The modifications are in the extraction and focusing optics and steering fields and have resulted in an increase in source brightness. Using computer control and optimization algorithms, the output parameters of the source can be rapidly optimized. The maximum half-angle of divergence of the Cs beam from the source was measured to be $\sim$9 mrad. The laboratory

emittance of the source was determined to be $\cong 7 \times 10^{-4}$ π-cm-mrad. For the 10-keV Cs ions, this corresponds to a normalized emittance of $\cong 7 \times 10^{-5}$ π-cm-mrad $(\text{MeV})^{1/2}$. This parameter indicates that, in the absence of aberrations, the theoretical spot size, without significant current loss, is ≅40 μm in diameter. In reality, the smallest spot size attained thus far at UNT is ≅60 μm in diameter. The use of reducing apertures can further reduce the spot size at the expense of intensity. This beam spot is continuously variable up to ≅2 mm in diameter for use in bulk analysis. Sie et al. (1998) and Ender et al. (1997c) have developed primary-ion microbeams that are on the order of 1 μm to a few micrometers in diameter.

The 10-keV Cs^+ ion beam is magnetically analyzed with a double-focusing 90° electromagnet to remove other alkali metals and Mo and Fe. These elements are produced from the components in the Cs and the ion gun. The ion source magnetic spectrometer helps to remove ambiguity and chance of contamination. Therefore, the Cs beam striking the sample is ^{133}Cs. Contamination has been shown to be well below 1 ppt after magnetic filtering.

Following the magnetic spectrometer, the primary ions enter an electrostatic zoom lens column that forms an intense microprobe. For depth-profiling measurements, the Cs^+ is scanned across the sample in a spiral digital meander pattern. The sample region scanned is ~1 mm × 1 mm in area and is divided into a 128 × 128 pixel integrated image, which is regenerated for each frame. The scanning system is digitally controlled from a digital input-output (I/O) card resident in a personal computer. Because of the number of active focusing and steering elements in the primary-ion beamline and the secondary-ion beamline, the ion source operation components are under computer control.

A microbeam may be formed from the momentum-analyzed ^{133}Cs ion beam with a combination of two quadrupole/octupole quadruplets. The first quadrupole is presently designed to produce a stigmatic image with a magnification $M_{Q1} = -1/50$ or less. This demagnification is followed by a long-working-distance quadrupole/octupole quadruplet objective to produce an elliptical spot at near-unity magnification in the vertical direction and a magnification of ~0.707 in the horizontal direction. Therefore, when the beam impinges on the sample surface at near 45°, the spot will appear almost circular. In practice, the strong horizontal electric field between the extraction plate and the sample bends the primary beam so that the angle of impact is ≅25° for ≅ −20-keV sample bias. The final optical element in the primary column of the ion source is a two-stage, eight-electrode dipole element that produces a translation of the beam of ±4 mm without introducing any variation in the angle of impact upon the sample. This feature is important when sputtering a uniformly deep crater for depth profiling. The dipole element scans the ion beam in a raster pattern. By correlating electronically the position of the primary-ion beam spot and the detection of the secondary ion at the other end of the apparatus, one can build an image of impurities distributed across the surface of the sample. The lateral resolution of the image, however, is limited by the beam spot size, which is presently ≅60 μm at the UNT facility. By acquiring the secondary ions versus time and knowing the rate at which the ion beam erodes the surface of the sample, one can build a profile of the impurity distribution versus depth. Therefore, in principle, a three-dimensional impurity map is possible. The limitation of available impurity ions in a sputtered volume will set a lower limit on the size of the beam spot for a given concentration of impurity atom. For example, for a typical Cs sputtering rate of ~2 Å/s (2×10^{-4} μm/s) and a 100 × 100 μm analyzed area, the sample volume of 2 μm^3 would contain 10^{11} Si atoms. One impurity atom detected in 1 s in this volume would correspond to a sensitivity of 10^{-11}, or 10 ppt. Higher impurity concentrations would require a smaller sample volume while lower impurity concentrations would require a larger sample volume.

Because of the developments in SIMS instruments that allow for a primary-ion microbeam and secondary electron imaging of a sample, some laboratories are using existing SIMS instruments to inject into the tandem accelerator and analysis system for the TEAMS facility (e.g., Grabowski et al., 1997). To improve the depth resolution of the data, the primary-ion energy and the sample bias should be as low as possible, while the sample bias should still be high enough to ensure adequate transmission through the remainder of the accelerator and beam transport system.

Secondary electron images permit accurate and rapid positioning of the primary Cs^+ ion beam on the sample. The sample is also viewed by a long-working-distance microscope with a field of view equal to the sample size (≅3 mm) and a resolution of a few micrometers. A remote high-resolution TV monitor views the optical image. The position of the sample is adjusted by a four-axis manipulator (x-y-z-θ) to allow three translations and one rotation about the secondary ion axis which is perpendicular to the sample surface.

Throughout the ion source, all first surfaces, which can be "seen" by the primary or secondary ions, are Si. The use of ultrapure and ultraclean Si on the first surfaces of the focusing and steering lenses, slits, Faraday cups, and apertures reduces the contamination of the sample from ion-wall scattering. These specialized components may be fabricated from relatively inexpensive materials (Kirchhoff et al., 1994). To further reduce sample contamination, the vacuum system is all metal sealed and has a base pressure, without baking, of $\cong 2 \times 10^{-9}$ torr. The sample holder, which will hold up to 19 individual samples 3 mm in diameter, is admitted and removed from the sample chamber via a sample interlock and manipulator. Samples are positioned on the sample holder and prepared for analysis in a class 100, portable clean room.

For samples that are insulating, a charge compensation system is required. This may be an electron gun that sprays electrons upon the sample to reduce sample charging in much the same way as with a standard SIMS instrument.

Calibration of TEAMS Magnetic and Electrostatic Analyzers

With TEAMS, in most cases, one is attempting to measure a trace quantity of an impurity element in a matrix, as

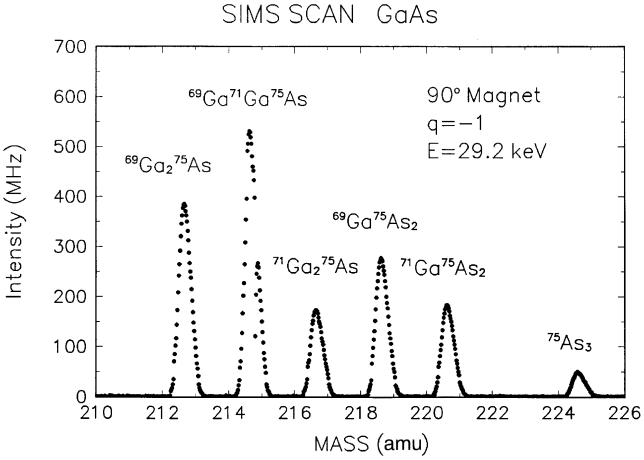


Figure 3. SIMS mass scan of the 90° preinjection magnet showing the characteristic signature spectrum of the trimer molecules formed in a charge state $q = 1^-$ from a GaAs matrix sample in the ion source. The peak heights are related to the natural abundances of the ^{69}Ga (60.1%) and ^{71}Ga (39.9%) isotopes. For each mass in the spectrum, a magnetic rigidity Br_b is determined, which is used to build a calibration file for the mass 212- to 225-amu region. Atomic and molecular ions of different masses are used for other mass regions.

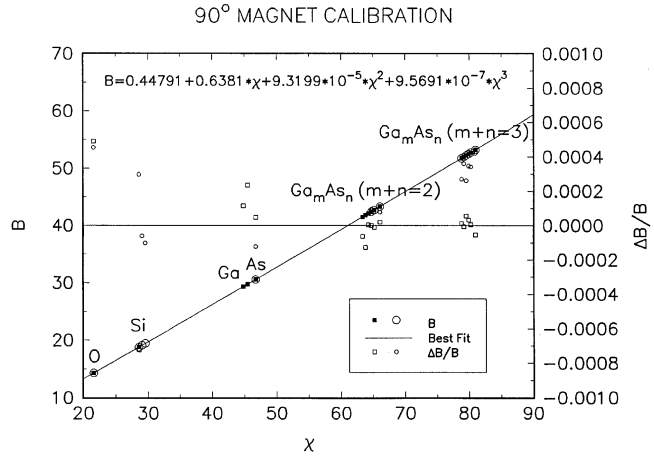


Figure 4. Graph of the 90° preinjection magnet calibration file showing magnetic field strength B versus $\chi_B = Br_b = mv/q$, the magnetic rigidity, which is dependent on the m, E, and q of the analyzed ion. The calibration file was determined from spectra like the one shown in Figure 3. The right-hand axis is a measure of the goodness of fit, the residual $\Delta B/B$, and is less than 0.0006 for all χ. The calculated fit, which is given at the top of the graph, is interpolated to find magnetic field strengths for other masses.

either a bulk elemental mass scan or an elemental depth profile of one or two elements. The small concentration of the impurity element requires that the TEAMS magnetic and electric fields used for analysis be precisely calibrated. The TEAMS system is calibrated by measuring mass scans of elements of known concentrations, i.e., the elements of the matrix for the sample being measured. Figure 3 shows a characteristic signature spectrum of the trimers (e.g., $(^{69}Ga^{71}Ga^{75}As)^{1-}$) that occurs with the 90° preinjection magnet mass scan of a GaAs sample. The calibrations are run sequentially through the 90° preinjection magnet, the 40° postacceleration magnet, and the 45° ESA for a fixed terminal voltage. A calibration file is constructed for each magnet, and the ESA from the isotopic mass and the magnetic ($\chi_B = Br_b = mv/q$) or electrostatic ($\chi_\mathcal{E} = \mathcal{E}r_e = 2E/q$) rigidity. Atomic and molecular ions as well as dissociated ions may be used for calibration. After the calibration files have been constructed, cubic polynomial fits to the data for each magnet were found to be necessary while a straight-line fit was found to be sufficient for the ESA. Figure 4 shows a representative calibration for the 90° preinjection magnet. The 40° postacceleration magnet (HVEC) is similarly calibrated. The electrostatic voltage for the ESA is plotted versus the electrostatic rigidity, $\chi_\mathcal{E} = \mathcal{E}r_e = 2E/q$, for different known masses. The calculated fits are interpolated to set magnetic and electric field strengths for future mass analyses. The secondary ions from the sample are analyzed through the entire spectrometer under computer control. A calibration precision of 0.001 in $\Delta B/B$ and 0.002 in $\Delta V/V$ is required for preservation of isotopic ratios, which is the true test of an accurate calibration of the system. For example, for elements with more than one stable isotope, the stable isotopes should appear in the correct proportions in a mass scan, as shown in Figure 3.

Existing TEAMS Facilities in Operation or Under Construction

The characteristics of a number of TEAMS facilities are discussed below. The information was provided to the author by personnel from the different TEAMS laboratories. (See the Internet Resources for contact information.)

Paul Scherrer Institute (PSI)/ETH Zurich Accelerator SIMS Laboratory. The PSI/ETH tandem accelerator facility serves as a national and international center for AMS. This very sensitive method is currently the leading technique for the detection of long-lived radionuclides such as ^{10}Be, ^{14}C, ^{26}Al, ^{36}Cl, and ^{129}I at isotopic ratios between 10^{-10} and 10^{-15}. These isotopes are intensively studied as natural tracers in earth sciences and in climate history research. In addition to radionuclide AMS, the PSI/ETH Zurich laboratory is actively involved in TEAMS research.

The TEAMS facility in Zurich includes a sputter chamber with an ideal extraction of secondary ions into the already existing beamline of the AMS facility in Zurich. At the same time, the chamber design keeps the contamination due to sputtering of secondary ions in the vicinity of the sample as low as possible. This was achieved with the extraction geometry shown in Figure 5.

The finely focused Cs^+ primary-ion beam is produced with a commercial Cs431 SIMS ion source from ATOMIKA in Munich, Germany (Wittmaack, 1995). It can produce beam currents from 0.2 to 800 nA with corresponding beam spot sizes of 1 to 150 μm. The Cs beam is analyzed with an $E \times B$ filter and sent through a 1° electrostatic deflector to remove impurities and neutral components and to keep contamination from the primary beam as low as possible. Also, the primary beam can be scanned over the sample to perform imaging of trace elements on the sample surface.

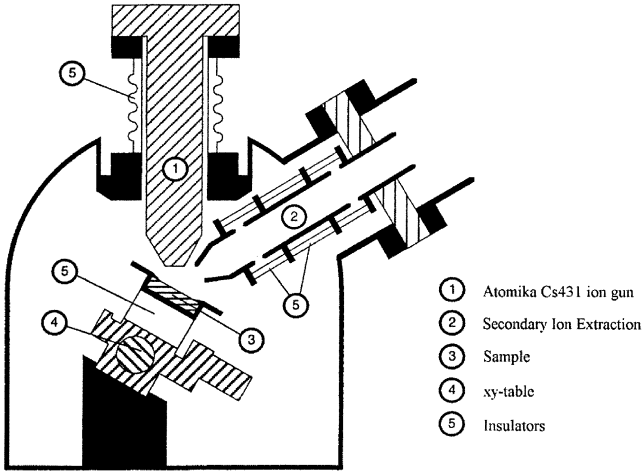


Figure 5. Schematic cross section through the TEAMS-sputtering chamber at the AMS facility in Zurich.

Using specific extraction voltages on the electrodes in the vicinity of the sample, the positive secondary ions are kept on the sample as much as possible and negative ions are extracted as much as possible so that they cannot be accelerated toward electrodes and produce tertiary ions and neutrals that would contaminate the sample. However, since this process cannot be suppressed completely, all the electrodes of the source have been coated with a 20-μm-thick layer of gold so that contamination of the sample due to the sputtering processes is kept low for all elements except gold (Döbeli et al., 1994; Ender et al., 1997a–c).

The extracted negative ions are deflected into the main beamline of the AMS facility and analyzed as in routine measurements. The EN6-Van de Graaff tandem accelerator is usually run at a terminal voltage of 5 MV, and a two-electrode gas ionization detector is used for the separation of isobaric interferences (Ender, 1997).

The facility has been tested with respect to detection limits, imaging, and depth profiling on trace elements in Si wafers. Table 1 and Figure 6 show a summary of the results for the TEAMS facility in Zurich for the bulk detection of trace elements in Si (Ender et al., 1997d).

Ongoing projects on the TEAMS facility are the measurement of isotopic ratios of trace elements in sedimen-

Table 1. Bulk Detection Limits for Trace Elements in Si Measured at PSI-ETH Zurich[a]

Element	Atomic Ratio	Atomic Density (atoms/cm^3)
B	1×10^{-10}	5×10^{12}
Al	7×10^{-11}	3.5×10^{12}
P	2×10^{-10}	1×10^{13}
Fe	2×10^{-9}	1×10^{14}
Ni	2×10^{-9}	1×10^{14}
Cu	2×10^{-9}	1×10^{14}
As	9×10^{-12}	4.5×10^{11}
Sb	2×10^{-10}	1×10^{13}

[a] The bulk detection limits are given as atomic ratios of trace element to Si and as atomic densities.

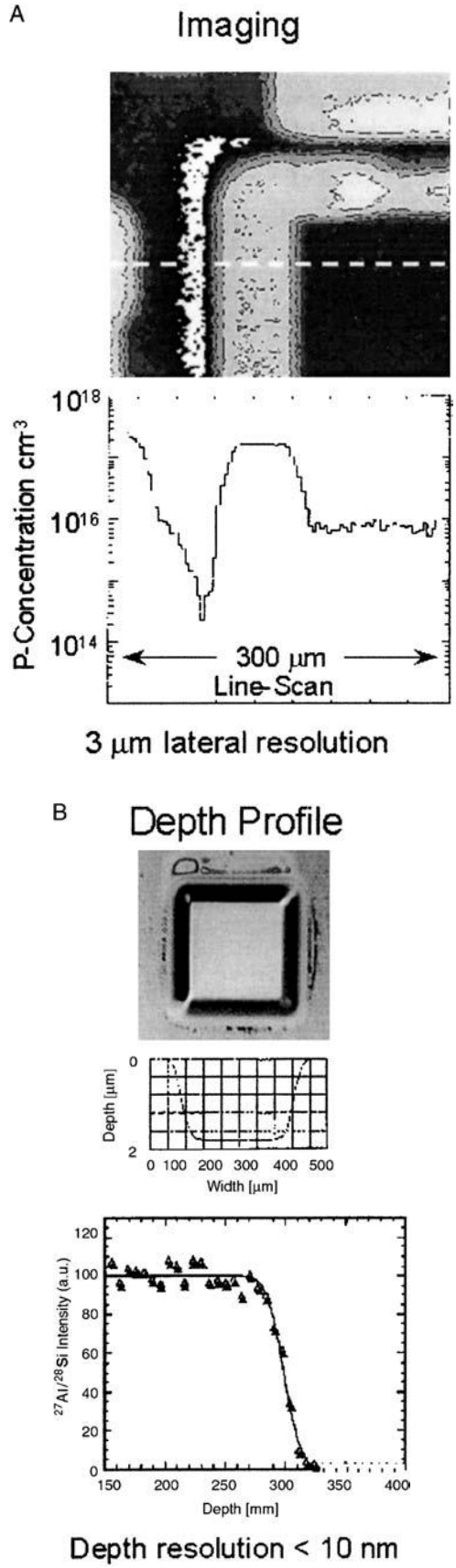


Figure 6. (**A**) Lateral resolution and (**B**) depth resolution of the TEAMS facility in Zurich.

tary layers and extraterrestrial matter with the goal to solve problems in geology and nuclear astrophysics. The main interest lies in the measurement of the heavy platinum group elements (PGEs). With the finely focused Cs beam, one can measure samples as small as ~200 ng. This also might have potential for the analysis of radioisotopes (Maden et al., 1999, 2000).

University of Toronto IsoTrace Laboratory. The University of Toronto is well known as one of the first laboratories to do radionuclide AMS. The existing AMS facility has also been used for a number of years for *in situ* TEAMS measurements of Au, Ag, and PGEs in mineral grains and meteorites (Rucklidge et al., 1981, 1982; Wilson et al., 1990; Rucklidge et al., 1990a,b; Kilius et al., 1990; Wilson et al., 1995; Rucklidge et al., 1992; Wilson et al., 1997a–c; Wilson, 1998; Sharara et al., 1999; Litherland et al., 2000, 2001). The "heavy-element" program developed at IsoTrace had its beginnings in the fertile period of collaborative research at the University of Rochester, circa 1977 to 1981, which saw trial analyses of PGEs such as Ir and Pt in pressed, powdered targets of rock and mineral samples using a broad-beam Cs^+ source on the tandem accelerator. Further development at IsoTrace was slow but sped up in the late 1980s with the evolution of a multitarget sample-mount format that involved the drilling of small (4-mm-wide) cylindrical "minicores" of samples and reference materials that were then mounted in sets of 12 in a 25-mm circular holder and polished to a fine finish, permitting mineralogical or metallographic tests prior to analysis. It should be noted that the sample chamber and 25 × 25-mm micrometer-resolution stage came first, in 1984, to combine unrestrained three-dimensional mobility and adequate sample-viewing optics, essential features for analysis of targets such as rocks. Rocks are almost always heterogeneous mixtures of one or more mineral species of variable shapes and grain sizes. The initial work conducted on the new stage was again of the pressed-powder variety, but the move to solid targets promoted a spate of measurements and rapid progress from circa 1989 to 1997, as summarized below. The principal limitations of the technology (aside from questions of cost and availability) that have thus far restricted adoption of these methods are perceived to be (1) the need to use conducting targets, whereas common rock-forming minerals are mostly good insulators, and (2) the broad Cs^+ beam size (maximum range may be 250 to 1500 μm, but commonly ~1000 μm), which limits application to relatively coarse grained samples. If any group succeeds in going beyond these constraints, which are common to ^{14}C-oriented accelerator laboratories, the next question will probably return to quantitation and the development of reference materials. In the course of this work, IsoTrace has addressed this by fabricating "fire assay" beads based on dissolution of well-characterized samples into a conducting nickel sulfide matrix.

Possible applications of TEAMS are almost limitless if the questions of surface charge build-up and spatial resolution can be addressed at a facility with a flexible analytical repertoire. Analytical capability evolved at IsoTrace due to the efforts of the late Linas Kilius and the other laboratory staff but did not exist in the very first incarnation of the laboratory, a straightforward ^{14}C system. Although test measurements were made on semiconductors, most trace element data at IsoTrace have involved conducting minerals and metals in rocks, meteorites, and archaeological artifacts. In addition, a large majority of the measurements have centered on just eight elements, the precious metals, namely Au and Ag, and the six PGEs (Pt, Ir, and Rh, which form negative ions readily, and Pd, Ru, and Os, for which sensitivity is significantly lower). Typical bulk detection levels achieved in short (e.g., 10 to 200 s) counting times are given in Table 2. Approximately 40 elements have been the subject of basic research at IsoTrace (see the short essay on heavy element research on the web page), but the precious metals have dominated the applications work.

Table 2. Standard Bulk Detection Levels (±30%) for Some Trace Elements for Analysis Times Less Than 200 s at the University of Toronto[a]

Element	Bulk Detection Level (ppb)	Atoms/cm^3
Ru	4	1.2×10^{14}
Rh	0.5	1.5×10^{13}
Pd	4	1.1×10^{14}
Ag	0.05	1.4×10^{12}
Os	20	3.2×10^{14}
Ir	0.25	3.9×10^{12}
Pt	0.1	1.5×10^{12}
Au	0.005	8×10^{10}

[a] If required, lower detection levels can be obtained with increased analysis time, limited by the availability of material.

It is important to note that in-house research has emphasized the importance of collecting TEAMS data as part of a broad spectrum of research on a topic. This ideally includes collection of samples in the field, classic petrographic (microscopic) studies, and sometimes whole-rock analyses of bulk samples plus collection of mineral-chemical data for major to trace elements using a full range of methods—some combination of electron, proton, and ion microprobe analyses as well as the broad-beam TEAMS.

The PGEs and Au typically behave as siderophile or chalcophile (iron- or sulfur-loving) metals, present at low-ppb levels or below in the earth's crust but at much higher levels in many meteorites and (by inference) in the cores of the differentiated terrestrial planets. A factor of roughly 1000 is necessary to enrich typical crystal rocks from ppb levels to the ppm grades, which constitute ore in most hard-rock gold and platinum mines today. The TEAMS of precious metals has largely been directed to the following materials: common Fe-Cu-Ni-As sulfides, Fe-Ti-Cr oxides, copper (both natural native copper and refined metal), and the common Fe-Ni-Co metal phases in iron meteorites and graphite. These materials have been analyzes in three broad contexts:

1. Distribution of the metals among coexisting ore minerals in various classes of mineral deposit. All the precious metals may occur both in concentrated form (rare minerals such as Au, $PtAS_2$, and PdSb)

and "invisibly" within the crystal structures of common ore minerals at levels from <1 ppb to 100 ppm or more—much more in the case of Ag.

2. Chemistry of iron meteorites, where the PGE and Au must occur dispersed among a generally small number of reduced phases: Ni-Fe alloys, phosphides, carbides, FeS, and graphite.
3. Provenance of archaeological metals. Thus, smelted Cu of European derivation has been shown by instrumental neutron activation analysis (INAA), AMS, and humble petrographic methods to be distinct from the purer native copper recovered for perhaps seven millennia by the inhabitants of the Great Lakes region. As with natural materials, samples must be characterized by context and microscopic properties prior to analyses, which may include ^{14}C dating of suitable components.

In terms of trace element work, the most important research at IsoTrace since 1997 has been method and equipment development by graduate students (see web site). In the geological context, the work of Ilia Tomski and Jonathan Doupé promises to broaden the range of available analytical strategies, while Jenny Krestow's project addresses the fundamental problem of analyzing insulating targets, including common minerals such as olivine and quartz.

Technical University Munich Secondary Ion AMS Facility. The Munich Tandem Accelerator Laboratory at the Technical University Munich, Garching, Germany, has developed a radionuclide AMS facility and has recently developed an ultraclean ion source for stable isotope impurities (Massonet et al., 1997).

At the Munich accelerator laboratory an ultraclean injector has been built that is used for SIAMS. It utilizes a magnetically analyzed Cs sputter beam and a large-volume ultrahigh vacuum target chamber. In the first stage apertures, electrodes, and target shielding were made of copper; in the second stage these parts were gold plated; and in the third stage the target aperture and the target shielding were made of high-purity Si. In the third stage, the best results were obtained. See, e.g., Fe and Cu in Table 3. The other elements were measured with copper- or gold-plated parts. The work was the doctoral thesis of Stefan Massonet (Massonet, 1998). Currently the AMS system is being used for the measurement of depth profiles. Table 3 presents bulk detection limits for trace elements in Si.

Table 3. Bulk Detection Limits for Trace Elements in Si Measured with the UltraClean Injector at the Munich Accelerator Laboratory with SIAMS[a]

Element	Atomic Ratio	Atomic Density (atoms/cm^3)
B	1.1×10^{-10}	5.5×10^{12}
N	8.2×10^{-8}	4.1×10^{16}
O	1.2×10^{-6}	6.1×10^{15}
Na	2.6×10^{-6}	1.3×10^{17}
Al	6.4×10^{-7}	3.2×10^{16}
P	1.7×10^{-8}	8.5×10^{14}
Cl	9.0×10^{-6}	4.5×10^{17}
K	4.4×10^{-8}	2.2×10^{15}
Ti	4.6×10^{-6}	2.3×10^{17}
V	2.8×10^{-8}	1.4×10^{15}
Cr	3.8×10^{-7}	1.9×10^{16}
Fe[b]	2.0×10^{-10}	1.0×10^{13}
Co	1.3×10^{-8}	6.6×10^{14}
Ni	4.8×10^{-7}	2.4×10^{16}
Cu[b]	6.5×10^{-10}	3.3×10^{13}
Ga	7.0×10^{-8}	3.5×10^{15}
Ge	1.9×10^{-8}	9.3×10^{14}
As	4.6×10^{-9}	2.3×10^{14}
Se	1.8×10^{-10}	8.8×10^{12}

[a] The bulk detection limits are given as atomic ratios of trace element to Si and as atomic densities.
[b] Target aperture and target shielding made of high-purity Si.

CSIRO Heavy Ion Analytical Facility (HIAF), Sydney, Australia. Another specialized TEAMS system is the AUSTRALIS (AMS for Ultra Sensitive Trace Element and Isotopic Studies) system, a microbeam TEAMS system that utilizes a 3-MV Tandetron at CSIRO HIAF, Sydney, Australia (Sie and Suter, 1994; Sie et al., 1997a–c, 1998, 1999; Sie et al., 2000a,b, 2001). Commissioned in 1997, the system is aimed at enabling *in situ* microanalysis of geological samples for radiogenic and stable isotope data free from molecular and isobaric interferences. The microbeam ion source, developed from a HICONEX source, produces a 30-μm-diameter Cs^+ beam routinely and includes a high-magnification sample viewing system in the reflected geometry, facilitating sample positioning and tuning of the primary beam. In measurements of Pb and S isotopes, high-precision isotope ratios of better than 1:1000 have been achieved for geochronological applications. The high precision is made possible by a fast bouncing system at both the low- and high-energy ends to counter the effect of instabilities in the ion source and beam transport system. Measurements were conducted at a terminal voltage of 1.5 to 2 MV. Lead isotopes are measured as Pb^{4+} ions from PbS^- injected ions. Sulfur isotopes are measured as S^{2+} or S^{3+} ions from injected S^- ions. The trace element detection capability is demonstrated in measurements of precious metals, the PGEs, and Au from an assortment of geological and meteoritic samples. Detection sensitivity as low as sub-ppb has been obtained for Au in sulfides, which will be of great benefit in studies of ore deposit mineralogy and mineral processing. The Australia group has also demonstrated the facility for Os isotope measurements in meteorite samples, opening up the possibility of widespread use of the Re-Os system in exploration programs.

Naval Research Laboratory TEAMS Facility. One of the most ambitious programs is at the NRL, Washington, DC (Grabowski et al., 1997). A large portion of the novel TEAMS facility at the NRL is now installed. When completed, this facility will provide for parallel mass analysis over a broad mass range for conducting and insulating samples and offer 10 μm lateral image resolution, depth

profiling, and sensitivity down to tens of ppt of trace impurities. The facility will use a modified commercial SIMS system as the source of secondary ions to provide these capabilities. This ion source is being installed. After the ion source, the facility uses a Pretzel magnet (Knies et al., 1997a) to act as a unique recombinator to simultaneously transmit from 1 to 200 amu ions but attenuate intense matrix-related beams. Following acceleration, a single charge state is selected by a 3° electrostatic bend; then the selected ions are electrostatically analyzed for energy/charge (E/q) by a 2.2-m-radius, 30° spherical ESA with $E/\Delta E \cong 800$. Finally, a split-pole mass spectrograph with a 1.5-m-long focal plane provides parallel analysis over a broad mass range ($m_{max}/m_{min} = 8$) with high mass resolution ($m/\Delta m \cong 2500$). Micro-channel-plate-based position-sensitive detector modules are used to populate the focal plane.

Currently, vacuum and beam optics hardware are in place, and testing is underway with a multicathode ion source in place of the SIMS ion source. This more intense source simplifies diagnostic testing and initial research efforts. All twelve position-sensitive detector modules for the focal plane of the spectrograph have been received, and their testing is progressing well. Programmatically, the NRL is actively involved in the study of gas hydrates present under the ocean floor, which includes plans to analyze cycling between various carbon pools present there. Since ^{14}C analysis is an important part of this work, the TEAMS facility has been modified to include a switching ESA to choose between the new multicathode ion source and the SIMS ion source for injection into the Pretzel magnet. It is anticipated that operations can be easily switched between standard radioisotope AMS and TEAMS efforts. Preliminary ^{14}C analysis is currently underway.

University of North Texas Ion Beam Modification and Analysis Laboratory TEAMS Facility. The UNT TEAMS facility has been described in detail in other parts of this unit and will only be briefly summarized here. The UNT TEAMS facility is a dedicated trace element AMS that is part of the IBMAL.

The facility is characterized by an ultraclean ion source with magnetic (mv/q) analysis of the primary Cs^+ sputter beam to remove contaminants. The Cs^+ beam may be raster scanned over the sample for depth-profiling measurements with a spot size of 60 to 90 μm or enlarged to ~2 mm diameter for bulk impurity measurements. The samples are analyzed in a high-vacuum chamber with all components that could be hit by the beam coated with ultrapure Si to reduce sample contamination. The secondary ions extracted from the sample are electrostatically analyzed at 45°, magnetically analyzed at 90°, and injected into the 3-MV National Electrostatics tandem accelerators. After the accelerator, the energetic secondary ions are magnetically analyzed at 40° and electrostatically analyzed at 45°. The secondary ions are detected in a Faraday cup for matrix ions and in a surface barrier detector or ionization chamber for trace impurity ions.

The TEAMS system has been used mainly to demonstrate depth-profiling capability for a number of trace elemental species in a variety of semiconductor materials, e.g., Si, GeSi, GaAs, GaN, and SiC (Datar et al., 1997a,b; McDaniel et al., 1998a,b; Datar et al., 2000). This is noteworthy because depth profile measurements are the primary concern of the semiconductor industry. Depth profile measurements can also be compared directly with SIMS results.

Table 4. Detection Limits in Si in Depth-Profiling Mode at the University of North Texas[a]

Element	Matrix	Depth-Profiling Detection Limits (atoms/cm^3)
B	Si	4×10^{14}
F	Si	2×10^{13}
P	Si	3×10^{13}
Ni	Si	1×10^{14}
Cu	Si	2×10^{14}
Zn	Si	1×10^{15}
As	Si	2×10^{14}
Mo	Si	1×10^{15}
Sb	Si	1×10^{15}
Se	Si	1×10^{12}
As	Ge_xSi_{1-x}	1×10^{15}

[a] The depth window is ~50 Å.

The TEAMS facility can be operated in a bulk analysis mode or depth-profiling mode. The bulk analysis mode is generally more sensitive, since one can count the impurity ions for a longer time. In the depth-profiling mode, there is a limited amount of material available at a particular sputtering depth. Table 4 gives the detection limits in Si in the depth-profiling mode in atoms per cubic centimeter.

Engineering Advances That Made TEAMS Possible

Some of the advances that made TEAMS feasible include voltage-stable tandem accelerators, specially designed low-contamination ion sources, the use of all-electrostatic focusing and steering elements, double-focusing spherical electrostatic energy analyzers, the development of fast-switching magnet power supplies, the development of nuclear detectors for mega-electron-volt energy ions. The operation and voltage control of tandem accelerators allows the production of energetic ions with less energy uncertainty, thereby enabling magnetic and electrostatic analysis.

Historical Evolution of TEAMS

Accelerator mass spectrometry was first used by Alvarez and Cornog (1939a,b) to measure natural abundances of ^3He. It was reincarnated when Muller (1977) suggested that a cyclotron could be used for detection of ^{14}C, ^{10}Be, and other long-lived radioisotopes. Independently in 1977, the University of Rochester group (Purser et al., 1977) demonstrated that AMS could be used to separate ^{14}C from the isobar ^{14}N because of the instability of the $^{14}N^-$ ion. Later, Nelson et al. (1977) at McMaster University and Bennett et al. (1977) at the University of Rochester reported accelerator measurements of natural ^{14}C.

Purser et al. (1979) reported that accelerator SIMS is a very promising technique for the detection of trace elements in ultrapure materials. A lot of the early work

done at Rochester and Toronto concentrated on the *in situ* TEAMS measurements of PGEs, Au, and Ag in mineralogical and geological samples (Rucklidge et al., 1981, 1982). Since then, the Toronto group has continued this TEAMS research in mineral grains and meteorites (Wilson et al., 1997a,b). The Rochester group measured quantitative TEAMS depth profiles of N and Cl in semiconductor-grade Si wafers (Elmore et al., 1989; Gove et al., 1990).

Anthony and Thomas (1983) first suggested the use of AMS for the detection of stable isotopes present as contaminants or impurity dopants in ultrapure semiconductor materials. Preliminary measurements were made on the carbon-dating AMS facility at the University of Arizona (Anthony et al., 1985, 1986). Anthony and Donahue (1987) showed that AMS could help provide solutions for semiconductor problems. Following this lead, a TEAMS system dedicated to broad-band stable isotope measurements over the entire range of the periodic table was designed and constructed at IBMAL at the UNT in collaboration with Texas Instruments (Anthony et al., 1989, 1990a,b, 1991, 1994a,b; Matteson et al., 1989, 1990; McDaniel et al., 1990a,b, 1992, 1993, 1994, 1995, 1998a,b; Kirchhoff et al., 1994; Datar et al., 1996, 1997a,b, 2000; Zhao et al., 1997). Other groups soon followed with modifications of existing radionuclide AMS facilities to perform stable isotope measurements at PSI/ETH Zurich (Döbeli et al., 1994; Ender, 1997; Ender et al., 1997a-d) and the Technical University Munich (Massonet et al., 1997) or construction of new facilities at CSIRO in Australia (Sie and Suter, 1994; Niklaus et al., 1997; Sie et al., 1997a–c) and at the NRL (Grabowski et al., 1997; Knies et al., 1997a,b).

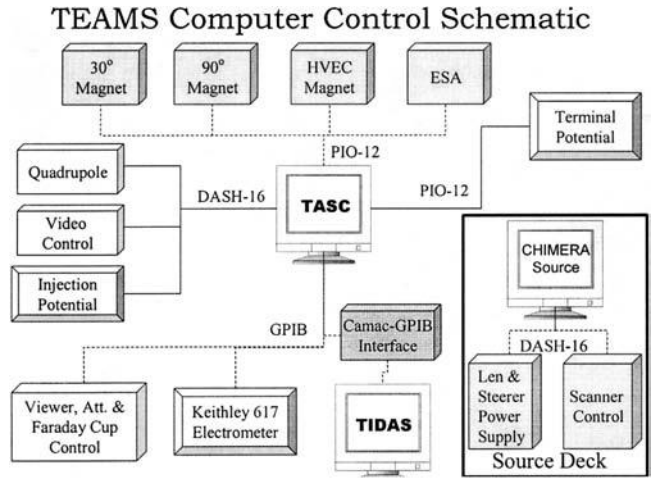


Figure 7. Block diagram of the TEAMS computer control and data acquisition system at UNT. The ion source, accelerator, and data acquisition system are controlled by three personal computers. The tandem accelerator system control (TASC) computer controls the magnets, electrostatic analyzer, quadrupole, terminal potential, ion source injection potential, viewers, Faraday cups, and attenuators and is connected to the CAMAC modular data-handling system crate through a GPIB interface. The TI data acquisition system (TIDAS) is connected to a CAMAC crate through a GPIB interface. Detector signals from either the ionization chamber or Si particle detector are passed into the CAMAC crate amplifier and stored in the TIDAS. The voltages (or currents) for the ion source Cs source bias, 90° magnet, eight Einzel lenses, six *x-y* steerers, Cs beam scanner, sample bias, extraction bias, and 45° ESA are controlled by the ion source computer, which can be operated in the accelerator control room or at the source.

METHOD AUTOMATION

Due to the extreme complexity of the ion source at UNT, computer control is essential for steering and focusing all elements in both primary- and secondary-ion beamlines. By systematically varying the control voltage and measuring the beam current, the optimum control voltage can be determined for each parameter. Ideally, if all beamlines are properly aligned, the steering voltages should be zero. Practically, as long as the steering voltages are near zero, near-optimum tuning can be assumed. The accelerator, beamline, and analyzing elements are also under computer control and can be set to the appropriate settings by keyboard entry. Figure 7 shows a block diagram of the computer control and data acquisition system at UNT.

For depth profile measurements, one simply clicks on the depth profile icon. A pop-up menu enables one to enter all the necessary parameters, and the computer automatically acquires the raw depth profile data. The operator determines the endpoint and stops data acquisition. The raw data are stored in a file of the operator's choice. Sample changes have to be done manually. The measurement should not be left unattended. For bulk measurements, the computer control program is also used, except that the entries in the pop-up menu are different. Due to the unique nature of the computer control system, most software was initially developed in-house and later converted to a LabView-based system.

DATA ANALYSIS AND INITIAL INTERPRETATION

Bulk Impurity Measurements

For bulk measurements, where a constant, very low level concentration is to be measured, the acquisition time can be increased and the tuning can be done on each sample and not just the blank, to increase transmission. Furthermore, if a standard sample with a constant low-level impurity doping is available, additional fine tuning of the TEAMS system can be done, leading to enhanced impurity transmission with a corresponding increase in sensitivity. This procedure is essentially the same as radioisotope AMS measurements, where following the use of a stable isotope as a pilot beam, the final fine tuning is done using a known standard. Bulk analysis measurements are identical to that used for isotope ratio measurements in AMS. The matrix ion count rate is measured briefly as a current, while the impurity count rate is measured for a longer time in a particle detector or ionization chamber. The time difference is because there are fewer impurity ions compared to matrix ions. By repeatedly cycling between the two measurements, good statistics can be obtained for both matrix

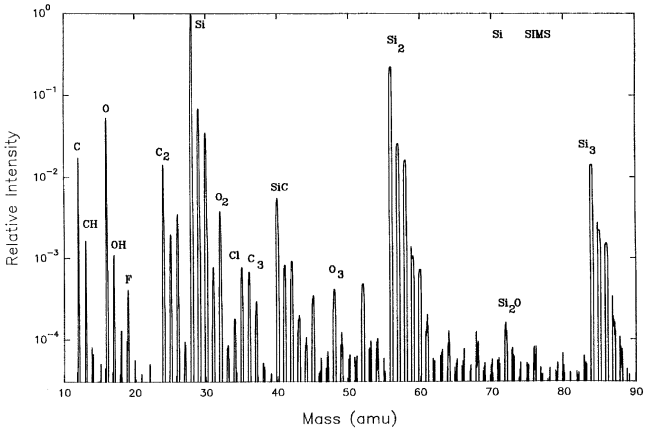


Figure 8. Bulk analysis SIMS mass scan of a heavily contaminated Si sample. All peak intensities are relative to the ^{28}Si peak, which is set equal to 1.0.

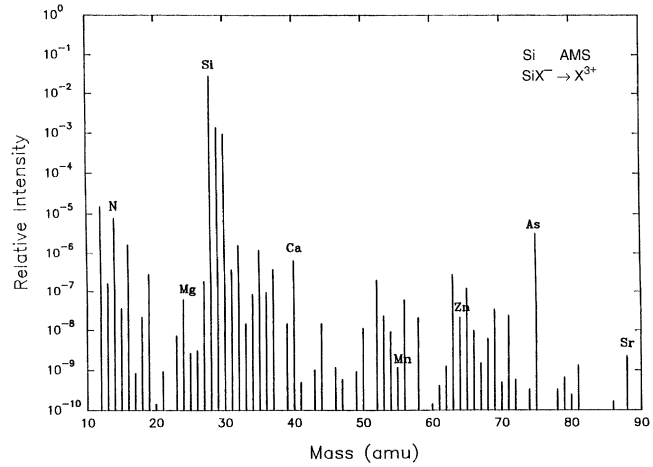


Figure 10. Bulk analysis TEAMS mass scan of the same sample similar to Figure 9, except a molecular mass was chosen to accelerate to the terminal of the accelerator. The signals have been normalized to the ^{28}Si signal in Figure 9 for direct comparison.

and impurity ions. Elmore et al. (1984) has given a detailed exposition of the technique.

The TEAMS bulk measurements can also provide a quick mass scan of elements present in a sample in much the same way as SIMS, except the molecular interferences can be suppressed. For example, Figures 8 to 10 show representative bulk analysis mass scans of a heavily contaminated Si sample over an extended mass range of 10 to 90 amu. All isotopic relative intensities are normalized to the ^{28}Si peak, which is set equal to 1.0 in Figures 8 and 9. The pre-tandem-acceleration analysis shown in Figure 8 represents a conventional negative ion SIMS spectrum with contributions to many mass regions from molecular ions. The atomic and molecular ions are extracted from the sample, pass out of the ion source chamber through the 45° electrostatic analyzer and the 90° magnetic spectrometer, and are measured as a current in a Faraday cup, as shown in Figure 1. Detection before the accelerator is limited to current measurements because of the high currents, but molecular ions are still present. For example, some of the molecular ion interferences shown in the SIMS mass scan in Figure 8 are given in Table 5 with their intensity relative to $^{28}Si^-$. In many cases the molecular ion intensity is much greater than an atomic ion at the same mass.

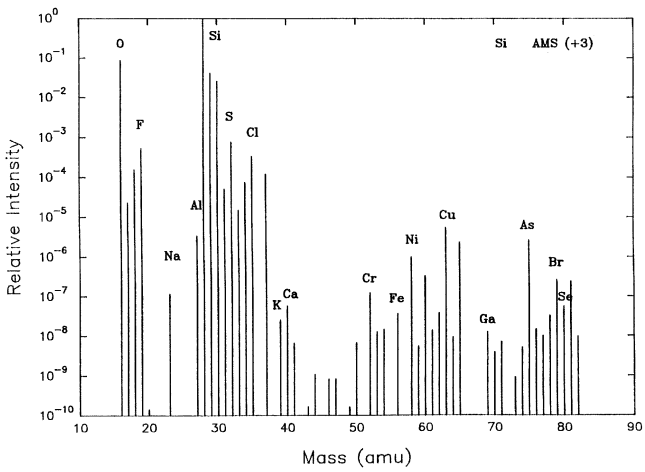


Figure 9. Bulk analysis TEAMS mass scan of the same sample as in Figure 8. All peak intensities are relative to ^{28}Si, which is set equal to 1.0.

Figure 9 is a TEAMS bulk analysis mass scan of the same Si sample after passage through the tandem accelerator and high-energy magnetic (mv/q) and electrostatic (E/q) analysis with charge state $q = 3^+$ and total energy (E) detection. The relative intensities of the isotopes shown are normalized to the intensity of the $^{28}Si^{3+}$ peak, which is set equal to 1.0. Most of the lower intensity measurements shown require direct ion counting in a particle detector as opposed to higher intensity current measurements. Counting time was 5 s per mass interval for all measurements or ~400 s for the entire mass scan. Since all isotopic ions analyzed were in a $q = 3^+$ charge state, interfering molecular ions have been rejected by breaking up the molecules in the terminal of the accelerator. In addition to removal of the molecular interferences, detection of ions in high-energy particle detectors has removed detector background noise as well as most m/q interferences. Signals from break-up products that still exhibit m/q interferences at masses 36, 42, 45, 48, 51, 57, and 84 are not shown in Figure 9. Elements with zero electron affinity (e.g., Mg, Mn, Zn) will not form negative ions, and the TEAMS mass scan does not show any signal for these masses. As seen by comparison of the relative intensities in Figures 8 and 9 and Table 5, the molecular interferences are reduced, in some cases, by many orders of magnitude. For example, the relative intensity of $(^{28}Si)_2$ at mass 56 (^{56}Fe) is reduced by almost seven orders of magnitude in the TEAMS data compared to the SIMS data. The $^{28}Si^{12}C$ at mass 40 (^{40}Ca) is reduced by almost five orders of magnitude in the TEAMS data compared to the SIMS data. The molecular peaks $(^{29}Si)_2$, $^{29}Si^{28}Si^{1}H$, etc. at mass 58 (^{58}Ni) are reduced by over four orders of magnitude in the TEAMS data compared to the SIMS data.

Table 5. Molecular Ion Impurities in SIMS Mass Scan in Figure 8.1 Relative to $^{28}Si^-$ Intensity and Atomic Ion Intensities from TEAMS Mass Scan in Figure 9 Relative to $^{28}Si^{3+}$ Intensity[a]

Molecular Ion	Mass (amu)	Intensity/$^{28}Si^-$ (Figure 8)	Intensity/$^{28}Si^{3+}$ (Figure 9)	Elemental Ion
$^{16}O^1H$	17	1.2×10^{-3}	3.0×10^{-5}	^{17}O
$(^{12}C)_2$	24	1.6×10^{-2}	$<10^{-10}$	^{24}Mg
$^{30}Si^1H$	31	7.0×10^{-4}	7.0×10^{-5}	^{31}P
$(^{16}O)_2$	32	4.0×10^{-3}	9.0×10^{-4}	^{32}S
$^{28}Si^{12}C$	40	6.0×10^{-3}	7.0×10^{-8}	^{40}Ca
$^{24}Mg^{28}Si$, etc	52	5.0×10^{-4}	1.3×10^{-7}	^{52}Cr
$(^{28}Si)_2$	56	2.5×10^{-1}	4.0×10^{-8}	^{56}Fe
$^{28}Si^{29}Si^1H$, $^{29}Si_2$, etc.	58	2.0×10^{-2}	1.0×10^{-6}	^{58}Ni
$^{28}SiO_2$, $^{30}Si_2$, etc.	60	7.0×10^{-4}	4.0×10^{-7}	^{60}Ni
$(^{28}Si)_2{}^{16}O$	72	1.8×10^{-4}	$<10^{-10}$	^{72}Ge

[a] Break-up of the molecular interferences reduces the intensities by orders of magnitude in many cases.

Because of the removal of the molecular interferences in the TEAMS data, the correct isotopic abundances are found for O, Si, Cl, Ni, and Cu, which indicates that calibration of the magnetic and electrostatic analyzers is correct.

Even though all the molecular ions are dissociated in the stripping canal of the tandem accelerator, there still may be some m/q interferences. During a mass scan, the preacceleration magnet is tuned for a charge state $q = 1^-$, while the postacceleration magnet is tuned for the charge state of interest, which is $q = 3^+$ for Figure 9. Even though direct molecular interferences are removed, molecular fragments may still cause an ambiguity. For example, $((^{12}C)_n)^-$ molecules can be produced from the vacuum system during the sputtering process (as seen in Figure 8). Any $((^{12}C)_3)^-$ molecular ions will dissociate after stripping but will result in a large production of $^{12}C^+$ ions. The $^{12}C^+$ ions at mass 12 have the same m/q as the atomic ions at mass 36 (e.g., $^{36}S^{3+}$) and therefore will pass into the final total energy detector. The $^{12}C^-$ ions created during the sputtering process will not be an interference for mass 36 because it will not pass through the preacceleration magnet. Even though the total energy spectroscopy will separate the peaks, the tails from high levels of ^{12}C may lead to saturation of the detector and incorrect results. Several of the masses with known interferences described above (36, 42, 45, 48, 51, 57, and 84) are not included in the mass scan in Figure 9. These interferences can be removed by analyzing a different charge state and therefore a different m/q.

Some elements do not form negative ions directly, e.g., N, Mg, Mn, and Zn. In these cases, molecular ions of the form SiX-, where X represents the element of interest, are injected into the accelerator. The break-up products X^{3+} are accelerated down the high-energy side of the accelerator, magnetically and electrostatically analyzed, and detected in the total energy detector. Figure 10 shows such a TEAMS spectrum in which the intensity of each signal has been normalized to the maximum ^{28}Si signal in Figure 9 for direct comparison. Several ions that were not detectable in the TEAMS mass scan in Figure 9 are now measured in the correct isotopic abundances, e.g., N, Mg, Mn, and Zn. Signals from break-up products at masses 36, 42, 45, 48, 51, 57, and 84 are not removed in Figure 10.

Data Analysis for Bulk Measurements

This procedure for bulk analysis provides a relative normalization of the impurity element to the amount of matrix element present in the sample, which is well known. Of course, this procedure does not account for the differences in negative ion formation probabilities of different elements or molecules or charge state fractionation of the ion in the terminal of the accelerator. In fact, some elements will not form a negative ion at all (e.g., N, Mg, Mn, Zn). For these elements that do not form a negative ion, in almost all cases, a molecule with this element can be found that will form a negative ion, therefore allowing analysis of essentially every element in the periodic table. A distribution of charge states is produced during electron stripping in the terminal of the accelerator. Because the magnetic (mv/q) and electrostatic (E/q) analyses pass only one charge state from this distribution, only a fraction of the original ions from the sample are analyzed and detected. Fortunately, the $q = 3^+$ charge state most used in the analysis is the most populated for masses in the range of ~30 to 60 amu when stripped at an energy of ~1.5 to 3 MeV. In addition, this procedure requires that the Cs sputter beam be constant over the length of the mass scan and that the calibration of the magnetic and electrostatic analyzers be accurate over the mass range of the scan. The calibration is normally checked before and after the measurement, as discussed earlier with Figure 3, by running a mass scan of the masses present for the matrix elements to see the characteristic signature spectrum.

Mass scans over a range of masses are useful to confirm isotopic natural abundances of an element of interest as well as to check for interferences that may be present at a particular mass. In the TEAMS analysis beamline, the UNT facility has a pneumatically controlled Faraday cup followed by a Si charged particle detector that may be inserted into the beam or withdrawn and then an ionization chamber. For a mass scan as shown in Figure 9 or 10,

the magnetic and electrostatic fields are set to the correct values to pass the initial mass of interest with the Faraday cup in the beam path. If the current measured in the Faraday cup for that mass is below a preset value, the Faraday cup is removed from the beam path and the ions of that mass are detected in the Si detector or ionization chamber for a preset time. If not, then the mass is measured as a current in the Faraday cup as Coulombs per second and converted to particles per second, taking into account the charge state of the ion. After completing the preset measuring time, the magnetic and electrostatic fields are incremented to the next mass of interest, and a measurement is made for that mass, and so on for the entire mass range of interest. Usually the time of measurement is of the order of a few seconds for each mass value. Counting for a longer time period can reduce the minimum background count and therefore improve the detection limit. For example, for a 5-s count at each mass value, the minimum detectable count rate is 1 count in 5 s or 0.2 counts/s (cps). For a 100-s count at each mass value, the minimum detectable count rate is 1 count in 100 s or 0.01 cps, which is 20 times more sensitive.

Depth-Profiling Impurity Measurements

To measure a depth profile of Cu in Si, for example, the key parameter is the terminal voltage on the accelerator. The terminal voltage should be chosen such that the most abundant charge state for Cu is 3^+ or higher. There is some flexibility in the choice of impurity ion species injected into the accelerator. For some elements, the ion MX^-, where M is a matrix atom and X is an impurity atom, is more prolific than the ion X^-. As far as possible, then the most prolific negative ion should be chosen. The RSF tables in Wilson et al. (1989) provide a good guide to the best possible negative ion to be injected. Care must be exercised in the choice of the positive charge state selected for postaccelerator analysis to avoid m/q degeneracies with the matrix ion(s) since it can swamp the detector. For example, during ^{58}Ni measurements, charge state 4^+ has the same m/q as $^{29}Si^{2+}$ and the Ni signal can be swamped in the detector by the tail of the Si signal, even though they are at different energies. The m/q interferences from nonmatrix species can be separated by energy in the detector, as the count rate is usually not very high. If the impurity concentration to be measured is high as for a dopant, then a less abundant charge state can be chosen to avoid swamping the detector. However, in the rest of the present discussion, it will be assumed that one is interested in only measuring very low concentration levels, which is the strength of TEAMS.

As previously mentioned, the nuclear-type particle detector has essentially zero dark current because the ion is at mega-electron-volt energies. Thus, the lowest counting rate of the detector is determined by the acquisition time for that particular data point in the depth profile, which in turn is determined by the volume of the available material. For example, with a sputtering rate of 2 Å/s, an acquisition time of 20 s implies that the measurement represents an average concentration for a slice of material 40 Å thick. The lowest measurable concentration in 20 s is 0.05 cps (one count in the 20 s). By increasing the acquisition time to 50 s, the lowest measured concentration now corresponds to 0.02 cps, but now it represents an average over 100 Å. However, one must remember that the sputtering process in the ion source causes depth profile broadening due to the penetration of the Cs ions into the sample. This effect is of the order of 100 Å (Vandervorst and Shepherd, 1987). Therefore, the Cs penetration depth limits the ultimate depth sensitivity.

For the example of a Cu depth profile in a Si sample, the experimental parameters were as follows:

Sample bias: 11.17 keV. The lower the sample bias, the better, as it reduces the Cs penetration depth (Vandervorst and Shepherd, 1987).

Secondary ions injected into the tandem accelerator: $^{63}Cu^-$, $^{28}Si^-$.

Terminal voltage: 1.7 MV.

Ions transported to the end of the beamline: 6.8 MeV $^{63}Cu^{3+}$, 6.8 MeV $^{28}Si^{3+}$.

Cu ions were measured in an ionization chamber with an acquisition time of 20 s, and Si ion current was measured in a Faraday cup just in front of the detector.

The tuning was done using a nominally blank pure Si sample. The focusing and steering elements were adjusted to get maximum $^{28}Si^{3+}$ current in the final Faraday cup. It was assumed that the analyzing elements (the magnets and the electrostatic analyzer) were already calibrated. This calibration is periodically checked. For depth profile measurements, one has to specify the isotopes to be measured along with the charge states and the acquisition time. In the case of m/q interferences, the total energy gate width in the ionization chamber has to be specified also. Typically, in depth profile measurements, one sequentially cycles between a matrix ion count rate measurement and an impurity isotope count rate measurement. Switching the system between different masses involves dead time since the magnets cannot be switched instantaneously. By suitable choice of matrix charge state, this dead time can be reduced. The matrix ion count rate measurement serves as a basis for normalization of unknown depth profiles to standard implanted depth profiles as well as serving as a check on system stability (very much similar to both AMS and SIMS).

Before measuring an unknown depth profile, a standard implant is measured. Thus, for the Cu profile, after the blank measurement, an implanted standard was measured. This serves as an additional check on the measurement process. Finally, the unknown sample was measured with exactly the same parameters as the standard implant. Figure 11 shows an example of a Cu impurity depth profile measurement for a 10-keV, 1×10^{16} atoms/cm^2 As-implanted Si wafer. The Cu concentration decreases from the surface but then shows an unexpected peak around 0.1 μm into the wafer before finally tailing off at a concentration level of 4×10^{14} atoms/cm^3, which is a factor of 6 lower than the SIMS measurement. A nominally blank sample of Si is run after measurements of an unknown sample to get an estimate of cross-contamination from other samples and also to make sure that running the unknown sample has not contaminated the source.

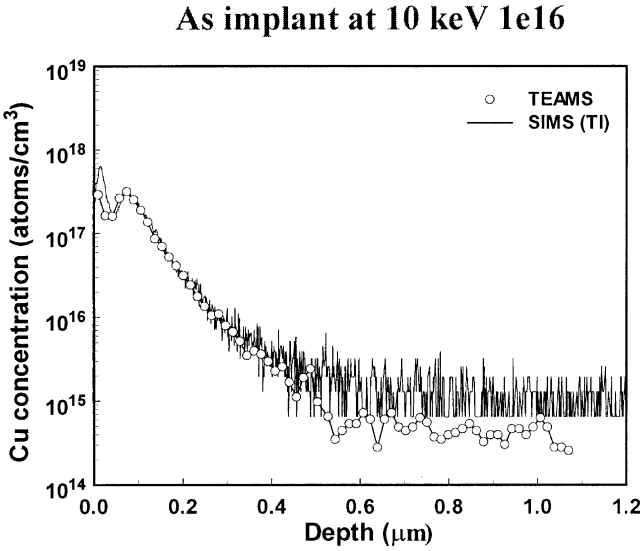


Figure 11. TEAMS Cu impurity depth profile in a 10-keV, 1×10^{16} atoms/cm^2 ^{75}As implant in Si. Also shown is a SIMS depth profile for comparison. The TEAMS and SIMS profiles are in good agreement until the SIMS detection limit is reached at ~0.4 μm in depth. After 0.4 μm, TEAMS is about a factor of 6 more sensitive than the SIMS measurements.

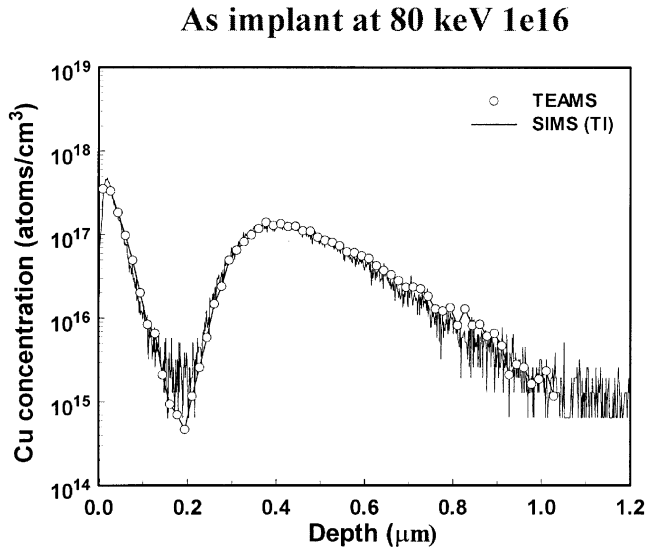


Figure 13. TEAMS Cu impurity depth profile in an 80-keV, 1×10^{16} atoms/cm^2 ^{75}As implant in Si. Also shown is a SIMS depth profile for comparison. The TEAMS and SIMS profiles are in good agreement. TEAMS can be seen to be more sensitive than the SIMS measurements in the dip in the spectrum at ~0.2 μm.

Figures 12 and 13 show TEAMS Cu impurity concentrations as a function of depth in As-implanted Si wafers. Figure 12 shows an impurity Cu depth profile for a 40-keV, 1×10^{16} atoms/cm^2 As implant. Figure 13 shows an impurity Cu depth profile for an 80-keV, 1×10^{16} atoms/cm^2 As implant. The Cu concentrations in both spectra decrease from the surface peak but then increase again to form a deeper secondary peak. The depth of the secondary peak scales with the As implant energy. SIMS measurements were done later to confirm the TEAMS data. The SIMS depth profile data are overlaid on the TEAMS data and the agreement is excellent. During data acquisition, both isotopes of Cu were measured with TEAMS, and the isotope ratios were found to be in agreement with their natural abundances. This is an indication that the Cu was not implanted along with the As and that the two Cu isotopes were not implanted separately. Only one mass can be implanted at a time.

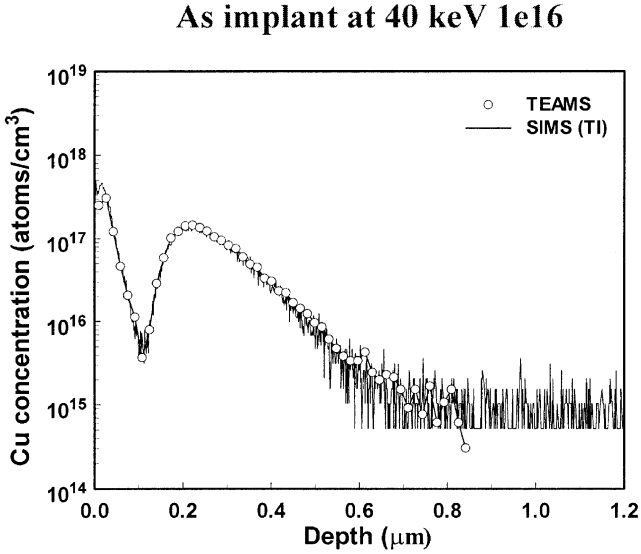


Figure 12. TEAMS Cu impurity depth profile in a 40-keV, 1×10^{16} atoms/cm^2 ^{75}As implant in Si. Also shown is a SIMS depth profile for comparison. The TEAMS and SIMS profiles are in good agreement until the SIMS detection limit for Cu is reached.

Data Analysis for Depth-Profiling Measurements

The raw data are measurements of count rate versus time. After analysis, the samples are removed from the ion source and the crater depths are measured with a surface profilometer. The measured depth gives a straightforward time-to-depth conversion assuming a constant sputtering rate, which is monitored via the matrix ion count rate. The count rate data are normalized to concentration (atoms per cubic centimeter). Normalization can be accomplished by scaling the peak count rate of the standard implant to the known concentration as well as by integrating the area under the standard profile and obtaining a yield factor. These methods are independent of each other. They have been described in detail by Wilson et al. (1989), Datar (1994), and Datar et al. (1997a, 1997b). The section on calibration of data acquisition below also has a sample demonstration of data normalization for an implant standard sample. The final depth profile is achieved by multiplying both axes of the raw data plot with the respective normalization factors. Typically, after a series of unknowns, another standard is run with the unknowns being sandwiched between standard samples. The final data are referred to the standards and the nominal blank. A nominally blank sample is run after measurements of an

unknown sample to get an estimate of cross-contamination from other samples and to make sure that running the unknown sample has not contaminated the source. Float-zone Si is a good blank sample.

Standard samples for depth-profiling measurements are implanted samples with known doses and known energies. However, any sample where the depth profile is known beforehand can be used as a standard. For bulk measurements, standard samples should be uniformly doped with the impurity of interest. Standards may be cross-checked in a number of TEAMS laboratories.

Interpretation of the Data. The final depth profile has exactly as much information as a SIMS profile, and in general, data interpretation has to take into account the same factors. Due to the extreme sensitivity, accuracy can only be compared to a known standard. Absolute measurements without reference to a standard are not possible. Also, for bulk measurements, one can only refer the data to a known standard in a manner similar to conventional AMS.

Calibration of Data Acquisition. Assuming a constant sputtering rate (checked by monitoring the Cs current before and after the measurement as well as by periodically measuring the matrix element current during the measurement), the time-to-depth conversion is simply done by measuring the depth of the crater after measurement using a profilometer. Converting the count rate to concentration is more involved. The SIMS RSFs cannot be used due to charge state fractionation as well as possible variations in transmission efficiency. In due course, TEAMS RSFs can be developed that take into account these factors as well. At present, calibration is done using implant standards. Assuming the depth profile is measured to exhaustion, the ratio of the integrated area under the depth profile to the implant dose gives the "yield" of the system. This information is used to normalize the unknown profile. The yield factor can also be used to normalize bulk impurity concentration measurements in the absence of uniformly doped standards. The concentration normalization factor (CNF) is given by Datar (1994) as

$$\mathrm{CNF} = \frac{\text{implanted dose (atoms / cm}^2\text{)}}{\text{area under profile} \times \text{sputtering rate (cm/s)}} \quad (3)$$

Then, the concentration at a given depth is given by

$$\text{Concentration} = \text{cps} \times \mathrm{CNF} \quad (4)$$

For example, Figure 14A shows a raw depth profile for a 100-keV, 2×10^{14} atoms/cm^2, ^{60}Ni implant in Si. The total number of Ni ions measured (obtained by integrating the area under the cps versus the time profile) was 4.287×10^6. The sputtering rate (obtained by measuring the depth of the crater after the measurement) was 2.6 Å/s. Since the implanted dose was 2×10^{14} atoms/cm^2,

$$\mathrm{CNF} = \frac{2 \times 10^{14}\ \text{atoms/cm}^2}{4.287 \times 10^6\ \text{atoms} \times 2.6 \times 10^{-8}\ \text{cm/s)}} = 1.82 \times 10^{15}\ \text{s/cm}^3 \quad (5)$$

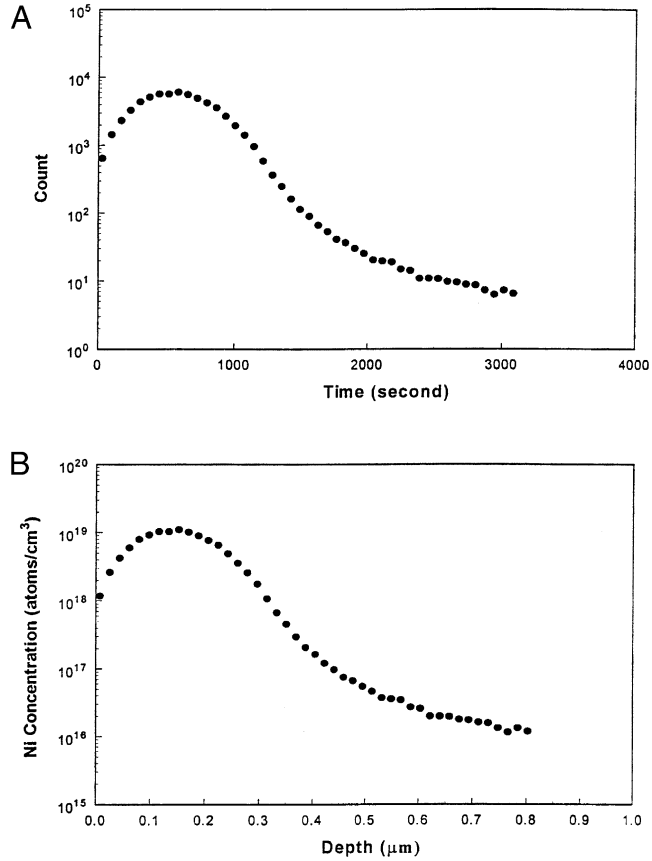


Figure 14. (**A**) Measured depth profile for a 100-keV, 2×10^{14} atoms/cm^2, ^{60}Ni implant in Si. (**B**) Normalized depth profile for Figure 14A showing concentration versus depth.

This means that a count rate of 1 cps corresponds to a concentration of 1.82×10^{15} atoms/cm^3. The detection limit is the least count multiplied by the normalization factor. Figure 14B shows the normalized depth profile as concentration (atoms per cubic centimeter) versus depth (micrometers). The two graphs are identical; just the x and y axes have been renormalized.

For the unknown sample, which has been run under the same conditions, the count rate numbers can be multiplied by the CNF to convert the depth profile into a concentration profile. Thus, the normalization is critically dependent upon the use of implant standards for depth profile measurements. For bulk measurements, an even better standard would be samples with known constant impurity levels. But basically, as in radioisotope AMS, the final numbers should always be referred to the standard used.

SAMPLE PREPARATION

Extreme care has to be taken during sample preparation to avoid surface particulate impurities. For depth profile measurements, it is desirable that the sample be flat and at least somewhat conductive. Typically, a TEAMS sample is a small square at least 4×4 mm. The samples are

loaded into the sample holder in a class 100 clean room. Care is taken to ensure that different kinds of samples do not come in contact with each other, essentially the same procedures as with SIMS.

SPECIMEN MODIFICATION

Trace element AMS is a destructive technique, and the samples cannot be reused for other measurements. In addition, the sputtering process itself affects the depth distribution of the impurity being measured due to ion beam mixing. Here, the sputtering ion actually drives the impurity further into the sample. This effect is more pronounced for higher Cs impact energies and causes depth profile broadening. This is the primary reason to attempt to use the lowest Cs impact energy possible that allows good transmission through the accelerator and analysis beamline. At UNT, the Cs impact energy is the sum of the 10 keV from the Cs oven and the energy from the negative sample bias, which has been operated as high as 30 kV, for a total impact energy of 40 keV. Recently, the sample bias has been reduced to 11 kV, which makes the Cs impact energy 21 keV, without significant loss of transmission through the entire system. The ranges of 21- and 40-keV Cs ions in Si are ~178 Å and ~266 Å, respectively. Therefore, 40-keV Cs ions penetrate Si ~1.5 times as far as 21-keV Cs ions.

PROBLEMS

Accidental contamination from particulates, elements in the vacuum system, and from sputtering neighboring samples is a hazard. Running samples in order from low concentration to high concentration if approximate concentration levels are known can minimize contamination. Sometimes, due to improper sample positioning, the transmission is unusually low. The matrix ion current measurement is usually a good indicator of this problem.

In general, sample positioning with respect to the extraction optics is the most critical factor affecting the reproducibility of the data. The ion source at UNT is being modified to improve the reproducibility in sample positioning.

Destroying a particle detector is possible by placing too much ion current on it due to unexpected matrix-based signal. Use of an ionization chamber for tune-up and initial measurements essentially solves this problem. The surface barrier particle detector can then be used to make higher resolution measurements.

A long measurement can take as long as an hour. Accelerator terminal voltage drift has to be monitored since this can affect transmission. Once again, periodically running standard samples or measuring matrix currents can be used to monitor the terminal voltage.

Over time, the analyzing magnet calibrations may change and will have to be recalibrated. Periodic measurements of known samples are the best safeguard against erroneous readings due to a number of different effects.

ACKNOWLEDGMENTS

The author would like to acknowledge the many funding sources that were required for the development of the TEAMS facility in the Ion Beam Modification and Analysis Laboratory (IBMAL) at the University of North Texas (UNT), including the State of Texas Coordinating Board Advanced Technology Program, the University of North Texas, Texas Instruments (TI), the Office of Naval Research, the National Science Foundation, and the Robert A. Welch Foundation. The TEAMS system at UNT would not have been possible without the technical support of Mark Anthony, Roy Beavers, and Tommy Bennett at TI.

The author would like to thank the many colleagues that have worked with him over the years on the design and development of the TEAMS system at UNT. Jerry Duggan, Sam Matteson, Duncan Weathers, Dwight Maxson, Jimmy Cobb, and Bobby Turner of UNT have been involved with the TEAMS development from the beginning. Graduate students and postdoctoral research scientists that have worked long and hard on TEAMS include Danny Marble, Joe Kirchhoff, György Vizkelethy, Sameer Datar, Yong-Dal Kim, Zhiyong Zhao, Steve Renfrow, Elaine Roth, Eugene Reznik, Baonian Guo, Karim Arrale, and David Gressett.

The author has benefited from many discussions with numerous people around the world that are experts in this field, including Soey Sie, Martin Suter, Ken Grabowski, Graham Hubler, David Knies, Barney Doyle, Ken Purser, Drew Evans, David Elmore, Harry Gove, Liam Kieser, Graham Wilson, Eckehart Nolte, Colin Maden, Hans-Arno Synal, John Vogel, Jay Davis, Doug Donahue, and Greg Norton. I especially thank Sameer Datar, who helped write parts of this unit and for his critical reading, and my wife Linda McDaniel for grammatical corrections. I particularly want to thank Mark Anthony for his belief in this project, for his continued support over the past few years, and for reading and commenting on the manuscript.

LITERATURE CITED

Alvarez, L. W. and Cornog, R. 1939a. ^{3}He in helium. *Phys. Rev.* 56:379–379.

Alvarez, L. W. and Cornog, R. 1939b. Helium and hydrogen of mass 3. *Phys. Rev.* 56:613–613.

Anthony, J. M. and Donahue, D. J. 1987. Accelerator mass spectrometry solutions to semiconductor problems. *Nucl. Instrum. Methods Phys. Res.* B29:77–82.

Anthony, J. M. and Thomas, J. 1983. Frontier analysis of stable isotopes present as contaminants or impurity dopants in ultrapure semiconductor materials. *Nucl. Instrum. Methods* 218:463–467.

Anthony, J. M., Donahue, D. J., Jull, A. J. T., and Zabel, T. H. 1985. Detection of semiconductor dopants using accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B10/11: 498–500.

Anthony, J. M., Donahue, D. J., and Jull, A. J. T. 1986. Super SIMS for ultra-sensitive impurity analysis. *Mat. Res. Soc. Symp. Proc.* 69:311–316.

Anthony, J. M., Matteson, S., McDaniel, F. D., and Duggan, J. L. 1989. Accelerator mass spectrometry at the University of North Texas. *Nucl. Instrum. Methods Phys. Res.* B40/41: 731–733.

Anthony, J. M., Matteson, S., Marble, D. K., Duggan, J. L., McDaniel, F. D., and Donahue, D. J. 1990a. Applications of accelerator mass spectrometry to electronic materials. *Nucl. Instrum. Methods Phys. Res.* B50:262–266.

Anthony, J. M., Matteson, S., Duggan, J. L., Elliott, P., Marble, D. K., McDaniel, F. D., and Weathers, D. L. 1990b. Atomic mass spectrometry of materials. *Nucl. Instrum. Methods Phys. Res.* B52:493–497.

Anthony, J. M., Beavers, R. L., Bennett, T. J., Matteson, S., Marble, D. K., Weathers, D. L., McDaniel, F. D., and Duggan, J. L. 1991. Materials characterization using accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B56/57: 873–876.

Anthony, J. M., Kirchhoff, J. F., Marble, D. K., Renfrow, S. N., Kim, Y. D., Matteson, S., and McDaniel, F. D. 1994a. Accelerator based secondary ion mass spectrometry for impurity analysis. *J. Vac. Sci. Tech.* A12:1547–1550.

Anthony, J. M., McDaniel, F. D., and Renfrow, S. N. 1994b. Molecular free SIMS for semiconductors. *In* Diagnostic Techniques for Semiconductor Materials and Devices (D. K. Schroder, J. L. Benton, and P. Rai-Choudhury, eds.). Proceedings of the Electrochemical Society 94:349–354.

Arrale, A. M., Matteson, S., McDaniel, F. D., and Duggan, J. L. 1991. Energy loss corrections for MeV ions in tandem accelerator stripping. *Nucl. Instrum. Methods Phys. Res.* B56/57:886–888.

Baumann, H. and Pavel, J. 1989. Determination of trace impurities in electronic grade quartz: Comparison of inductively coupled plasma mass spectroscopy with other analytical techniques. *Mikrochim. Acta [Wien]* 111:413–422.

Bennett, C. L., Beukens, R. P., Clover, M. R., Gove, H. E., Liebert, R. B., Litherland, A. E., Purser, K. H., and Sondheim, W. E. 1977. Radiocarbon dating using electrostatic accelerators: Negative ions the key. *Science* 198:508–510.

Benninghoven, A., Rüdenauer, F.G., and Werner, H. W. 1987. Secondary Ion Mass Spectrometry: Basic Concepts, Instrumental Aspects, Applications, and Trends. John Wiley & Sons, New York.

Betz, H. D. 1972. Charge state and charge-changing cross sections of fast heavy ions penetrating through gaseous and solid media. *Rev. Mod. Phys.* 44:465–507.

Bigelow, A. W., Li, S. L., Matteson, S., and Weathers, D. L. 1999. Sputter-initiated resonance ionization spectroscopy at the University of North Texas. *AIP Conf. Proc.* 475:569–572.

Bigelow, A. W., Li, S. L., Matteson, S., and Weathers, D. L. 2001. Measured energy and angular distributions of sputtered neutral atoms from a Ga-In eutectic alloy target. *AIP Conf. Proc.* 576:108–111.

Burdick, G. W., Appling, J. R., and Moran, T. F. 1986. Stable multiply charged molecular ions. *J. Phys. B* 19:629–641.

Datar, S. A. 1994. The Diffusion of Ion-Implanted Cl in Si. Ph.D. thesis, University of Rochester.

Datar, S. A., Gove, H. E., Teng, R. T. D., and Lavine, J. P. 1995. AMS studies of the diffusion of Cl in Si wafers. *Nucl. Instrum. Methods* B99:549–552.

Datar, S. A., Renfrow, S. N., Anthony, J. M., and McDaniel, F. D. 1996. Trace-element accelerator mass spectrometry: A new technique for low-level impurity measurements in semiconductors. *Mat. Res. Soc. Symp. Proc.* 406:395–400.

Datar, S. A., Renfrow, S. N., Guo, B. N., Anthony, J. M., Zhao, Z. Y., and McDaniel, F. D. 1997a. TEAMS depth profiles in semiconductors. *Nucl. Instrum. Methods Phys. Res.* B123:571–574.

Datar, S. A., Zhao, Z. Y., Renfrow, S. N., Guo, B. N., Anthony, J. M., and McDaniel, F. D. 1997b. High sensitivity impurity measurements in semiconductors using trace-element accelerator mass spectrometry (TEAMS). *AIP Conf. Proc.* 392:815–818.

Datar, S. A., Wu, L., Guo, B. N., Nigam, M., Necsoiu, D., Zhai, Y. J., Smith, D. E., Yang, C., El Bouanani, M., Lee, J. J., and McDaniel, F. D. 2000. High sensitivity measurement of implanted As in the presence of Ge in Ge_xSi_{1-x}/Si layered alloys using trace element accelerator mass spectrometry. *Appl. Phys. Lett.* 77:3974–3976.

Döbeli, M., Nebiker, P. W., Suter, M., Synal, H.-A., and Vetterli, D. 1994. Accelerator SIMS for trace element detection. *Nucl. Instrum. Methods* B85:770–774.

Elmore, D. and Phillips, F. M. 1987. Accelerator mass spectrometry for measurement of long-lived radioisotopes. *Science* 236:543–550.

Elmore, D., Conard, N., Kubik, P. W., and Fabryka-Martin, J. 1984. Computer controlled isotope ratio measurements and data analysis. *Nucl. Instrum. Methods B*5:233–237.

Elmore, D., Hossain, T. Z., Gove, H. E., Hemmick, T. K., Kubik, P. W., Jiang, S., Lavine, J. P., and Lee, S. T. 1989. Depth profiles of nitrogen and Cl in pure materials through AMS of the neutron activation products ^{14}C and ^{36}Cl. *Radiocarbon* 31:292–297.

Ender, R. M. 1997. Analyse von Spurenelementen mit Beschleuniger-Sekundärionen-Massenspektrometrie. Dissertation, Eidgenössiche Technische Hochschule, Zurich.

Ender, R. M., Döbeli, M., Suter, M., and Synal, H.-A. 1997a. Accelerator SIMS at PSI/ETH Zurich. *Nucl. Instrum. Methods Phys. Res.* B123:575–578.

Ender, R. M., Döbeli, M., Suter, M., and Synal, H.-A. 1997b. Characterization of the accelerator SIMS setup at PSI/ETH Zurich. *AIP Conf. Proc.* 392:819–822.

Ender, R. M., Döbeli, M., Nebiker, P., Suter, M., and Synal, H.-A., Vetterli, D., 1997c. A new setup for accelerator SIMS. *Proc. SIMS* 10:1018 –1022.

Ender R. M., et al. 1997d. *Applications of Accelerator SIMS.* PSI Annual Report Annex IIIA, p. 43.

Gillen, G., Lareau, R., Bennett, J., and Stevie, F. 1998. Secondary Ion Mass Spectrometry, SIMS XI. John Wiley & Sons, New York.

Gove, H. E., Kubik, P. W., Sharma, P., Datar, S., Fehn, U., Hossain, T. Z., Koffer, J., Lavine, J. P., Lee, S.-T., and Elmore, D. 1990. Applications of AMS to electronic and silver halide imaging research. *Nucl. Instrum. Methods Phys. Res.* B52:502–506.

Grabowski, K., Knies, D. L., Hubler, G. K., and Enge, H. A. 1997. A new accelerator mass spectrometer for trace element analysis at the Naval Research Laboratory. *Nucl. Instrum. Methods Phys. Res.* B123:566–570.

Houk, R. S. 1986. Spectrometry of inductively coupled plasmas. *Anal. Chem.* 58:97A-105A.

Hubbard, K. M., Weller, R. A., Weathers, D. L., and Tombrello, T. A. 1989. The angular distribution of atoms sputtered from a Ga-In eutectic alloy target. *Nucl. Instrum. Methods* B36: 395–403.

Jull, A. J. T., Beck, J. W., and Burr, G. S. 1997. *Nucl. Instrum. Methods Phys. Res.* B123:1–612.

Kilius, L. R., Baba, N., Garwan, M. A., Litherland, A. E., Nadeau, M.-J., Rucklidge, J. C., Wilson, G. C., and Zhao, X.-L., 1990. AMS of heavy ions with small accelerators. *Nucl. Instrum. and Methods in Phys. Res. B*52:357–365.

Kim, Y. D., Jin, J. Y., Matteson, S., Weathers, D. L., Anthony, J. M., Marshall, P., and McDaniel, F. D. 1995. Collision-induced interaction cross sections of 1-7 MeV B_2 ions incident on an N_2 gas target. *Nucl. Instrum. Methods Phys. Res.* B99:82–85.

Kirchhoff, J. F., Marble, D. K., Weathers, D. L., McDaniel, F. D., Matteson, S., Anthony, J. M., Beavers, R. L., and Bennett, T. J. 1994. Fabrication of Si-based optical components for an ultra-clean accelerator mass spectrometry ion source. *Rev. Sci. Instrum.* 65:1570–1574.

Knies, D. L., Grabowski, K. S., Hubler, G. K., and Enge, H. A. 1997a. 1-200 amu tunable Pretzel magnet notch-mass-filter and injector for trace element accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B123:589–593.

Knies, D. L., Grabowski, K. S., Hubler, G. K., Treacy Jr., D. J., DeTurck, T. M., and Enge, H. A. 1997b. Status of the Naval Research Laboratory trace element accelerator mass spectrometer: Characterization of the pretzel magnet. *AIP Conf. Proc.* 392:783–786.

Kruger, P. 1971. Principles of Neutron Activation Analysis. Wiley-Interscience, New York.

Litherland, A. E., Beukens, R. P., Doupe, J., Kieser,W. E., Krestow, J., Rucklidge, J. C., Tomski, I., Wilson, G. C., and Zhao, X.-L. 2000. Progress in accelerator mass spectrometry (AMS) research at IsoTrace. *Nucl. Instrum. Methods Phys. Res.* B172:206–210.

Litherland, A. E., Doupe, J. P., Tomski, I., Krestow, J., Zhao, X.-L., Kieser, W. E., and Beukens, R. P. 2001. Ion preparation systems for atomic isobar reduction in accelerator mass spectrometry. *AIP Conf. Proc.* 576:390–393.

Lu, S. 1997. Secondary ion mass spectrometry (SIMS) applications on characterization of RF npn bipolar polySi emitter process. *AIP Conf. Proc.* 392:823–826.

Maden, C., Döbeli, M., and Hoffman, B. 1999. Investigation of platinum group elements with accelerator SIMS. PSI Annual Report, Vol. 1, p. 199 (available at *http://ihp-power1.ethz.ch/IPP/tandem/Annual/1999.html*).

Maden, C., Frank, M., Suter, M., Kubik, P. W., and Döbeli, M. 2000. Investigation of natural ^{10}Be/Be ratios with accelerator mass spectrometry. PSI Annual Report, Vol. 1, p. 164 (available at *http://ihp-power1.ethz.ch/IPP/tandem/Annual/2000.html*).

Massonet, S., Faude, Ch., Nolte, E., and Xu, S. 1997. AMS with stable isotopes and primordial radionuclides for material analysis and background detection. *AIP Conf. Proc.* 392:795–797.

Massonet, S. 1998. Beschleunigermassenspektrometrie mit stabilen Isotopen und primordialen radionukliden. Ph.D Thesis, Technical University, Munich

Matteson, S., McDaniel, F. D., Duggan, J. L., Anthony, J. M., Bennett, T. J., and Beavers, R. L. 1989. A high resolution electrostatic analyzer for accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B40/41:759–761.

Matteson, S., Marble, D. K., Hodges, L. S., Hajsaleh, J. Y., Arrale, A. M., McNeir, M. R., Duggan, J. L., McDaniel, F. D., and Anthony, J. M. 1990. Molecular-interference-free accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B45:575–579.

McDaniel, F. D., Matteson, S., Weathers, D. L., Marble, D. K., Duggan, J. L., Elliott, P. S., Wilson, D. K., and Anthony, J. M. 1990a. The University of North Texas atomic mass spectrometry facility for detection of impurities in electronic materials. *Nucl. Instrum. Methods Phys. Res.* B52:310–314.

McDaniel, F. D., Matteson, S., Marble, D. K., Hodges, L. S., Arrale, A. M., Hajsaleh, J. Y., McNeir, M. R., Duggan, J. L., and Anthony, J. M. 1990b. Molecular dissociation in accelerator mass spectrometry for trace impurity characterization in electronic materials. *In* Proceedings of the High Energy and Heavy Ion Beams in Materials Analysis Workshop, Albuquerque, NM, June 14–16, 1989 (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). pp. 269–278. Materials Research Society, Pittsburgh, PA.

McDaniel, F. D., Matteson, S., Weathers, D. L., Duggan, J. L., Marble, D. K., Hassan, I., Zhao, Z. Y., and Anthony, J. M. 1992. Radionuclide dating and trace element analysis by accelerator mass spectrometry. *J. Radioanal. Nucl. Chem.* 160:119–140.

McDaniel, F. D., Matteson, S., Anthony, J. M., Weathers, D. L., Duggan, J. L., Marble, D. K., Hassan, I., Zhao, Z. Y., Arrale, A. M., and Kim, Y. D. 1993. Trace element analysis by accelerator mass spectrometry. *J. Radioanal. Nucl. Chem.* 167:423–432.

McDaniel, F. D., Anthony, J. M., Kirchhoff, J. F., Marble, D. K., Kim, Y. D., Renfrow, S. N., Grannan, E. C., Reznik, E. R., Vizkelethy, G., and Matteson, S. 1994. Impurity determination in electronic materials by accelerator mass spectrometry. *Nucl. Instrum. Methods Phys. Res.* B89:242–249.

McDaniel, F. D., Anthony, J. M., Renfrow, S. N., Kim, Y. D., Datar, S. A., and Matteson, S. 1995. Depth profiling analysis of semiconductor materials by accelerator mass spectrometry. *Nucl. Instrum. Methods.* B99:537–540.

McDaniel, F. D., Datar, S. A., Guo, B. N., Renfrow, S. N., Matteson, S., Zoran, V. I., Zhao, Z. Y., and Anthony, J. M. 1998a. A method to reduce molecular interferences in SIMS: Trace-element accelerator mass spectrometry (TEAMS). *In* Proceedings of the Eleventh Inter. Conf. On Secondary Ion Mass Spectrometry, Orlando, FL, September 7–12, 1997 (G. Gillen, R. Lareau, J. Bennett, and F. Stevie, eds.). pp. 167–170. John Wiley & Sons, New York.

McDaniel, F. D., Datar, S. A., Guo, B. N., Renfrow, S. N., Zhao, Z. Y., and Anthony, J. M. 1998b. Low-level copper concentration measurements in Si wafers using trace-element accelerator mass spectrometry (TEAMS). *Appl. Phys. Lett.* 72:3008–3010.

Morvay, L. and Cornides, I. 1984. Triply charged ions of small molecules. *Int. J. Mass Spectrom. Ion Process.* 62:263–268.

Muller, R. A. 1977. Radioisotope dating with a cyclotron. *Science* 196:489–494.

Nelson, D. E., Korteling, R. G., and Stott, W. R. 1977. Carbon-14: Direct detection at natural concentrations. *Science* 198:507–508.

Niklaus, Th.R., Sie, S. H., and Suter, G. F. 1997. Australis: A microbeam AMS beamline. *AIP Conf. Proc.* 392:779–782.

Pellin, M. J., Young, C. E., Calaway, W. F., Burnett, J. W., Jorgensen, B., Schweitzer, E. L., and Gruen, D. M. 1987. Sensitive, low damage surface analysis using resonance ionization of sputtered atoms. *Nucl. Instrum. Methods* B18:446–451.

Pellin, M. J., Young, C. E., Calaway, W. F., Whitten, J. E., Gruen, D. M., Blim, J. D., Hutcheon, I. D., and Wasserburg, G. J. 1990. Secondary neutral mass spectrometry using 3-colour resonance ionization: Osmium detection at the ppb level and ion-detection in Si at the less than 200 ppt level. *Philos. Trans. R. Soc. Lond. A*333:133–146.

Purser, K. H., Liebert, R. B., Litherland, A. E., Beukens, R. P., Gove, H. E., Bennett, C. L., Clover, M. E., and Sondheim,

W. E. 1977. An attempt to detect stable atomic nitrogen (–) ions from a sputter source and some implications of the results for the design of tandems for ultra-sensitive carbon analysis. *Rev. Appl. Phys.* 12:1487–1492.

Purser, K. H., Litherland, A. E., and Rucklidge, J. C. 1979. Secondary ion mass spectrometry using dc accelerators. *Surf. Interface Anal.* 1:12–19.

Rucklidge, J. C., Evensen, N. M., Gorton, M. P., Beukens, R. P., Kilius, L. R., Lee, H. W., Litherland, A. E., Elmore, D., Gove H. E., and Purser, K. H. 1981. Rare isotope detection with tandem accelerators. *Nucl. Instrum. Methods* 191:1–9.

Rucklidge, J. C., Gorton, M. P., Wilson, G. C., Kilius, L. R., Litherland, A. E., Elmore, D., and Gove, H. E. 1982. Measurement of Pt and Ir at sub-ppb levels using tandem accelerator mass spectrometry. *Can. Mineral.* 20:111–119.

Rucklidge, J. C., Wilson, G. C., and Kilius, L. R. 1990a. AMS advances in the geosciences and heavy-element analysis. *Nucl. Instrum. Methods Phys. Res.* B45:565–569.

Rucklidge, J. C., Wilson, G. C., and Kilius, L. R. 1990b. In situ trace element determination by AMS. *Nucl. Instrum. Methods Phys. Res.* B52:507–511.

Rucklidge, J. C., Wilson, G. C., Kilius, L. R., and Cabri, L. J. 1992. Trace element analysis of sulfide concentrates from Sudbury by accelerator mass spectrometry. *Can. Mineral.* 30:1023–1032.

Saunders, W. A. 1989. Charge exchange and metastability of small multiply charged gold clusters. *Phys. Rev. Lett.* 62: 1037–1040.

Sharara, N. A, Wilson, G. C., and Rucklidge, J. C. 1999. Platinum-group elements and gold in Cu-Ni mineralized peridotite at Gabbro Akarem, Eastern Desert, Egypt. *Can. Mineral.* 37:1081–1097.

Sie, S. H. and Suter, G. F. 1994. A microbeam AMS system for mineralogical applications. *Nucl. Instrum. Methods Phys. Res.* 92:221–226.

Sie, S. H., Niklaus, T. R., and Suter, G. F. 1997a. Microbeam AMS: Prospects of new geological applications. *Nucl. Instrum. Methods Phys. Res.* B123:112–121.

Sie, S. H., Niklaus, T. R., and Suter, G. F. 1997b. The AUSTRALIS microbeam ion source. *Nucl. Instrum. Methods Phys. Res.* B123:558–565.

Sie, S. H., Niklaus, T. R., and Suter, G. F. 1997c. Prospects of new geological applications with a microbeam-ams. *AIP Conf. Proc.* 392:799–802.

Sie, S. H., Niklaus, T. R., Suter, G. F., and Bruhn, F. 1998. A microbeam Cs source for accelerator mass spectrometry. *Rev. Sci. Instrum.* 69:1353–1358.

Sie, S. H., Sims, D. A., Bruhn, F. Suter, G. F., and Niklaus, T. R. 1999. Microbeam AMS for trace element and isotopic studies. *AIP Conf. Proc.* 475:648–651.

Sie, S. H., Sims, D. A., Suter, G. F., Cripps, G. C., Bruhn, F., and Niklaus, T. R. 2000a. Precision Pb and S isotopic ratio measurements by microbeam AMS. *Nucl. Instrum. Methods* B172:228–234.

Sie, S. H., Sims, D. A., Niklaus, T. R., and Suter, G. F. 2000b. A fast bouncing system for the high energy end of AMS. *Nucl. Instrum. Methods* B172:268–273.

Sie, S. H., Sims, D. A., and Suter, G. F. 2001. Microbeam AMS measurements of PGE and Au trace and osmium isotopic ratios. *AIP Conf. Proc.* 576:399–402.

Vandervorst, W. and Shepherd, F. R. 1987. Secondary ion mass spectrometry profiling of shallow implanted layers using quadrupole and magnetic sector instruments. *J. Vac. Sci. Technol.* A5:313–320.

Weathers, D. L., McDaniel, F. D., Matteson, S., Duggan, J. L., Anthony, J. M., and Douglas, M. A. 1991. Triply-ionized B_2 molecules from a tandem accelerator. *Nucl. Instrum. Methods Phys. Res.* B56/57:889–892.

Wiebert, A., Erlandsson, B., Hellborg, R., Stenström, K., and Skog, G. 1994. The charge state distributions of carbon beams measured at the Lund pelletron accelerator. *Nucl. Instrum. Methods Phys. Res.* B89:259–261.

Wilson, G. C. 1998. Economic applications of accelerator mass spectrometry. *Rev. Econ. Geol.* 7:187–198.

Wilson, G. C., Rucklidge, J. C., and Kilius, L. R. 1990. Sulfide gold content of skarn mineralization at Rossland, British Columbia. *Econ. Geol.* 85:1252–1259.

Wilson, G. C., Kilius, L. R., and Rucklidge, J. C. 1995. Precious metal contents of sulfide, oxide and graphite crystals: Determinations by accelerator mass spectrometry. *Econ. Geol.* 90:255–270.

Wilson, G. C., Rucklidge, J. C., Kilius, L. R., Ding, G.-J., and Zhao, X.-L. 1997a. Trace-element analysis of mineral grains using accelerator mass spectrometry—from sampling to interpretation. *Nucl. Instrum. Methods Phys. Res.* B123:579–582.

Wilson, G. C., Rucklidge, J. C., Kilius, L. R., Ding, G.-J., and Cresswell, R. G. 1997b. Precious metal abundances in selected iron meteorites: In-situ AMS measurements of the six platnium-group elements plus gold. *Nucl. Instrum. Methods Phys. Res.* B123:583–588.

Wilson, G. C., Pavlish, L. A., Ding, G.-J., and Farquhar, R. M. 1997c. Textural and in-situ analytical constraints on the provenance of smelted and native archaeological copper in the Great Lakes region of eastern North America. *Nucl. Instrum. Methods Phys. Res.* B123:498–503.

Wilson, R. G., Stevie, F. A., and Magee, C. W. 1989. Secondary Ion Mass Spectrometry: A Practical Handbook for Depth Profiling and Bulk Impurity Analysis. John Wiley & Sons, New York.

Winograd, N., Baxter, J. P., and Kimock, F. M. 1982. Multiphonon resonance ionization of sputtered neutrals: A novel approach to materials characterization. *Chem. Phys. Lett.* 88:581–584.

Wittmaack, K. 1995. Small-area depth profiling in a quadrupole based SIMS instrument. *Int. J. Mass Spectrum Ion Process.* 143:19.

Zhao, Z. Y., Mehta, S., Angel, G., Datar, S. A., Renfrow, S. N., McDaniel, F. D., and Anthony, J. M. 1997. Use of accelerator mass spectrometry for trace element detection. *In* Proceedings of the Eleventh International Conference on Ion Implantation Technology, Austin, TX, June 16–21, 1996, Vol. 11 (E. Ishidida, S. Banerjee, S. Mehta, T. C. Smith, M. Current, L. Larson, A. Tasch, and T. Romig, eds.). pp. 131–134. IEEE, Piscataway, NJ.

KEY REFERENCES

Benninghoven et al., 1987. See above.

An excellent reference for SIMS.

Datar et al., 1997a. See above.

Discusses TEAMS techniques to depth profile impurities in materials.

Datar et al., 2000. See above.

A reference to using TEAMS to measure As in the presence of Ge in Ge_xSi_{1-x} layers where a number of mass interferences exist.

Kirchhoff et al., 1994. See above.

Discusses how to construct an ultraclean ion source for TEAMS measurements.

McDaniel et al., 1998b. See above.

A reference to TEAMS measurements of low-level Cu impurities in Si.

Sie et al., 1998. See above.

Discusses a microbeam ion source for TEAMS.

Wilson et al., 1989. See above.

A complete reference to SIMS measurement techniques and contains many useful tables and figures to help determine negative ion formation probabilities and relative sensitivity factors.

INTERNET RESOURCES

http://mstd.nrl.navy.mil/6370/ams.html

The Naval Research Laboratory, Washington, DC. Currently constructing a TEAMS facility by mating a SIMS instrument to a tandem accelerator. The NRL facility features parallel injection of a range of masses with a pretzel magnet and detection using a split-pole spectrograph. The contact person is Dr. Kenneth E. Grabowski. Email: grabowski@nrl.navy.mil.

http://www.ams.physik.tu-muenchen.de

The AMS group at the Technical University Munich. Performs radionuclide AMS and has recently built an ultraclean ion source for stable isotope impurities for materials analysis. The project leader is Dr. Eckehart Nolte. Email: eckehart_nolte@physik.tu-muenchen.de.

http://www.cea.com/tutorial.htm

Charles Evans and Associates, specialists in materials characterization. Provides an excellent SIMS tutorial.

http://www.phys.ethz.ch/ipp/tandem

The AMS facility at PSI/ETH Ion Beam Physics Institute of Particle Physics HPK, ETH Hönggerberg, CH-8093, Zurich, Switzerland. Mainly involved in radionuclide AMS and has built a separate ion source for TEAMS measurements. Dr. Martin Suter is project leader. Email: suter@particle.phys.ethz.ch.

http://www.physics.utoronto.ca/~isotrace

The IsoTrace Laboratory in the Department of Physics at the University of Toronto, Toronto, Canada. One of the first laboratories to make radionuclide AMS measurements and also to make stable isotope measurements of the platinum group elements, Au, and Ag in mineral grains and meteorites. The project leader is Dr. W. E. "Liam" Kieser. Email: liam.kieser@utoronto.ca.

http://www.phys.unt.edu/ibmal

The TEAMS facility at the Ion Beam Modification and Analysis Laboratory, Department of Physics, University of North Texas, Denton, Texas. Used to routinely analyze bulk materials as well as depth profiles of impurities. Dr. Floyd Del McDaniel is the project leader of the TEAMS research at UNT. Email: mcdaniel@unt.edu.

http://www.syd.dem.CSIRO.AU/research//hiaf/AUSTRALIS

The AUSTRALIS project at CSIRO, Sydney, Australia. An AMS system with a microprobe ion source. This system is designed primarily for in situ microanalysis of geological samples. The web site is case sensitive. Dr. Soey Sie is project leader at CSIRO in Australia. Email: soey.sie@dem.csiro.au.

APPENDIX A: GLOSSARY OF TERMS

AMS	Accelerator mass spectrometry
Charge state	The amount of charge on an ion. Usually refers to positive ions since only singly charged negative ions have been observed. Obviously, the maximum charge state of an elemental ion corresponds to its atomic number.
Class-100 clean room	A room with atmospheric particulate concentration of particles greater than 0.5 μm diameter not exceeding 100 particles/ft^3 (or 3500 particles/m^3)
CNF	Concentration normalization factor
cps	Counts per second
Depth profile	Plot of elemental concentration as a function of depth from the surface
Detection limit	Lowest meaningful concentration level that can be measured
ICPMS	Inductively coupled plasma mass spectrometry
Ionization chamber	An instrument used to measure the energy of a particle. It is basically a parallel-plate capacitor filled with a gas. Energetic particles entering the chamber lose energy by ionizing the gas. The amount of ionization produced is proportional to the energy. It is a very sensitive instrument capable of detecting single particles.
ion source	Produces ions for injection into accelerator
INAA	Instrumental neutron activation analysis
NAA	Neutron activation analysis
NA-AMS	Neutron activation followed by acceler- ator mass spectroscopy
NRA	Nuclear reaction analysis
PGEs	Platinum group elements (Ru, Rh, Pd, Os, Ir, Pt)
PIXE	Particle-induced X-Ray emission
ppb	parts per billion ($1/10^9$)
ppm	parts per million ($1/10^6$)
ppt	parts per trillion ($1/10^{12}$)
Primary ion	Ion incident on sample surface that sputters the sample. Typically for TEAMS it is Cs^+.
RBS	Rutherford backscattering spectroscopy
RSFs	Relative sensitivity factors
Secondary ion	Ion sputtered from the sample by the primary ion. Can be positive or negative. Neutral atoms or molecules can also be sputtered from the sample surface.
Sensitivity	Lowest measurable concentration or detection limit in atoms per cubic centimeters
SIMS	Secondary-ion mass spectrometry
SIRIS	Sputter-initiated resonance ionization spectrometry
Solid-state or surface barrier	A semiconductor detector that works in a manner similar to an ionization

detector	chamber but is substantially more compact. Energetic particles incident on the detector create electron-hole pairs in the depletion region of a p-n junction. The electrons and holes are swept off in opposite directions by an applied electric field, constituting a current pulse that generates a signal. The signal strength is roughly proportional to the energy of the incident particles.
Tandem accelerator	Electrostatic accelerator based on the Van De Graaff principle, with a high positive potential at the center terminal. The center terminal is charged to a high potential by charges carried from ground by a belt or a chain. The accelerating column is enclosed in a pressure vessel filled with SF_6 or other insulating gases.
TEAMS	Trace element accelerator mass spectrometry or trace element atomic mass spectrometry
TOF-SIMS	Time of flight–secondary ion mass spectrometry

APPENDIX B: INSTRUMENT SPECIFICATIONS AND SUPPLIERS

1. Cs gun: General Ionex, model 133 surface ionization Cs source. This gun is no longer made. Very similar guns are sold by Peabody Scientific, Atomika.
2. Cs analyzing magnet: Made by Magnecoil, Peabody, MA. Supplied by National Electrostatics Corp.
3. Cs Analyzing Magnet Power supply, TCR 20T125: Made by Electronic Measurements.
4. Sample chamber: Made to specification by Huntington Mechanical Laboratories, 1040 L'Avenida, Mountain View, CA 94043. Chamber features four-axis manipulator and load-lock for sample changing.
5. Electron flood gun for charge compensation; model EFG-8 with EGPS-8 power supply: Kimball Physics, 311 Kimball Hill Rd. Wilton NH, 03086.
6. 45° secondary-ion electrostatic analyzer: In-house fabrication.
7. 90° analyzing magnet: Supplied by National Electrostatics Corp.
8. 3MV, electrostatic, tandem pelletron accelerator, model 9SDH-2: Part of the accelerator beamline, supplied by National Electrostatics Corp., 7540 Graber Rd. Box 620310, Middleton, WI 53562-0310.
9. 40° bending magnet: HVEE Europa BV PO Box 99, 3800 AB Amersfoort, The Netherlands. Mass-Energy product $(mE/q^2) \sim 60$.
10. 45° double-focusing spherical electrostatic analyzer: In-house fabrication.
11. Si surface-barrier particle detector: Ortec, 801 S. Illinois Ave., Oak Ridge, Tenn. 37831-0895.
12. Total energy detector ionization chamber: In-house development.
13. Data acquisition and control software: Developed in-house using LabWindows and LabView 4.1 (National Instruments) for PCs using a CAMAC crate with a GPIB interface.

FLOYD DEL MCDANIEL
Ion Beam Modification and Analysis Laboratory
University of North Texas
Denton, Texas

INTRODUCTION TO MEDIUM-ENERGY ION BEAM ANALYSIS

INTRODUCTION

The earliest applications of ion beam techniques for materials analysis, typified by the work of Tollestrup et al. (1949), were carried out with mega-electron-volt beams in the course of basic research in nuclear physics, and the mega-electron-volt beam techniques (described in Section 12a of this chapter) still form the indispensable core of the subject. In the broadest sense, the importance of these techniques lies in the fact that the information they yield depends upon nuclear properties and interactions and is, therefore, relatively insensitive to perturbation by local properties such as the chemical environment and electronic structure. The relevant nuclear physics and scattering theory are well known experimentally and often theoretically, and in consequence, important parameters such as interaction cross-sections are known from first principles (PARTICLE SCATTERING). This is in sharp contrast to the situation for some other important techniques (e.g., secondary ion mass spectrometry) where the fundamental measurement interactions may be very vulnerable to the unique idiosyncrasies of a specimen. It is essential to stress the unity of ion beam analysis even as one seeks to differentiate the relatively newer medium-energy techniques from their more mature high-energy predecessors.

First it is important to understand what is meant by "medium energy." There are two answers, one historical and one based upon the physics of scattering. Physically, ion beam techniques may be termed medium energy if the energy of the beam is low enough that the deviation of the scattering cross-section from the unscreened Rutherford value is greater than ~1%. In addition, the energy must be high enough to assume both that the beam interacts with the target by binary elastic collisions (except when it is traveling in a channeling direction) and that most backscattered particles remain ionized. (The latter constraint differentiates medium- from low-energy

ion scattering, where it is assumed that the ionized component of backscattered particles is strongly correlated with scattering events at the specimen's surface.) For practical purposes, medium energy encompasses ion beam energies from a few tens of to a few hundreds of kilo-electron-volts, and even a bit higher for heavy ions being used in forward-recoil work. Below this range, ion scattering becomes increasingly surface specific. Above it lies the full range of conventional elastic scattering and nuclear reaction analysis.

Historically, medium energy has been defined by technology. The field of ion beam analysis owes its ubiquity, if not its existence, to a single device, the silicon surface-barrier particle detector (see the review by McKenzie, 1979, and references therein). A surface-barrier detector is essentially a large-area Schottky-barrier diode. It has approximately unit efficiency for detecting particle radiation, produces an output signal that is highly linear with the energy of the particle striking it, and is remarkably inexpensive to produce and operate. Along with the lead pencil and the digital compact disk, it is an example of a technology that, from its inception, was so elegant in design, so well suited to its purpose, and so demonstrably superior in function and economics to competitive alternatives that it at once revolutionized a field of activity. However, surface-barrier detectors experience diminished relative energy resolution when used for α particle radiation below $\sim$500 keV. As a result, for years almost all ion beam analysis has been conducted using projectiles with energy greater than this.

At the low end of the energy scale, a technique called low-energy ion scattering (known both by the acronym LEIS and, in early literature, as ISS, for ion surface scattering) was used to study surface structure. Typically, this work was done by pulsing the accelerator beam and using an electron multiplier with exposed cathode to detect scattered ions. Formidable technical problems with producing adequately short beam pulses, as well as other fundamental issues, placed a practical upper limit on this technique of 20 to 30 keV. Thus, the historic range of medium energy began above LEIS and ranged up to the domain of the surface-barrier detector.

As is the case for high-energy ion beam analysis, most of the tools needed for medium-energy ion beam analysis have been developed and perfected for nuclear research and are widely available commercially. This is notably not the case for spectrometers, which still fall largely in the class of research instrumentation and are available, but less readily so.

The units that follow describe some medium-energy ion beam analytical techniques that are optimized for different problems. Time-of-flight medium-energy backscattering and forward-recoil spectrometry, described in MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY, are quite similar to the high-energy techniques of Section 12a but offer increased depth resolution, sensitivity, and surface specificity at the expense of total analyzable depth and ease of use. Forward-recoil measurement of hydrogen at medium energies can be particularly effective, especially when it is important to measure hydrogen and other light elements simultaneously.

Heavy-ion backscattering spectrometry, described in HEAVY-ION BACKSCATTERING SPECTROMETRY, is a variant of time-of-flight backscattering that is optimized for the highest possible sensitivity, specifically for the measurement of ultra-low-level metallic contaminants on device-grade Si wafer surfaces. Along with total-reflection x-ray fluorescence spectrometry, it is the definitive method currently available for this measurement.

Medium-energy ion scattering, which uses a toroidal electrostatic spectrometer to analyze both energy and angle of scattered particles, is the oldest of the medium-energy techniques and is a definitive method for the precise lattice location of surface atoms by a detailed analysis of channeling and blocking.

If you are reading this, you are probably trying to decide if a medium-energy ion beam technique is right for your problem. If you have not done so, you should first familiarize yourself with the general principles of ion scattering in PARTICLE SCATTERING. If you can solve your problem with one of the high-energy techniques described in Section 12a, then you should probably do so. Medium-energy techniques are appropriate for specialized problems and circumstances as described in the pages that follow.

LITERATURE CITED

McKenzie, J. M. 1979. Development of the semiconductor radiation detector. *Nucl. Instrum. Methods* 162:49–73.

Tollestrup, A. V., Fowler, W. A., and Lauritsen, C. C., 1949. Energy release in beryllium and lithium reactions with protons. *Phys. Rev.* 76:428–450.

ROBERT A. WELLER
Vanderbilt University
Nashville, Tennessee

MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY

INTRODUCTION

Medium-energy backscattering spectrometry is a variation of Rutherford backscattering spectrometry (RBS) (ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS) using ions with energies ranging from a few tens to a few hundreds of kilo-electron-volts. The ion beam, frequently He^+, is directed onto a specimen and the number of backscattered ions per incident ion is measured as a function of scattered-ion energy at one or more angles relative to the beam direction. The fundamental physical process being used is that of a binary collision between the nuclei of an ion in the beam and a near-stationary atom in the target. In such a collision, the energy of the backscattered ion is determined uniquely by its original energy, the scattering angle (which is the angle between the ion's trajectories before and after the collision), and the ratio of the masses. Consequently, a measurement of the energy of a backscattered ion at a specific angle may be used to infer the mass

of the target atom with which it collided. The number of backscattered ions is proportional to the collision cross-section and the total number of available scattering centers. Since collision cross-sections are known *a priori* in this energy range, the number of backscattered ions with a given energy is a direct measure of the number of atoms of the corresponding mass in the target.

The method achieves additional richness as the result of a second physical phenomenon, the gradual loss of kinetic energy through collisions as an ion penetrates through matter. This is measured by stopping power (also called linear energy transfer), which is the average energy lost, typically in electron volts per unit of length (measured in atoms per square centimeter) of trajectory in the target. If an ion passes through the surface of a specimen and experiences a collision at some depth within it, it will lose energy on both the inward and outward legs of its trajectory. Consequently, a second factor controlling the energy of backscattered ions is the depth within the target at which the collision occurs. As a result, it is usually possible to extract quantitative information about the distribution of specific species within the near-surface region of the specimen. There are a number of excellent reviews of ion penetration in the literature that include both theoretical background and algorithms for computing stopping power (Ziegler et al., 1985; Rauhala, 1995; Nastasi et al., 1996). For the purposes of this unit, it is sufficient to note that for medium-energy ions the stopping power is dominated by collisions between the nucleus of the projectile and the electrons of the target, so-called electronic stopping, which results in relatively uniform loss of energy with minimal angular divergence of the beam. Widely accepted semiempirical procedures are available for computing electronic stopping power.

Additional information can often be derived from backscattering spectra for specimens that are crystalline or multilayer structures with some crystalline and other amorphous layers by exploiting the phenomenon of particle channeling (Feldman et al., 1982; Swanson, 1995). At medium energies, the angular widths of particle channels, which vary as the inverse square root of particle energy, are of the general order of a degree. This can be exploited through precise control of target orientation in order to obtain information about target crystallinity or additional information about target structure or to deliberately suppress scattering from one region of a specimen in order to emphasize that from another. Channels are so wide at medium energies that accidental channeling occurs frequently and can adversely affect the precision of measurements.

Backscattering spectrometry is most formidable for the analysis of higher atomic mass constituents of a sample, because the spectral features attributable to lower-atomic-mass surface constituents such as carbon or oxygen will occur at the same energy as features attributable to higher-atomic-mass constituents such as silicon located more deeply in the sample. Thus, a heavier substrate produces a background that complicates the analysis of light elements on the surface. Forward-recoil spectrometry, particularly when done by time of flight, addresses this situation.

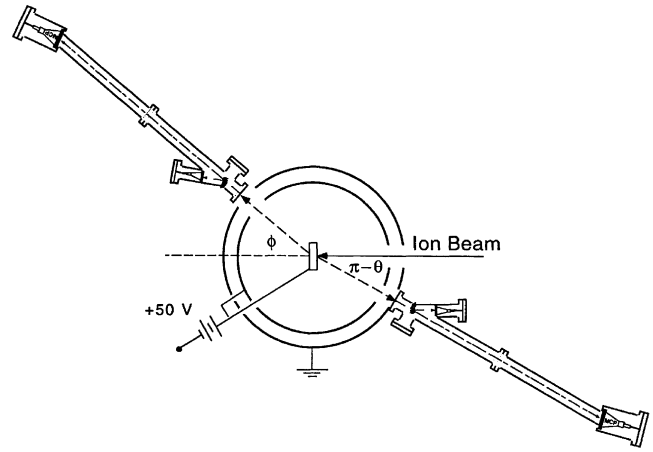


Figure 1. Schematic of the vacuum chamber of a medium-energy ion scattering apparatus showing the target, a forward-scattering and backscattering spectrometer, and their relative orientations with respect to the incident ion beam. The target is biased at +50 V and is situated in a large Faraday cup.

Forward-recoil and backscattering spectrometry, as the names imply, differ primarily in the geometry of the basic ion beam–target collision (Fig. 1). In medium-energy forward-recoil spectrometry, in the laboratory the detector of scattered ions is located at a direction that is typically $\sim 30^\circ$ with respect to the original direction of the ion beam. In binary collisions between beam ions and atoms of the target, the latter are forbidden from having laboratory scattering angles exceeding 90°, so that the features of backscattering spectra are all attributable to scattered ions from the beam. In forward-recoil spectrometry, both target and beam species are present. The optimum situation for light-element detection is achieved by a combination of time-of-flight spectrometry and the use of a beam projectile with higher atomic mass than the target elements of interest. When a heavy projectile strikes a light target, the latter moves away at a speed that, in the laboratory, can be nearly twice that of the former. As a result, at the fixed angle of the detector, of the common light elements hydrogen will be moving most swiftly, followed by carbon, nitrogen, and so on. By measuring the time that these particles require to traverse a fixed distance, they may be sorted by element in a spectrum devoid of the background that would be present in either a backscattering spectrum (where hydrogen would be unobservable) or a forward-recoil measurement using an energy-dispersive detector. Medium-energy forward-recoil spectrometry is among the most sensitive techniques available for the quantitative measurement of surface hydrogen, particularly when it is important to measure other light elements, such as carbon, nitrogen, or oxygen, simultaneously.

Particle detection technology is an extremely important issue in medium-energy ion beam analysis. The remainder of this unit will focus on time-of-flight spectrometry optimized for high-resolution measurements. Other methods, including magnetic and electrostatic spectrometers, pulsed-beam accelerator systems, and time-of-flight spectrometry optimized for other measurements, will be discussed

briefly (see Complementary and Alternative Techniques, below), along with several other methods for depth profiling, trace-element identification, and hydrogen measurement. Throughout, special attention will be paid to the relationship between medium-energy backscattering and conventional high-energy RBS.

Complementary and Alternative Techniques

Time-of-flight medium-energy backscattering is a variation of conventional (high-energy) RBS that has been optimized for high-resolution depth profiling of the first few tens of nanometers of the surface of multilayer planar structures. When devising an ion beam analytical strategy for a specific specimen, it is best to determine first if high-energy (≈2-MeV) backscattering will provide satisfactory information. If it will, then considerations of cost, availability, and ease of interpretation of the data all argue that it should be used. Medium-energy backscattering should be considered if the region of interest is within a few tens of nanometers of the surface and if the depth resolution of conventional RBS appears to be inadequate. Medium-energy backscattering can routinely achieve near-surface depth resolutions better by about a factor of 4 than high-energy RBS with a surface-barrier detector. For the special case of profiling oxygen in SiO_2 films on Si, one should also consider medium-energy proton scattering using a toroidal electrostatic analyzer.

In general, it is always best to use an ion beam technique for elemental depth profiling, if one is available. This is because absolute cross-sections are known, and therefore measurements are quantitative, and because ion beam techniques typically look at the specimen in depth. An alternative procedure is to combine a highly surface-specific technique with the gradual erosion of the specimen by either chemical erosion or sputtering. Examples would include secondary ion mass spectrometry (SIMS) and sputter Auger profiling (see AUGER ELECTRON SPECTROSCOPY). SIMS is extremely sensitive to trace elements. When it is performed in conjunction with slow near-threshold sputtering of a sample, it is termed "static" and can be used for elemental depth profiling by observing secondary ion yields as a function of duration of erosion. This procedure can produce excellent depth profiles of many elements simultaneously, but like all variations of SIMS, absolute values are not readily obtained. In sputter Auger electron spectrometry, one or more specific Auger lines are monitored as a function of erosion. Both static SIMS and sputter Auger require an independent measurement of erosion to assign an absolute depth scale. With ion beam techniques, the depth scale follows from a knowledge of the stopping power.

Medium-energy forward-recoil spectrometry is particularly well suited to measure surface hydrocarbons, especially when they are on metallic or other high-atomic-mass substrates. Alternative techniques include high-energy forward recoil (Barbour and Doyle, 1995; Tirira et al., 1996) and, for specific cases such as hydrogen, nuclear reaction analysis. At mega-electron-volt energies, forward-recoil spectrometry has been carried out both by time-of-flight spectrometry and by the use of surface-barrier detectors in conjunction with thin foils that function essentially as filters to reject forward-scattered beam ions (Barbour and Doyle, 1995). Whenever possible, all forward recoil, regardless of beam energy, should be carried out using time-of-flight techniques. This is because velocity and not energy is the appropriate natural parameter for dispersive analysis. Medium-energy forward-recoil spectrometry should be considered when very high sensitivity is a goal of the experiment. It may also be appropriate when a specimen is very fragile and must receive the minimum possible radiation dose consistent with analysis. Since damage cross-sections often vary as E_0^{-1} while scattering cross-sections vary as E_0^{-2}, the lowest energy suitable for analysis is optimum. Medium-energy forward-recoil measurements can be made on thin organic foils with as few as 10^{11} total ions striking the target (Arps and Weller, 1995).

The preferred method for hydrogen measurement on and near the surfaces of materials is by nuclear resonant reaction analysis using $^1H(^{15}N, \alpha\gamma)^{12}C$ or $^1H(^{19}F, \alpha\gamma)^{16}O$ (PARTICLE-INDUCED X-RAY EMISSION). However, neither of these methods is capable of simultaneously returning information about other components of the specimen. Medium-energy forward-recoil spectrometry provides information about all low-atomic-mass constituents at once and, therefore, can be used in situations where the picture obtained by nuclear reaction analysis is incomplete. Medium-energy forward-recoil spectrometry has been shown to be capable of achieving a depth resolution and sensitivity for hydrogen of ~6 nm and $10^{13}\,cm^{-2}$, respectively (Arps and Weller, 1996). SIMS is also used successfully for hydrogen measurement, but is not as predictable as nuclear reaction or forward-recoil analysis because the cross-section for secondary hydrogen ion production is a function of the state of the surface and is not well known.

PRINCIPLES OF THE METHOD

The definitive treatment of backscattering spectrometry is the book by Chu et al. (1978). The subject has been reviewed recently by Leavitt et al. (1995) and also in ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS of this work, which describes conventional mega-electron-volt methods. What follows is a brief review of the basic principles of backscattering spectrometry with particular emphasis on those aspects in which medium-energy techniques differ from their high-energy counterparts. Additional material specific to forward-recoil spectrometry is also provided. The basic geometry of a medium-energy backscattering experiment with a time-of-flight spectrometer is shown in Figure 1.

When a particle with energy E_0 and mass M_1 strikes a stationary target particle with mass $M_2 > M_1$ and is observed to be scattered by an angle θ in the laboratory, then the energy of the scattered particle will be

$$E_1 = K \cdot E_0 = E_0 \left(\frac{\{x \cdot \cos(\theta) + [1 - x^2 \sin^2(\theta)]^{1/2}\}^2}{(1 + x^2)^2} \right) \quad (1)$$

where $x \equiv M_1/M_2$. This expression defines the kinematic factor K, which relates the incident energy of a particle to its energy after the collision. This is the fundamental relationship by which the mass M_2 of a surface constituent is obtained from the measured backscattered energy E_1.

When an ion beam is normally incident on a surface covered by a thin layer of thickness t, the observed number of backscattering events Y_i attributable to the ith constituent is given by the relationship

$$Y_i = \eta(K \cdot E_0) \cdot Q \cdot (N_i \cdot t) \cdot \Omega \cdot \sigma_i(E_0, \theta) \qquad (2)$$

where N_i is the density of the ith component in atoms per cubic centimeter, Q is the total number of incident projectiles, Ω is the solid angle of the spectrometer, and σ_i is the differential cross-section for the collision (PARTICLE SCATTERING). Strictly speaking, one should integrate the cross-section over the solid angle subtended by the detector, but in practice the solid angle is usually so small that the approximation of Equation 2 is satisfactory. The remaining term, $\eta(K \cdot E_0)$, is the spectrometer quantum efficiency, which is the probability that a particle incident upon the detector will result in a measurable event (Weller et al., 1994). The efficiency is a function of the energy of the backscattered particle and, implicitly, of the particle species. The presence of the efficiency term in Equation 2 is one of the features that distinguishes medium-energy backscattering from its higher-energy analog. At mega-electron-volt energies, the particle detector of choice is the silicon surface barrier detector. The quantum efficiency of surface barrier detectors is so near unity that it is almost universally ignored, and linearity of the output signal with particle energy is also assumed (Lennard et al., 1986). As a result, $N_i t$ can be directly inferred from a measurement of Y_i, Q, and Ω, the latter being, at least in principle, a simple exercise in geometry. This, of course, supposes a knowledge of σ_i, but this is known from theory for mega-electron-volt scattering to be the Rutherford cross-section, given in the laboratory reference frame (Darwin, 1914; Chu et al., 1978) by

$$\sigma_R(E_0, \theta) = \left(\frac{Z_1 Z_2 e^2}{2E_0}\right)^2 \left(\frac{\{\cos(\theta) + [1 - x^2 \sin^2(\theta)]^{1/2}\}^2}{\sin^4(\theta)[1 - x^2 \sin^2(\theta)]^{1/2}}\right) \qquad (3)$$

Here Z_1 and Z_2 are the atomic numbers of the projectile and the ith target constituent, respectively, and e^2 is the square of the electron's charge, which may be conveniently expressed as 1.44 eV·nm.

In medium-energy work, both the unit efficiency and Rutherford cross-section approximations fail, adding significantly to the complexity of interpreting medium-energy backscattering data. It is perhaps some consolation, however, that they only appear as a product $\eta \cdot \sigma$ (see Equation 2) and therefore can be treated as a single entity. Discussion of the efficiency will be postponed until after the discussion of the general features of a time-of-flight spectrometer (see Practical Aspects of the Method, Spectrometer Efficiency). The scattering cross-section will be considered first.

A general rule for the range of validity of the Rutherford cross-section has been given by Davies et al. (1995). They assert that the above expression should be within ±4% between the limits of $0.03 Z_1 \cdot {Z_2}^2$ keV and $0.3 Z_1 \cdot {Z_2}^{2/3}$ MeV. Above this limit, nuclear reactions begin to occur. Below it, in the range of medium-energy collisions, the presence of electrons significantly screens the collision. For the common case of a He^+ particle beam, the lower threshold for silicon, $Z_2 = 14$, is ~ 12 keV, while for gold, $Z_2 = 79$, it is ~370 keV. The departure of the cross-section from the Rutherford value at low energies has been studied extensively (L'Ecuyer et al., 1979, and references therein) and found to be well described by the Lenz-Jensen (Andersen et al., 1980) screened Coulomb potential. For reference purposes, algorithms have been developed (Mendenhall and Weller, 1991) that are sufficiently accurate and efficient to permit the classical cross-section for this potential to be computed in real time. However, in view of the uncertainty in the spectrometer efficiency, the following simple expression for the ratio of the screened to the Rutherford cross-section proposed by Andersen et al. (1980), which is typically accurate to a few percent, is recommended for most practical calculations:

$$\frac{\sigma}{\sigma_R} = \left(1 + \frac{48.73 Z_1 Z_2 (1 + x)(Z_1^{2/3} + Z_2^{2/3})^{1/2}}{E_0(\text{eV})}\right)^{-1} \qquad (4)$$

where, as indicated, E_0 is expressed in electron volts, and the mass ratio x has been defined above. Figure 2A shows the ratio of the Lenz-Jensen screened cross-section to the Rutherford cross-section for He^+ ions on Si, Cu, and Au. Figure 2B, which shows a comparison of the approximation of Equation 4 with the exact Lenz-Jensen cross-section for the extreme case of He^+ on gold, serves to illustrate the quality of the approximation.

Figure 3 presents a graph of the stopping power of He^+ ions in Si, Cu, and Au in the range of Rutherford and medium-energy backscattering. As described above, stopping power is the physical phenomenon that gives backscattering spectrometry its depth-profiling ability. From this graph alone, one would conclude that, because the stopping power is smaller, medium-energy depth profiles would, in principle, contain less information. However, the loss in intrinsic depth differentiation is more than made up by the gains in spectrometer resolution made possible by exchanging a surface barrier detector for a time-of-flight spectrometer.

Backscattering data consist of spectra made by producing a histogram of the number of backscattered particles observed within energy bins of equal width. Figures 4A–D contain computer-simulated backscattering spectra as they would be observed with a surface barrier detector (Figs. 4A–C, E_0 equals 2 MeV, 1 MeV, and 500 keV, respectively) and a time-of-flight spectrometer (Fig. 4D, $E_0 = 250$ keV). In all cases the specimen is assumed to be a device-grade silicon wafer covered with 20 nm of SiO_2 (density 2.22 g/cm^3), 10 nm of $TiN_{0.8}$ (molecular density same as stoichiometric TiN), 15 nm of TiN (density 4.94 g/cm^3), and a trace contamination of W on the surface with an areal density of 4×10^{14} cm^{-2}. In all cases, the He^+

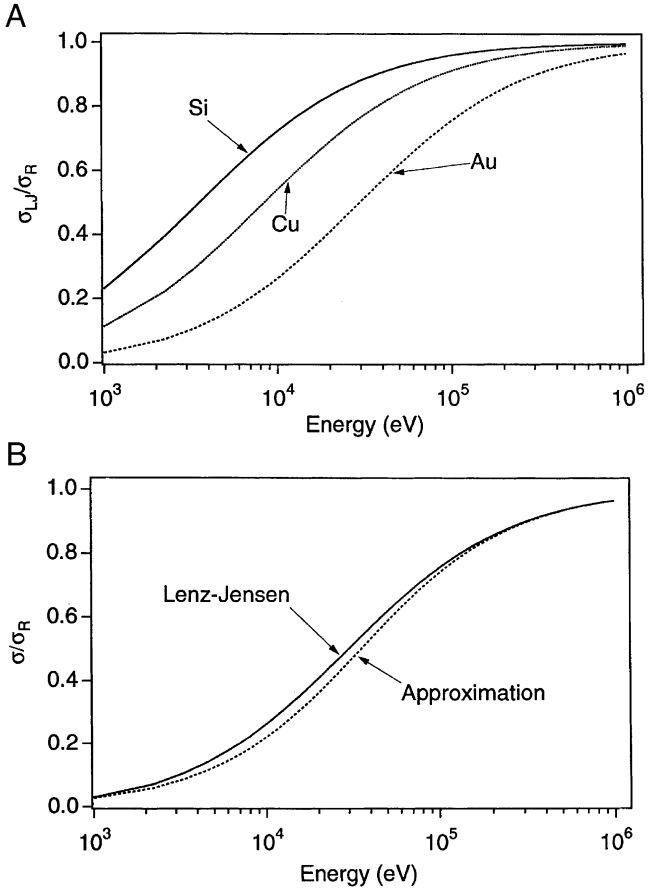


Figure 2. Medium-energy scattering cross-sections. (**A**) The ratio of the Lenz-Jensen cross-section to the Rutherford cross-section, as a function of energy, for He^+ scattering from several target nuclei at a laboratory angle of 150°. (**B**) A comparison of the ratio of the Lenz-Jensen cross-section to the Rutherford cross-section and the approximation to this ratio suggested by Andersen et al. (1980). The approximation is given here as Equation 4. The comparison is for He^+ scattering from Au at 150° in the laboratory.

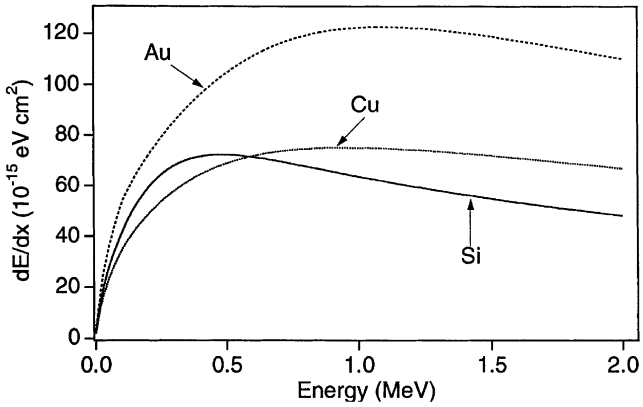


Figure 3. The stopping power of He ions in Si, Cu, and Au as a function of energy. These were computed as described in Ziegler et al. (1985).

beam is incident normally on the surface and the scattering angle is 150°. The amount of deposited charge is 100 μC and the solid angle of both detectors is 1 msr. Both spectrometers are assumed to have unit efficiency as a function of energy. The resolution of the surface barrier detector is taken to be 13 keV, which is typical of commercial devices when new. The resolution of the medium-energy system is more complicated since it is energy dependent, but it has been taken to be the experimentally determined resolution of a typical time-of-flight spectrometer, expressed in the energy domain. For the simulation, all isotopes of all the elements present have been included explicitly, as has energy straggling, using the results of Yang et al. (1991). Stopping power has been computed using the methods of Ziegler et al. (1985) and cross-sections using the algorithms of Mendenhall and Weller (1991). It is assumed that the spectra have been taken in such a way that channeling is not significant. In practice, this is probably best accomplished by aligning the beam a few degrees from a known channel or by rotating the target continuously during the measurement. The plots have been scaled vertically to mimic data taken with a multichannel analyzer adjusted for a conversion gain of 1 keV per channel for the surface barrier spectra and 250 eV per channel for the time-of-flight spectrum. The effects of counting statistics have been omitted for clarity. These figures help to illustrate graphically the relative merits of mega-electron-volt and kilo-electron-volt backscattering.

The small isolated peak at the right of each of the spectra shown in Figures 4A–D is attributable to the trace contamination of W on the sample surface. To quantify the surface concentration, one would simply use Equation 2 with the total yield taken as the integrated number of counts comprising the peak. Because the W is not distributed in depth but instead is concentrated on the surface, the width of the peak is a measure of the system resolution. For the surface barrier spectra this energy resolution is approximately constant as a function of energy. For the time-of-flight spectrum the resolution is poorer at larger energies. For the spectra of Figures 4A and B the large isolated peak is attributable to Ti, and its width is also narrow. However, it shows some broadening relative to the W peak. In Figure 4C it is clearly apparent that the Ti peak has width, but no structure is visible. In all these spectra, the total Ti can be easily measured by the same procedure used for quantifying W.

The pronounced step feature near the center of the spectra of Figures 4A–D is attributable to silicon in the SiO_2 layer and in the silicon substrate, and is referred to as the silicon edge. Because this target is thick, the silicon signal occupies the full range of the spectrum below (to the left of) the edge. Notice the two small peaks visible above the silicon background in Figure 4A. These are attributable to nitrogen and oxygen, with the oxygen peak being at higher energy. For beam energies of 1 and 2 MeV these are fully resolved, but at 0.5 MeV, the surface barrier detector is unable to separate these peaks.

The spectrum of Figure 4D shows the view of the specimen, again omitting statistical fluctuations, that would be obtained by time-of-flight medium-energy backscattering. (See Practical Aspects of the Method for a method to derive an energy spectrum from raw time-of-flight data.) The

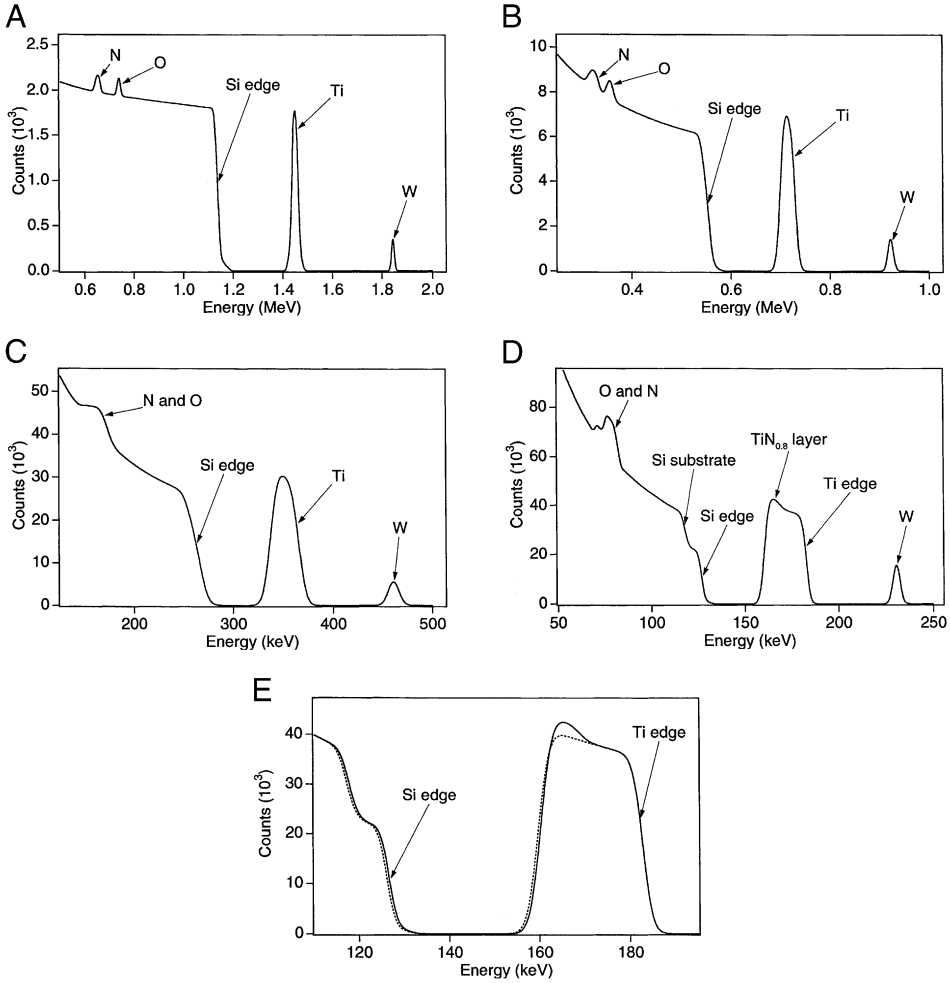


Figure 4. A computer simulation comparing backscattering at various energies using a surface barrier detector (**A–C**) and a time-of-flight spectrometer (**D–E**). The target is a TiN film, with a nitrogen-deficient region, on SiO_2 on Si. There is a trace of W contamination on the surface. (**E**) An expanded view of the Ti peak from (**D**) along with the result (dotted line) that would be expected from a fully stoichiometric TiN film, demonstrating that the nitrogen-deficient region is clearly discernible.

important points to notice are that the Ti peak now clearly contains structure to the extent that it is higher on the left than the right side, and that the silicon edge now contains definite evidence of a step. From these observations alone, it is possible to conclude that the TiN layer does not have uniform composition and that the oxygen is associated with silicon in the form of an SiO_2 film whose thickness can be inferred from the spectrum. This additional insight has been achieved at the expense of the loss of resolution of the nitrogen and oxygen features, which are now fully merged. This is demonstrated clearly in Figure 4E, which compares the Ti peak with a portion partially depleted in N against a simulation of the same target with a single stoichiometric layer of TiN 25 nm thick.

The method of inferring the stoichiometry of a thin film from backscattering data is the same for both medium-energy and high-energy backscattering and is described at length by Chu et al. (1978) and by Leavitt et al. (1995) as well as in ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS. Similarly, procedures have been devised to infer depth profiles from measured backscattering data. It is important to emphasize, however, that the problem of inverting a backscattering spectrum to produce a target structure does not, in general, have a unique solution and is further complicated by statistical fluctuations in the data, small- and large-angle multiple scattering of the beam and scattered ions, channeling, uncertainties in density, and the surface topography of the specimen, in addition to factors such as energy straggling and measurement resolution, which have been simulated in Figures 4A–E.

It is generally true that the more *a priori* knowledge that one has about a specimen, the more precisely one can infer its structure. For example, in the semiconductor industry, where fabricating thin-film structures is the *sine qua non* of integrated circuit manufacture, engineers ordinarily know the order and approximate composition of individual layers in a multilayer thin-film structure. In this case, it is possible, working from approximate starting values, to use a simulation procedure such as the one used to produce Figures 4A–E in a closed-loop algorithm to find the values of parameters such as film thickness and composition that provide the best fit to measured data in the sense of χ^2 minimization or some other measure of goodness of fit. For mega-electron-volt data there are computer algorithms in the published literature to implement such procedures (Doolittle, 1984, 1985, 1990) as well as proprietary computer programs available from commercial

sources (e.g., Strathman, 1990). (See Leavitt et al., 1995, for a more complete list of published simulation algorithms.) However, as of this writing, these have not yet been adapted to explicitly handle medium-energy backscattering. The general problem of inversion of backscattering spectra continues to be an area of active research as the advance of computer power continues to widen the range of methods that can be undertaken practically. The goal of this research, which is to extract all statistically significant information contained in backscattering spectra, is still elusive, although important progress is being made (Barradas et al., 1997).

PRACTICAL ASPECTS OF THE METHOD

The previous section discussed issues relating to medium-energy ion scattering for the most part without reference to any specific spectrometric technique. This is appropriate because there are a number of methods, including electrostatic and magnetic spectrometers, pulsed-beam and foil-based time-of-flight spectrometers, and even cryogenic solid-state detectors, by which the experiments can be performed, and novel variations frequently appear in the literature. Each method has advantages and disadvantages and continues to benefit from time to time by advances in technology. The choice of which method to use is not abstract, but is closely related to the nature of the problem at hand. The discussion in this section will be restricted to a particular style of foil-based time-of-flight spectrometer that has a straightforward design and an attractive set of operating characteristics. Foil-based time-of-flight spectrometers were originally introduced for heavy-ion spectrometry in nuclear physics (see Pfeffer et al., 1973, and Betts, 1979, and references therein) and later adapted for backscattering spectrometry using mega-electron-volt heavy ions (Chevarier et al., 1981). They were first applied to medium-energy ion beam analysis by Mendenhall and Weller (1989, 1990). Several time-of-flight configurations have been reviewed critically by Whitlow (1990).

Time-of-Flight Spectrometry

An energy-dispersive particle spectrometer such as a silicon surface-barrier detector functions by converting the energy of a kinetic particle to an approximately proportional quantity of electric charge. A time-of-flight spectrometer functions by measuring the time that it takes for a particle to traverse a fixed distance. This requires two signals representing a "start" and a "stop" that are associated with, but possibly offset by a fixed time from, the beginning and end of the traversal. In time-of-flight spectrometry, the issue is how best to obtain a start signal, since a stop signal can always be produced by simply collecting the particle in any of several detectors. In pulsed-beam time of flight, the start signal is derived by knowing that all ions of the beam arrive at the target in a narrow pulse. In foil-based time of flight, a very thin foil, usually of carbon, is placed at the beginning of the measured flight path. This foil is so thin that an ion hitting it passes through with minimal loss of energy and perturbation of its trajectory.

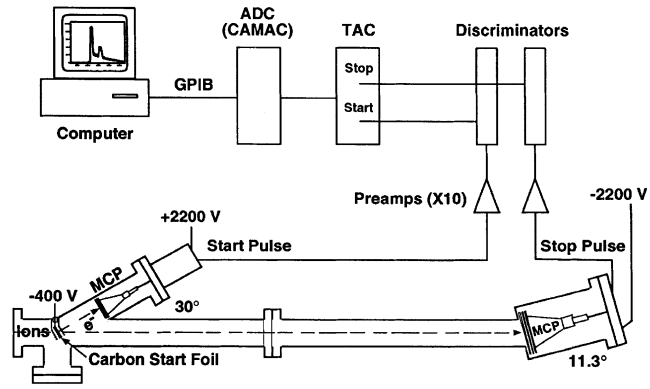


Figure 5. Principal components of a medium-energy time-of-flight spectrometer system. This system is computer controlled and CAMAC based (Weller, 1995a). The angle of the stop detector is chosen to optimize resolution as described in the text.

As it exits the foil, the ion produces secondary electrons with significant (but not unit) probability, and these are accelerated away and detected to produce a start signal. A time-of-flight spectrometer of this type is shown in Figure 5.

Spectrometer Efficiency

The characteristics of the spectrometer enter into the determination of the concentrations of constituents via the energy-dependent efficiency term η in Equation 2. It is possible to understand the physical origin of this efficiency by considering individually the mechanisms by which events may be lost. For example, while the perturbation of the trajectory by the start foil is small, it is not zero, and a predictable number of events will be lost because small-angle multiple scattering deflects some particles enough to miss the stop detector (see Fig. 5). Similarly, since the production of a start pulse requires secondary electrons, and since the production of secondary electrons is inherently statistical, there will be instances in which a particle will pass through the start foil without creating a start pulse. Additional processes that may disrupt the normal sequence of events in the spectrometer include scattering of ions or electrons from various meshes, the intrinsic dead area on the surface of microchannel plates, and the statistical nature of the development of pulses in these detectors. The cumulative effect of these various events has been summarized in a model of time-of-flight spectrometer performance that is, in principle, capable of predicting the energy-dependent performance of a time-of-flight spectrometer from first principles (Weller et al., 1994). However, because many input parameters to the model, such as secondary electron emission yields, are highly variable and dependent upon vacuum conditions and age and composition of the components, among other things, the model is best used in conjunction with measurements of efficiency, so that these parameters need not be known *a priori*. This semiempirical approach is demonstrated in Figure 6, which uses experimentally determined parameters along with the equations of Weller et al. (1994) to obtain efficiency curves

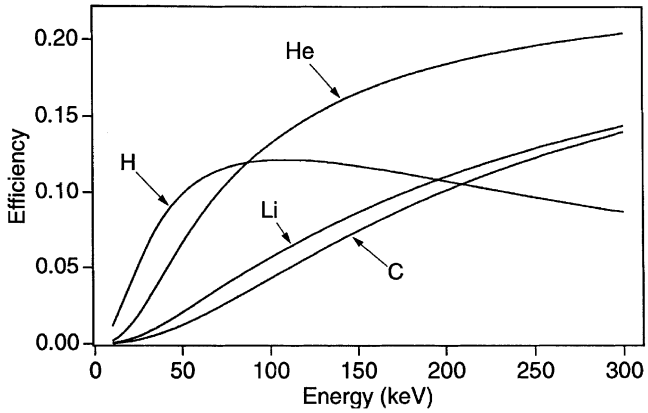


Figure 6. The efficiency of the time-of-flight spectrometer shown in Figure 5 for hydrogen, helium, and carbon ions as a function of energy, computed using empirical parameters in conjunction with the formalism described by Weller et al. (1994).

for a time-of-flight spectrometer like that shown in Figure 5. Note that the curve for H, unlike the others, has a maximum in the region of interest. This is attributable to the dependence of the secondary electron yield on electronic stopping power. It is important to emphasize that the efficiency of a spectrometer for a given ion is a property of the measuring device and not of the target being measured. Thus, while it is necessary to use a standard for the initial calibration of the instrument, it is not necessary to employ standards that mimic the properties of individual targets in order to achieve quantitative results. In this sense, the method is to be contrasted with SIMS, where the characteristics of a surface may have a strong impact on the relative intensities of the signals from various surface constituents.

Sensitivity

In general, elastic scattering spectrometries are more sensitive at lower incident beam energies, all other things being equal, because the collision cross-section decreases with increasing energy (see Equation 3). Thus, medium-energy backscattering and forward-recoil spectrometry enjoy an intrinsic advantage of as much as two orders of magnitude when compared with their high-energy counterparts. However, a portion of this is forfeited in practice because of the efficiency of the spectrometer, as discussed in the previous paragraph. See HEAVY-ION BACKSCATTERING SPECTROMETRY for a discussion of a system that has been optimized for extremely high sensitivity.

The trace-element sensitivity of medium-energy backscattering is limited by two factors, the gradual sputter removal of the constituent of interest and multiple scattering of the beam within the target. If, instead of scattering once at an angle θ in order to enter the spectrometer, a particle scatters twice at an angle $\theta/2$, then its energy will be considerably higher. Ordinarily, these multiple collisions (including all possible pairs of events that sum to an angle θ) are sufficiently rare that they are ignored in analyzing backscattering data. However, for heavy trace elements such as, for example, iron on silicon, multiple scattering in the silicon substrate can lead to the production of counts in the region of the iron peak that is kinematically forbidden to binary collisions with individual substrate atoms. These counts constitute a background that limits the signal-to-noise ratio of true iron counts. This multiple-scattering background, which has been studied extensively by Brice (1992), is most significant for heavy-ion backscattering (HEAVY-ION BACKSCATTERING SPECTROMETRY).

For light-ion backscattering, as for example with He^+, the predominant factor controlling sensitivity is sputtering (Weller, 1993). Sputtering, the erosion of a material by ion impacts, typically results in from 10^{-3} to as many as 10 or more surface atoms being removed for each ion impact (Yamamura and Tawara, 1996). In the absence of sputtering, a trace-element backscattering measurement could proceed until all atoms of the element of interest are involved in a collision. With sputtering, it is convenient to define the sensitivity by the constraint that the fractional amount of the trace element removed is equal to the fractional uncertainty that is acceptable in the final answer. The only difficulty with this is that the *a priori* sputtering yields of trace elements in a foreign host are not well known (Pedersen et al., 1996), so that plausible assumptions about these yields must be made. Doing so results in the following formula for the minimum areal concentration ρ of a surface constituent that can be measured with a fractional uncertainty ε (Weller, 1993):

$$\rho = \frac{1 + \sqrt{1 + 8\varepsilon^2 B}}{2\varepsilon^3} \cdot \frac{Y_s}{\rho_s A_s \Omega \sigma \eta} \tag{5}$$

where A_s is the area of the target that is irradiated by the beam and B is the number of background counts in the region of the spectrum that contains valid counts. The quantities σ and η are the cross-section and spectrometer efficiency, respectively, and Ω is the spectrometer's solid angle. This expression is based upon the assumption that the probability P that a beam ion removes a target atom is related to the (dimensionless) sputtering yield of the substrate Y_s through the relation

$$P \approx Y_s \cdot \frac{\rho}{\rho_s} \tag{6}$$

where ρ_s is taken to be the areal density of substrate atoms in the region from which sputtered particles are drawn. This is typically the first two or three atomic layers of the target (Dumke et al., 1983). Equation 6 expresses the assumption that trace elements are sputtered randomly from the surface with a probability equal to their fractional representation in the volume from which sputtered particles can emerge. Estimates of sputtering yields may be obtained using the equations or tables of Yamamura and Tawara (1996). For trace quantities of copper on silicon analyzed by He^+ with a general-purpose spectrometer like that shown in Figure 5, Equation 6 predicts a sensitivity of $\sim 10^{11}$ cm^{-2} in the absence of background ($B = 0$). However, achieving this figure would require exceptionally long measuring times. For trace elements with concentrations below $\sim 10^{13}$ cm^{-2}, a system optimized

for sensitivity should be used (HEAVY-ION BACKSCATTERING SPECTROMETRY).

Resolution

The energy resolution of a spectrometer is important because it is a direct measure of the depth resolution of backscattering spectrometry. It is important to emphasize, however, that resolution is not a unique concept. Most often, in analogy to the Rayleigh criterion familiar from optics, two isolated peaks in a backscattering spectrum are said to be resolved when their separation is equal to the full width at half-maximum (FWHM) of the peaks. In this case, any additional separation results in an observable bimodal structure. This definition certainly captures the essence of the physics for cases, such as superlattices, where the objective is to discern changes in the concentration of an element as a function of depth. However, there is another closely related concept, that of layer thickness, for which this definition of resolution results in an unduly pessimistic estimate of the quality of the data. In this second type of measurement, one is typically interested in small variations in the thickness of a layer rather than in the details of its structure. Essentially, the problem is that of locating the edge of a spectral feature, which is approximately equivalent to locating the centroid of an isolated peak. This can typically be done with an uncertainty of $\approx$10% of a standard deviation, σ, of the peak, as compared with the FWHM, which is 2.35σ. Thus, it is possible to measure thickness variations that are more than an order of magnitude smaller than the standard definition of resolution would suggest. There is no contradiction here, simply an inconsistency in what exactly is meant by "resolution." In the paragraphs that follow, the usual Rayleigh concept will be adopted, with the understanding that it may not be appropriate in all circumstances.

If δE is the uncertainty (FWHM) from all sources in the measurement of a backscattered particle's energy, then the corresponding uncertainty in depth of the target atom is (Chu et al., 1978)

$$\delta x = \frac{\delta E}{[S]} \tag{7}$$

where the denominator is known as the energy loss or stopping power factor and is given by

$$[S] = \left[K\left(\frac{dE}{dx}\right)_{\text{in}} \cdot \frac{1}{\cos(\theta_1)} + \left(\frac{dE}{dx}\right)_{\text{out}} \cdot \frac{1}{\cos(\theta_2)}\right] \tag{8}$$

Here K is the kinematic factor defined above, dE/dx is the stopping power, and θ_1 and θ_2 are angles made with the target surface normal by the inward and outward trajectories of the particle, respectively (Leavitt et al., 1995). As a particle penetrates a target, the uncertainty in its energy grows as a result of straggling, which is the intrinsic uncertainty associated with the statistics of energy loss (Yang et al., 1991). Thus, depth resolution is always best at the surface of a target, where it is limited by the uncertainty of the beam energy E_0 and the properties of the spectrometer.

Three primary design characteristics govern the resolution of a time-of-flight spectrometer such as is shown in Figure 5: minor differences in the length of the flight paths of different ions, straggling in the carbon start foil, and nonuniformity of the foil either in thickness or density. Of these, foil nonuniformity is the most serious (McDonald et al., 1999). An additional variation due to the natural kinematic broadening associated with the variation of θ across the finite acceptance solid angle of the spectrometer is also significant (Equation 1), as is a charge-exchange process that produces small parasitic features offset from the true value by plus or minus the value of the start foil bias, or $\sim$300 eV (Weller et al., 1996).

The velocity of a 100-keV He^+ ion is $\approx$2.2 mm/ns. As a result, it is necessary to design a spectrometer so that most trajectories differ by no more than a millimeter or so. This is the reason for the relative angles between the start and stop detectors in Figure 5. If all ions entered the spectrometer along parallel trajectories, then the offset angles measured from the spectrometer axis should be equal. However, for particles emerging from a point source, there is another angle that produces minimum path length dispersion, which is dependent upon the distance between the target and the start foil and the length of the drift section. Interestingly, a further adjustment of this angle conveys an advantage unique to time-of-flight methods. Ions with larger scattering angles move more slowly as a result of the natural variation of the kinematic factor K with angle (Equation 1). As a result, it is possible to adjust the relative angles of the start and stop detectors to minimize kinematic dispersion from the finite solid angle of the spectrometer. This adjustment, although it amounts to only about a degree, produces a clear improvement in performance for the ion for which it is optimized (Weller et al., 1996).

Considering fundamental physical processes such as straggling and limitations imposed by geometry, time-of-flight spectrometers like that shown in Figure 5 should have a resolution of $<$1%, or perhaps 1 keV for a 150-keV ion (McDonald et al., 1999). With favorable target orientation (see Equation 8), this should produce a depth resolution of $\approx$1 nm at the target surface. This is approximately a factor of 2 better than has been achieved in the laboratory because of the microscopic roughness of available carbon start foils, but still somewhat inferior to results obtainable in proton backscattering using high-resolution electrostatic analyzers. It compares quite favorably with results obtained using surface barrier detectors simply because of the improved resolution (Williams and Möller, 1978).

Instrumentation

The electronic instrumentation needed to implement a time-of-flight system is standard (Weller, 1995a) and available from commercial vendors. The essential components are the start and stop detectors, timing discriminators, and time-to-amplitude or time-to-digital converters. The fundamental, active, particle-detecting elements in time-of-flight spectrometers are microchannel plates (Fig. 5). These are high-density arrays of microscopic, continuous-dynode electron multipliers fabricated in millimeter-thick glass wafers (Wiza, 1979). New microchannel

plates operated at rated bias can produce pulses that are a significant fraction of a volt. However, they are so fast that most general-purpose oscilloscopes cannot display them.

With proper signal-handling techniques (Weller, 1995a, and references cited therein), it should be possible to couple the start and stop pulses directly to timing discriminators, whose function it is to produce uniform time markers that are insensitive to pulse amplitude. However, in practice, it may be necessary to add a stage of voltage amplification between the microchannel plates and timing discriminators. Amplifiers with bandwidth of at least 1 GHz are needed for this purpose. The timing discriminator outputs are used as the definitive start and stop signals are fed to a timing module.

Two styles of timing modules are in common use. One, called a time-to-amplitude converter (TAC), produces an output pulse whose height is proportional to the time difference between the start and stop signal. This is very convenient because the resulting linear pulse may be directly substituted for the amplified and shaped output of a surface-barrier detector. In a system of this type, the pulse from the TAC goes to a multichannel pulse-height analyzer that completes the analysis by producing a histogram of the number of pulses as a function of amplitude. This scheme may sound laboriously indirect, but it is capable of achieving resolutions of ≈50 ps when configured for medium-energy scattering and, until very recently, was clearly superior to directly counting ticks of a clock (Porat, 1973). The alternative time-to-digital converter (TDC) scheme in its purist form does, indeed, directly count the ticks of a clock between the start and stop signals. In this approach, the number of clock ticks is used as a pointer into a memory bank (a histogramming memory) and one count is recorded for each valid interval measurement. With countable clock frequencies now in the tens of gigahertz, producing correspondingly high resolution, this method is a practical alternative to time-to-amplitude conversion. However, it is worth noting that, at the time of this writing, commercial modules called TDCs are often, in reality, TACs and pulse analog-to-digital converters packaged as a unit.

Examples

Figures 7A–C show the result of measuring an SiO_2 thin film nominally 13.9 nm thick grown on crystalline Si by medium-energy backscattering using a beam of 270-keV He^+ ions. Figure 7A shows the raw time-of-flight spectrum as accumulated. In Figure 7B, this spectrum has been transformed to the energy domain using the procedure described below. Figure 7C contains the final version of this spectrum, which has been corrected for the efficiency of the spectrometer. These data were taken with the specimen oriented so that the beam is aligned with the ⟨110⟩ direction of the crystalline Si. This causes the number of counts from the bulk Si region to be suppressed, while Si and O counts from the amorphous SiO_2 layer are unaffected. An analysis of the oxygen peak reveals that the oxide is ~13.1 nm thick, in good agreement with the nominal thickness. The analysis of the Si peak is more complicated because it contains constituents from both the SiO_2 layer and the first few layers of the crystalline Si. This channeling surface peak is discussed by Feldman et al. (1982) and Swanson (1995).

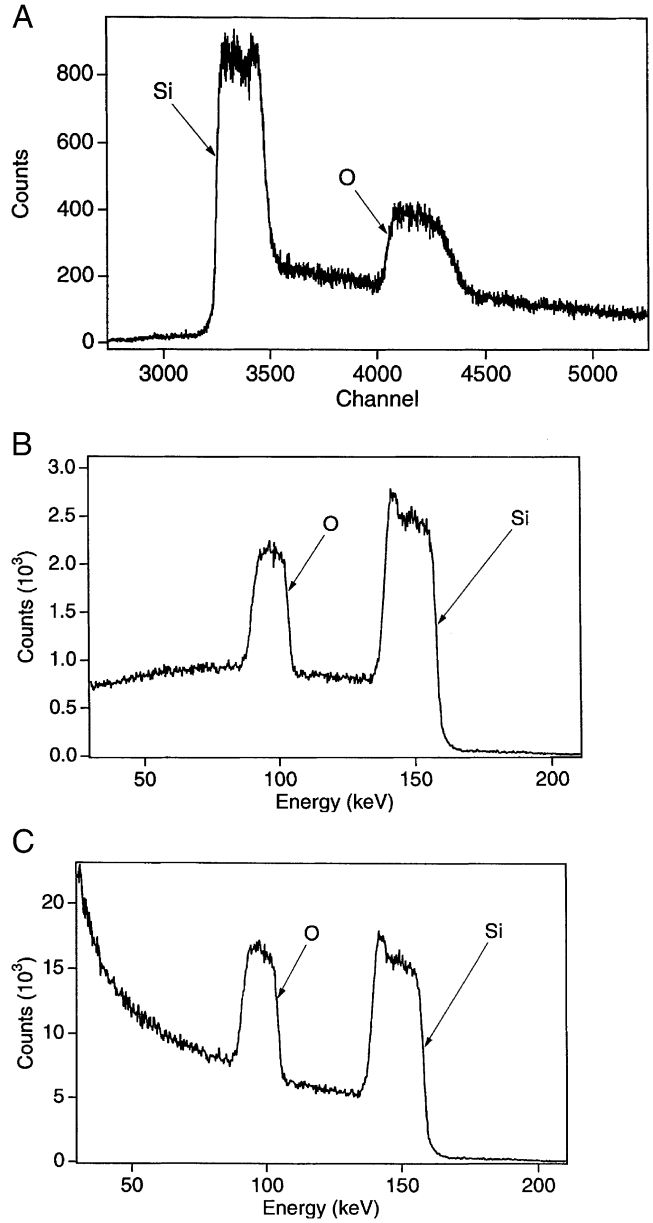


Figure 7. Time-of-flight medium-energy backscattering spectrum of a 13-nm SiO_2 film on Si. The ion beam was He^+ at 270 keV. The beam is channeled in the ⟨110⟩ direction of the Si substrate in order to accentuate the Si and O in the amorphous oxide layer. The scattering angle was 150°. (**A**) Raw data showing number of counts versus channel in the multichannel analyzer. Each channel has a width of 122 ps. (**B**) The data rendered in the energy domain using the transformation described in the text. (**C**) The energy spectrum corrected for spectrometer efficiency using the efficiency function plotted in Figure 6. Note in each case, but particularly in the energy spectra, that the channeling surface peak attributable to the first few atomic layers of Si in the substrate is clearly visible on the low-energy side of the Si spectral feature.

Figure 8 shows the result of measuring the surface of an Si wafer, as delivered by the vendor, for hydrocarbon

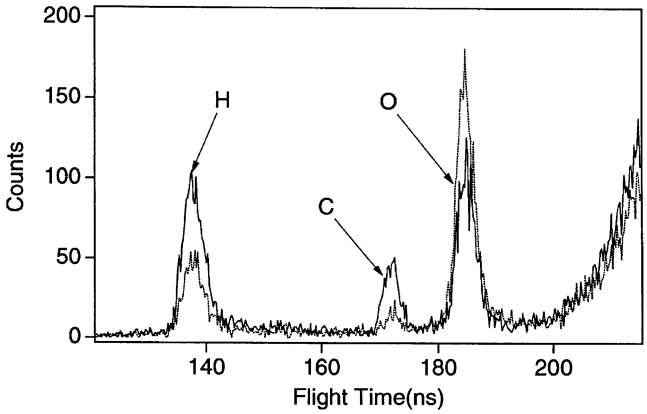


Figure 8. Time-of-flight forward-recoil spectrum of Si wafer showing surface H, C, and O before and after cleaning with ozone. The solid line is the spectrum of the as-delivered wafer. The spectrum shown as a dotted line was made after cleaning. The ozone exposure removed H and C but increased the thickness of the surface SiO_2 layer as indicated by the increased O signal. The ion beam was Ar^{3+} at 810 keV.

contamination using medium-energy forward-recoil spectrometry. The beam was 810 keV Ar^{3+}. The spectrum reveals three distinct peaks that are attributable to surface H, C, and O. While the hydrogen could have been measured by, for example, nuclear reaction analysis (Lanford, 1995; also see NUCLEAR REACTION ANALYSIS AND PROTON-INDUCED GAMMA RAY EMISSION), and the oxygen separately by backscattering, this measurement has revealed both in a way that makes it straightforward to determine the relative areal densities of the three constituents simultaneously.

Safety Considerations

Medium-energy backscattering spectrometry poses no special laboratory safety hazards. General safety considerations are the same as for most classes of surface analytical equipment. The most serious concern is shock hazard from any of several high-voltage sources that are needed to carry out the experiments. Ion beam irradiation by kilo-electron-volt ions is essentially nondestructive for most targets and produces no lingering effect to alter the intrinsic toxicity of a specimen. Particle accelerators used in the technique are often capable of producing x rays, especially when operated improperly, and must therefore be supported by appropriate radiation safety staff under the laws of the jurisdiction in which they are located.

A subtle safety issue in this regard is less well known. When a highly insulating material such as silica is irradiated with kilo-electron-volt ions, the surface can become charged to many tens of kilovolts. Eventually, the surface will break down electrically and a significant spark will occur in the vacuum chamber. These sparks are easily capable of producing measurable pulses of x rays that may sometimes even be observed outside the vacuum system, especially through windows. Although it is quite unlikely that this source would ever generate enough x rays to be an issue for personal safety, it could affect the results of radiation surveys and might also impact the need for shielding or placement of the target chamber in the laboratory.

METHOD AUTOMATION

In principle, the acquisition of backscattering spectra may be highly automated. A full range of NIM and CAMAC modular electronics is available commercially, and multichannel pulse-height analysis is now usually done by computer (Weller, 1995b). Specialized equipment, such as goniometers for channeling studies, may easily be synchronized with data collection under computer control. However, medium-energy backscattering is still a specialized technique, and production-scale systems are not available as they are for conventional mega-electron-volt Rutherford backscattering.

The analysis of backscattering data is, in general, not highly automated, at least not in the sense of, for example, gamma ray spectroscopy, where peak search-and-fit algorithms have been developed that are capable of nearly independently extracting the identity and intensity of all the gamma rays in a spectrum. Although algorithms have existed for many years to assist in the reduction of backscattering data (Doolittle, 1984, 1985, 1990) and new methods continue to be explored that further reduce the need for operator intervention (Barradas et al., 1997), human judgment still plays a role in most ion scattering data analysis. These principles are discussed next (see Data Analysis and Interpretation), along with considerations that are specific to the handling of medium-energy ion scattering data.

DATA ANALYSIS AND INTERPRETATION

Backscattering Data

The steps for analyzing time-of-flight medium-energy backscattering data are, in order, to convert the spectrum from the time to the energy domain, to apply a correction for the efficiency of the spectrometer as a function of energy, and finally to extract information on elemental concentrations and depth profiles using techniques identical to those used for high-energy Rutherford backscattering spectra. In the final stage, it is essential to use cross-sections and stopping powers that are appropriate for the energy range.

The process of converting a time-of-flight spectrum to an equivalent representation in the energy domain is often called rebinning, since it represents a sorting of collected counts from equal-width time bins to equal-width energy bins. Both time-of-flight and energy spectra are essentially differential quantities, since they are histograms of the number of events whose defining parameter, time or energy, falls within a given narrow range defined by a single channel. Thus, to convert from a time-domain spectrum $P_t(t)$ to an equivalent energy-domain spectrum $P_E(E)$ requires use of the fundamental relationship

$$P_E(E) = P_t(t(E))\left|\frac{dt(E)}{dE}\right| \tag{9}$$

where t is the flight time that corresponds with backscattered energy E, and the parameters are related through the mass M_1 of the backscattered particle and the drift length d of the spectrometer by

$$E = \frac{1}{2} M_1 \left(\frac{d}{t}\right)^2 \tag{10}$$

Although Equation 9 can be used directly in order to perform the conversion, it is not the preferred method. A more direct procedure is to use the integral form of this equation:

$$\int_0^E P_E(E')dE' = \int_{t(T)}^{\infty} P_t(t')dt' \tag{11}$$

To apply Equation 11, first it is necessary to create an integral time spectrum by iteratively summing adjacent channels of the time-of-flight spectrum from the longest flight time (rightmost, highest array index) to the shortest. This can be accomplished on an array in place. At the $N-1$ channel of an N-channel spectrum, the value of channel $N-1$ is added to the value of channel N and the sum is put in location $N-1$. This is repeated with N replaced by $N-1$ iteratively until N is zero. This is the integral time spectrum. To obtain an integral energy spectrum, an array is created and each location is filled with the value of the integral time spectrum at the time corresponding to the various equally spaced energies. It will, of course, rarely be the case that an energy and corresponding time value both fall exactly on channel boundaries in their respective spectra, so normally one must interpolate linearly between two values in the integral time spectrum to obtain the energy value. After the integral energy spectrum has been created, the process must be inverted by subtracting adjacent channels to recover a differential energy spectrum. This too can be done on an array in place. This conversion from a time to an energy spectrum has the virtue that it very accurately preserves total numbers of counts through the transformation and is very economical to compute.

It is important to issue a caution at this point. It is generally assumed in the analysis of ion scattering data that the variance associated with each channel of a multichannel spectrum is the Poisson value, that is, that the variance equals the number of counts and *the uncertainties of adjacent channels are uncorrelated*. This is true in raw time-of-flight spectra but may not be true in the corresponding rebinned energy spectra. The problem lies with the widths of the channels. For proper statistics, the width of time channels must be less than the time-equivalent width of the highest-energy channel in the energy spectrum. If this condition fails, then it is clear from the above description of rebinning that, in the high-energy region of the spectrum, channel values will be created by extracting two or more energy values by interpolation between the same two points in the integral time spectrum. This introduces strong correlation into adjacent channel values and changes the statistical character of the data. It is good practice to choose the width of time bins so that, even at the highest energies of interest, each energy bin corresponds to two or more time bins. In this way, the statistics of rebinned time-of-flight spectra will be practically equivalent to those of energy spectra that have been measured directly by an energy-dispersive spectrometer.

After time-of-flight spectra have been rebinned to the energy domain, they should be corrected for spectrometer efficiency as described above (see Practical Aspects of the Method, Spectrometer Efficiency). This consists of a straightforward channel-by-channel multiplication of the energy spectrum by a value appropriate to the ion and energy. It is important to emphasize again that this correction is completely independent of the specimen and only a function of spectrometer characteristics. As such, it does not introduce sample-to-sample variation. It does, however, change the statistics of the spectrum, since in any given channel statistical uncertainly is based upon the actual number of measured counts and not the corrected number of counts. Strictly speaking, one should also include at this point an estimate of error produced by the correction itself. With careful measurements, however, this can be made quite small. Also, the efficiency correction produces its greatest effect in the lower-energy portion of the spectrum, where its impact on the appearance of a spectrum is large but where there are seldom features of quantitative interest.

After rebinning and efficiency correction, time-of-flight spectra, along with their associated arrays of variances, may be treated with the tools and techniques that have been developed for elastic scattering spectrometry with energy-dispersive spectrometers (ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS).

Forward-Recoil Data

The rebinning process described above is of limited usefulness for one-dimensional forward-recoil data. Since forward-recoil spectra contain events attributable to different species, it is only possible to do rebinning of a one-dimensional spectrum sensibly in a spectral region known to contain only one species of ion, such as hydrogen or carbon. Furthermore, in forward recoil, time is the proper distinguishing characteristic, the one with the greatest power of differentiation of species. Thus, with the exception of near-surface hydrogen profiling, one-dimensional forward-recoil spectra should be used directly to infer properties of the specimen.

The mass of surface constituents can be identified by their energy E_2, computed using Equation 10, along with the recoil kinematic factor K_R:

$$E_2 = K_R \cdot E_0 = \frac{4M_1M_2}{(M_1+M_2)^2}\cos^2(\phi) \tag{12}$$

where ϕ is the angle of the recoil measured with respect to the initial beam direction. Absolute concentrations of low-atomic-mass surface constituents can be computed directly using Equation 2 along with spectrometer efficiencies, the peak areas (total counts), and the forward-recoil cross-section σ_r, given here in terms of the center-of-mass scattering angle $\theta_c = \pi - 2\phi$ and center-of-mass cross-section

$\sigma_c(\theta_c)$ by a standard kinematic transformation (e.g., Weller, 1995b):

$$\sigma_r(\phi) = 4\sigma_c[\theta_c(\phi)]\cos(\phi) = 4\sigma_c(\pi - 2\phi)\cos(\phi) \qquad (13)$$

Most previous research on forward-recoil spectrometry has used high-energy ions and surface-barrier detectors in conjunction with range foils. These are self-supporting thin films placed between the specimen and the detector to act as filters to the passage of certain ions. Procedures have been developed to analyze these data (Barbour and Doyle, 1995) but are of limited applicability for medium-energy time-of-flight work.

The most advanced technology for forward-recoil spectrometry combines time-of-flight with energy spectrometry in an experiment in which both parameters are measured for each event (Tirira et al., 1996). This gives rise to two-dimensional histograms with distinct spectral features that correspond to specific elements. This method works best for heavy ions with energies of many (often tens of) mega-electron-volts, and as a result, application of this method has been limited to a relatively small number of laboratories. Multiparameter forward-recoil spectrometry at medium energies has not been well studied.

PROBLEMS

The definitive enumeration of the myriad of ways that ion scattering measurements can go astray has been compiled by Davies et al. (1995) and should be read by all serious students and practitioners of the field. Issues important for medium-energy time-of-flight work are emphasized here.

For quantitative measurements, it is essential to be able to integrate the beam charge accurately, since it is the measure of the total number of incident ions (Equation 2). This is done with a specially designed charge-collection structure known as a Faraday cup (England, 1974; Davies et al., 1995). Nature confounds the casual designer of a Faraday cup with a dizzying array of processes that complicate what would seem to be a straightforward measurement task. Currents used in time-of-flight medium-energy ion beam analysis range from a maximum of a few tens of nanoamperes, when analyzing for trace elements on a low-atomic-mass substrate, to <1 nA for forward-recoil measurements of hydrogen. Making a 1% measurement in the latter case could be compromised by a mere millivolt of stray voltage (or instrumental offset) across a resistance of 100 MΩ. More likely to cause trouble, however, are processes intrinsic to ion beam experiments such as secondary electron emission at the target or ion beam neutralization through charge exchange with the residual vacuum. With a good design and careful attention to detail (Davies et al., 1995), charge integration at the level of 5% should be easily achieved.

With good charge integration, the efficiency of the time-of-flight spectrometer can be measured. The simplest approach is to obtain a target with a known areal density of a heavy trace element such as gold at the surface and to measure this target with different incident beam energies so that the range of backscattered energies covers the range of interest in general measurements. The effectiveness of this approach is dependent upon the accuracy of beam-current integration. The quality of the measurement can be improved by using a target with two isolated peaks that are well separated in energy. The beam energy is then raised or lowered in such a way that, through successive runs, the low-energy peak moves to the position formerly occupied by the high-energy peak, or the reverse. This maintains an internal check of charge integration and provides data to normalize from run to run. In this way, a curve of spectrometer efficiency as a function of energy can be produced from the calculated cross-section and reliably interpolated between measured points, using either a semiempirical model (Weller et al., 1994) or convenient analytic functions of energy. Since spectrometer efficiency must be measured for every ion to achieve the highest-accuracy results, it is sufficient to use the approximate screened cross-section given by Equation 4.

At medium energies, the critical angles for channeling in crystalline targets are very large (Swanson, 1995). This can be of very great help in suppressing background counts so that, for example, the oxygen content of a thin oxide on Si can be measured more accurately (Fig. 7C). However, accidental channeling can distort a spectrum and create a misleading impression about the structure of the sample. It is good practice to make critical measurements at several angles in order to be able to check for this kind of distortion. It is even better to continuously rotate the sample during a measurement to approximate the appearance of an amorphous target.

Another effect that varies strongly with target orientation is loss of resolution due to surface roughness. Ion scattering measurements at all energies work best on planar targets. Moreover, depth resolution can be enhanced by tilting the target so that ion trajectories, and consequently energy loss, are maximized in layers of interest. Detailed studies by Williams and Möller (1978) suggest that for 2-MeV ions the best depth resolution can be obtained using grazing incidence and exit angles of 5° to 10°. However, at these extreme angles, even minor surface imperfections can produce large variations in the cumulative energy loss of ions on different trajectories. For this reason, discretion should be exercised when using target angle to enhance resolution, and under normal circumstances, trajectories closer than <15° to the plane of the target should be used with considerable caution. It is possible to obtain ~2-nm resolution using conventional RBS with grazing exit-angle geometry provided the target surface is (atomically) smooth. Similar resolution can be obtained using medium-energy backscattering without the need for grazing exit angles.

Finally, it is important to keep in mind that the beam that is striking the target may not be what it is thought to be. This is because most accelerators use magnets to select a specific beam from the many that are produced by the ion source. The relevant parameter, known as magnetic rigidity, is $(2M \cdot E/q^2)^{1/2}$, where M, E, and q are, respectively, the mass, energy, and charge of the ions in the beam. In medium-energy accelerators, E is almost

always equal to $q \cdot V$, where V is the accelerating voltage, so that any time two ions have the same M/q, they cannot be distinguished. The most important case for medium-energy backscattering is that of He^{2+} and H_2^+. The omnipresence of the hydrogen molecular ion makes the use of doubly charged helium, to obtain higher energies, a dangerous proposition. This mass interference taken together with the difficulty of producing He^{2+} in the first place argues strongly against the use of this ion for routine analytical work.

ACKNOWLEDGMENTS

The author would like to thank Kyle McDonald and Len C. Feldman for helpful comments during the preparation of this unit. The primary support for the development of time-of-flight medium-energy ion beam analysis has been provided by the U.S. Army Research Office under contracts DAAL 03-92-G-0037, DAAH 04-94-G-0148, and DAAH 04-95-1-0565.

LITERATURE CITED

Andersen, H. H., Besenbacher, F., Loftager, P., and Möller, W. 1980. Large-angle scattering of light ions in the weakly screened Rutherford region. *Phys. Rev. A* 21:1891–1901.

Arps, J. H. and Weller, R. A. 1995. Time-of-flight elastic recoil analysis of ion-beam modified nitrocellulose thin films. *Nucl. Instrum. Methods Phys. Res. B* 100:331–335.

Arps, J. H. and Weller, R. A. 1996. Determination of hydrogen sensitivity and depth resolution of medium-energy, time-of-flight, forward-recoil spectrometry. *Nucl. Instrum. Methods Phys. Res. B* 119:527–532.

Barbour, J. C. and Doyle, B. L. 1995. Elastic recoil detection: ERD. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 83–138. Materials Research Society, Pittsburgh.

Barradas, N. P., Jeynes, C., and Webb, R. P. 1997. Simulated annealing analysis of Rutherford backscattering data. *Appl. Phys. Lett.* 71:291–293.

Betts, R. R. 1979. Time-of-flight detectors for heavy ions. *Nucl. Instrum. Methods* 162:531–538.

Brice, D. K. 1992. Screened Rutherford multiple scattering estimates for heavy ion backscattering applications. *Nucl. Instrum. Methods Phys. Res. B* 69:349–360.

Chevarier, A., Chevarier, N., and Chiodelli, S. 1981. A high resolution spectrometer used in MeV heavy ion backscattering analysis. *Nucl. Instrum. Methods* 189:525–531.

Chu, W. K., Mayer, J. W., and Nicolet, M.-A. 1978. Backscattering Spectrometry. Academic Press, New York.

Darwin, C. G. 1914. Collision of α particles with light atoms. *Philos. Mag.* 27:499–506.

Davies, J. A., Lennard, W. N., and Mitchell, I. V. 1995. Pitfalls in ion beam analysis. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 343–363. Materials Research Society, Pittsburgh.

Doolittle, L. R. 1984. Algorithms for the rapid simulation of Rutherford backscattering spectra. *Nucl. Instrum. Methods Phys. Res. B* 9:344–351.

Doolittle, L. R. 1985. A semiautomatic algorithm for Rutherford backscattering analysis. *Nucl. Instrum. Methods Phys. Res. B* 15:227–231.

Doolittle, L. R., 1990. High energy backscattering analysis using RUMP. *In* High Energy and Heavy Ion Beams in Materials Analysis: Proceedings High Energy and Heavy Ion Beams in Materials Analysis Workshop (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). pp. 175–182. Materials Research Society, Pittsburgh.

Dumke, M. F., Tombrello, T. A., Weller, R. A., Housley, R. M., and Cirlin, E. H. 1983. Sputtering of the gallium-indium eutectic alloy in the liquid phase. *Surf. Sci.* 124:407–422.

England, J. B. A. 1974. Techniques in Nuclear Structure Physics. John Wiley & Sons, New York.

Feldman, L. C., Mayer, J. W., and Picraux, S. T. 1982. Materials Analysis by Ion Channeling. Academic Press, New York.

Lanford, W. A. 1995. Nuclear reactions for hydrogen analysis. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 193-204. Materials Research Society, Pittsburgh.

Leavitt, J. A., McIntyre, Jr., L. C., and Weller, M. R. 1995. Backscattering spectrometry. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 37–81. Materials Research Society, Pittsburgh.

L'Ecuyer, J., Davies, J. A., and Matsunami, N. 1979. How accurate are absolute Rutherford backscattering yields? *Nucl. Instrum. Methods* 160:337–346.

Lennard, W. N., Geissel, H., Winterbon, K. B., Phillips, D., Alexander, T. K., and Forster, J. S. 1986. Nonlinear response of Si detectors for low-z ions. *Nucl. Instrum. Methods Phys. Res. A* 248:454–460.

McDonald, K., Weller, R. A., and Liechtenstein, V. Kh. 1999. Quantitative evaluation of the determinants of resolution in time-of-flight spectrometers for medium energy ion beam analysis. *Nucl. Instrum. Methods Phys. Res. B* 152:171–181.

Mendenhall, M. H. and Weller, R. A. 1989. A time-of-flight spectrometer for medium energy ion scattering. *Nucl. Instrum. Methods Phys. Res. B* 47:193–201.

Mendenhall, M. H. and Weller, R. A. 1990. Performance of a time-of-flight spectrometer for thin film analysis by medium energy ion scattering. *Nucl. Instrum. Methods Phys. Res. B* 40/41:1239–1243.

Mendenhall, M. H. and Weller, R. A. 1991. Algorithms for the rapid computation of classical cross sections for screened Coulomb collisions. *Nucl. Instrum. Methods Phys. Res. B* 58:11–17.

Nastasi, M., Mayer, J. W., and Hirvonen, J. K. 1996. Ion-Solid Interactions: Fundamentals and Applications. Cambridge University Press, Cambridge.

Pedersen, D., Weller, R. A., Weller, M. R., Montemayor, V. J., Banks, J. C., and Knapp, J. A. 1996. Sputtering and migration of trace quantities of transition metal atoms on silicon. *Nucl. Instrum. Methods Phys. Res. B* 117:170–174.

Pfeffer, W., Kohlmeyer, B., and Schneider, W. F. W. 1973. A fast zero-time detector for heavy ions using the channel electron multiplier. *Nucl. Instrum. Methods* 107:121–124.

Porat, D. I. 1973. Review of sub-nanosecond time-interval measurements. *IEEE Trans. Nucl. Sci.* NS-20(5):36:51.

Rauhala, E. 1995. Energy loss. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 3–19. Materials Research Society, Pittsburgh.

Strathman, M. D. 1990. SCATT. *In* High Energy and Heavy Ion Beams in Materials Analysis: Proceedings High Energy and Heavy Ion Beams in Materials Analysis Workshop (J. R. Tesmer,

C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). pp. 183–188. Materials Research Society, Pittsburgh.

Swanson, M. L. 1995. Channeling. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 231–300. Materials Research Society, Pittsburgh.

Tirira, J., Serruys, Y., and Trocellier, P. 1996. Forward Recoil Spectrometry. Plenum Press, New York.

Weller, R. A. 1993. Instrumental effects on time-of-flight spectra. *Nucl. Instrum. Methods Phys. Res. B* 79:817–820.

Weller, R. A. 1995a. Instrumentation and laboratory practice. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 301–341. Materials Research Society, Pittsburgh.

Weller, R. A. 1995b. Scattering and reaction kinematics. *In* Handbook of Modern Ion Beam Materials Analysis (J. R. Tesmer and M. Nastasi, eds.). pp. 412–416. Materials Research Society, Pittsburgh.

Weller, R. A., Arps, J. H., Pedersen, D., and Mendenhall, M. H. 1994. A model of the intrinsic efficiency of a time-of-flight spectrometer for keV ions. *Nucl. Instrum. Methods Phys. Res. A* 353:579–582.

Weller, R. A., McDonald, K., Pedersen, D., and Keenan, J. A. 1996. Analysis of a thin, silicon-oxide, silicon-nitride multilayer target by time-of-flight medium energy backscattering. *Nucl. Instrum. Methods Phys. Res. B* 118:556–559.

Whitlow, H. J. 1990. Time of flight spectroscopy methods for analysis of materials with heavy ions: A tutorial. *In* High Energy and Heavy Ion Beams in Materials Analysis: Proceedings High Energy and Heavy Ion Beams in Materials Analysis Workshop (J. R. Tesmer, C. J. Maggiore, M. Nastasi, J. C. Barbour, and J. W. Mayer, eds.). pp. 243–256. Materials Research Society, Pittsburgh.

Williams, J. S. and Möller, W. 1978. On the determination of optimum depth-resolution conditions for Rutherford backscattering analysis. *Nucl. Instrum. Methods* 157:213–221.

Wiza, J. L. 1979. Microchannel plate detectors. *Nucl. Instrum. Methods* 162:587–601.

Yamamura, Y. and Tawara, H., 1996. Energy dependence of ion-induced sputtering yields from monatomic solids at normal incidence. *Atomic Data Nucl. Data Tables* 62:149–253.

Yang, Q., O'Connor, D. J., and Wang, Z. 1991. Empirical formulae for energy loss straggling of ions in matter. *Nucl. Instrum. Methods Phys. Res. B* 61:149–155.

Ziegler, J. F., Biersack, J. P., and Littmark, U. 1985. The Stopping and Range of Ions in Solids. Pergamon Press, New York.

KEY REFERENCES

Chu et al., 1978. See above.

The definitive work on conventional mega-electron-volt backscattering with surface barrier detectors.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. North-Holland, Elsevier Science Publishing, New York.

Introductory text containing excellent descriptions of many surface and thin-film analytical techniques.

Gibson, W. M. and Teitelbaum, H. H. 1980. Ion-Solid Interactions, a Comprehensive Bibliography, Vols. 1–3. INSPEC, Institution of Electrical Engineers, London.

Exhaustive compilation of the literature of the field of ion beam interactions with solids up to approximately the date of its publication.

Knoll, G. F. 1989. Radiation Detection and Measurement, 2nd ed. John Wiley & Sons, New York.

A standard text on all aspects of radiation detection.

Nastasi et al., 1996. See above.

Excellent text reviewing fundamentals of ion beam interactions with materials.

Nuclear Instruments and Methods in Physics Research, Section B. Elsevier Science Publishers, Amsterdam.

In recent years a very significant portion of the literature of the field of ion beam interactions with materials has been published in this journal. Much additional material may be found in Section A of the journal and in the original journal Nuclear Instruments and Methods from which the modern series are derived.

Tesmer, J. R. and Nastasi, M. (eds.). 1995. Handbook of Modern Ion Beam Materials Analysis. Materials Research Society, Pittsburgh.

Detailed treatment of all aspects of ion beam analysis with numerous tables and tutorial material. The standard reference.

Tirira et al., 1996. See above.

A recent, comprehensive treatment of mega-electron-volt forward-recoil spectrometry.

ROBERT A. WELLER
Vanderbilt University
Nashville, Tennessee

HEAVY-ION BACKSCATTERING SPECTROMETRY

INTRODUCTION

Heavy-ion backscattering spectrometry (HIBS) is a new technique for nondestructively analyzing ultratrace levels of impurities on the surface of very pure substrates. Although any high-Z contaminant ($Z \geq 18$) on a pure low-Z substrate can be measured, recent practical applications for HIBS have focused on measuring contaminants found on silicon wafers used in semiconductor manufacturing. Therefore, discussions of HIBS will focus in this area. The technique is based on the same principles described elsewhere (see ELASTIC ION SCATTERING FOR COMPOSITION ANALYSIS and MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY) for high-energy Rutherford backscattering spectrometry (RBS) and medium-energy ion backscattering spectrometry. The approach was invented and patented at Sandia National Laboratories and led to the construction of a HIBS User Facility, built in collaboration with SEMATECH member companies and Vanderbilt University. The facility was opened in June 1995 for use by U.S. industry, national laboratories, and universities in conducting contamination studies and is located at Sandia.

The motivation for HIBS is to detect metallic contamination at levels significantly below those that can be detected by RBS or medium-energy ion backscattering, both of which have a limit of $\sim 1 \times 10^{13}$ atoms/cm^2 for near-surface impurities. By the year 2001, as reported in the National Technology Roadmap for Semiconductors

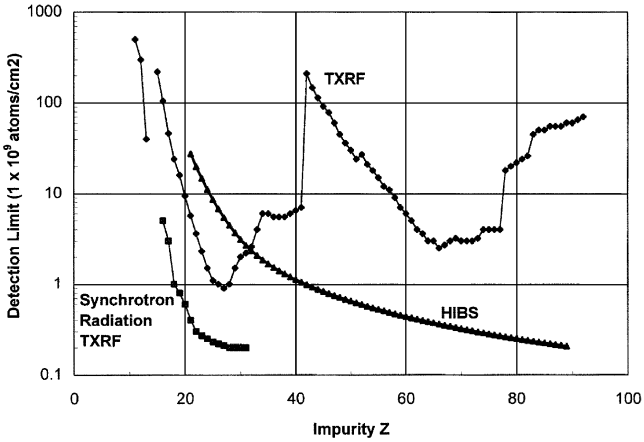


Figure 1. HIBS vs. total-reflection x-ray fluorescence (TXRF) detection limits on pure silicon. TXRF limits are shown using traditional x-ray and newer synchrotron radiation sources.

(Semiconductor Industry Association, 1997), contamination control in the microelectronics industry will require tools that can measure $\sim 1 \times 10^{10}$ atoms/cm^2 for transition metals in starting material. A sensitivity curve, shown in Figure 1, demonstrates how HIBS meets this requirement on silicon. The detection limit for well-separated elements on a clean Si surface ranges from $\sim 6 \times 10^9$ atoms/cm^2 for Fe to $\sim 3 \times 10^8$ atoms/cm^2 for Au, without the use of vapor phase decomposition (VPD) to preconcentrate the impurities. Using VPD would improve the sensitivity by at least an order of magnitude.

The HIBS technique can be illustrated by following the ion trajectory shown in Figure 2. An 120-keV C^+ ion beam is typically used, with the beam focused onto an $\sim$3-mm^2 spot at the sample surface. The relatively large beam spot size, which reduces sputtering and sample heating, allows the use of a higher beam current ($\sim$20 nA). Ion beam effects such as sputtering, sample heating, and ion beam enhanced diffusion are discussed later. It should be noted that sputtering, although limiting the ultimate sensitivity of HIBS, is not as serious a limitation as once thought (Pedersen et al., 1996). Carbon ions backscattered from the sample are collected by a time-of-flight detector array having a large solid angle maximized for high sensitivity. The screening foil in each detector has a thickness chosen to filter out carbon ions scattered from the substrate but to pass ions scattered from heavier atoms, such as impurities on the surface. Measuring the yield versus flight time of the backscattered carbon ions allows both the mass and areal density of the near-surface impurities to be determined from basic physical backscattering concepts.

Other surface analysis techniques for detection of trace element contamination are briefly discussed next, with the strengths and weaknesses of each compared to those of HIBS. A detailed discussion of the basic theoretical and practical aspects of HIBS will then be provided.

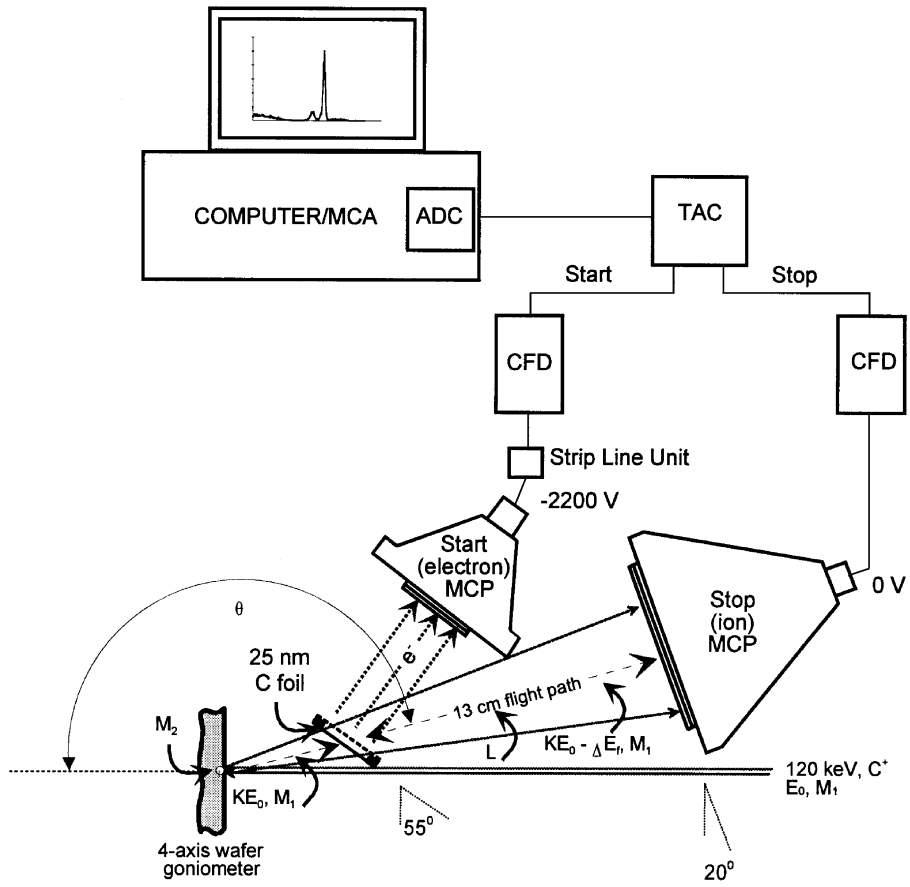


Figure 2. Typical HIBS time-of-flight detector with associated data acquisition instrumentation. Backscattering kinematics are also illustrated.

Competitive, Complementary, and Alternative Techniques

A number of other techniques for trace element analysis are available, with total-reflection x-ray fluorescence (TXRF) and secondary-ion mass spectroscopy (SIMS) being the most widely used. All the techniques differ from HIBS in one very important aspect: they require standard samples for quantitative analysis. Since HIBS is a standardless measurement, one of its primary applications is to calibrate standard samples for use by the other analysis techniques. Another difference is that all but TXRF are destructive; that is, the sample may not be reused or subjected to an alternative analysis in the same area. The HIBS technique is nondestructive for further wafer analysis, but the wafer cannot be returned to production without further processing.

Total-reflection x-ray fluorescence using traditional x-ray sources is both a competitive and a complementary technique. Recently, a version of TXRF using synchrotron radiation has been introduced with the potential for higher sensitivity to many elements than can be achieved with any other method. Neither HIBS nor TXRF require sample preparation, and both provide spatial information about the impurities on the test sample. For an accurate measurement, TXRF requires an optically flat surface, such as the front side of a wafer, whereas HIBS can also measure rough surfaces, such as the backside of a wafer. Typically, TXRF is used to measure only the transition metals, with an anode change required to measure high-Z metals, while HIBS surveys all elements with $Z \geq 18$ in a single experiment. As shown in Figure 1, TXRF has relatively poor sensitivity for high-Z metallic contamination ($Z > 30$), so HIBS complements TXRF for measuring high-Z contamination. On the other hand, TXRF, using traditional x-ray sources, has better mass resolution than HIBS and is used extensively by the semiconductor industry, being the tool of choice when measuring production wafers. Mass resolution in HIBS ranges from ~2 amu for Fe to ~20 amu for Pb, but its survey capability and accuracy make it an excellent tool for precise quantification of known contaminants. The analysis depth for TXRF is ~5 nm on an unpatterned bare silicon wafer or SiO_2, whereas HIBS measures contamination in samples to a depth of ~10 nm. Even when measuring the same wafer for the concentration of contaminants, TXRF instruments may report widely varying areal densities. Because of its high accuracy and standardless configuration, HIBS has been used to nondestructively verify the areal density of contamination on TXRF standards and calibrate TXRF tools (Werho et al., 1997; Banks et al., 1998). High concentrations of one element can cause interference in detecting other elements, for both TXRF and HIBS. The TXRF instruments have been highly developed, providing automated wafer exchange, data acquisition, and analysis. The HIBS system at Sandia is partially automated, but it requires up to 90 min per point when analyzing for $\sim 1 \times 10^9$ atoms/cm^2 levels, while TXRF measurements take ~15 min per point for the same level.

Secondary-ion mass spectrometry has been used for many years and is a valuable tool for many problems. Based on the sputter erosion of the surface being analyzed, SIMS is inherently a depth profiling technique with good bulk sensitivity (<1 ppm). However, its sensitivity to a thin surface contaminant is not as good as other techniques and its sensitivity to various elements can vary by orders of magnitude in the same substrate. Even the sensitivity for a single element varies with different substrates.

Destructive alternative techniques such as inductively coupled plasma mass spectroscopy (ICP-MS), graphite furnace atomic absorption mass spectroscopy (GFAA-MS), glow discharge mass spectroscopy (GD-MS), and a variation of TXRF called microsample x-ray analysis (MXA) are used to measure contamination, mostly in liquids, to the ppt level. They can be used, indirectly, to measure surface contamination by concentrating the elements through dissolution of the sample surface using VPD. The VPD technique exposes the surface of a wafer to an acidic HF vapor for a set time, and then the etched products are collected by a droplet of ultraclean deionized water rolled around the surface. Finally, the liquid droplet is analyzed by one of the above techniques. The droplet can also be placed on a separate clean low-Z film and allowed to dry for x-ray analysis (or HIBS). Collection efficiencies between 95% and 99% have been reported using the VPD process, except for Cu, for which it is ~20%. There is some concern about the use of this technique as wafer sizes increase. All the techniques using VPD are labor intensive; however, automated VPD equipment has recently been introduced and could increase throughput. The availability of these techniques and their exceptional sensitivity, ranging to 1×10^8 atoms/cm^2, make them competitive.

PRINCIPLES OF THE METHOD

The four principles of traditional RBS on which HIBS is based (see MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY) are (1) the kinematic factor K, which describes the energy transfer between a projectile and target nucleus in an elastic collision; (2) the scattering cross-section σ, defined as the probability of a scattering event occurring; (3) the stopping cross-section ε, which determines the energy loss of a projectile as it undergoes interactions with target nuclei and electrons; and (4) the energy straggling Ω_{Bohr}, i.e., a statistical fluctuation in the energy loss of a projectile.

The yield equation $Y = Q\sigma N t \Omega_D \eta$ gives the yield of backscattered particles as a function of the number of ions (Q) hitting the target layer, the atomic density (N) and thickness (t) of the layer, and the probability (σ) of the projectile ions being backscattered from the layer into a detector with a finite solid angle (Ω_{D}) and detection efficiency (η).

It follows from the yield equation that if the factors Q, σ, Ω_{D}, and η are increased, then the corresponding yield increase allows for the detection of increasingly lower areal densities, where the areal density is the product of the target atomic density and thickness, Nt, normally expressed in atoms per square centimeter. Detection of increasingly lower areal densities is commonly referred to as increasing the sensitivity.

Although HIBS is based on the same fundamental concepts as RBS and medium-energy ion backscattering, the emphasis is placed on maximizing sensitivity to trace elements on the surface of a lower Z substrate. This increased sensitivity is accomplished in two ways. First, since the backscattering cross-section (σ) scales as Z^2/E^2, where Z and E are the atomic number and energy of the analysis beam, the HIBS scattering yield from all elements is dramatically increased by using a high-Z analysis beam with a relatively low energy, such as C^+ at 120 keV.

Of course, this increase in yield also increases the backscattering from the substrate, so a means of eliminating the ions scattered from substrate atoms is needed. This is accomplished by the second, key aspect of the technique: a screening foil incorporated in the design of the time-of-flight detectors eliminates essentially all ions backscattered from the substrate while still allowing those backscattered from heavy impurities on the surface to be detected. The foil is already required in the time-of-flight detector for producing a start pulse for the timing. By making the foil somewhat thicker than required, only ions backscattered with energies above a given threshold are passed through the foil to the second part of the detector, thus eliminating most of the ions backscattered from the substrate (such as silicon), which would otherwise comprise an overwhelmingly large background in the spectrum. It should be noted that simply using electronic discrimination of this background cannot be used because the microchannel plates would have to process the recoiled ions from substrate atoms, saturating the detectors.

Some ions scattered from the substrate will still be passed through the foil because of straggling in the foil and multiple small-angle scattering in the substrate, so further improvements can be made by use of ion channeling. When allowed by the sample configuration, aligning the analyzing ion beam with the crystalline channels in the substrate will further reduce the residual background from the substrate, helping to increase sensitivity.

Further improvements in sensitivity are obtained in HIBS by using multiple time-of-flight detectors, each with a large solid angle (Ω_D), and a relatively large ($\sim$3-mm^2) beam spot. The large beam allows higher beam current and longer counting time, increasing the number of projectile ions (Q) while minimizing sputter erosion of the sample. Although the efficiency of each time-of-flight detector is lower than that of the surface barrier detector (SBD) commonly used in RBS measurements, the gain in mass resolution offsets the lower efficiency. Another advantage of an ion beam analysis such as HIBS is that measurement of areal density of surface impurities on a target is largely insensitive to chemical bonding or electronic configuration in the target (Feldman and Mayer, 1986; Ziegler and Manoyan, 1988).

Depth information about the sample is contained in the energy loss of the projectile ion and can be extracted using the known stopping cross-section (ε), with straggling (Ω_{Bohr}) limiting the resolution of this depth information. The energy loss is due to two mechanisms: projectile interactions with electrons in the target, referred to as electronic stopping, and small angle scattering with target atoms, called nuclear stopping. Since HIBS uses a low-energy beam of carbon ions, depth information is limited to very near the surface, and thus the technique is generally not useful for depth profiling.

Basic Theory

The fundamentals of HIBS are based on the four basic physical concepts of traditional RBS, as previously noted and found in several excellent reference books (e.g., Chu et al., 1978; Feldman and Mayer, 1986; Tesmer et al., 1995). These references discuss RBS fundamentals using a He ion beam and an SBD to obtain a yield-versus-energy spectrum. Early HIBS experiments also used an SBD (Knapp et al., 1994a). However, limitations of the SBD, discussed later, caused its discontinued use. The HIBS technique now uses the carbon foil-based time-of-flight detection scheme, which is described in the literature (Chevarier et al., 1981; Mendenhall and Weller, 1989; Dobeli et al., 1991; Knapp et al., 1994b; Anz et al., 1995).

The four basic Rutherford backscattering physical concepts are kinematic factor, scattering cross-section, stopping cross-section, and straggling, with the information gained being that of mass identity, atomic composition, depth, and the limits to mass and depth resolution, respectively. Equations based on these concepts allow analysis of the raw HIBS backscattering data.

Referring to Figure 2, when a projectile ion with initial energy E_0 and mass M_1 undergoes a simple elastic collision with a stationary atom M_2 at the surface of the substrate, the backscatter energy E_1 at an angle θ in the laboratory frame of reference can be determined from the kinematic factor, defined as $K \equiv E_1/E_0$, where $M_2 > M_1$. Thus the backscatter energy is

$$E_1 = KE_0 = E_0\left[\frac{\{x\cos(\theta) + [1 - x^2\sin^2(\theta)]^{1/2}\}}{1+x}\right]^2 \quad (1)$$

where $x \equiv M_1/M_2$. Note that the backscattered particle energy E_1 is only dependent on the initial energy E_0, the ratio of the projectile and target masses, M_1/M_2, and the laboratory scattering angle θ. After backscattering, the particle energy can also be described by the product of its mass and velocity, $E_1 = \frac{1}{2}M_1V_1^2$, where the projectile velocity can be found from $V_1 = L/\tau$ by measuring the time τ to travel the fixed path length L the ion traverses in the detector. These basic equations allow a determination of the mass of the atom from which the ion was backscattered, which is the general principle on which the time-of-flight technique is based.

The carbon projectile loses energy when penetrating and exiting the target and when passing through the carbon foil. Carbon ions that are not immediately backscattered from surface contaminant atoms penetrate the surface of the target, lose a finite amount of energy, are backscattered, and then lose more energy on their exit path, with the total energy loss denoted as ΔE_s. Many of the ions, of course, are not backscattered and come to rest in the sample. (The probability of backscattering, or the scattering cross-section, will be discussed later.) The backscattered ions are collimated, to reduce stray ion

scattering, and then pass through the carbon foil, losing additional energy ΔE_f. Most energetic ions that pass through the carbon foil eject electrons that are accelerated into the start microchannel plate (MCP) and create a start timing pulse τ_0. As described earlier, the cumulative energy losses occur through electronic and nuclear stopping, with the losses used to advantage by stopping most projectile ions scattered from the substrate in the carbon foil, reducing unnecessary counts from these scattering events. However, due to straggling in the foil and multiple scattering in the substrate, the ions backscattered from silicon can never be eliminated completely.

Energy lost in a layer can be calculated from the stopping power of the material in the layer. This loss is given as $\Delta E_s = [\varepsilon]Nt$, where $[\varepsilon]$ is the stopping cross-section factor, N is the atomic density of the layer, and t is the thickness of the layer. The stopping cross-section is defined as $\varepsilon \equiv (1/N)(dE/dx)$, where dE/dx is the stopping power or change in energy per unit distance. The stopping cross-section typically has units expressed as eV/10^{15} atoms/cm^2. The stopping cross-section factor is given by the formula (Chu et al., 1978)

$$[\varepsilon] = \left[\frac{K}{\cos\theta_1}\varepsilon_{in} + \frac{1}{\cos\theta_2}\varepsilon_{out}\right] \tag{2}$$

where θ_1 and θ_2 are the angles that the projectile ion makes with respect to the surface normal as it enters and leaves the target and ε_{in} and ε_{out} are the stopping cross-sections along the inward and outward paths of the projectile ion, respectively. Although the low-energy carbon projectile ions lose a small amount of energy while penetrating and exiting the target, the loss is neglected in HIBS analyses because the surface energy loss KE_0 is approximately equal to the total loss ΔE_s.

The energy of the ions escaping the substrate or the carbon foil exhibits a statistical energy distribution referred to as straggling. This straggling limits both the energy (and therefore the time) and mass resolution of the system. Straggling is commonly described using Bohr's equation,

$$\Omega^2_{Bohr} = 4\pi Z_1^2 e^4 N Z_2 t \tag{3}$$

which shows that for a projectile ion with atomic number Z_1 passing through a layer with atomic density N, atomic number Z_2, and thickness t, the projectile ion energy straggling (Ω_{Bohr}) increases with the square root of the foil electron areal density, NZ_2t. Straggling exhibits a projectile ion velocity dependence, not shown in Bohr's equation, and therefore, modified versions of Bohr's theory are now used (Lindhard and Scharff, 1953; Bonderup and Hvelplund, 1971; Chu, 1976; and Besenbacher et al., 1980). It should be noted that no theory exists for slow heavy ions in a light matrix. Bohr's equation predicts a symmetrical Gaussian distribution of energy equal to $2.355\,\Omega_{Bohr}$ at full width at half-maximum (FWHM) due to projectile ion interactions with electrons in the foil. For most backscattering spectra, straggling appears as the integral of the Gaussian distribution and is described by a complementary error function. Empirically, this is seen as a slope in the backscattered energy edge for each target mass, with the energy distribution of $2.355\Omega_{Bohr}$ at FWHM corresponding to the 12% to 88% range of the slope.

After the ion traverses the start foil and produces the start timing signal, it continues its flight until being stopped by the ion MCP, where it produces a stop signal τ_{stop}. After signal processing, a time-to-amplitude converter is used to give a digital pulse, with a height proportional to the time difference between the start and stop timing signal, $\tau_{stop} - \tau_0$. These signals are binned in a multichannel analyzer, giving a yield-versus-time spectrum. The yield for the ith element, Y_i, is dependent on several factors, according to the equation

$$Y_i = \eta(KE_0 - \Delta E_f)Q(N_i t)\Omega_D \sigma_i(E_0, \theta) \tag{4}$$

where $\eta(KE_0 - \Delta E_f)$ is the efficiency of the detector at the energy of the projectile ion after passing through the foil, Q is the total number of projectiles hitting the target, N_i is the atomic density in atoms per cubic centimeter of the ith contaminant element, t is the thickness of the contaminant layer, Ω_D is the detector solid angle, and $\sigma_i(E_0, \theta)$ is the screened cross-section for the collision with the ith element. Determining the yield is complicated by the energy dependence of the detector efficiency and the reduction of the scattering cross-section due to electron screening as discussed in detail by elsewhere (see MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY). However, the yield can be determined with reasonable accuracy by a computer analysis algorithm, described later. The screened scattering cross-section (Anderson et al., 1980) is given in the equation

$$\sigma_i = \sigma_R\left(1 + \frac{0.04873\,Z_1 Z_2 (Z_1^{2/3} Z_2^{2/3})^{1/2}(1+x)}{E_0}\right)^{-1} \tag{5}$$

where E_0 is the initial projectile ion energy, Z_1 and Z_2 are the projectile and the ith target element atomic numbers, respectively, and $x \equiv M_1/M_2$. This value differs by only a few percent from the more precise Lenz-Jensen screened cross-section (see MEDIUM-ENERGY BACKSCATTERING AND FORWARD-RECOIL SPECTROMETRY). The Rutherford cross-section σ_R in the laboratory frame of reference is given (Darwin, 1914; Chu et al., 1978) as

$$\sigma_R(E_0, \theta) = \left(\frac{Z_1 Z_2 e^2}{2E_0}\right)^2 \left(\frac{\{\cos(\theta) + [1 - x^2\sin^2(\theta)]^{1/2}\}^2}{\sin^4(\theta)[1 - x^2\sin^2(\theta)]^{1/2}}\right) \tag{6}$$

where E_0, Z_1, Z_2, and x are described in Equation 5, θ is the laboratory angle through which the projectile ion is scattered, and e^2 is the square of the electron's charge, which may be expressed as 1.4398×10^{-10} keV·cm.

PRACTICAL ASPECTS OF THE METHOD

The first HIBS experiments used an SBD. However, limitations of the SBD (e.g., low count rate leading to pulse

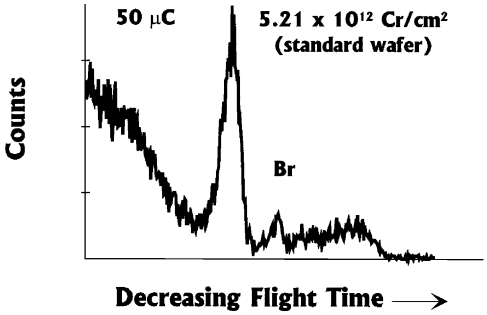


Figure 3. An important use of HIBS is to calibrate other trace element analysis tools. The HIBS measurements on Cr (above), Fe, and Ni contamination standard wafers were used to calibrate TXRF instruments at Motorola.

pileup, degraded energy resolution due to electronic noise, and radiation damage from high-Z ions) warranted changing to a time-of-flight detection scheme using a carbon foil. Since the time-of-flight HIBS detectors give a yield-versus-flight-time spectrum, it looks similar to one acquired by RBS, with spectra showing increasing mass numbers to the right (this assumes that flight time is shown decreasing to the right). Typical spectra can be seen in Figure 3 and Figure 4. Portions of the spectra in Figure 4 show the physical phenomena that limit the mass resolution and sensitivity of the time-of-flight technique.

Time-of-Flight Spectrometer

A schematic of the time-of-flight spectrometer used in HIBS is shown in Figure 2 and consists of a collimator, a thin carbon foil (~25 nm) placed in the path of the backscattered carbon projectiles, an electron accelerating grid, and two MCPs. Secondary electrons are produced when the projectile passes through the foil, with the number ejected based on statistics. These electrons are then accelerated by the grid into the first MCP to provide a start timing pulse. As discussed previously, the foil also serves as a filter to eliminate most of the carbon ions scattered from the substrate, suppressing the substrate background. Carbon ions that pass through the foil continue their flight until being stopped by the second MCP, where they produce a stop timing pulse.

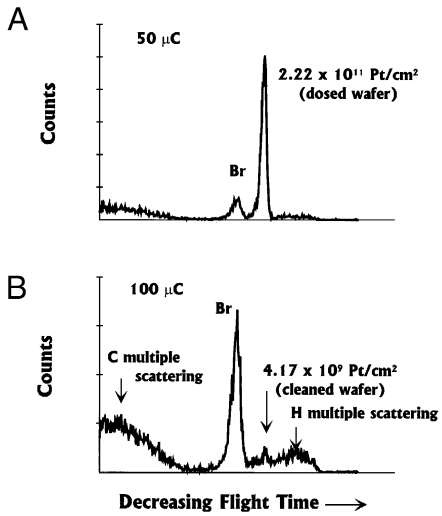


Figure 4. Another important use of HIBS is for contamination-free semiconductor manufacturing research. Texas Instruments used HIBS analyses to determine the dose of Pt contaminant on a wafer (**A**) and to show the efficacy of a HNO_3:HF (95:5) mix to clean the wafer (**B**). Scales for (**A**) and (**B**) are identical. Spectra background (**B**) is due to multiple scattering of the C analysis beam in the Si substrate and H recoiled from the wafer's surface.

Sensitivity and Mass Resolution

A number of factors influence the detection limits and mass resolution that can be achieved with HIBS: detector efficiency, choice of beam species and energy, sputtering, and background due to multiple scattering and random coincidences (Knapp et al., 1996). The detection limits can be reduced in a number of ways: increasing the solid angle by using multiple detectors, using a larger beam spot, channeling the incident beam along a crystal axis of the substrate, and matching the choice of beam to the screening foil thickness. All of these were incorporated in the design of the HIBS system at Sandia. Of course, sputtering of the sample surface by the analysis beam is the ultimate limit to the statistics that can be obtained, and hence the sensitivity. Beam effects due to sample heating are considered negligible. Sputtering and ion beam–enhanced diffusion can be potential problems when measuring surface contaminants in a metal-silicon system because erosion and diffusion can cause a decrease in the backscattering signal. If it is suspected that sputtering or beam-enhanced diffusion might be a problem, the areal density as a function of beam flux can be measured to determine if the problem exists and, if it does, extract the starting areal density. A study using a large N^+ ion flux to measure metallic surface contaminants (Fe, Cu, Mo, W, and Au on silicon) showed that sputtering is not a significant limitation to the sensitivity of HIBS for these elements and that diffusion was limited to <10 nm (Pedersen et al., 1996).

We define the detection limit as $DL = 2 \times (\text{background})^{0.5} \times (\text{calibrationfactor})$ where the background is taken under the full width of the peak for the element of interest. A number of alternative definitions could be used, and many are discussed by Currie (1960). Heavy-ion backscattering spectrometry cannot be considered to have a "well-known blank" (a sample with no contaminants, to be measured as a background), since we have found that no Si wafer is completely free of heavy-element contamination at levels HIBS can detect. However, since the peaks due to contaminants in a spectrum are generally surrounded on each side by wide regions of well-defined background, the background under the peaks can be fit fairly well and its standard deviation reduced to a low level. In this case, the criteria of Currie (1960) would be $DL = 1.64 \times (\text{background})^{0.5} \times (\text{calibrationfactor})$, or 18% lower than our definition.

The efficiency of the MCPs for detecting electrons, ions, and neutral particles in the sub-100-keV energy range is not well known, nor is the electron-producing efficiency of the carbon screening foil. The MCPs also have a count

rate limitation that may in turn limit the amount of beam current that can be used. Straggling in the foil is also an important effect, dominating the observed time resolution and limiting the mass resolution that can be achieved, through broadening of the peak widths. The mass resolution of the Sandia HIBS system, with its present 5-$\mu g/cm^2$ carbon foil, is about 1 amu for Ti-V, 3 amu for Cu-Zn, and 18 to 20 amu for the heaviest elements. That is, for elements near Cu, two peaks of equal magnitude can be separated if the masses of the two elements differ by 3 amu or more.

Scattering in the foil is also a serious consideration for the detector sensitivity, since it results in some ions exiting at angles that do not intercept the ion MCP, reducing the effective solid angle of the detector. This effect is stronger at lower energy, offsetting the gains in overall sensitivity that can be obtained by going to lower beam energy. The problem has been illustrated by TRIM (a code applied to the transport of ions in matter; Ziegler and Biersack, 1995) calculations of the distribution of trajectories of N particles passing through a foil at two different energies. The percentage of particles that remain in the acceptance angle of the ion MCP is much lower at low energy: the simulation shows 85% of the 160-keV particles reach the MCP, but only 22% do so at 40 keV.

The effect of foil scattering will vary with beam species and energy and of course will be different for different foil thicknesses. For a given foil, measurements of the detector efficiency will show that some ions are detected more efficiently than others, primarily because of the relative amounts of foil scattering, but also due to details of electron production efficiency and MCP detection efficiency. An analytical model of the efficiency of time-of-flight ion detectors of this general configuration has been developed elsewhere (Weller et al., 1994).

Background in the spectra is the major limit to the sensitivity that can be achieved. We have found that background can come from several sources: scattering of ions and electrons around the chamber, scattering from the substrate "leaking" through the foil, and beam contamination posing major problems. The first problem can be minimized by proper containment of the beam and shielding of the detectors from unwanted ions and electrons. The second problem is due not just to straggling in the foil but also to multiple scattering of the ion beam in the substrate: for a very low number of incoming ions, multiple scattering events can steer the ions back out of the substrate with higher than expected energy, making the particles indistinguishable from those scattered from heavy atoms on the surface. This low-level background cannot be eliminated, but it can be minimized by channeling of the analysis beam.

While using N^+ and O^+ ion beams for HIBS analysis, we also identified beam contamination by NH_2 and CH_2 and forward scattering of H as sources of background in the Sandia HIBS system, often at such low levels that they had remained unnoticed in an earlier HIBS system with less sensitivity. To minimize beam contamination problems, we settled on a C^+ beam for all high-sensitivity measurements. Using the Sandia HIBS system with three detectors and 25-nm foils, we have observed background levels implying detection limits to impurities on Si ranging monotonically from $\sim 6 \times 10^9$ atoms/cm^2 for Fe to $\sim 3 \times 10^8$ atoms/cm^2 for Au.

Instrumentation

Figre 2 shows a typical time-of-flight detector and associated instrumentation used for HIBS measurements. The detector and instruments are commercially available and consist of MCP assemblies for the start and stop MCPs, a strip line unit for the start MCP, constant fraction discriminators (CFDs), a time-to-amplitude converter (TAC), an analog-to-digital converter (ADC) board, and a multichannel analyzer (MCA). The start MCP uses a 20-mm-diameter chevron and the stop MCP uses a 40-mm-diameter chevron. Since HIBS uses three time-of-flight detectors to maximize the detector solid angle, three each of the units mentioned above are needed, with the exception of the CFDs and MCA. Two CFDs (with quad inputs) can be used to process all start and stop timing signals, and the MCA has the capability of processing data from all three detectors. The strip line unit is used to block the -2200-V output level of the electron MCP and allow the timing signal to pass while providing impedance matching to reduce signal reflections. A fast oscilloscope (bandwidth $\geq$ 400 MHz) is needed to monitor the timing pulse outputs from the MCP assemblies and to set up the CFDs.

A CFD that provides a zero-crossing timing pulse output for use by the TAC module gives better time resolution than the leading-edge type. A short coaxial cable, whose length is determined experimentally, must be used to set the zero-crossover point and adjust the CFD walk, which can be labor intensive. At least one manufacturer (Ortec) has recently offered a picosecond timing discriminator that does not require this cable but processes only one signal.

Recent advances in MCP assembly technology (Galileo) have resulted in low-profile versions being made available, with a tenfold height reduction, which could allow a more compact HIBS system to be designed. This new generation of MCPs has an extended dynamic range, a higher gain, a shorter output pulse width, and lower noise. These features could allow a higher count rate, greater efficiency, and better mass resolution. These low-profile MCPs also have integral resistor biasing of the plates and feature a quick-change module for replacement of the plates, reducing both installation and maintenance time.

A small 200-keV accelerator (Peabody Scientific, model PS-250) using a Duoplasmatron gas source is used to obtain the C^+ projectile ions measured by the HIBS instrumentation described above. However, implanters or other accelerators with good energy stability ($\sim$250 eV), a high-resolution mass analyzing magnet ($\leq$0.5 amu), and other features similar to the Peabody accelerator would also work well.

Examples

Measurements by two major semiconductor manufacturers, Motorola and Texas Instruments, serve as examples of using HIBS to solve integrated circuit manufacturing

problems. Motorola has used the HIBS facility to help solve problems in the calibration of TXRF instruments (Werho et al., 1997; Banks et al., 1998). The TXRF instruments are used for detecting trace levels of transition metals introduced by semiconductor processing procedures and equipment. These instruments are typically calibrated using the manufacturer's calibration data. This independent calibration has led to inaccuracies and a resulting lack of interchangeability in TXRF data. To illustrate the problem caused by independent calibration, the average error in measuring the same contamination standard wafers by different laboratory sites during a round robin conducted at Motorola showed errors in excess of 350%. To solve this problem, HIBS was used to nondestructively verify the areal density of a contaminant on commercially obtained contamination standard wafers used to calibrate these instruments. Figure 3 shows a typical spectrum taken on a Cr standard wafer. The HIBS values measured on this Cr-contaminated wafer, as well as other standard wafers contaminated with Fe and Ni, generally verified the vendor's value to within $\pm 10\%$. The HIBS data were then used to create a universal sensitivity factor curve for recalibration of the TXRF instruments used in the study. This recalibration generally reduced the error between TXRF instruments to $< \pm 10\%$. Using this universal method for calibrating TXRF instruments has led to improved quality control during integrated circuit (IC) manufacturing. The calibration method has international applicability, since the lack of interchangeability of data between TXRF instruments is a problem that has been reported worldwide (Hockett et al., 1992).

Texas Instruments has used HIBS technology to evaluate the handling and processing of new materials that can be used in the manufacture of ultralarge-scale integrated (ULSI) dynamic random-access memories (DRAMs). Texas Instruments used the HIBS facility at Sandia National Laboratories in the development of cleaning processes for the removal of high-Z metallic contamination (Banks et al., 1998). To decrease the size of capacitors used in ULSI DRAMs, Texas Instruments has been studying the use of a high-k dielectric (where k is the relative dielectric constant), barium-strontium-titanate (BST) between platinum electrodes (Pt/BST/Pt). Because of the concern over cross-contamination of wafers and processing tools with these materials, several questions needed to be answered, including how effective their present cleaning methods were at removing these heavy metals. To begin answering this question, wafers were uniformly dosed with the contaminants and the areal density was measured. The wafers were then subjected to various cleaning methods, the results of one of which is shown in Figure 4, and measured again. This comparison shows the efficacy of the cleaning process and demonstrates the excellent sensitivity of HIBS.

Safety Considerations

The two primary safety hazards associated with HIBS measurements are exposure to (1) high voltage and (2) x radiation. These hazards are mitigated through the use of stringent safeguards. A physical barrier must be set up around the accelerator with an appropriate door interlock used to disconnect power to the high-voltage supply in the event that high voltage is inadvertently left on when attempting to enter the accelerator enclosure. In addition, it is a good practice to use a radiation monitor interlocked with the high-voltage supply, shutting the supply off when radiation above acceptable limits is detected. Another feature that should be incorporated in the accelerator enclosure design is to provide an OSHA-approved grounding rod for discharging high voltages that might exist on various isolated accelerator components. The accelerator should always be grounded before conducting maintenance or repairs. Finally, easily accessible panic switches should be located both inside and outside an enclosure, providing for emergency shutdown of the accelerator. All applicable regulations of local, regional, and national authorities should be adhered to for installation and operation of the accelerator. General radiological guidelines for conducting ion beam measurements and a listing of regulating agencies can be found in Tesmer et al. (1995).

METHOD AUTOMATION

Sample positioning and data collection in HIBS can be semiautomated so that little operator intervention is needed during data acquisition, with the operator only needing to make minor adjustments of the ion accelerator parameters. In the HIBS system at Sandia, motor control hardware and software using the IEEE-488 protocol are used in conjunction with Microsoft Visual Basic® to control a four-axis sample goniometer. This control is needed so that a wafer (or sample) may be positioned at precise x and y coordinates and also allow channeling. The software automatically converts x and y coordinates (given in millimeters) into polar coordinates for translation by the stepping motors. To mimic the spot size of TXRF instruments, HIBS steps the beam spot ($\sim$3 mm^2) around in a $\sim$1-cm^2 pattern. This allows direct comparison of the areal density collected and reported by TXRF instruments to that of HIBS. This large pattern also further reduces the sputtering of contaminants being measured.

The raw time-of-flight data are collected by three ADC boards installed in a personal computer, with the three spectra being processed by a commercially available MCA software package (Oxford) that interfaces with the boards. Data are normally acquired until a preset amount of charge has been collected on a sample. An automated beam stop incorporated into the beam line blocks the beam after reaching a preset condition. With this minimal degree of automation, data collection can proceed while data are being analyzed.

DATA ANALYSIS AND INITIAL INTERPRETATION

The HIBS data analysis provides information about the identity and concentration of heavy-metal surface impurities on very pure substrates. Computer algorithms, derived from fundamental equations and discussed below,

are used to determine areal density and mass for a given peak in the time-of-flight spectrum. It should be noted that this spectrum is the sum of three separate but simultaneously acquired spectra from the HIBS detectors. Spectra manipulation, such as shifting and scaling prior to the analysis, is not discussed. However, this spectra manipulation and subsequent analysis have been semi-automated with software written in Microsoft Visual Basic®. Copies of the program, called HIBWIN© (Knapp, 1995), can be obtained by contacting the author at Sandia National Laboratories.

Areal Density

To find the areal density, Equation 4 can be used. To use this equation, however, the detector efficiency η, which is a function of the backscattered ion energy, must be determined empirically (Weller, 1993). For the HIBS system at Sandia, a standard target consisting of three elements, Fe, Ag, and Pt, evaporated onto a clean silicon substrate is used to determine η. The concentration of each element is approximately one monolayer so that standard RBS can be used to calibrate the standard. Using Equation 4, it follows that the areal density of an unknown element i can be found from

$$Y_i = \eta(KE_0 - \Delta E_f)Q(N_i t)\Omega_D \sigma_i(E_0, \theta) \tag{7}$$

and for a known standard,

$$Y_s = \eta(KE_0 - \Delta E_f)Q(N_s t)\Omega_D \sigma_s(E_0, \theta) \tag{8}$$

so that, solved simultaneously, the two equations give

$$N_i t = \frac{Y_i}{Q\sigma_i(E_0, \theta)} f \tag{9}$$

where f is the factor $1/(a + bE + cE^2)$. The coefficients of this second-order polynomial equation are determined from the known areal densities (from RBS measurements) and HIBS yields for Fe, Ag, and Pt measured on the standard by substituting them into the equation

$$N_n t = \frac{Y_n}{Q\sigma_n(E_0, \theta)}(a + bE_n + cE_n^2) \tag{10}$$

where $n = 1, 2, 3$ for Fe, Ag, and Pt, respectively. The values needed for the energy E_n of the carbon particle after scattering from the standard elements and passing through the range foil can be approximated from the equation (Doyle et al., 1989)

$$E_n = [KE_0^{(1-a)} - C_1]^{1/(1-a)} \tag{11}$$

where $n = 1, 2, 3$ and the terms K and E_0 have been described. The exponential terms $1 - a$ and $1/(1 - a)$ allow a fit of the E-versus-mass curve, and C_1 is an energy constant proportional to the thickness of the carbon foil. Once the unknown coefficients, exponential, and constant terms from Equations 10 and 11 are found, subsequent areal densities for other elements are easily determined.

Background and Sensitivity

From the previous discussion on HIBS sensitivity, it follows that the sensitivity is found by substituting $2\sqrt{\text{Bkg}}$ for the yield in Equation 9. The value of the background is found by performing a fit to the spectral background using the equation

$$y = de[-a(x - c_2)]^m + be[-a(x - c_1)]^m \tag{12}$$

The variable d is found from the average height of the spectral background in the high-velocity (fast-time) part of the spectrum and b from the average height in the slow-time portion of the spectrum. A best fit to the background is obtained in the software by the user inputting a value for m and setting four markers at strategic locations on the spectral background to define the variables d and b as well as the exponential terms involving c_1 and c_2.

Mass

When performing the areal density calibration, as discussed above, the flight times for a carbon ion backscattered from the standard masses Fe, Ag, and Pt can also be determined by using the beam energy E_n after passing through the foil, calculated from Equation 11 and using $V_n = (L/\tau_n)$. The velocity is found from

$$V_n = \left(\frac{2E_n(\text{keV})}{M_1(\text{amu})}k\right)^{1/2} \tag{13}$$

where E_n and M_1 were previously described and $k = 9.648 \times 10^{14}$ for units of centimeters per second, $L = 13$ cm, and n is 1, 2, 3 for Fe, Ag, and Pt, respectively. Knowing the time-to-amplitude converter time range, the number of channels in a spectrum, and the calculated flight times from the standard masses allows calibration of the time/channel. The peak channel position for each standard element is found by calculating the peak median (counts on either side of the peak are equal). The calibration is then used for subsequent time-to-mass conversions.

SAMPLE PREPARATION

The HIBS measurements require no sample preparation; however, special precautions should be taken to prevent any extraneous contamination. Unpacking and insertion of samples for HIBS analyses into the chamber loadlock should be done in at least a class 10 environment. The loadlock is required to prevent damaging the thin foils used by the HIBS detectors. A small footprint class 10 minienvironment is ideal for space-limited laboratories and only requires wearing clean room gloves and a laboratory coat. Samples or wafers should only be handled with low-Z tweezers or a wafer wand.

PROBLEMS

There are four main areas where the potential for misapplication or interpretive errors might occur when using the

HIBS technique. First, HIBS is an ultratrace element surface analytical technique, with the depth of analysis being ~10 nm. Use for analysis of samples with impurities extending well below the surface is difficult. Other techniques more suited for trace element depth profiling, such as SIMS, are discussed under Competitive, Complementary, and Alternative Techniques, above. Second, HIBS mass resolution is relatively poor and mass dependent, as compared to TXRF and other techniques, with HIBS mass resolution ranging from ~2 amu for Fe to ~20 amu for Pb. This can lead to interpretive errors, for example, in assigning a mass to a single high-Z peak or overlapping peaks that are close in mass, as discussed under Sensitivity and Mass Resolution, above. Also, as with other techniques, high concentrations of one element can interfere in the detection of other elements. Third, sputtering and ion beam-enhanced diffusion can be potential problems when measuring surface comtaminants in a metal-silicon system because erosion and diffusion can cause a decrease in the backscattering signal and reduced sensitivity. However, these effects do not appear to be a significant problem for the HIBS technique (Pedersen et al., 1996). If it is suspected that sputtering or beam-enhanced diffusion might be a problem, the areal density as a function of beam flux can be measured to determine if the problem exists and, if it does, extract the starting areal density. Fourth, background in the spectra is a major limit to the sensitivity that can be achieved with the HIBS technique, which ranges from $\sim 6 \times 10^9$ atoms/cm^2 for Fe to $\sim 3 \times 10^8$ atoms/cm^2 for Au when measured on very pure substrates. Careful design eliminates many factors contributing to this background, as discussed under Sensitivity and Mass Resolution, above. However, multiple scattering of the ion beam in the sample and forward scattering of H on the sample surface cannot be entirely eliminated. Multiple scattering is greatly reduced by channeling of the analysis beam, leaving H scattered by the ion beam into the detectors as a problem. This is discussed in more detail under Sensitivity and Mass Resolution and is shown in the spectrum of Figure 4B. The analysis software used for HIBS, discussed above (see Data Analysis and Initial Interpretation), was written to minimize errors introduced by background.

ACKNOWLEDGMENTS

The authors would like to thank B. L. Doyle for initiating the HIBS concept and providing technical oversight during the research and development phases of HIBS. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract no. DE-AC04-94AL85000.

LITERATURE CITED

Andersen, H. H., Besenbacher, F., Loftager, P., and Möller, W. 1980. Large-angle scattering of light ions in the weakly screened Rutherford region. *Phys. Rev. A* 21:1891–1901.

Anz, S. J., Felter, T. E., Daley, R. S., Roberts, M. L., Williams, R. S., and Hess, B. V. 1995. Recipes for high resolution time-of-flight detectors. Sandia Report, SAND94–8251, Albuquerque, N.M.

Banks, J. C., Doyle, B. L., Knapp, J. A., Werho, D., Gregory, R. B., Anthony, M., Hurd, T. Q., and Diebold, A. C. 1998. Using heavy ion backscattering spectrometry (HIBS) to solve integrated circuit manufacturing problems. *Nucl. Instrum. Methods B* 136/138:1223–1228.

Besenbacher, F., Andersen, J. U., and Bonderup, E. 1980. Straggling in energy loss of energetic hydrogen and helium ions. *Nucl. Instrum. Methods* 168:1–15.

Bonderup, E. and Hvelplund, P. 1971. Stopping power and energy straggling for swift protons. *Phys. Rev. A* 4:562.

Chu, W. K. 1976. Calculation of energy straggling for protons and helium ions. *Phys. Rev. A* 13:2057.

Chevarier, A., Chevarier, N., and Chiodelli, S., 1981. A high resolution spectrometer used in MeV heavy ion backscattering analysis. *Nucl. Instrum. Methods* 189:525–531.

Currie, L. A. 1960. *Anal. Chem.* 40:586.

Darwin, C. G. 1914. Collision of α particles with light atoms. *Philos. Mag.* 27:499–506.

Döbeli, M., Haubert, P. C., Livi, R. P., Spicklemire, S. J., Weathers, D. L., and Trombrello, T. A. 1991. A time-of-flight detector for heavy ion RBS. *Nucl. Instrum. Methods* B56/57:764.

Doyle, B. L., Knapp, J. A., and Buller, D. L. 1989. Heavy ion backscattering spectrometry (HIBS)—an improved technique for trace element detection. *Nucl. Instrum. Methods* B42:295–297.

Feldman, L. C. and Mayer, J. W. 1986. Fundamentals of Surface and Thin Film Analysis. Prentice-Hall, Englewood Cliffs, N.J.

Hockett, R. S., Ikeda, S., and Taniguchi, T., 1992. TXRF round robin results. *In* Proceedings of the Second International Symposium on Cleaning Technology in Semiconductor Device Manufacturing (J. Ruzyllo and R. E. Novak, eds.), pp. 324–337. Electrochemical Society, Pennington, N.J.

Knapp, J. A. 1995. HIBWIN© computer software for HIBS analysis. Sandia National Laboratories, Dept. III, Albuquerque, N. Mex.

Knapp, J. A., Banks, J. C., and Doyle, B. L. 1994a. Time-of-flight detector for heavy ion backscattering spectrometry. Sandia National Laboratories, publication SAND94–0391.

Knapp, J. A., Banks, J. C., and Doyle, B. L., 1994b. Time-of-flight heavy ion backscattering spectrometry. *Nucl. Instrum. Methods* B85:20–23.

Knapp, J. A., Brice, D. K., and Banks, J. C. 1996. Trace element sensitivity for Heavy Ion Backscattering Spectrometry. *Nucl. Instrum. Methods* B108:324–330.

Lindhard, J. and Scharff, M. 1953. Approximation method in classical scattering by screened Coulomb fields. *Mat. Fys. Medd. Dan. Vid. Selsk.* 27(15).

Mendenhall, M. H. and Weller, R. A. 1989. A time-of-flight spectrometer for medium energy ion scattering. *Nucl. Instrum. Methods* B47:193–201.

Pedersen, D., Weller, R. A., Weller, M. R., Montemayor, V. J., Banks, J. C., and Knapp, J. A. 1996. Sputtering and migration of trace quantities of transition metal atoms on silicon. *Nucl. Instrum. Methods* B117:170–174.

Semiconductor Industry Association. 1997. The National Technology Roadmap for Semiconductors: Technology Needs. SEMATECH, Inc., San Jose, Calif.

Weller, R. A. 1993. Instrumental effects on time-of-flight spectra. *Nucl. Instrum. Methods* B79:817–820.

Weller, R. A., Arps, J. H., Pedersen, D., and Mendenhall, M. H. 1994. A model of the intrinsic efficiency of a time-of-flight spectrometer for keV ions. *Nucl. Instrum. Methods* A353:579–582.

Werho, D., Gregory, R. B., Schauer, S., Liu, X., Carney, G., Banks, J. C., Knapp, J. A., Doyle, B. L., and Diebold, A. C. 1997. Calibration of reference materials for total-reflection X-ray fluorescence analysis by heavy ion backscattering spectrometry. *Spectrochim. Acta B* 52:881–886.

Ziegler, J. F. and Biersack, J. P. 1995. Transport of Ions in Matter (version TRIM-95) software. IBM-Research, Yorktown, N.Y.

Ziegler, J. F. and Manoyan, J. M., 1988. The stopping ions in compounds. *Nucl. Instrum. Methods B* 35:215.

KEY REFERENCES

Chu, W. K., Mayer, J. W., and Nicolet, M-A. 1978. Backscattering Spectrometry. Academic Press, New York.

Definitive reference for traditional MeV RBS using surface barrier detectors, with the fundamentals applicable to HIBS.

Feldman and Mayer, 1986. See above.

Excellent introductory text for ion and electron beam techniques used for surface and thin film analyses.

Tesmer, J. R., and Nastasi, M. (eds.). 1995. Handbook of Modern Ion Beam Materials Analysis. Materials Research Society, Pittsburgh, Pa.

Provides an invaluable overview of ion beam analysis fundamentals and techniques, including sections on traditional RBS and medium-energy ion beam analysis, from which the HIBS technique is derived.

JAMES C. BANKS
JAMES A. KNAPP
Sandia National Laboratories
Albuquerque, New Mexico

This page intentionally left blank

NEUTRON TECHNIQUES

INTRODUCTION

This part and its supplements explore the wide range of applications of elastic and inelastic neutron scattering to the study of materials. Neutron scattering provides information complementary to several other techniques found in this volume. Perhaps most closely related are the x-ray scattering methods described in X-RAY TECHNIQUES. The utility of both x-ray and neutron scattering methods in investigations of atomic scale structure arises from the close match of the wavelength of these probes to typical interatomic distances (a few Angstroms). The differences between these two techniques bear some further discussion.

The absorption of neutrons by most elemental species is generally quite weak, while the x-ray absorption cross-sections, at typical energies (around 8 keV), are much larger. Therefore, neutron scattering measurements probe bulk properties of the system while x-rays generally sample only the first few microns of the bulk. The principal x-ray scattering interaction in materials involves the atomic electrons, so the scattering power for x rays is proportional to the electron density and, therefore, scales with the atomic number, Z, of elements in the sample. Neutron scattering lengths, on the other hand, do not exhibit any simple scaling relationship with Z and can vary significantly between neighboring elements in the periodic table. For this reason, neutron and x-ray diffraction measurements are often combined to provide elemental contrast in complicated systems.

The principal neutron scattering interactions involve both the nuclei of constituent elements and the magnetic moment of the outer electrons. Indeed, the cross-section for scattering from magnetic electrons is of the same order as scattering from the nuclei, so this technique is of great utility in studying magnetic structures of magnetic materials. Indeed, neutron scattering is generally the probe of choice for the determination of magnetic structure and microscopic investigations of magnetic properties of materials.

Finally, thermal neutron energies are typically on the order of a few meV to tens of meV, an energy scale that is comparable to many important elementary excitations in solids. Inelastic neutron scattering measures the gain or loss in energy of the scattered neutron as compared to the incident neutron energy. The energy loss or gain corresponds to energy transferred to or from the system in the form of excitations. Therefore inelastic neutron scattering has become a critical probe of elementary excitations including phonon and magnon dispersion in solids.

Facilities for neutron scattering are found at both nuclear reactors and, more recently, at accelerator-based spallation neutron sources. The steady flux at thermal energies at reactor sources is well-suited to instruments such as triple-axis spectrometers, which can probe elementary excitations at a particular point in reciprocal space. Spallation sources are, by design, best suited to time-of-flight studies of inelastic scattering because of the inherent time structure of the source.

ALAN I. GOLDMAN

NEUTRON POWDER DIFFRACTION

INTRODUCTION

Historical Aspects of the Method

The first powder diffraction studies on simple materials such as iron metal using x rays were done independently by Debye and Scherrer (1916) in Germany and by Hull (1917) in the United States. For a long time x-ray powder diffraction was primarily used for qualitative purposes such as phase identification and the assessment of crystallinity. The earliest structure determinations using only powder diffraction data was demonstrated when Zachariasen (1949) solved the structures of α- and β-UF_5 by an intuitive trial-and-error approach. An important subsequent step was taken when Zachariasen and Ellinger (1963) solved the monoclinic structure of β-plutonium using manual direct-method procedures (see X-RAY POWDER DIFFRACTION). The use of neutrons for such powder diffraction work has advanced significantly over the past 50 years.

The first neutron powder diffractometer was built at Argonne National Laboratory in 1945 (Zinn, 1947). The first generation of neutron powder diffractometers were subsequently built almost simultaneously by Wollan and Shull (1948) at Oakridge National Laboratory, Hurst and co-workers (1950) at Chalk River, Canada, and Bacon et al. (1950) at Harwell, England. They were the prototype of the so-called constant-wavelength angle-dispersive two-axis diffractometer whose basic features have not changed over the last 50 years. What has changed and is largely responsible for the rapid development of neutron powder diffraction was the increased flux of neutrons made available by the controlled fission processes of uranium. The highest neutron fluxes at a nuclear research reactor source can be found at the Institute Laue-Langevin (ILL) in Grenoble, France, with its 60-MW high-flux reactor (HFR). High-flux neutron sources allow the use of tighter neutron beam collimation and larger "take-off" angles, thereby increasing the resolution of diffraction experiments by reducing the line widths of Bragg reflections. Today, the required mechanical precision to move a massive detector bank is routinely somewhere between 0.01° and 0.02° in 2θ. The size and type of the detector also have undergone dramatic changes. Instead of the BF_3 counters used in the early days, today either position-sensitive detectors based on microstrip technology (Convert et al., 1997) or multicounter arrays typically with up to 64 individual ^3He counters are state-of-the-art. And last but not least, improvements in large,vertically focusing monochromators

with a reproducible aniso-tropic mosaic microstructure yielding well-defined spectral peak shapes have allowed the potential of high-resolution neutron powder diffraction to be fully developed (Axe et al., 1994).

Investigations of the magnetic susceptibility of materials such as MnO led to the concept of antiferromagnetism proposed by Néel (1932). The first experimental proof of this concept was provided by the powder neutron scattering experiments of Shull and Smart (1949). They demonstrated that, by cooling below the transition temperature, new Bragg diffraction peaks ("magnetic peaks") appeared that arise because the magnetic unit cell is twice the size of the "chemical" unit cell.

The seminal achievement of Rietveld in the late 1960s opened the door for refinements of neutron and x-ray powder diffraction patterns of complex structures with up to 50 atoms in the asymmetric unit cell (Rietveld, 1967, 1969). In the early 1970s Rietveld refinements were still performed predominantly to refine neutron powder diffraction data due to the simpler peak shape of the Bragg reflections. More complex peak shapes were developed subsequently for x-ray and especially synchrotron powder diffraction data. Since then, the attempt to solve more and more complex structures from powder diffraction data alone (*ab initio*) by using and adapting the tools developed for single-crystal crystallography was pursued by a steadily growing community of powder diffraction users. Today, structure solution and refinement are no longer the exclusive domain of single-crystal diffraction. The availability of highly collimated monochromatic neutron and x-ray beams, the development of the neutron time-of-flight powder diffraction technique used at neutron spallation sources, and the advances in instrumentation and computational methods make it possible to tackle the problems of *ab initio* structure determination in an almost routine and systematic manner (Cheetham, 1986; Cheetham and Wilkinson, 1992; see KINEMATIC DIFFRACTION OF X RAYS and DYNAMICAL DIFFRACTION). In fact, powder diffraction is the only tool capable of obtaining crystallographic information from samples such as catalytically active zeolites and low- and high-temperature phases where single crystals are simply not available. At the same time, information concerning the mesoscopic properties of materials such as texture, particle size, and stacking faults can also be obtained. The continuing efforts made in unraveling the structural details of superconducting oxides and the recently discovered oxides exhibiting colossal magnetoresistance have demonstrated the indisputable usefulness of high-resolution neutron powder diffraction. Modern powder diffraction using neutrons, x rays, and synchrotron x-ray radiation has developed from a qualitative method some 20 years ago to a quantitative method now used to detect new phases and determine their atomic, magnetic, and mesoscopic structure as well as their volume fractions if they are present in a mixture.

Neutron powder diffraction is a complementary technique to x-ray powder diffraction and electron diffraction. The greater penetration depth of neutrons, the fact that the neutron-nucleus interaction is a point scattering process implying no variation of the nuclear scattering length with scattering angle, the independence of the scattering cross-section from the number of electrons (Z) of an element, and therefore the stronger interaction of neutrons with "light" elements such as oxygen and hydrogen, and its isotope specificity as well as its interaction with unpaired electrons ("magnetic scattering") make neutrons a unique and indispensable probe for structural condensed matter physics and chemistry. However, x-ray powder diffraction and, in particular, synchrotron x-ray powder diffraction can investigate samples many thousand times smaller and have an intrinsic resolution at least one order of magnitude better than the best neutron powder diffractometer. A major drawback for neutron scattering is that it can presently only be done at large facilities (research reactors and spallation sources), whereas laboratory-based x-ray scattering equipment provides the same flux much more conveniently. Current governmental policies will not permit the construction or upgrade of present-day research reactor sources in the United States. The future of U.S. neutron scattering therefore relies solely on accelerator-based spallation sources. Electron diffraction is a very powerful tool to obtain both real and reciprocal space images of minute amounts of powders. In combination with other supplementary techniques such as solid-state nuclear magnetic resonance (NMR), which allows, e.g., the determination of Wyckoff multiplicities and interatomic distances by exploiting the nuclear Overhauser effect (NOE), physicists, chemists, and materials scientists have a powerful arsenal to elucidate condensed matter structures.

PRINCIPLES OF THE METHOD

The Neutron—A Different Kind of Probe

Independent of the specific interaction probes such as electrons, x rays, neutrons, or positrons have with matter, there is a general formalism for scattering phenomena upon which we will rely in the following. We distinguish two types of scattering. In elastic scattering, the probing particle is deflected without energy loss or gain. In inelastic scattering, the probing particle loses or gains energy. This energy gain or loss is called energy transfer. In both cases the probe is scattered by an angle 2θ and the scattering event is defined by the scattering vector $\mathbf{Q} = \mathbf{K}' - \mathbf{K}$ (Bragg's law), where $\mathbf{K}$ and $\mathbf{K}'$ are the wave vectors before and after diffraction. In the elastic case, $\mathbf{Q} = 4\pi \sin \theta/\lambda$, with λ being the wavelength. The scattering vector $\mathbf{Q}$ multiplied by $h/2\pi$ is called the momentum transfer, where h is Planck's constant. If the probing particles interact with the scattering centers in such a way that the scattered waves have a phase relationship with each other, the case is referred to as coherent scattering and interference is possible between scattered waves with different amplitudes and phases. Diffraction or Bragg scattering is the simplest form of coherent scattering. If the probes interact in an independent, random-phase-related manner among the different scattering centers, the case is referred to as incoherent scattering. No interference is possible in the incoherent scattering case. The intensities rather than

the amplitudes originating from the different scattering centers only add up. This may simply increase the background of a diffraction experiment or be used to provide information about a specific scattering species at different times and positions.

The neutron is an invaluable probe for condensed matter research. When neutrons interact elastically with matter, there are important differences from x rays.

Neutrons do not primarily interact with the electrons but with the nuclei of atoms. Due to the short range of nuclear forces, the size of these scattering centers are on the order of 10^5 times smaller than the distances between them. The nucleus is effectively a point scatterer, and the neutron is therefore scattered isotropicially. The interaction potential is a delta function, called Fermi's pseudopotential, and can be expressed as

$$V(r) = \frac{2\pi h^2}{m} b\delta(r) \tag{1}$$

h being Planck's constant and m the mass of the neutron. The Fermi length or scattering length b has the dimension of a length and is often given in femtometers. It can be expressed using the Breit-Wigner (1936) equation

$$b = R - \frac{\Gamma}{2kE_r} \tag{2}$$

where R is the nuclear radius, k the Boltzman constant, E_r a resonant energy for the neutron-nucleus system, and Γ the width of the nuclear energy level.

In contrast, x rays are not scattered isotropically since the size of the "electron cloud" is comparable to the wavelengths used to probe them. As a result, one observes a decrease of the atomic form factor and therefore of the Bragg intensities at high scattering angles. In the case of neutrons, however, the scattering length b is constant at all scattering angles. The interaction of the neutron with the nucleus is weak, especially when compared to electrons involved in very strong electrostatic potential interactions or x rays interacting via their electromagnetic radiation field with electrons. The energy of a 1-Å-wavelength neutron is 82 meV, whereas the energy of a 1-Å x-ray photon is 12 keV. A 0.037-Å electron has an energy of 100 keV. An often-neglected experimental consequence of this weak neutron-nucleus interaction is that neutrons have a very high penetration depth into matter, typically on the order of centimeters. Neutrons are a true bulk probe. The penetration depth of x rays (typically 0.5 mm) and electrons (typically 100 Å) is low enough that one has to take into consideration the possibility of near-surface effects, especially when the sample contains high-Z atoms, Z being the atomic number. The linear absorption coefficients in Al for neutrons and x rays both with a wavelength of 1.79 Å (i.e., comparable to a lattice spacing) are 0.014 and 212 cm^{-1}, respectively. In other words, the intensity of the neutron beam is reduced by half after going through roughly 50 cm of Al, whereas for x rays, it is already halved after going through 0.003 cm. This is why, from an experimentalist's point of view, sample environments (e.g., cryostats, magnets, pressure cells) are a lot easier to engineer and use in neutron than in x-ray scattering experiments. However, even with the highest currently available neutron fluxes, one will only be able to work with monochromatic beams that have fluxes of only $\sim 10^8$ neutrons/cm^2/s impinging on the sample. This is equivalent to what an ordinary sealed x-ray tube generates. Typical fluxes at synchrotron sources are more than 6 orders of magnitude higher. As a result of this intrinsically low flux, the typical neutron powder diffraction experiment requires gram quantities of material in order to obtain data with good statistics within a reasonable amount of time. This is in marked contrast to conventional and synchrotron x-ray powder diffraction experiments where milligram quantities are sufficient.

Another important difference from x rays is that the interaction of the neutron with the nucleus does not vary systematically with the atomic number Z. Thus, in certain cases, low-Z elements such as deuterium ($Z = 1$, $b = 6.671$ fm) or oxygen ($Z = 8$, $b = 5.80$ fm) interact more strongly with neutrons than do high-Z elements such as Ce ($Z = 58$, $b = 4.84$ fm) or W ($Z = 74$, $b = 4.86$ fm). In oxides, the scattering contribution of the oxygen to the Bragg reflections is in general higher with neutrons than with x rays, especially when high-Z cations are present. This is of tremendous advantage when investigating oxide structures, where the subtleties of oxygen displacements and tilts of oxygen coordination polyhedra are structural signatures of materials with significantly different physical properties. This oxygen-sensitive scattering power combined with the fact that there is no "fall-off" of the scattering length at high scattering angles when scattering neutrons led to the enormous number of neutron powder diffraction studies performed over the last decades on oxides and in particular the high-temperature cuprate superconductors.

Furthermore, even neighboring elements can have very different scattering lengths such as Mn ($Z = 25$, $b = -3.73$ fm), Fe ($Z = 26$, $b = 9.45$ fm), and Co ($Z = 27$, $b = 2.5$ fm). Without the need to perform x-ray diffraction experiments at various wavelengths and to use anomalous dispersion to distinguish between Mn and Fe or between Fe and Co, one neutron measurement can establish a particular cation distribution. This is very important when investigating the magnetic properties in solid solutions of, e.g., $Fe_{3-x}Co_xO_4$ oxides with the spinell structure. The distribution of Fe and Co among the tetrahedral and octahedral sites of the oxide structure determines the magnetism in these compounds.

In another illustrative example, the compound $NaMnFeF_6$ can be described as an ordered distribution of Na^+, Mn^{2+}, and Fe^{3+} cations located in the hexagonal close-packed (hcp) lattice of fluorine. X rays have no problem distinguishing between Na^+ and the Mn^{2+} or Fe^{3+} sites. However, there are two possible ordering schemes for Mn and Fe in the space group P321 where there are three Wyckoff positions available for these metals: 1a, 2d, and 3f. The difference between the model where Mn is located in 3f and Fe in 1a and 2d and the model where Fe is located in 3f and Mn in 1a and 2d is striking. This remarkable difference between the two models for the cation distribution is shown in Figure 1.

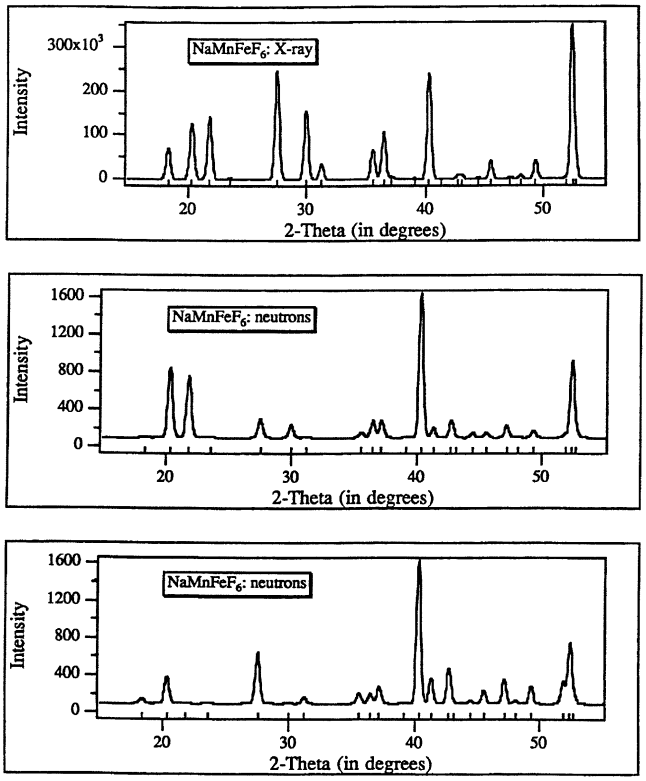


Figure 1. Powder diffraction pattern of $NaMnFeF_6$ calculated for two different ordering models. Top: X-ray diffraction pattern: cation ordering is not distinguishable since Fe^{3+} and Mn^{2+} have the same form factor far from the absorption edge. Middle: Neutron powder diffraction pattern calculated with Mn in 3*f* and Fe in 1*a* and 2*d*. Bottom: Neutron powder diffraction pattern calculated with Mn in 1*a* and 2*d* and Fe in 3*f*.

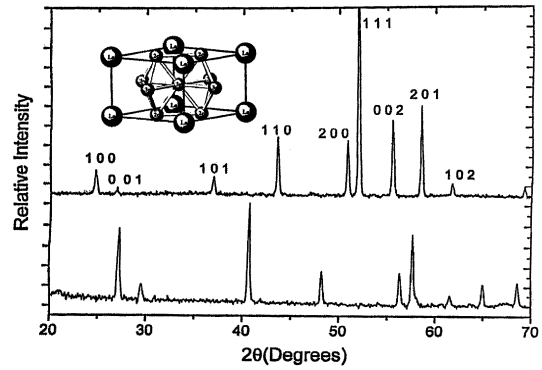


Figure 2. Neutron powder diffraction pattern of $LaNi_{3.55}Co_{0.75}Al_{0.3}Mn_{0.4}$ ($\lambda = 1.9$ Å) and $La(^{58}Ni_{0.376}\ ^{62}Ni_{.0.623})_{3.55}Co_{0.75}Al_{0.3}Mn_{0.4}$ ($\lambda = 2.1$ Å).

The neutron-nucleus interaction depends on the atomic number Z as well as on the atomic weight. Isotopes, which have the same Z but different weight, can vary substantially in their neutron scattering length. The neutron-nucleus interaction also allows for a negative scattering length b (e.g., H, Mn, ^{62}Ni, ^{48}Ti). This corresponds to a phase change of p radians during the scattering process.

Exploiting these last two properties of the neutron-nucleus interaction allows contrast variation experiments to be done by altering the scattering length b of a given atomic species by isotope substitution. The most known example is hydrogen ($b = -3.74$ fm) and deuterium ($b = 6.674$ fm). Varying the ratio of, e.g., H_2O/D_2O changes the "visibility" of water in diffraction experiments. This technique is extensively used in small-angle neutron scattering (SANS). In general, neutron powder diffraction experiments are performed with deuterated samples due to the large incoherent scattering of hydrogen. However, if only small amounts of hydrogen are present, hydrogenated samples can be used, as will be discussed below.

In certain favorable cases the net elastic scattering contribution of certain elements can be reduced to zero by varying the isotopic ratios within the sample: Ni has three isotopes with positive scattering lengths and one with a negative scattering length (^{58}Ni, $b = 14.4$ fm; ^{60}Ni, $b = 2.8$ fm; ^{61}Ni, $b = 7.60$ fm; and ^{62}Ni, $b = -8.7$ fm); Ti has two isotopes, namely, ^{48}Ti with $b = -6.08$ fm and ^{50}Ti with $b = 6.18$ fm. Certain members of intermetallic alloys with the AB_5 structure are replacing the cadmium electrode in nickel cadmium batteries due to their benign environmental impact and high-energy density. The AB_5 structure contains planes of A and B cations in the basal plane and only B cations in the midplane, as shown in the inset of Figure 2. To reduce corrosion and thus enhance the cycle life of the battery while maintaining good energy storage capacity, commercial electrodes have the stoichiometry $LaNi_{3.55}Co_{0.75}Mn_{0.4}Al_{0.3}$. However, due to the high cost of cobalt, current research efforts are focused on reducing its amount to make the battery economically competitive. It is therefore important to know the location of the cobalt within the alloy. When using metals with isotopes present in their natural abundance, the scattering is dominated by nickel, and only very little stems from cobalt since nickel has a scattering length of 10.3 fm, which is four times higher than that of cobalt ($b = 2.5$ fm). If, however, one makes an alloy using a mixture of the isotopes ^{58}Ni and ^{62}Ni in the respective proportions 37.6% to 62.3%, the picture changes dramatically; nickel no longer contributes to the Bragg scattering, and cobalt dominates the scattering from the B_5 sublattice. The tremendous changes in the diffraction pattern are shown in Figure 2, where the first nine Bragg reflections of an alloy containing metals with a natural isotopic abundance are compared to one containing the zero-scattering "isotopic alloy" $^{58}Ni_{0.376}\ ^{62}Ni_{0.623}$. Rietveld refinements of this data set together with data of samples with natural isotope abundance revealed that cobalt prefers to be located in the midplane and not in the basal plane of this structure.

Another important application of neutron scattering is possible because neutrons have a magnetic moment of $1.9132\mu_n$ that can interact via dipole-dipole interactions with the magnetic moments of unpaired electrons. The magnitudes of magnetic and nuclear scattering are of the same order of magnitude, but now the unpaired electron density can no longer be approximated as a point scatterer, as is done for the nucleus. Magnetic scattering of neutrons is not isotropic, and angle-dependent form factors similar

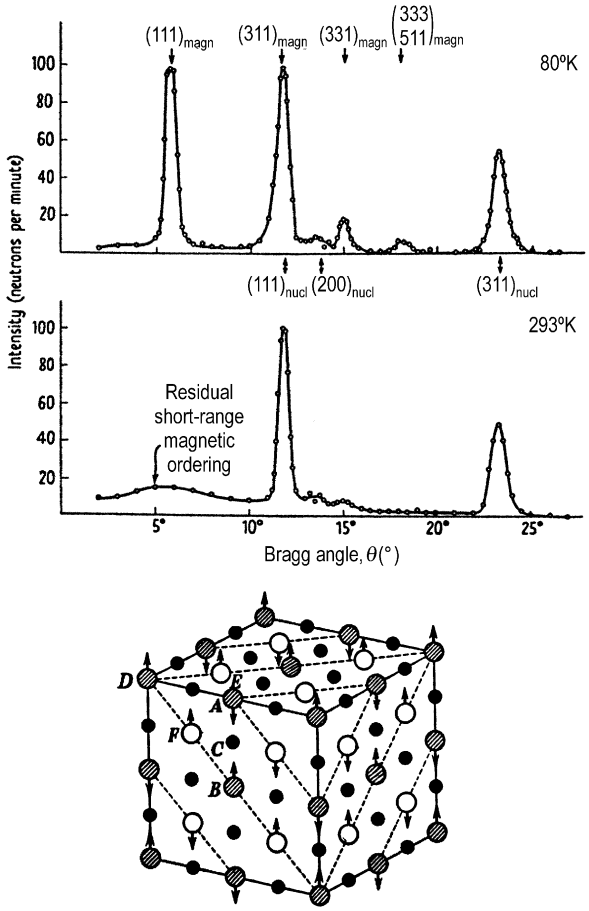


Figure 3. Neutron powder diffraction pattern of MnO at 80 and 293 K. The low-temperature pattern reveals extra antiferromagnetic reflections that can be indexed within a magnetic unit cell twice the size of the chemical unit cell. The magnetic unit cell below the Curie point is shown as an inset.

to the ones in x-ray scattering are observed (see MAGNETIC NEUTRON SCATTERING).

As mentioned above, the theory of antiferromagnetism was developed by Néel (1932), but it took 17 years until neutron powder diffraction provided its first experimental proof. Shull and Smart (1949) measured the neutron powder diffraction pattern of MnO at 80 and 293 K, which is below and above its antiferromagnetic phase transition at 120 K. As shown in Figure 3, the pattern at 80 K revealed the appearance of additional diffraction peaks, which could only be accounted for by a doubling of the "chemical" cell of MnO. The first reflection is thus the 111 of the "magnetic" unit cell. Subsequent experiments by Shull et al. (1951) revealed that the intensity of this peak decreases and becomes zero as one approaches the phase transition. This was the first observation of magnetic Bragg scattering, and the observed magnetic intensities were compared with model calculations provided by Néel (1948). The model has a magnetic unit cell made up of eight chemical unit cells. These four sublattices are arranged so that they have sheets of atoms with their spin pointing in the same direction parallel to (111) planes. These sheets of ferromagnetically coupled spins are coupled antiferromagnetically along (111). This antiferromagntic coupling between second nearest neighbors is achieved via a superexchange mediated by the oxygen. This was postulated by Kramers (1934). Figure 3 depicts the magnetic unit cell. A model that gives equally good agreement with the intensities of the magnetic Bragg reflections is one where the magnetic moments of the neighboring atoms in the individual sublattices vary randomly from crystallite to crystallite; in other words, the sublattices are not coupled. If the sublattices are coupled, the structure is no longer cubic but rhombohedral. This was confirmed by low-temperature x-ray measurements by Tombs and Rooksby (1950)—a historical first to illustrate the valuable complementary use of x-ray and neutron scattering. In the rhombohedral symmetry, the four pairs of reflections that make up the {111} magnetic Bragg reflections are no longer equivalent. Only the (111) and $(-1-1-1)$ planes will give rise to magnetic intensities. The $(-11-1)$ plane has as many atoms with their magnetic moment pointing in one way as in the opposite and thus will contribute no magnetic intensity to the {111} reflection. The orientation of the magnetic moment parallel to the [100] axis is not arbitrary. If one calculates the intensity of the {111} magnetic Bragg reflection with the magnetic moments perpendicular to (111), no magnetic intensity results since only the perpendicular component of the magnetic moment with respect to the scattering vector contributes to magnetic Bragg scattering. If one would align the magnetic moments along the (111) sheets, an exceedingly strong magnetic intensity results. That distinction between these various models is a direct consequence of the reduction of symmetry from cubic to rhombohedral.

The neutron also has a spin angular momentum of $\frac{1}{2}$. One can separate neutrons into beams of spin-up and spin-down neutrons having equal moments of $\pm\frac{1}{2}\hbar$ (where $\hbar = h/2\pi$), respectively. This is done by, e.g., using the (111) reflection of the Heusler alloy Cu_2MnAl, whose nuclear and magnetic scattering contributions compensate each other exactly for one of the two spin states. These neutrons are referred to as polarized neutrons and can be used to distinguish between nuclear and magnetic scattering since only the perpendicular component of the atomic magnetic moment with respect to the scattering vector will contribute to magnetic scattering. Polarized neutron scattering is severely hampered and has difficulties to live up to its potential since the available present-day sources are all too weak to provide high-flux polarized neutron beams using current techniques: polarized neutron scattering is currently a truly signal-limited technique. Further details concerning this neutron diffraction technique are found in Williams (1988).

PRACTICAL ASPECTS OF THE METHOD

The Angle-Dispersive Constant-Wavelength Diffractometer

This is the most familiar experimental set-up for neutron powder diffraction experiments and resembles the standard laboratory-based x-ray powder diffractometer. In a

nuclear reactor, neutrons are created by controlled fission of ^{235}U. Neutrons released by nuclear fission have a kinetic energy of ~5 MeV. A so-called moderator made in most cases of either H_2O or D_2O surrounds the reactor core. Its purpose is to "slow down" the neutrons via inelastic neutron-proton or neutron-deuterium collisions. Using the wave-particle formalism and de Broglie's equation $\lambda = h/mv$, with v being the velocity, the energy of a neutron can be expressed as

$$E(\text{meV}) = 0.08617T = 5.227v^2 = \frac{81.81}{\lambda^2} \qquad (3)$$

where the temperature of the moderator T is in kelvins, the velocity ν is in kilometers per second, and the wavelength λ is in angstroms. If water is used ($T = 300$ K), the mean wavelength is 1.78 Å and the neutrons have an energy of 25.8 meV. This moderation process is rather imperfect and results in a broad distribution of wavelengths available for experiments. If the moderator is sufficiently thick, the neutrons will have a Maxwellian energy distribution, their average kinetic energy being $\frac{3}{2}kT$, where k is the Boltzmann constant. If one selects a monochromatic beam with a certain wavelength resolution $\Delta\lambda/\lambda$ by using a single-crystal monochromator and the scattering angle 2θ is varied by stepping the detector, then the measurement mode is referred to as angle dispersive. The reflectivity of a perfect single crystal is low due to the small angular misalignment of the mosaic blocks—the so-called mosaic spread. In neutron powder diffraction, germanium is the natural choice for monochromation, especially since it has the diamond structure in which lattice planes whose Miller indices (hkl) are all odd are systematically absent. Therefore, λ/2 contamination will not occur from Bragg scattering. However, the mosaic spread has to be increased to make monochromators that can provide a reasonable flux. Recently, a significant advancement was reported by Axe et al. (1994). Instead of "squashing" single crystals, thin wafers are plastically deformed and then reassembled into composites. The reproducible creation of a spatially homogeneous but anisotropic mosaic spread led to a well-defined neutron energy profile and symmetrical spectral peak shapes, which are crucial for high-resolution neutron powder diffraction. The angle at which the neutrons are reflected off their flight path by the monochromator toward the sample is called the take-off angle. The higher this take-off angle, the better the angular resolution $\Delta d/d$ will be. Differentiating Bragg's law, one obtains

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta d}{d} + \Delta\theta \cot\theta \qquad (4)$$

The resolution is given by

$$\frac{\Delta d}{d} = \Delta\theta \cot\theta \qquad (5)$$

with $\Delta\theta$ being the mosaic spread of the monochromator reflecting at an angle θ, the take-off angle. To obtain a high angular resolution (small $\Delta d/d$), a large take-off angle has to be chosen. However, the intensity also depends on $\Delta/\lambda\lambda$ and will thus decrease as the take-off angle is increased. This is why only in high-flux reactors can the gain in resolution be afforded without paying the penalty of unreasonably small fluxes at the samples (resulting in prohibitive long counting times) in order to achieve good signal-to-noise ratios. The resolution of an angle-dispersive set-up depends on the narrow band $\Delta/\lambda\lambda$ selected by the monochromator and is further enhanced by using Soller collimators before the monochromator and detector and optionally between the monochromator and sample. Soller collimators (Soller, 1924) consist of a number of thin sheets bound together to form a series of long and narrow channels. These sheets are painted with a neutron-absorbing paint containing gadolinium to absorb diverging neutrons. The resolution of an angle-dispersive neutron powder diffractometer is further discussed in Appendix A.

Time-of-Flight Diffractometers

For didactic reasons, most arguments and examples that follow will be made using the constant-wavelength or angle-dispersive diffractometers found at research reactors. However, there is another type of neutron source, the spallation source and associated with it a second type of diffractometer, the time-of-flight diffractometer. The spallation source relies on a process whereby highly accelerated particles such as protons with energies of ~800-MeV impact on a target made of uranium or lead. Thereby neutrons are "kicked out" of the nuclei. The experimental requirements for diffraction at a spallation source are different. The time structure of the neutron beam created by the proton beam on the target is exploited in these measurements. In the time-of-flight experiment the detector is located at a fixed value of 2θ while the sample is irradiated by a pulsed, "white" neutron beam containing a very large wavelength distribution $\Delta/\lambda\lambda$. All wavelengths created in the spallation process are then moderated and impinge onto the sample without being selected by a monochromator. A coarse wavelength selection is achieved by placing choppers in the beam. These are rotating discs of neutron-absorbing material with an opening cut into it. By varying the speed, one can select neutrons with different speeds and therefore energies and wavelengths. For a given reflection, the Bragg condition $\lambda = 2d \sin\theta$ is satisfied at a specific wavelength. This wavelength can be determined by measuring the time it takes the neutron to reach the detector after hitting the sample. The resolution of a time-of-flight powder diffractometer is given by

$$\frac{\Delta Q}{Q} = \left[\left(\frac{\Delta t}{t}\right)^2 + \left(\frac{\Delta L}{L}\right)^2 + \left(\frac{\cot\theta}{\Delta\theta}\right)^2\right]^{1/2} \qquad (6)$$

where t is the time of flight of the neutrons and Δt its uncertainty mainly due to the uncertainty in the moderation time, L the flight path length and ΔL its uncertainty due to a finite target and moderator size. The angle θ is the scattering angle and $\Delta\theta$ is the angle subtended at the source by the active area of the detector. From this equa-

tion, one can deduce that the highest resolution is obtained using a long flight path L and a detector close to $\theta = 90^\circ$, referred to as the backscattering position. As an example, the high-resolution neutron powder diffractometer located at the spallation source ISIS at the Rutherford Laboratory in the United Kingdom has a 96-m-long flight path and a detector in the backscattering position. If one would use only a detector array covering a small angular range close to the backscattering position, the largest d spacing using a neutron pulse with a wavelength distribution $\lambda_1 < \lambda < \lambda_2$ that could be observed would be $\lambda_2/2$. At the $2\theta = 90^\circ$ position, a d spacing of $\lambda_2\sqrt{2}$ can be observed. One has to keep this in mind because the observation of high d spacings is of crucial importance for the determination and refinement of magnetic structures as well as for indexing unknown structures. Each time neutrons are created in a spallation process, typically every 20 ms, a complete diffraction pattern is collected. Successive pulses have to be accumulated to improve the signal-to-noise ratio. The resolution curve of a time-of-flight diffractometer is almost constant with scattering angle 2θ. Thus, minute lattice constant deviations can be resolved from high-order reflections without suffering the instrumental broadening that angle-dispersive instruments suffer beyond the minimum at the take-off angle (see Appendix A). A disadvantage of time-of-flight neutron powder diffraction is that wavelength-dependent corrections for the incident intensity, absorption, extinction, and detector efficiencies have to be made and the peak shape is highly asymmetrical. Despite these tedious corrections, sophisticated software packages exist and the time-of-flight method is used as routinely as the angle-dispersive procedure. It appears that in the future new neutron sources will probably be spallation sources rather than reactors. Time-of-flight neutron powder data have been used to determine structures from powder data alone (Cheetham, 1986). David et al. (1993) used high-resolution time-of-flight neutron powder diffraction data collected as the spallation source ISIS to study in detail the structure of the soccer ball–shaped molecule C_{60} as a function of temperature. Subtle aspects of an orientational glass transition and precursor effects of the order-disorder transition at 260 K were observed. Detailed analysis revealed that, despite the lower precision when compared to single-crystal measurements and the resulting larger errors, time-of-flight neutron powder diffraction data can be used to reproduce all the essential details found in a single-crystal synchrotron radiation experiment. This illustrates the remarkable amount of information that can be extracted using state-of-the-art neutron time-of-flight data and data analysis.

SAMPLE PREPARATION

Samples should resemble as much as possible an ideal random powder consisting of particles with sizes between 1 and 5 μm. The size distribution should be as smooth and Gaussian as possible. If larger crystallites dominate the sample, the measured intensities will have severe and nonsystematic deviations from the "true" values based on crystallography. The intensities of certain d spacings are then no longer evenly distributed on the Debye-Scherrer cones but rather are clustered in spots depending on the orientation of the bigger crystallites. There are no corrections that can be applied to these type of nonrandom samples. If too small particles are present, line-broadening effects will occur that can be corrected for (see The particle size effect). The minute amount of sample present in the beam and the highly collimated x-ray beam in a synchrotron x-ray powder diffraction experiment will exacerbate problems with nonrandomness. If the nonrandomness is systematic, as in samples with preferred orientation, it can be corrected for in Rietveld refinements. Recently, Wessels et al. (1999) used preferred orientation to solve structures from powder data alone. In neutron powder diffraction experiments, the larger size of the beam, its larger divergence, and the required gram quantities of sample to obtain useful signal-to-noise ratios within reasonable time result in a better sampling and averaging of the powder sample. Therefore, fewer problems with nonrandomness are encountered. If, however, a low-temperature structure of a compound is investigated that is a liquid or a gas at room temperature, it is advisable to cool and reheat the sample several times in order to obtain a "good" powder sample. The cooling time might become a crucial parameter since cooling too slow might result in crystallites that are too large, and therefore not all crystallite orientations will be equally represented.

DATA ANALYSIS AND INITIAL INTERPRETATION

Positions of the Bragg Reflections

For constant-wavelength diffractometers the d spacings of a crystalline sample are obtained from the measured angular position 2θ using Bragg's law:

$$\lambda = 2d \sin\theta \tag{7}$$

In a time-of-flight neutron powder diffractometer, the d spacings can be obtained by converting the time-of-flight t via

$$d = \frac{ht}{2mL \sin\theta} \tag{8}$$

where the detectors are located so that $L \sin\theta$ is a constant and m is the mass of the neutron.

The obtained d spacings allow the determination of the lattice constants with an accuracy of a few hundredths of angstroms in the case of neutrons. This accuracy is crucial when attempting to use lattice constants to determine phases in mixtures, index unknown phases, or determine thermal expansion coefficients and residual stress in materials. Accurate d spacings can only be obtained by carefully aligning the diffractometer, thus minimizing systematic errors (zero shift of the detector in 2θ), and using the proper peak shape when fitting the diffraction peaks. An adequate correction to the asymmetry due to axial divergence is especially important (see Peak Asymmetry Due to Axial Divergence). Figure 4 shows a diffraction pattern

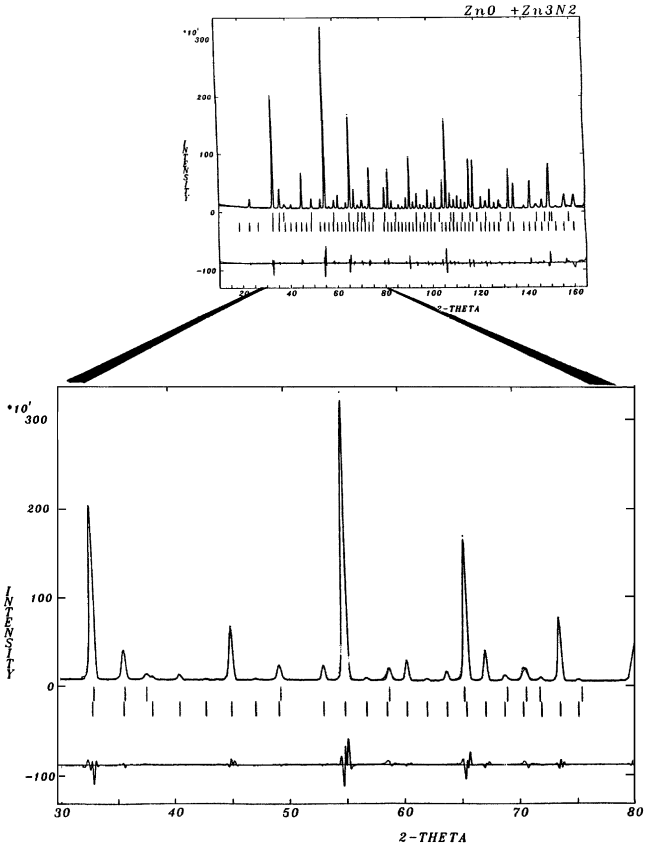


Figure 4. Neutron powder diffraction pattern of Zn_3N_2 containing ZnO as an impurity. The first row of tick marks below the pattern indicates the position of the Bragg reflections stemming from Zn_3N_2 [$a = b = 3.255(1)$ Å, $c = 5.2112(1)$ Å, space group $P63mc$], the second from cubic ZnO [$a = 9.7839(2)$ Å, space group $I\alpha3$].

of a sample of Zn_3N_2 that contains ZnO as an impurity phase. It is apparent that due to the superposition of the Bragg reflections of these two phases, only a high angular resolution allows the separation of these two phases.

The Peak Shape

The diffraction profile function or peak shape describes the intensity distribution about the Bragg position 2θ. Ultimately, the width of Bragg reflections is determined by the Darwin width. For an ideal crystal, Darwin (1914) found that the width of a reflection depends only on the structure factor and the number of interfering d spacings. Its finite width is due to the weakening of the primary beam inside the crystal due to multiple Bragg reflection—also called primary extinction. In addition to the Darwin width, various effects broaden the diffraction lines. The most commonly used quantifier for the width of a reflection is the full width at half-maximum (FWHM). The experimentally observed peak shape in the case of a constant-wavelength diffractometer is the result of the convolution of two components: (1) the optical characteristics of a diffractometer such as the wavelength distribution $\Delta\lambda/\lambda$ diffracted off the monochromator at a given take-off angle θ_m and the various collimators (see Appendix A) and (2) the actual physical contributions of the sample to the profile.

Instrumental Contributions to the Peak Shape

The resolution of a constant-wavelength neutron powder diffractometer can be calculated according to a theory first developed by Caglioti et al. (1958). The most important result of this simple theory is that the neutron peak shapes are to a very good first approximation Gaussian:

$$\Omega(2\theta) = \frac{2}{\mathrm{FWHM}}\left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left(\frac{-4 \ln 2(2\theta - 2\theta_0)^2}{\mathrm{FWHM}^2}\right) \quad (9)$$

The intensity distribution $\Omega(2\theta)$ about their Bragg position $2\theta_0$ is fully described by the FWHM. This is a very good approximation for low- and medium-resolution neutron powder diffractometers, and the rapid development of the Rietveld method was largely due to this simple mathematical description of the peak shape. For high-resolution neutron powder diffraction data and especially laboratory-based and synchrotron x-ray powder diffraction data, more sophisticated peak shapes had to be developed. In general, the resolution curve of a neutron powder diffractometer that describes the angular variation of the FWHM of the Bragg reflections can be given according to Cagliotti et al. as a simple expression (see Appendix A):

$$\mathrm{FWHM}^2 = U \tan^2\theta + V \tan\theta + W \quad (10)$$

However, there are significant deviations from this simple description that are especially apparent in high-resolution neutron powder diffraction.

Peak Asymmetry Due to Axial Divergence

An adequate peak-shape function is of utmost importance in the analysis of crystalline structures when using the Rietveld profile refinement technique. This is especially true for the low-angle region of a constant-wavelength high-resolution neutron powder diffraction pattern where severe asymmetries due to axial divergence occur. The appropriate asymmetry correction in these low-angle 2θ regions might determine success or failure when using, e.g., auto-indexing routines to determine the unit-cell dimensions of unknown structures. With a nondivergent point source and a truly randomized powder, the radiation scattered at a given Bragg reflection will lie on the surface of a Debye-Scherrer cone with semiangle 2θ. The detector itself lies on the surface of a cylinder with its axis parallel to the 2θ axis of the diffractometer. The intersection of the Debye-Scherrer cone with the detector cylinder is an ellipse. The center of this ellipse is at 0° for Bragg reflections with $2\theta < 90^\circ$ and at 180° for reflections with a $2\theta > 90^\circ$. The further from 90°, the more the ellipticity will increase. As Figure 5 shows, this departure leads to a peak asymmetry because diffracted neutrons from the ends of the intercepted part of the Debye-Scherrer cone will hit the detector on the side of the peak closer to the center of the ellipse first. This is known as axial divergence. Increasing the height of the detector and the size of

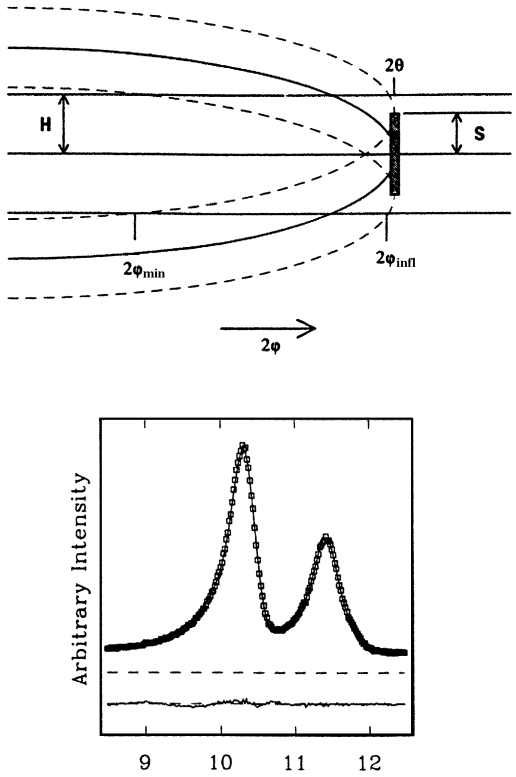


Figure 5. Band of intensity, diffracted by a sample with height $2S$, as seen by a detector with opening $2H$ and a detector angle 2φ moving in the detector cylinder. For angles below $2\varphi_{min}$, no intensity is detected. For angles between $2\varphi_{infl}$ and 2θ, the entire sample can be seen by the detector. Observed and calculated profiles are for neutron powder diffraction peaks of a zeolite sample measured at the instrument HRNPD at Brookhaven National Laboratory.

the sample will lead to greater counting rates and shorter measurement times—an important factor for high-resolution neutron powder diffraction due to the intrinsically weaker sources available in neutron scattering, as discussed earlier. However, these combined measures to increase the counting rates will also increase the asymmetry. The integrated intensities of affected reflections are biased since a larger part of the Debye-Scherrer ring is collected at lower angles. Rietveld (1969) introduced a refineable asymmetry factor providing a purely empirical correction. A widespread and commonly used empirical correction is the split–Pearson VII function described by Brown and Edmonds (1980) and Toraya (1986). Another approach is to take into account the essential optical features of the diffraction optics to correct for the asymmetry. Cooper and Sayer (1975) introduced an asymmetry correction for neutron powder diffraction based on the calculation of the resolution function in reciprocal space. However, this correction was not widely used due to the difficulties of deriving asymmetric profile functions from the source, sample, and detector contribution. Howard (1982) used the approach of Eastabrook (1952) and applied it to neutron diffraction with a point source and sample as well as a finite detector height. The correction term contains one variable parameter depending on the ratio of the detector height to the distance between sample and detector. This is currently incorporated into the Rietveld refinement programs LHPM (Hill and Howard, 1987), DBW (Wiles and Young, 1981), and older versions of GSAS (Larson and Von Dreele, 1990) and works quite well for the predominantly Gaussian peak shapes occurring in neutron powder diffraction. A more rigorous approach was advanced by van Laar and Yelon (1984). Their correction includes a finite sample size and detector height when the latter is bigger and a weight function takes into account edge effects when the detector first intercepts the Debye-Scherrer cone up to the point where it "sees" the whole sample. Finger et al. (1994) generalized the van Laar and Yelon approach, and it is currently encoded in the latest versions of GSAS. The latest generation of high-resolution neutron powder diffractometers has benefited tremendously from this correction. These diffractometers use large, vertically focusing monochromators and have detector apertures with axial divergences of several degrees. The asymmetry of low-angle peaks from materials with large unit cells such as zeolites can be described without using any free parameters.

Mesoscopic Properties Influencing the Peak Shape

When using a "real" diffractometer, the measured resolution will always be worse than the design specifications. Besides neutron optical effects stemming from nonideal collimators and monochromators that degrade the flux and resolution, there are so-called mesoscopic properties of the sample that influence the peak shape. It is important to account for these structural variations at length scales of tens to thousands of angstroms in high-resolution neutron powder diffraction data sets. We will mainly concentrate on four of them: the particle size effect, microstrain, extinction, and stacking faults.

The Particle Size Effect. The particle size effect is due to coherent diffraction emanating from finite domain sizes within the grains of a powder. Kinematical diffraction theory assumes an infinite lattice. In this case, the Bragg reflections are δ functions. The smaller the particle from which diffraction emanates, the more this δ function will be "smeared out." The broadening of all points in reciprocal space is uniform. Using Bragg's law, one can derive that

$$\frac{\Delta d}{d^2} = -\frac{\Delta\theta}{\tan\theta} = -\frac{\Delta(2\theta)\cos\theta}{\lambda} \tag{11}$$

This reveals that particle size broadening $\Delta(2\theta)$ varies with $1/\cos\theta$ in 2θ space. Analysis of size broadening is often performed using the Scherrer equation:

$$\Delta(2\theta) = \frac{K\lambda}{A\cos\theta} \tag{12}$$

where A is the thickness of the coherently diffracting domain and K is a dimensionless crystal shape constant close to unity, which is also referred to as the Scherrer constant.

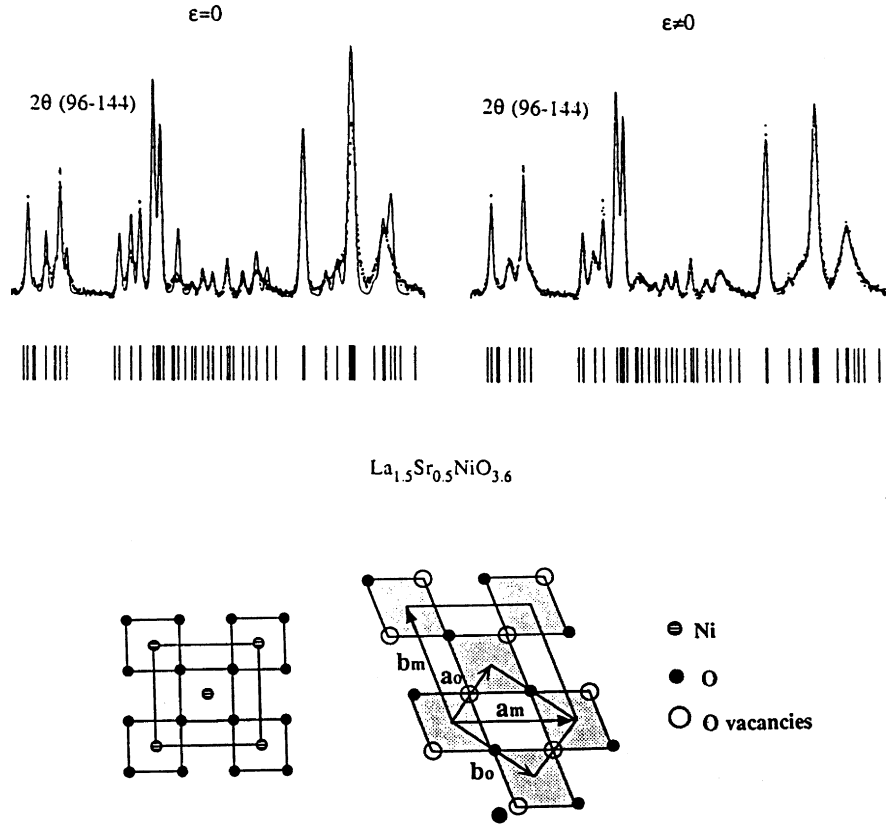


Figure 6. Portions of the neutron powder diffraction pattern of $La_{1.5}Sr_{0.5}NiO_{3.6}$ showing the effect of microstrain due to the partial ordering of the oxygen vacancies as depicted in the lower part.The fit depicted in the left part does not take microstrain into account.

Microstrain Broadening. Microstrain ε is due to a local variation $\Delta d/d$ of the average d spacing. These variations can be due to external stresses leading to anisotropic strain in the crystallites, lattice defects, or local compositional fluctuations described by a constant $\varepsilon = \Delta d/d$. In contrast to particle size broadening, the effect of strain varies in reciprocal space. Differentiating Bragg's law with respect to the diffraction angle θ, one obtains

$$\mathrm{d}\theta = -\tan\theta \frac{\mathrm{d}d_{hkl}}{d_{hkl}} \tag{13}$$

In this case, we have the compositional fluctuations as constant and not the size A, and therefore

$$\Delta\theta = -\tan\theta \frac{\Delta d}{d} \tag{14}$$

This term is no longer wavelength dependent. Therefore, if we change wavelength, the width of a line (corrected for instrumental contribution!) should not vary in the case of only microstrain broadening. The variations in d spacings $\Delta \mathrm{d}/d$ manifest themselves in a broadening of the diffraction line centered at 2θ:

$$\Delta(2\theta) = -2\varepsilon\tan\theta \tag{15}$$

These two mesoscopic properties of materials, particle size and strain, are increasingly studied due to their important consequences for the physical properties of materials such as resistivity, magnetism, and the occurrence of phase transitions. Medarde and Rodriguez-Carvajal (1997) investigated the oxygen vacancy ordering in $La_{2-x}Sr_xNiO_{4-\delta}(0 \leq x \leq 0.5)$. In La_2NiO_4, the crystallographic structure is locally orthorhombic. However, the local fluctuations are correlated in such a way that their macroscopic average is tetragonal. This is due to the stress induced by the size mismatch between NiO_2 and La_2O_2 layers. The (hhl) reflections are narrow, whereas the $(h0l)$ and $(0kl)$ reflections reveal an increasing broadening with increasing values of h and k (Rodriguez-Carvajal et al., 1991). When substituting La^{3+} by Sr^{2+}, the broadening diminishes since the size mismatch between the NiO_2 and $(La,Sr)_2O_2$ layers is reduced. Above a critical $x \sim; 0.135$, the oxygen vacancies order and give rise to an extra broadening of reflections. This microstrain created by ordered oxygen vacancies along the orthorhombic a axis corresponds to ~1.7% of its total length. Figure 6 shows two Rietveld refinements of $La_{1.5}Sr_{0.5}NiO_{3.6}$, one of them refining a model for the microstrain, the other assuming no microstrain.

Microstrain and particle size effects can be separated by their q dependency when both are present in a material. There are various techniques to extract these types of mesoscopic properties from powder diffraction data. The most commonly used is based on the Williamson-Hall plot. Combining the Scherrer equation with the equation describing microstrain broadening leads to

$$\mathrm{FWHM}_{corr}\frac{\cos\theta}{\lambda} = \frac{1}{A} + \frac{C\varepsilon\sin\theta}{\lambda} \tag{16}$$

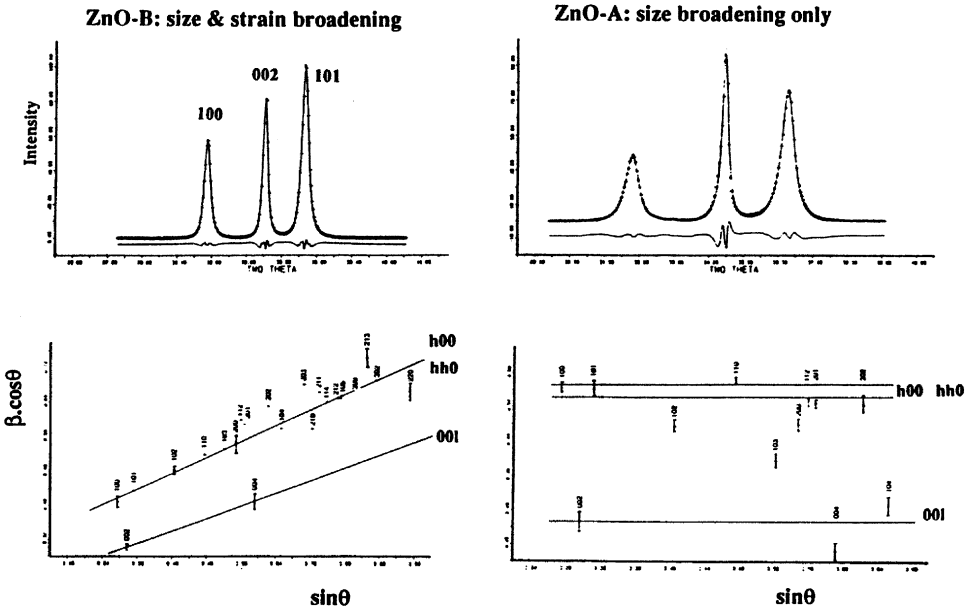


Figure 7. Williamson-Hall plots for two samples of ZnO prepared in different conditions: sample A is strain free but made of small ansiotropic crystallites. Sample B reveals both particle size and microstrain effect. (Adapted from Langford and Louër 1986.)

where $\mathrm{FWHM}_{\mathrm{corr}}$ is corrected for instrumental resolution and C varies with ε and the FWHM. The Scherrer constant K is assumed to be unity. When plotting $\mathrm{FWHM}_{\mathrm{corr}} \cos\theta$ of Bragg reflections versus $\sin\theta$, one can extract the domain size A as the y intercept and ε from the slope. Both T and ε can be anisotropic. One can therefore determine the crystalline shape and the direction of microstrain in crystallites. Figure 7 compares two samples of ZnO, one of them revealing broadening due to particle sizes, the other where both a broadening due to particle size and strain occur. The Williamson-Hall plots for both cases show how powerful this simple analysis can be.

Extinction is defined as the reduction of the ideal kinematical intensity due to depletion of the primary beam and rescattering of the diffracted beam. Sabine et al. (1988) demonstrated that in time-of-flight neutron powder diffraction primary extinction can be substantial when highly crystalline samples are measured. The extinction coefficient E is a function of the scattering angle and can be expressed as

$$E = E_{\mathrm{L}} \cos^2\theta + E_{\mathrm{B}} \sin^2\theta \tag{17}$$

$$E_{\mathrm{L}} = 1 - \frac{x}{2} + \frac{x^2}{4} - \frac{5x^3}{48} + \frac{7x^4}{192} \tag{18}$$

$$E_{\mathrm{B}} = \frac{1}{(1+x)^{1/2}} \tag{19}$$

$$x = (KN_{\mathrm{c}}\lambda FD)^2 \tag{20}$$

Here, D is a refinable parameter and K a shape factor that is unity for a cube and $\frac{3}{4}$ for a sphere of diameter D; N_{c} is the number of unit cells, F the structure factor, and λ the wavelength.

Stacking Faults. Due to the high angular resolutions achievable with modern synchrotron and neutron powder diffractometers, stacking faults are more and more noticeable. The diffracted intensity from a sample containing stacking disorder is the weighted incoherent sum of diffraction pattern emanating from each crystallite orientation and defect arrangement. Any single stacking fault can be included into the calculations in a rather straightforward fashion, but the number of permutations of stacking faults increases the computational effort substantially. Stacking faults were originally discovered in cold-worked metals. They represent disorder in the stacking sequence of layers along a particular direction. In the face-centered cubic (fcc) structure layers of atoms can be seen stacking along the [111] direction, whereas in the hcp structure the layers stack along [001]. The possible atomic layer positions in the hcp lattice is ABABAB... along [001], and in the fcc lattice it is ABCABCABCABC... along [111]. A so-called deformation fault occurs in the fcc structure when the sequence ABCABCABC is altered to, e.g., ABCA**C**ABCA This deformation fault corresponds to a shift of all layers after the fault (bold **C**) to the right. A twin fault in the fcc structure is represented by a sequence change from ABCABCABC to ABCA**B**ACBA. The stacking sequence is reversed to the left and right of the fault. Twin faults in fcc structures along [111] lead to diffraction line broadening, whereas deformation faults lead to line broadening as well as line positional shifts. This can be understood in a simple qualitative picture: if one writes down three possible deformation fault sequences and compares them to the original sequence, one observes that all A, B, and C layers following the fault (in bold) are now closer to the original A^* layer than in the unfaulted fcc sequence:

A^* B C A **C** A
A^* B C A B **A**
A^* B C **B** C A
A^* B C A B C

This change of the translational periodicity leads to a broadening due to local variations and a peak shift. If one does the same for possible twin-fault structures, one observes that the A, B, and C layers following the fault

(in bold) are shifted forward or backward or remain in the same sequence position as the unfaulted fcc sequence, leading only to an increased dispersion along c resulting in peak broadening:

A* B C A **B** A
A* B C **A** C B
A* B C A B **C**
A* B C A B C

An exemplary study by Berliner and Werner (1986) showed that the structure of lithium at 20 K has a so-called 9R structure with the nine-layer repeating sequence ABCBCACAB, which is highly faulted below 20 K. Their stacking fault analysis revealed deformation stacking-fault probabilities of 7%.

Wilson (1942, 1943) developed a difference equation method, where the probability of the occurrence of a given layer is related to the probability that it occurred in previous layers. Hendricks and Teller (1942) generalized this approach by using correlation probability matrices including effects of the nearest-neighbor layer-layer correlations and different stacking vectors.

Cowley (1976a,b, 1981) came up with another approach by analyzing single-layer terms in the Patterson function and separating intensity contributions into "no-fault" terms, which were summed as geometric series, and "fault" terms, which were summed explicitly. This led to quicker convergence in low-fault cases. Recently, Treacy et al. (1991) implemented a general recursion algorithm to calculate the diffraction pattern with coherent planar faults. Their DIFFaX program is rather easy to apply to even complex structures such as zeolite beta (Newsam et al., 1988) and intergrowths.

Delmas and Tessier (1997) correlated the broadening observed in $Ni(OH)_2$ to stacking faults. The occurrence of fcc domains in the hcp of oxygen can be related to the electrochemical behavior of this material used as electrodes in batteries and explains unexpected Raman bands. The work by Roessli et al. (1993) emphasizes the fact that when diffracting with neutrons magnetic stacking faults can also occur. In their low-temperature measurements of $HoBa_2Cu_3O_7$, the magnetic superstructure reflections show remarkably different widths and shapes: asymmetric, e.g., $0, \frac{1}{2}, \frac{1}{2}$; broad but symmetric, $0, \frac{1}{2}, \frac{3}{2}$; and narrow and resolution limited, $1, \frac{1}{2}, \frac{1}{2}$. They can be explained by an interplay of the q dependence of the magnetic form factor and finite magnetic correlations on the order of 30 Å along the c axis. The latter is the equivalent to magnetic stacking faults along the c axis.

The Rietveld Method

The Rietveld method makes use of the fact that the peak shapes of Bragg reflections can be described analytically and the variation of their widths (FWHM) with the scattering angle 2θ can be expressed as a convolution of optical characteristics depending on the diffractometer (take-off angle of the monochromator and collimation) and sample-related effects such as strain and particle size broadening effects, as discussed above.

This allows least-squares refinements of crystal structure parameters without explicitly extracting structure factors or integrated intensities. For details with respect to reliability factors and estimated standard deviations, see Appendices A and B. Less frequently, one finds studies relying on a two-stage approach where individual intensities are extracted (Jansen et al., 1988).

The Rietveld algorithm fits the observed diffraction pattern using as variables the optical characteristics and the structural parameters. A function

$$M = \sum w_i (y_i - y_{c,i})^2 \qquad (21)$$

is minimized, where $w_i = 1/\sigma_i^2$ is the weight assigned to an individually measured intensity y_i at step i, σ is the variance assigned to the observation y_i, and $y_{c,i}$ is the calculated intensity at step i. The calculated intensities $y_{c,i}$ are obtained by adding the contributions stemming from the contributions of the overlapping Bragg reflections to the background $y_{b,i}$:

$$y_{c,i} = y_{b,i} + \sum_i s_\phi \sum j_{\phi k} \mathrm{Lp}_{\phi k} O_{\phi k} |F_{\phi k}|^2 \Omega_{i\phi k} \qquad (22)$$

The first sum adds all the various contributions from different phases ϕ; the second sum adds all contributions from different reflections k. The scale factor S_ϕ is proportional to the volume fraction of the phase, $j_{\phi k}$ is the multiplicity of the kth reflection, $\mathrm{Lp}_{\phi k}$ is the Lorentz factor, $O_{\phi k}$ is a factor describing any mesoscopic effects (particle size, strain, preferred orientation), $F_{\phi k}$ is the nuclear and magnetic structure factor including the displacement contributions, and $\Omega_{i\phi k}$ describes the peak profile function of reflection k. A well-defined and symmetrical peak profile function that can be parameterized by a small number of variables and whose 2θ dependency can be described by Cagliotti et al. (1958) equation or similar functions is advantageous. It will result in fewer correlations between the so-called machine parameters and the structural parameters (positional coordinates and displacement parameters). A not well-defined peak shape leads to "unstable" least-squares refinements and larger errors of the refined structural parameters. The density of reflections increases tremendously with increasing 2θ, and the quality of the Rietveld fit relies on the separation of close d spacings at higher scattering angles. This separation will benefit from the highest resolution available at high scattering angles and a well-understood peak shape. Furthermore, as shown above, certain mesoscopic effects manifest themselves as deviations of the peak shape at high scattering angles.

Quantitative Phase Analysis

In the absence of severe absorption, the intensity diffracted by a crystalline phase is proportional to the amount of irradiated material. The Rietveld method can be used to extract the structures and volume fractions of individual phases when phase mixtures are present (Hill and Howard, 1987; Bish and Howard, 1988). One just has to sum over the contributions of all phases present represented by their individual scale factor proportional to the volume fraction of the phase. When measuring in

the Debye-Scherrer geometry, the scale factor for each phase is proportional to

$$a_j \propto \frac{m_j}{(ZMV_c)_j} \tag{23}$$

with M being the mass per formula unit of the phase, Z the number of formula units per unit cell, and V_c the volume of the unit cell. By constraining the sum of the weight fractions to unity, one can determine the relative weight of any component of a mixture. This is an implicit constraint since one assumes no amorphous phase to be present. Adding an internal standard to the sample allows determination of absolute weight fractions. Even in severe overlap one can determine the weight fractions of phase mixtures. The high-resolution neutron powder diffraction pattern of the Zn_3N_2 and ZnO mixture depicted in Figure 4 was refined as a two-phase mixture revealing that the sample contains 10% ZnO as an impurity phase. Any refinements that did not take into account ZnO resulted in poor fits and unreliable Zn_3N_2 parameters.

Partial Structure Determination

In many structural investigations using neutron powder diffraction, a large portion of the crystallographic structure is already known but the location of a few atoms remains elusive. When solving and refining structures using x-ray powder diffraction data, often low-Z elements cannot be located with high enough precision and accuracy. Then neutron powder diffraction becomes essential to fully determine and understand the structure. This is the case when hydrogen or lithium is present as extra framework atoms in zeolites or silicates (Paulus et al., 1990; Vogt et al., 1990). The location of organic molecules in zeolites opened up an important application for neutron powder diffraction. In such cases the known fragment of the structure is refined against the diffraction data using the Rietveld refinement technique, and a difference Fourier map based on the difference between experimentally observed and calculated structure factors of the known fragment using the calculated phases is constructed:

$$\Delta(hkl) = |F_{obs}(hkl)| - |F_{calc}(hkl)| \tag{24}$$

This map will in many cases reveal the positions of the missing atoms as peaks. Rotella et al. (1982) used this approach to locate the deuterium atoms in $DTaWO_6$. In certain cases even nondeuterated samples can be investigated: Harrison et al. (1995) measured the neutron powder diffraction pattern of $(NH_4)_2(WO_3)_3SeO_3$ and using difference Fourier techniques were able to determine the four hydrogen atoms as negative peaks (due to their negative scattering length). One hydrogen reveals larger displacement parameters than the other three. Examining the structure revealed that this was the only hydrogen not involved in hydrogen bonding. The interpretation of difference Fourier maps is not always as straightforward as described above due to noise introduced by series termination errors (Prince, 1994). An alternative to locate atoms, especially extra framework cations in zeolites, is the maximum-entropy technique (Papoular and Cox, 1995). This technique is based on Bayesian analysis and the use of entropy maximization in crystallography (Bricogne, 1991). The use in powder diffraction was pioneered by Antoniadis et al. (1990).

Ab Initio Structure Determination from Power Data

The development of high-resolution neutron powder diffractometers with their high angular 2θ resolution resulted in a dramatic reduction of peak overlap in the neutron powder diffraction pattern. Although solving structures from powder diffraction data alone is by no means as straightforward as when using single-crystal diffraction data, more and more structures of important materials are determined this way. Many times a combination of electron, x-ray, and neutron diffraction is used. The complexities of structures being solved is impressive: Morris et al. (1992) solved and refined the structure of $Ga_2(HPO_3)_3 \cdot 4H_2O$, a novel framework structure with 29 atoms in the asymmetric unit cell and 171 structural parameters by combining the use of synchrotron x-ray and neutron powder diffraction data. From high-resolution neutron powder diffraction data alone, Buttrey et al. (1994) determined the structure of the high-temperature polymorph of Bi_2MoO_6. This monoclinic structure has 36 atoms in the asymmetric unit cell and was refined using 146 structural parameters. In another structure determination using synchrotron x-ray and neutron powder diffraction data, Morris et al. (1994) solved the structure of $La_3Ti_5Al_{15}O_{37}$ with 60 atoms in the asymmetric unit and 183 structural parameters. The successive steps of such an *ab initio* structure determination are presented in the following paragraphs.

Indexing. Before intensities can be extracted, one has to determine the unit cell and space group. Using single peak-fitting routines, individual d spacings are determined. It is very important in this step that systematic errors are eliminated as much as possible. In neutron powder diffraction data, the two most serious errors that will determine success or failure when indexing are (1) the low-angle asymmetry due to axial divergence (see above) and (2) the zero-point shift of the detector. One way to determine the zero point of the detector is to mix the sample with a known standard such as silicon and thus "sacrifice" a certain amount of material for indexing purposes. Another approach is to determine the zero-point shift in a separate experiment with a known standard since the mechanical reproducibility is in general $<0.01^\circ$ with state-of-the-art equipment such as absolute encoders. If possible, one should also attempt to use larger wavelengths, if available, since the peak overlap will be further reduced in 2θ, as can be seen from Bragg's law. However, any wavelength with $\lambda/2$ components should be avoided. Any presence of even minute impurity phases will make the task of indexing the powder pattern more difficult and in many cases impossible, especially when dealing with low-symmetry structures. The pattern of the Zn_3N_2-ZnO mixture shown in Figure 4 shows the effect an impurity phase would have. Peak splitting and slight shoulders

in the low-angle 2θ region could lead to the wrong assumption that the structure actually has a lower symmetry. The significant better angular resolution of synchrotron x-ray powder diffraction data due to almost ten times smaller FWHMs can be of great help if one runs into problems with large unit cells and low-symmetry space groups. Once reliable d spacings have been extracted, several computer programs for indexing are available. They are based on various strategies ranging from trial and error, deductive determination, and semiexhaustive to exhaustive search algorithms. Certain programs work better for low-symmetry space groups, others for high-symmetry ones. An overview of these techniques is given by Shirley (1984). In general, indexing programs will give the user more than one solution each associated with a certain quality factor value. The most common quality factor used is M_{20} and is defined as

$$M_{20} = \frac{Q_{\text{calc}}^{\max}}{2N_{\text{calc}}\langle\delta\rangle} \tag{25}$$

using the first 20 peak positions where $Q = 1/(d_{hkl})^2$, N_{calc} is the number of potentially observable calculated lines up to the last indexed line, and $\langle\delta\rangle$ is the absolute mean difference in Q between experimental and calculated peak positions. As a general rule of thumb, any solution with an $M_{20} < 10$ is suspicious, whereas solutions with $M_{20} > 20$ are generally correct. Another common "quality factor" is F_N:

$$F_N = \frac{N}{\langle|\Delta(2\theta)|\rangle N_{\text{possible}}} \tag{26}$$

In this case, N_{possible} is the number of possible peak positions out to the Nth observed peak.

The most popular autoindexing programs are ITO (Visser, 1969), TREOR (Werner et al., 1985), and DIVCOL (Louër and Vargas, 1982).

The space group can be determined the "old fashioned" way by comparing the various extinction rules listed, e.g., in the International Tables (IUC, 1992). Certain program suites provide little search routines for this purpose (e.g., Byrom and Lucas, 1991).

The Extraction of Individual Structure Factors. Once the cell parameters and the space group are known, a full cell-constrained profile refinement can be performed. The individual intensities are treated as variables together with the unit cell parameters, the zero-point shift, and other "machine parameters" such as the U, V, and W coefficients (Appendix A), the asymmetry parameter, and any other parameters describing the peak shape. Cell-constrained refinements were first proposed by Pawley (1981). However, in the original procedure all peaks that occurred within a given angular range were generated from the starting cell parameters and then their intensities were refined in a least-squares procedure together with the machine parameters and the cell constants. These led to large correlations among the parameters and ill conditioning of the least-squares fit, often resulting in divergence. The use of the individual intensities in the least-squares refinement resulted in enormous normal equations and the major cause of instabilities (see Basics of Crystallographic Refinements). A computationally more stable algorithm was proposed by J.C. Taylor and first implemented by LeBail et al. (1988). The integrated intensities of the individual reflections are not used in the least-squares process but are instead calculated by iteration starting from crudely estimated intensities. However, due to this iterative process, the errors assigned to the individual intensities are no longer correct in the sense that they represent the counting statistics. These are crucial for the subsequent step in an *ab initio* structure determination. Therefore, according to Pawley, a cell-constrained refinement should be used as a last step to obtain correct errors. Furthermore, the errors of completely overlapping reflection should be assigned with the method derived by David (1987).

Structure Solution. All known standard structure solution techniques can be applied (Christensen et al., 1985, 1989; Cheetham, 1986; Cheetham and Wilkinson, 1991). Often trial-and-error attempts based on a good knowledge of crystal chemistry are successful. The two main methods are Patterson methods and direct methods. Patterson methods are based on the use of the autocorrelation function:

$$P(x,y,z) = \frac{1}{V^2}\sum\sum\sum F_{hkl}^2 \cos 2\pi(hx + ky + lz) \tag{27}$$

They are easy to calculate but often difficult to interpret since they represent interatomic vectors rather than atomic positions. If heavy scatterers are present, they will dominate the Patterson maps. Therefore, the success rate with x-ray data sets is higher than with neutron data sets, where the scattering power is in general more equally distributed among the atoms. Structures are frequently solved more easily using direct methods based on probabilistic relations between intensities and phases of the structure factors when neutron data are available. Other less-frequently used approaches to solve structures from powder diffraction only are (1) the MEDIC algorithm based on Patterson map deconvolution (Rius and Miravitles, 1988), (2) methods based on statistical mechanics such as simulated annealing using the R factor as a cost function (Semenovskaya et al., 1985), (3) image reconstruction techniques (Johnson and Davide, 1987), (4) the increasingly popular maximum-entropy approach (David, 1990, Gilmore et al., 1993), and (5) a recently very promising genetic algorithm for fitting trial structures against measured powder diffraction data (Shankland et al., 1997).

Time-Dependent Neutron Powder Diffraction

The combination of powerful neutron sources and the development of fast detectors allows diffraction experiments to be recorded within a few minutes and in some cases even in a few seconds. We can thus observe rapid changes. With the new high-intensity diffractometer D20

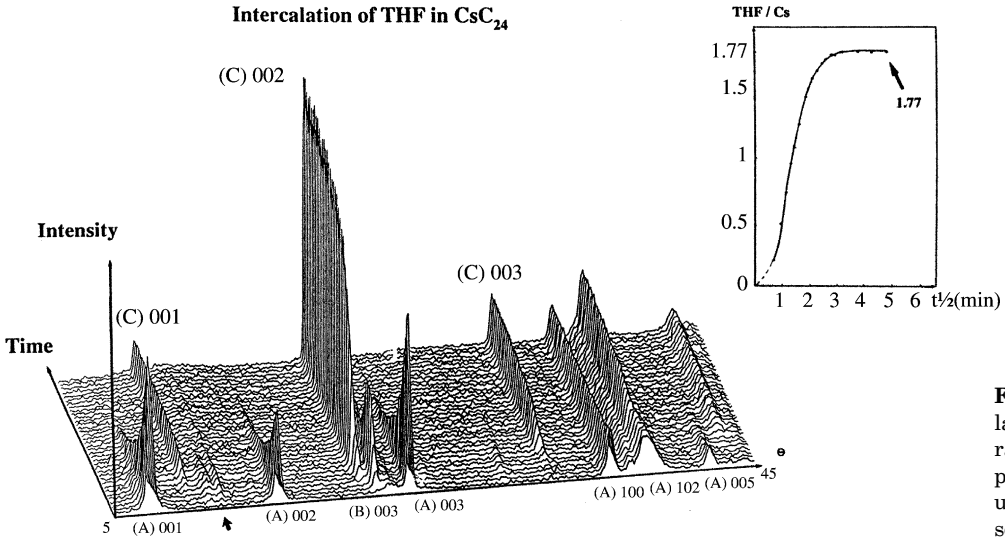


Figure 8. Evolution of the intercalation of deuterated tetrahydrofuran in CsC_{24} with time: a transient phase with a life time of a few minutes as indicated by arrow can be seen.

at the ILL (Convert et al., 1997) and the GEM project at the spallation source ISIS in the United Kingdom, neutron powder diffraction will open up the "time domain" for chemists and material scientists in an unprecedented manner. The temporal information gained during a chemical or physical process can be very important and opens up the study of chemical reactions by identifying intermediate crystalline species as well as the study of phase transformations in general. The notion of a static chemical structure "frozen in space" is being replaced by the dynamics of an evolving system. Time-dependent diffraction experiments are performed by externally perturbing a sample and then monitoring the relaxation of this sample back toward equilibrium. The time it takes to perform a single measurement will be shorter than the relaxation time of the process. The external perturbation can be temperature, pressure, magnetic, or electric field. It can be applied once or in a cyclic fashion. Two distinctions can be made with respect to the observed processes: First, in nonreversible processes time dependence is monitored by recording data sequentially. Second, in reversible processes the relaxation time of the process can be much shorter than the time it takes to record a pattern with good statistics. In such an experiment the perturbation time is divided into different time slices, and the data are recorded and accumulated over many cycles. Reversible processes in the range of 10 ms were observed using a special variable-time-delay electronics at the D20 prototype by Convert et al. (1990).

The use of position-sensitive detectors (PSDs) in thermal neutron scattering (Convert and Forsyth, 1983) in the late 1970s triggered these investigations. The low-resolution high-intensity diffractometer D1B at the ILL was increasingly being used by chemists and material scientists to identify intermediate and transient crystalline species during chemical reactions and phase transformations. Pannetier (1986, 1988) performed the groundbreaking work in this field. Besides qualitative studies, time-dependent studies allow in a quantitative way the determination of the kinetic laws governing the transformations. This is done by measuring the volume fraction of each phase as a function of time. By monitoring the normalized intensities of a few strong reflections, one can determine a time-dependent volume fraction $\alpha(t) = I(t) - I(0)/I(\infty) - I(0)$ for an appearing and $\alpha(t) = I(\infty) - I(t)/I(\infty) - I(0)$ for a disappearing phase. The temporal evolution of α can be described in very simple terms of nucleation and growth (Christian, 1975) using the so-called Avrami-Erofeev equation: $a(t) = 1 - \exp[-k_T(t)^n]$, where the exponent n is a constant related to the dimensionality of nucleation and growth [$3 \leq n \leq 4$ for a three-dimensional (3D) process] and $k_T(t)$ the kinetic constant of the reaction depending on the temperature.

Examples in this field are the pressure-induced NaCl-to-CsCl type structural transformation in RbI (Yamada et al., 1984) and the solid-gas reaction of tetrahydrofuran (THF) vapor with the second-stage graphite intercalation compound CsC_{24}. The latter reaction takes place over a few hours at room temperature. The temporal evolution of this chemical system is shown in Figure 8. Goldman et al. (1988) identified the distinct steps in this reaction: CsC_{24} vanishes rapidly as an intermediate $CsC_{24}(THF)_x$ with $x \approx 1$ evolves. This intermediate phase, indicated by the arrow in Figure 8, has a lifetime on the order of minutes until the final first-stage intermediate $CsC_{24}(THF)_{\sim 1.8}$ appears. There were no indications from the bulk properties that this structural intermediate exists.

Neutron powder diffraction studies, combined with the unique capabilities of neutrons to alter the scattering lengths by the right mixture of isotopes, was shown earlier to be a very powerful tool, especially in site-disordered alloys. In time-dependent studies, this can be used as a "marker" for the specific element during transformations, as the following example illustrates: the discovery of quasicrystalline Al-Mn alloys with long-range icosahedral symmetry (Schechtman et al., 1984) led to many intriguing questions. One of them was what happens when these quasicrystals transform to "normal" crystalline alloys? Substituting part of the Mn ($b_{coh} = -3.73$ fm) with various amounts of Fe ($b_{coh} = 9.54$ fm) allows again the formation

of a "zero-scattering" composition, namely $Mn_{0.72}Fe_{0.28}$ in the transition metal sublattice. Comparing $Al_{85}Si(Mn_{0.72}Fe_{0.28})_{14}$ with $Al_{85}SiMn_{14}$ revealed diffuse scattering at the onset of crystallization of the latter but not in the case of the zero-scattering alloy. It appears that the transformation is an order-disorder transition involving the transition metal sublattice and not the Al subnetwork (Pannetier et al., 1987).

Another increasingly popular time-dependent structural investigation is the *in situ* electrochemical experiment. Using electrochemical cells directly in the neutron beam, Latoche et al. (1992) investigated the structural changes taking place in a $LaNi_{4.5}Al_{0.5}Dx$ electrode during charge and discharge. Phases with out-of-equilibrium cell parameters were observed and related to the charge-discharge rate. In the near future more and more studies using high-intensity neutron powder diffractometers to probe the time dependency of chemical processes will become feasible due to new and powerful machines currently under construction or close to completion.

PROBLEMS

Basics of Crystallographic Refinements

An increasing sophistication in data collection and correction procedures for absorption, extinction, preferred orientation, strain, particle size, and asymmetry due to axial divergence has substantially enhanced the quality of data available nowadays from modern neutron powder diffractometers. However, the misuse of Rietveld refinement programs, perhaps through failure to grasp the principles, sometimes leads to poor refinements of high-quality data. With increasing complexity (large unit cell, low symmetry) of the investigated structures, the need to use restraints and constraints in the refinement to limit parameter space and achieve convergence is becoming increasingly important. The following is an attempt to shed some light onto the mystery of modern refinement strategies.

Any structural refinement program for x-ray, neutron, or electron diffraction data obtained from gases, liquids, or powder or single crystal will provide a model for the pattern of electron or nuclear density describing the atomic structure. We have well-established and proven theories of the interaction of x rays, neutrons (mainly kinematical scattering), and electrons (dynamical scattering) with the electrons (in the case of x rays and electrons) and nuclei (in the case of neutrons) of the investigated matter. Given high-quality observations of scattering amplitudes and estimates of phases, we can compute electron or nuclear density by using a 3D Fourier summation of structure factors. Crystallography is about defining and refining these models.

Detailed references for a full mathematical background of least-squares techniques can be found in Press (1986) and Prince (1994). Here is the "poor man's version":

We have a mathematical expression for Y_c (calculated intensities) as a function of variables x_i (positional and displacement parameters) whose values we wish to determine. We have experimental values Y_0 from our diffraction experiment. Expanding Y_c as a Taylor series in x_i, we determine shifts δx zthat will be applied to the initial values of x to improve Y_c, thus minimizing the difference between Y_0 and Y_c:

$$Y_{\text{old}} + \left(\frac{\partial Y}{\partial x_1}\right)\delta x_1 + \left(\frac{\partial Y}{\partial x_2}\right)\delta x_2 \cdots = Y_{\text{new}} \tag{28}$$

By minimizing,

$$Y_0 - Y_c = \sum \left(\frac{\partial Y}{\partial x_i}\right)\delta x_i \tag{29}$$

we express all individual equations in matrices as

$$\mathbf{A}\,\delta x = \mathbf{y} \tag{30}$$

Multiplying both sides by $\mathbf{A}'$, the transpose of $\mathbf{A}$, we obtain the so-called normal equations:

$$\mathbf{H}\,\delta x = \mathbf{A}'\mathbf{y} = \mathbf{r} \tag{31}$$

The normal matrix $\mathbf{H}$ is square and made up of terms $h_{ij} = \Sigma(\partial Y_m/\partial x_i)\,(\partial Y_m/\partial x_j)$ and $\mathbf{r}$ is a vector of terms

$$r_i = \sum \left(\frac{\partial y_m}{\partial x_i}\right)(Y_{\text{obs}} - Y_{\text{calc}}) \tag{32}$$

Solving these equations for δx leads to

$$\delta x = \mathbf{H}^{-1}\mathbf{r} \tag{33}$$

As is standard practice in crystallography, the least-squares errors, the estimated standard deviations, are obtained from the diagonal elements of the variance-covariance matrix and the goodness-of-fit index S:

$$\sigma_i = S_{\text{wp}} A^{-1/2} \tag{34}$$

with $S_{\text{wp}} = [\sum w_i (y_{i,\text{obs}} - y_{i,\text{calc}})^2/(N_0 - N_{\text{par}})^2]^{1/2}$, where N_0 is the number of profile points and N_{par} the number of parameters used. Prince (1981) has criticized this procedure. A recent summary of this highly controversial and statistically unsound procedure is given by Post and Bish (1989). The underlying assumption that a proper weighting scheme is being used is questionable. As a "rule of thumb," the calculated estimated standard deviations tend to underestimate the "true" error by a factor of 2 to 3 for positional parameters and even more in the case of displacement parameters and site occupancies (see Appendix C).

The most used refinement algorithm is a weighted least-squares algorithm that minimizes χ^2. For more details on statistics, see Prince (1994). This assumes that the distribution of the random variable is due to counting statistics and is Gaussian. If the errors of the observations are independent and normally distributed, one obtains the correct estimates of the parameters. If this is no longer true due to, e.g., low counting statistics, other better suited methods should be used. In this case, a Poisson distribution might better represent the experimental errors and

the maximum-likelihood method used (Antoniadis et al., 1990). Currently, one Rietveld program (FULLPROF) has this option incorporated. With good counting statistics, maximum-likelihood and minimum-χ^2 methods give the same results. The maximum-likelihood method is an iteratively reweighted least-squares fit. The weights are determined by the fit. This is in marked difference to the weighted least-squares fit where the weights determine the fit.

Constraints

Constraints and restraints are both assumptions made about the solutions to crystallographic problems. In the case of a restraint, the solution is more tightly restrained to satisfy the assumption the greater our confidence is in the assumption. When using a constraint, we insist that this assumption is obeyed. In the early literature (Rollet, 1965), restraints were often referred to as "slack constraints." We can subdivide constraints into explicit and implicit constraints. Implicit constraints are implied to be without appreciable errors, and our model will be implicitly constrained by the fact that we are using these corrections and assumptions. Examples of implicit constraints are the corrections for Lorentz polarization, absorption corrections, the scattering lengths used in neutron diffraction, and the form factors used in x-ray or magnetic neutron scattering. The assumption that no significant amorphous phase is present in a phase mixture is also an implicit constraint when refining multiphase mixtures and attempting to determine weight fractions. Any errors in these implicit constraints will lead to errors in the final model that are not accounted for by the estimated standard deviations of the refined parameters. In neutron powder diffraction refinements, these considerations are generally applicable to refinements where the choice of form factors will crucially influence the refined magnetic moment, or for instance an incorrect absorption correction will bias the nonharmonic expansion of the anisotropic displacement parameters. To extract physically meaningful information from displacement parameters, one has to deal with implicit constraints in an appropriate manner. Explicit constraints are choices we make in the course of a refinement. The peak shape is an important but often overlooked one. The choice of space groups is an obvious one. In many refinements, it is difficult to decide if a structure is better described in a high-symmetry space group than in a lower symmetry space group with more variables that are highly correlated and lead to an unstable refinement. One solution to this problem is to refine in the lower space group but apply symmetry constraints between certain atoms related by pseudomirror planes. Using the same parameters but assigning opposite signs to the two atoms will achieve this. Coupling the parameters of these atom pairs leads to only one pair of least-squares parameters, thus reducing correlations and leading to a well-converging refinement. Thus, although the structure is described in a lower symmetry space group, the use of explicit constraints introduces symmetry elements from a higher symmetry space group. Another "classical" pitfall in refinements is the need to impose a constraint when describing the model in certain space groups where the origin is not fixed by symmetry such as $P1$, Pc, and $P2$. If there is no fixed origin, the normal matrix will be singular. The solution is to fix a certain atom. This is done by not refining, e.g., the z coordinate in space group $P2$. However, this position will not have an estimated standard deviation and all estimated standard deviations of the z components of the other atoms will therefore be biased. Using the variance-covariance matrix is the appropriate way to extract the correct estimated standard deviations in such a case (Prince, 1994). Another problem often occurring is site disorder. This can be the case of two different atomic species occupying or partially occupying the same site. In both cases, the total occupancy can be an explicit constraint. Many problems in refinements are concerned with the displacement parameters. They are often loosely referred to as "temperature factors." The question arises if a given atom has an isotropic or anisotropic displacement parameter. This choice will influence the bond distances to the neighboring atoms. Sometimes, even anisotropic displacement parameters are not the solution to the problem. One might have to describe rigid units of the structure by using a TLS model (Prince, 1994). The T stands for translation, the L for libration, and the S for screw movement of the rigid group. The right choice of the rigid unit and explicit constraints (e.g., all translational components of the fragments are the same) is crucial and will influence the final R factors (see Appendix B). Often, rigid-body constraints are used to lead the refinement in the right direction. Certain structural coordination polyhedra such as octahedra or tetrahedra are initially constrained to a rigid unit with an "ideal" geometry. The individual atomic parameters are replaced by a total of six parameters, three defining a rotation, three a translation. The significant reduction of parameters from $3n$ to 6 for n atoms reduces the complexity of the normal matrix substantially and generally leads to a smoother refinement. In the initial states of an attempt to solve the structure of SF_6, Cockcroft and Firch (1988) applied such chemical constraints to preserve the SF_6 octahedra in refinements against high-resolution neutron powder diffraction data. This resulted in a smooth convergence to a model, which was then the starting model for further refinements.

Restraints

Restraints, or "slack or soft constraints" as they are referred to in the older literature (Waser, 1963; Rollet, 1965), are ways of influencing the refinement by introducing other observables Y_{obs}. These new observables are then subject to the same mathematical treatment as the "real" experimental observables (intensities). They are added into the normal matrix and the $\mathbf{r}$ vector. Additional variables can be as simple as a dummy atom needed to restrain planarity in a benzene molecule, a nonbonding restraint to avoid overly unphysical distances during the refinement, or the restraints with respect to refining combined x-ray and neutron diffraction data sets. It is important to consider the weights applied to these supplemental observations or subsidiary conditions, as they are also sometimes called, when refining with restraints. If there

are only few restraints, then we can assign a desired estimated standard deviation to them and scale them to the weighted residual for the intensities (Rollet, 1965). The most widely used restraint is the bond restraint:

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = d^2 \qquad (35)$$

The value for d is generally taken from the literature. Six parameters are coupled. In the refinement $d^2 = \Delta x' \mathbf{G} \Delta x$, with $\mathbf{G}$ being the metric tensor. Differentiating leads to

$$\frac{\partial d}{\partial x_1} = \frac{g_{11}\Delta x + g_{12}\Delta y + g_{13}\Delta z}{d} \qquad (36)$$

and

$$\frac{\partial d}{\partial x_2} = \frac{g_{11}\Delta x + g_{12}\Delta y + g_{13}\Delta z}{d} \qquad (37)$$

However, note that the position within the unit cell is not fixed. If used alone, this will lead to a singular matrix. For fairly rigid units, this type of constraint is quite appropriate since the assigned estimated standard deviation is symmetric. When trying to mimic, for example, van der Waals interactions, one might want to use an "asymmetric" estimated standard deviation. Using a "penalty function" can achieve this by calculating an energy using a van der Waals potential involving the restraints. This takes into account the fact that when dealing with van der Waals interactions, you accept longer but not shorter interactions as physically sensible.

LITERATURE CITED

Antoniadis, A., Berruyer, J., and Filhol, A. 1990. Maximum likelihood method in powder diffraction refinements. *Acta Crystallogr. A* 46:692–711.

Axe, J. D., Cheung, S., Cox, D. E., Passell, L., Vogt, T., and Bar-Ziv, S.-B. 1994. Composite germanium monochromators for high resolution neutron powder diffraction applications. *J. Neutron Res.* 2(3):85.

Bacon, G. E., Smith, J. A. G., and Whitehead, C. D. 1950. *J. Sci. Instrum.* 27:330.

Baerlocher, C. 1986. *Zeolites* 6:325.

Baerlocher, C. 1993. Restraints and constraints in Rietveld refinement. *In* The Rietveld Method (R. A. Young, ed.) pp. 186–196. Oxford University Press, New York.

Bendall, P. J., Fitch, A. N., and Fender, B. E. F. 1983. The structure of Na_2UCl_6 and Li_2UCl_6 from mulitiphase powder neutron profile refinement. *J. Appl. Crystallogr.* 16:164.

Berar, J. F. and Lelann, P. 1991. E. S. D.'s and estimated probable error obtained in Rietveld refinements with local correlations. *J. Appl. Crystallogr.* 24:1–5.

Berliner, R. and Werner, S. A. 1986. Effect of stacking faults on diffraction. The structure of lithium metal. *Phys. Rev. B* 34:3586–3603.

Bish, D. L. and Howard, S. A. 1988. Quantitative phase analysis using the Rietveld method. *J. Appl. Crystallogr.* 21:86–91.

Breit, G. and Wigner, E. 1936. *Phys. Rev.* 49:519.

Bricogne, G. 1991. Maximum entropy and the foundations of direct methods. *Acta Crystallogr. A* 47:803–829.

Brown, A. and Edmonds, J. W. 1980. *Adv. X-ray Anal.* 23:361–374.

Buttrey, D. J., Vogt, T., Wildgruber, U., and Robinson, W. R. 1994. The high temperature structure of Bi_2MoO_6. *J. Solid State Chem.* 111:118.

Byrom, P. G. and Lucas, B. W. 1991. POWABS: A computer program for the automatic determination of reflection conditions in powder diffraction patterns. *J. Appl. Crystallogr.* 24:70–72.

Cagliotti, G., Paoletti, A., and Ricci, F. P. 1958. Choice of collimators for a crystal spectrometer for neutron diffraction. *Nucl. Instrum. Methods* 3:223.

Cheetham, A. K. 1986. Structure determination by powder diffraction. *Mater. Sci. Forum* 9:103.

Cheetham, A. K., David, W. I. F., Eddy, M. M., Jakeman, R. J. B., Johnson, M. W., and Torardi, C. C. 1986. Crystal structure determination by powder neutron diffraction at the spallation source ISIS. *Nature (London)* 320:46–48.

Cheetham, A. K. and Wilkinson, A. P. W. 1991. Structure determination and refinement with synchrotron X-ray powder diffraction data. *J. Phys. Chem. Solids* 52:1199–1208.

Cheetham, A. K. and Wilkinson, A. P. 1992. Synchrotron X-ray and neutron diffraction studies in solid state chemistry. *Angew. Chem. Int. Ed. Engl.* 31:1557–1570.

Christensen, A. N., Cox, D. E., and Lehmann, M. S. 1989. A crystal structure determination of PbC_2O_4 from synchrotron X-ray and neutron powder diffraction data. *Acta Chem. Scand.* 43:19–25.

Christensen, A. N., Lehmann, M. S., and Nielsen, M. 1985. Solving crystal structures from powder diffraction data. *Aust. J. Phys.* 38:497–505.

Christian, J. W. 1975. The Theory of Transformations in Metals and Alloys. Pergamon Press, Oxford.

Cockcroft, J. K. and Firch, A. N. 1988. The solid phases of sulphur hexafluoride by powder neutron diffraction. *Z. Kristallogr.* 184:123–145.

Convert, P., Berneron, M., Gandelli, R., Hansen, T., Oed, A., Rambaud, A., Ratel, J., and Torregrossa, J. 1997. A large high counting rate one-dimensional position sensitive detector: The D20 banana. *Physica B* 234–236:1082.

Convert, P. and Forsyth, J. B. 1983. Position-Sensitive Detection of Thermal Neutrons. Academic Press, San Diego.

Convert, P., Hock, R., and Vogt, T. 1990. High-speed neutron diffraction—A feasibility study. *Nucl. Instrum. Methods A* 292:731–733.

Cooper, M. J. and Sayer, J. P. 1975. The asymmetry of neutron powder diffraction peaks. *J. Appl. Crystallogr.* 8:615–618.

Cowley, J. M. 1976a. Diffraction by crystals with planar faults. I. General theory. *Acta Crystallogr. A* 34:83–87.

Cowley, J. M. 1976b. Diffraction by crystals with planar faults. II. Magnesium fluorogermanate. *Acta Crystallogr. A* 32:88–91.

Cowley, J. M. 1981. Diffraction Physics. North-Holland Publishing, New York (see pp. 388–400).

Darwin, G. G. 1914. *Philos. Mag.* 27:315.

David, W. I. F. 1987. The probabilistic determination of intensities of completely overlapping reflections in powder diffraction patterns. *J. Appl. Crystallogr.* 20:316–319.

David W. I. F. 1990. Extending the power of powder diffraction for structure determination. *Nature (London)* 346:731.

David, W. I. F., Ibberson, R. M., and Matsuo, T. 1993. High resolution neutron powder diffraction: A case study of the structure of C_{60}. *Proc. R. Soc. London A* 442:129–146.

Debye, P. and Scherrer, P. 1916. *Z. Phys.* 17:277–283.

Delmas, C. and Tessier, C. 1997. Stacking faults in the structure of nickel hydroxides: A rationale of its high electrochemical activity. *J. Mater. Chem.* 7:1439–1443.

Eastabrook, J. N. B. 1952. *J. Appl. Phys.* 3:349–352.

Finger, L. W., Cox, D. E., and Jephcoat, A. P. J. 1994. A correction for powder diffraction peak asymmetry due to axial divergence. *Appl. Crystallogr.* 27:892–900.

Gilmore, G., Shankland, K., and Bricogne, G. 1993. Applications of maximum entropy method to powder diffraction and electron crystallography. *Proc. R. Soc. London A* 442:97–111.

Goldman, M., Pannetier, J., Beguin, F., and Gonzalez, F. 1988. *Synthetic Metals* 23:133.

Harrison, W. T. A., Dussack, L. L., Vogt, T., and Jacobson, A. J. 1995. Synthese, crystal structures and properties of new layered tungsten(VI)-containing materials based on the hexasgonal-WO_3 structure: $M_2(WO_3)_3SeO_3$, M = NH_4, Rb, Cs). *J. Solid State Chem.* 120:112.

Hendricks, S. and Teller, E. 1942 X-ray interference in partially ordered layer lattices. *J. Chem. Phys.* 10:147–167.

Hewat, A. W. 1973. UKAEA, Harwell, United Kingdom, Research Group Report R7350.

Hill, R. 1992. Rietveld refinement round robin. I. Analysis of standard X-ray and neutron data for $PbSO_4$. *J. Appl. Crystallogr.* 25:589–610.

Hill, R. J. and Flack, H. D. 1987. The use of the Durbin-Watson d-statistics in Rietveld analysis. *J. Appl. Crystallogr.* 20:356–361.

Hill, R. J. and Howard, C. J. 1987. Quantitative phase analysis fron neutron powder diffraction data using the Rietveld method. *J. Appl. Crystallogr.* 20:467–474.

Howard, C. J. 1982. The approximation of asymmetric neutron powder diffraction peaks by sums of gaussians. *J. Appl. Crystallogr.* 15:615–620.

Hull, A. W. 1917. *Phys. Rev.* 9:84.

Hurst, D. G., Pressesky, A. J., and Tunnicliffe, P. R. 1950. *Rev. Sci. Instrum.* 21:705.

International Union of Crystallography (IUC). 1992. International Tables of Crystallography. Dordrecht, Boston and London.

Jansen, E., Schäfer, W., and Will, G. 1988. Profile fitting and the two-stage method in neutron powder diffractometry for structure and texture analysis. *J. Appl. Crystallogr.* 21:228–239.

Johnson, M. W. and Davide, W. I. F. 1987. Rutherford Appleton Lab, United Kingdom, Report RAL-87-059.

Kramers, H. A. 1934. *Physica* 1:182.

Langford, J. B. and Louër, D. 1986. *Powder Diffraction* 1:211.

Larson, A. C. and Von Dreele, R. B. 1990. Unpublished GSAS Program Manual, Los Alamos National Lab.

Latoche, M., Pecheron-Guyan, A., Chabre, Y., Poinsignon, C., and Pannetier, J. 1992. Correlations between the structural and thermodynamic properties of $LaNi_5$ type hydrides and their electrode performances. *J. Alloys Compounds* 189:59–65.

LeBail, A., Duroy, H., and Fourquet, J. L. 1988. *Ab-initio* structure determination of $LiSbWO_6$ by X-ray powder diffraction. *Mater. Res. Bull.* 23:447–452.

Louër, D. and Vargas, R. 1982. Indexation automatique des diagrammes des poudres per dichotomies successives. *J. Appl. Crystallogr.* 15:542.

Medarde, M. and Rodriguez-Carvajal, J. 1997. Oxygen vacancy ordering in La_{2-x} Sr_xNiO_{4-d} ($0 \leq x \leq 0.5$): The crystal structure and defects investigated by neutron diffraction. *Z. Phys. B* 102:307–315.

Morris, R. E., Harrisson, W. T. A., Nicol, J. M., Wilkinson, A. P., and Cheetham, A. K. 1992. Determination of complex structures by combined neutron and synchrotron X-ray powder diffraction. *Nature (London)* 359:519–522.

Morris, R. E., Owen, J. J., Stalick, J. K., and Cheetham, A. K. 1994. Determination of complex structures from powder diffraction data: The crystal structure of $La_3Ti_5Al_{15}O_{37}$. *J. Solid State Chem.* 111:52–57.

Néel, L. 1932. *Ann. Phys.* 17:64.

Néel, L. 1948. *Ann. Phys.* 3:137.

Newsam, J. M., Treacy, M. M. J., Koetsier, W. T., and deGruyter, C. B. 1988. Structural characterization of zeolite beta. *Proc. R. Soc. London A* 420:375–405.

Pannetier, J. 1986. Time-resolved neutron powder diffraction. *Chem. Scr.* 26A:131–139.

Pannetier, J. 1988. Real time neutron powder diffraction. *In* Chemical Crystallography with Pulsed Neutrons and Synchrotron X-rays (M. A. Carrondo and G. A. Jeffrey, eds.) p. 313. D. Reidel Publishing, Oxford.

Pannetier, J., Dubois, J. M., Janot, C., and Bilde, A. 1987. *Philos. Mag. B* 55:435–457.

Papoular, R. J. and Cox, D. E. 1995. Model-free search for extra framework cations in zeolites using powder diffraction. *Europhys. Lett.* 32:337–342.

Paulus, H., Fuess, H., Müller, G., and Vogt, T. 1990. The crystal structure of β-quartz type $HALSi_2O_6$. *N. Jb. Miner. Mh.* 5:232–240.

Pawley, G. S. 1981. Unit cell refinement from powder diffraction data. *J. Appl. Crystallogr.* 14:357–361.

Post, J. E. and Bish, D. L. 1989. Modern powder diffraction. *Rev. Miner.* 20:277.

Press, W. H. 1986. Numerical Recipies—The Art of Scientific Computing. Cambridge University Press, Cambridge.

Prince, E. 1981. *J. Appl. Crystallogr.* 14:157.

Prince, E. 1994. Mathematical Techniques in Crystallography and Materials Science. Springer-Verlag, Heidelberg.

Rietveld, H. M. 1967. *Acta Crystallogr.* 22:151–152.

Rietveld, H. M. 1969. A profile refinement for nuclear and magnetic structures. *J. Appl. Crystallogr.* 2:65.

Rius, J. and Miravitles, C. 1988. Determination of crystal structures with large known fragments directly from measured x-ray powder diffraction intensities. *J. Appl. Crystallogr.* 21:224–227.

Rodriguez-Carvajal, J. 1993. Recent advances in magnetic structure determination by neutron powder diffraction. *Physica B* 192:55–69.

Rodriguez-Carvajal, J., Fernandez-Diaz, M. T., and Martinez, J. L. 1991. *J. Phys. Condensed Matter* 3:3215.

Roessli, B., Fischer, P., Staub, U., Zolliker, M., and Furrer, A. 1993. Combined electronic-nuclear magnetic ordering of the Ho^{3+} ions and magnetic stacking faults in the high-Tc superconductor $HoBa_2Cu_3O_7$. *Europhys. Lett.* 23:511–515.

Rollett, J. S. 1965. Computing Methods in Crystallography. Pergamon Press, Elmsford, N.Y.

Rotella, F. J., Jorgensen, J. D., Biefeld, R. M., and Morosin, B. 1982. Location of deuterium sites in the defect pyrochlore $DTaWO_6$ from neutron powder diffraction data. *Acta Crystallogr. B* 38:1697.

Sabine, T. M., Von Dreele, R. B., and Jorgensen, J. E. 1988. *Acta Crystallogr. A* 44:274.

Schechtman, D., Blech, I., Gratias, D., and Cahn, J. W. 1984. Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* 53:1951.

Semenovskaya, S. U., Khatchaturayan, K. A., and Khatchaturayan, A. L. 1985. Statistical mechanics approach to the structure determination of a crystal. *Acta Crystallogr. A* 41:268–276.

Shankland, K., David, W. I. F., and Csoka, T. 1997. Crystal structure determination from powder diffraction data by the application of a genetic algorithm. *Z. Kristallogr.* 112: 550–552.

Shirley, R. 1984. Measurements and analysis of powder data from single solid phases. *In* Methods and Applications in Crystallographic Computing (S. R. Hall and T. Ashida, eds.) pp. 414–437. Clarendon Press, Oxford.

Shull, C. G. and Smart, J. S. 1949. *Phys. Rev.* 76:1256.

Shull, C. G., Strauser, W. A., and Wollan, E. O. 1951. *Phys. Rev.* 83:333.

Soller, W. A. 1924. A new precision X-ray spectrometer. *Phys. Rev.* 24:158–167.

Tombs, N. C. and Rooksby, H. P. 1950. *Nature (London)* 165:442.

Toraya, H. J. 1986. Whole-powder-pattern fitting without reference to a structural model: Application to X-ray powder diffraction spectra. *Appl. Crystallogr.* 19:940–947.

Treacy, M. M. J., Newsam, J. M., and Deem, M. W. 1991. A general recursion method for calculating diffracted intensities from crystals containing planar faults. *Proc. R. Soc. London* 433:499–520.

van Laar, B. and Yelon, W. B. 1984. The peak in neutron powder diffraction. *J. Appl. Crystallogr.* 17:47–54.

Visser, J. 1969. A fully automatic program for finding the unit cell from powder data. *J. Appl. Crystallogr.* 2:89.

Vogt, T., Paulus, H., Fuess, H., and Muller, G. 1990 The crystal structure of $HAlSi_2O_6$ with a keatite-type framework. *Z. Kristallogr.* 190:7–18.

Waser, J. 1963. Least-squares refinement with subsidiary conditions. *Acta Crystallogr.* 16:1091–1094.

Werner, P.-E., Eriksson, L., and Westdahl, M. J. 1985. TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Crystallogr.* 18:367.

Wessells, T., Baerlocher, C., and McCusker, L. B. 1999. Single-crystal-like diffraction from polycrystalline materials. *Science* 284:477.

Wiles, D. B. and Young, R. A. 1981. A new computer program for Rietveld analysis of X-ray powder diffraction patterns. *J. Appl. Crystallogr.* 14:149–151.

Williams, W. G. 1988. Polarized Neutrons. Clarendon Press, Oxford.

Wilson, A. J. C. 1942. Imperfections in the structure of cobalt II. Mathematical treatment of proposed structures. *Proc. R. Soc. London A* 180:277–285.

Wilson, A. J. C. 1943. The reflection of X-rays from the 'anti-phase nuclei' of $AuCu_3$. *Proc. R. Soc. London A* 181:360–368.

Wollan, E. O. and Shull, C. G. 1948. *Phys. Rev.* 73:830.

Woodward, P. M., Sleight, A. W., and Vogt, T. 1995. Structure refinement of triclinic tungsten trioxide. *J. Phys. Chem. Solids* 56:1305–1315.

Yamada, Y., Hamaya, N., Axe, J. D., and Shapiro, S. M. 1984. Nucleation, growth, and scaling in a pressure-induced first-order phase transition: RbI. *Phys. Rev. Lett.* 53:1665.

Zachariasen, W. H. 1949. *Acta Crystallogr.* 2:296.

Zachariasen, W. H. and Ellinger, F. H. 1963. *Acta Crystallogr.* 16:369.

Zinn, W. H. 1947. *Phys. Rev.* 71:752.

KEY REFERENCES

Bish, D. L. and Post, J. E. 1989. Modern Powder Diffraction, Reviews in Mineralogy, Vol. 20. The Mineralogical Society of America, Washington, D.C.

Covers neutron powder diffraction as well as x-ray and synchrotron powder diffraction. Contains valuable information with respect to experimental procedures and data analysis.

Young, R. A. 1993. The Rietveld Method. International Union of Crystallography. Oxford University Press, New York.

Provides an excellent introduction to modern Rietveld refinement techniques and neutron powder diffraction using pulsed sources and reactor-based instruments. The 15 parts are written by acknowledged researchers who cover the field in theoretical as well as practical aspects. Highly recommended.

APPENDIX A: OPTICS OF A CONSTANT WAVELENGTH DIFFRACTOMETER

The Cagliotti coefficients U, V, and W are functions of the collimators α_1, α_2, and α_3, of the mosaic spread β of the monochromator crystals and the monochromator scattering angle ("take-off" angle) θ_m:

$$\mathrm{FWHM} = (U\tan^2\theta + V\tan\theta + W)^{1/2} \tag{38}$$

where

$$U = \frac{4(\alpha_1^2\alpha_2^2 + \alpha_1^2\beta^2 + \alpha_2^2\beta^2)}{\tan 2\theta_m(\alpha_1^2 + \alpha_2^2 + 4\beta^2)} \tag{39}$$

$$V = -\frac{4\alpha_2^2(\alpha_1^2 + 2\beta^2)}{\tan\theta_m(\alpha_1^2 + \alpha_2^2 + 4\beta^2)} \tag{40}$$

$$W = \frac{\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_2^2\alpha_3^2 + 4\beta^2(\alpha_2^2 + \alpha_3^2)}{\alpha_1^2 + \alpha_2^2 + 4\beta^2} \tag{41}$$

and with a minimum of the FWHM at

$$\left(W - \frac{V^2}{4U}\right)^{1/2} \tag{42}$$

Here, α_1, α_2, α_3 are the beam divergences of the in-pile, monochromator-sample, and sample-detector collimators, respectively. This FWHM can be rewritten as

$$\mathrm{FWHM} = \frac{N}{(\alpha_1^2 + \alpha_2^2 + 4\beta^2)^{1/2}} \tag{43}$$

where

$$N = \left(\frac{\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_2^2\alpha_3^2 + 4\beta^2(\alpha_2^2 + \alpha_3^2)}{(\alpha_1^2 + \alpha_2^2 + 4\beta^2) - 4a\alpha_2^2(\alpha_1^2 + 2\beta^2) + 4a^2(\alpha_1^2\alpha_2^2 + \alpha_1^2\beta^2 + \alpha_2^2\beta^2}\right)^{1/2} \tag{44}$$

$$a = \frac{\tan\theta}{\tan\theta_m} \tag{45}$$

θ, being the Bragg angle. An important result is that the FWHM decreases with N and N becomes smaller the closer θ is to the take-off angle θ_m. The best resolution of a neutron powder diffractometer will be near the take-off position. After the take-off position, the FWHM increases considerably. For high-resolution purposes, one will attempt to use a take-off angle as high as possible. The diffractometer D2B at the ILL has a θ_m of 135°, and the HRNPD at Brookhaven National Laboratory has a θ_m of 120°.

APPENDIX B: RELIABILITY FACTORS

One of the most debated areas of Rietveld refinement is of the the so-called reliability factors, often called "R factors," which assess the quality of the model with respect to the data. In a least-squares refinement, one minimizes R. Therefore, the choice of R is nontrivial. The two most-used R factors in a Rietveld refinement are the profile R factor R_p and the weighted profile R factor R_{wp}:

$$R_p = \frac{\sum |y_i - y_{i,\mathrm{calc}}|}{\sum y_i} \tag{46}$$

$$R_{wp} = \left(\frac{m}{\sum w_i y_I^2}\right)^{1/2} \tag{47}$$

where $M = \Sigma w_i (y_i - y_{i,\mathrm{calc}})^2$ is the actual function the refinement routine is minimizing and $w_i = 1/\sigma_i^2$ the weight assigned to the observed intensity y_i at step i, σ_i being the variance of observation y_i and $y_{i,\mathrm{calc}}$ the calculated intensity at the ith step of the pattern. Another important R factor is

$$R_{\mathrm{Bragg}} = \frac{\sum |I_i - I_{i,\mathrm{calc}}|}{\sum I_i} \tag{48}$$

which is a reliability factor commonly used in refinements of integrated intensity data. The term I_i is the integrated intensities at point i.

The best R factor to judge the agreement between experimental data and the structural model is R_{Bragg}. However, R_{Bragg} is deduced with the help of a model and is therefore biased in favor of the model. This is different when using model-free structure refinement techniques such as maximum-entropy techniques. The R_{Bragg} is also rather insensitive to inadequate modeling of the peak shape and the presence of peaks stemming from other phases. This is not the case with R_{wp}, which depends very much on how well the profile function used describes the peak shape, especially the low-angle asymmetry correction. Therefore, R_{wp} is less sensitive to the structural parameters than R_{Bragg}. Another problem with R_{wp} is that it can be misleadingly small if the refined background is high since it is easier to get a good fit of a slowly varying background than of the Bragg reflections that contain the structural information. This is an important consideration when measuring materials containing hydrogen, since the incoherent scattering will raise the background. Even minute residual hydrogen in deuterated compounds or materials synthesized from hydrogen-containing precursors will result in a higher background.

Other R factors are

$$\chi^2 = \frac{M}{N - P + C} \tag{49}$$

$$R_e = \frac{R_{wp}}{\sqrt{\chi^2}} \tag{50}$$

where N, P, and C are the numbers of data points, refined parameters, and constraints, respectively.

The frequently encountered Durbin-Watson statistic parameters d and Q (Hill and Flack, 1987) measure the correlation between adjacent residuals:

$$d = \frac{\sum [w_i (y_i - y_{i,\mathrm{calc}}) - w_{i-1}(y_{i-1} - y_{i-1,\mathrm{calc}})]^2}{\sum w_1^2 (y_i - y_{i,\mathrm{calc}})^2} \tag{51}$$

Serial correlation is tested at the 99.9% confidence level by comparing d to Q:

$$Q = 2\left(\frac{N-1}{N-P} - \frac{3.0902}{(N+2)^{1/2}}\right) \tag{52}$$

If $d < Q$, there is a positive serial correlation, which means successive residuals will have the same sign. If $Q < d < 4 - Q$, there is no correlation between neighboring points, and if $d > 4 - Q$, there is a negative correlation. The ideal value for d is 2.00. If the calculated and observed profile functions do not match well, there will be a strong correlation of the residuals of neighboring points and the Durbin-Watson d value will be far from 2.00.

APPENDIX C: ESTIMATED STANDARD DEVIATIONS

The question of what is an independent observation N is not only for purists, the reason being that N is used to calculate the estimated standard deviation (esd):

$$\mathrm{esd}_i = \left(\frac{M_{ii}^{-1} \sum w_i (y_i - y_{i,\mathrm{calc}})^2}{N - P + C}\right)^{1/2} \tag{53}$$

where M_{ii}^{-1} is the diagonal in the inverted least-squares matrix, P the number of refined variables, and C the total number of constraints. In Rietveld refinements one uses the number of points measured and not the number of individual reflections I_{hkl} as in single-crystal refinements. The estimated standard deviations are not the probable experimental error; rather they are the minimal possible probable error arising from random errors alone. There are certain systematic errors (e.g., peak shape, background) that arise from implicit constraints that are not considered in the estimated standard deviations. In other words, the experimental reproducibility will only then reflect the estimated standard deviations when the random errors ("counting statistics") are the dominant errors.

Berar and Lelann (1991) showed that the underestimation of the estimated standard deviations is due to serial correlations as assessed by the Durbin-Watson statistics and modified the summation to adjust the estimated standard deviations. An important problem arises when reflections completely overlap and therefore the error associated with the individual intensity has to be assigned. Using a probabilistic method, David (1987) derived a method to deal with these cases.

APPENDIX D: COMPUTER PROGRAMS FOR RIETVELD ANALYSIS

The best source for up-to-date Rietveld refinements programs and other useful programs for powder diffraction are available on the Commission for Powder Diffraction's homepage *www.za.iucr.org/iucr-top/comm/cpd/rietveld.html*. This site provides access to new and updated program versions and should be consulted by anyone looking for powder diffraction–related software.

The first Rietveld refinement code was written in ALGOL by H. M. Rietveld (1969). Hewat (1973) modified the code and converted it to FORTRAN. PROF, as it became known, was the standard program in the early days to analyze neutron powder pattern. A further offspring of the original Rietveld program was MPROF, which featured a modification to incorporate two phase refinements (Bendall et al., 1983).

GSAS Generalized Structure Analysis System, written by Allen C. Larson and Robert B. von Dreele, LANSCE, MS-H805, Los Alamos National Laboratory, Los Alamos, NM 87545. This is one of the most used Rietveld refinement programs in the world, even though, as its name implies, it is a multipurpose program suite. It runs under OPENVMS/VAX, OPENVW/ALPHA, Silicon Graphics IRIX, Ultrix, HPUX, and DOS. This program allows the refinement of fixed wavelength as well as time-of-flight data. It furthermore allows the combined refinement of various powder and single- crystal data sets as well as neutron and x-ray diffraction data and has all the "mesoscopic" features incorporated such as particle size and microstrain as well as a recently added feature for the refinement of texture using spherical harmonics.

Website: *http://www.mist.lansce.lanl.gov*
Author: *vondreele@lanl.gov*

FULLPROF Written by Juan Rodriguez-Carvajal (unpublished). It is also originally based on DBW3.2 but incorporates, among many extensions, the capability to refine microstrain and particle size in a generalized manner. It runs on Vax machines under VMS, on Suns, IBM Riscs, HP Unix, and Silicon Graphics as well as PCs. This program is especially useful for refining magnetic structures. The formula for the magnetic structure factor and its derivatives with respect to the parameters was programmed in a general fashion. The Fourier coefficients can be given either as standard crystallographic or spherical components. The program generates satellite reflections of up to 24 propagation vectors per phase and refines their components in reciprocal lattice units (Rodriguez-Carvajal, 1993). This program allows the user to perform many nonstandard refinements because of user-supplied executables that can be linked to the FULLPROF library. FULLPROF also allows refinements using the maximum-likelihood technique.

Internet address of author: *juan@bali.scalay.cea.fr*

UNIX-anonymous ftp areas where FULLPROF is stored: *www.bali.scalay.cea.fr/pub/diverse/fullp/freadme ftp://charybde.saclay.cea.fr/pub/divers/fullprof98*

LHPM Written by Hill and Howard. It is based on Wiles and Young's DBW3.2 program and features an improved peak asymmetry, absorption correction, multiphase refinement, and a Voigt peak shape (Hill and Howard, 1987).

Anonymous ftp: *atom.ansto.gov.au/pub/physics/neutron/rietveld*
Author: *bah@atom.ansto.gov.au* (Brett Hunter)

RIETAN Written by F. Izumi. The program runs on a UNIX platform. It is also a modern Rietveld refinement program that allows the refinement of incommensurate structures as well as superstructures. Furthermore, it also allows the choice of least-squares algorithms (Gauss-Newton, modified Marquardt, and conjugate direction) at the beginning of each refinement cycle, which is very helpful to assure smoother convergence especially in the early stages of a refinement.

Anonymous ftp: *nirim.go.jp/pub/sci/rietan*
Author: *izumi@nirim.go.ip*

XRS-82 Written by C. Baerlocher (1986). It is based on the X-Ray 72 system. This code is particularly interesting for scientist working on the structures of zeolites since it has elaborate constraints and restraints with various stereochemical aspects as well as a learned profile function (Baerlocher, 1993).

PROFIL Written by J. K. Cockcroft. It is a modern Rietveld refinement that was not based on any older versions of code but written completely from scratch. It is a very fast and easy-to-use program resembling SHELX in the way its free format input file is constructed. The output files are very detailed and contain very useful crystallographic information, such as conversion of the various displacement parameters, distances, and angles with estimated standard deviations, correlation matrices, and resolution curves converted in FWHM versus 2θ. This is a program suited for the nonspecialist. It allows the focus to remain on crystallography. Various output files feed into other programs. It also contains a full-profile-refinements option based on LeBails algorithm, which generates input files for use in SHELX-82 direct methods and Patterson techniques.

Anonymous ftp: *gordon.cryst.bbk.ac.uk*

It should be pointed out that refining neutron powder diffraction data with different Rietveld refinement programs leads to differences that sometimes exceed the determined errors. This was shown by Woodward et al. (1995) when comparing the results of refinements of the same data sets of triclinic and monoclinic WO_3 using the three Rietveld refinement programs GSAS, Rietan, and

PROFIL. Subtle differences in peak shapes, the different low-angle asymmetry corrections for axial divergence, the way the background is accounted for, and the way errors are calculated lead to these discrepancies. In another study, Hill (1992) reported on the results of a Rietveld round robin. The main conclusion was again that the refinement of neutron data showed a rather narrow distribution for the coordinates and anisotropic displacement parameters about the single-crystal values and that the errors are substantially smaller than those of interrefinement variations.

THOMAS VOGT
Brookhaven National Laboratory
Upton, New York

SINGLE-CRYSTAL NEUTRON DIFFRACTION

INTRODUCTION

Single-crystal neutron diffraction measures the elastic Bragg reflection intensities from crystals of a material, the structure of which is the subject of investigation. A single crystal is placed in a beam of neutrons produced at a nuclear reactor or at a proton accelerator-based spallation source. Single-crystal diffraction measurements are commonly made at thermal neutron beam energies, which correspond to neutron wavelengths in the neighborhood of 1 Å. For high-resolution studies requiring shorter wavelengths (~0.3 to 0.8 Å), a pulsed spallation source, or a high-temperature moderator (a "hot source") at a reactor may be used. When complex structures with large unit-cells are under investigation, as is the case in structural biology, a cryogenic-temperature moderator (a "cold source") may be employed to obtain longer neutron wavelengths (~4 to 10 Å). A monochromatic beam technique is generally used at reactors, which normally operate as continuous wave (CW) neutron sources. At spallation sources, which normally are pulsed, a time-of-flight Laue method with a broad range of neutron wavelengths is generally used, and the neutron wavelength is determined from the time-of-flight for neutrons to reach the detector. The type of moderator is again generally optimized for the particular application, in a manner similar to what is the common practice at reactors.

A single-crystal neutron diffraction analysis will determine the crystal structure of the material, typically including its unit cell and space group, the positions of the atomic nuclei and their mean-square displacements, and relevant lattice site occupancies. Because the neutron possesses a magnetic moment, the magnetic structure of the material can be determined as well from the magnetic contribution to the Bragg intensities (see MAGNETIC NEUTRON SCATTERING).

Instruments for single-crystal diffraction (single-crystal diffractometers or SCDs) are generally available among the facilities at the major neutron-scattering centers. Beam time on many of these instruments is available through a proposal mechanism. For a listing of neutron SCD instruments and their corresponding facility contacts, see Table 1.

Complementary and Related Techniques

A number of techniques give information on crystal structure complementary to that provided by single-crystal neutron diffraction. Some of the most important of these are listed below along with brief descriptions. For additional information on these techniques, and their advantages and disadvantages, see the corresponding chapters of *Characterization of Materials* in which they are described. It is worth noting that neutron diffraction is a very expensive proposition, first of all involving the often demanding and time-consuming preparation of large, high-quality crystals (at least 1 mm^3 in volume). Then it is generally necessary to travel to a neutron scattering research center to perform the measurements, which typically require from several days to several weeks. One therefore should be certain that neutron diffraction is the technique of choice for solving the problem at hand—otherwise one of the complementary methods should be considered first.

Single-crystal x-ray diffraction is the method most directly complementary to single-crystal neutron diffraction. The chief difference between the two methods is that, while neutron diffraction images the nuclear scattering density in the crystal, x-ray diffraction images the electron-density distribution. Neutron diffraction therefore has important advantages in studies where hydrogen or other light atoms must be located, or where isotopic substitutions may be of interest, and of course for magnetic structures as well. Magnetic structures can also be studied using x-rays, although there the effect is much smaller, so that synchrotron radiation is generally required. In the special case where the x-ray energy is close to an absorption edge of a magnetic center in the sample, resonance enhancement of the magnetic scattering can produce effects comparable in magnitude to those observed with neutrons. X-ray and neutron diffraction measurements can be combined to good effect in studies seeking detailed information about the valence electron-density distribution in a material, particularly when hydrogen is present (Coppens, 1997).

Powder neutron (NEUTRON POWDER DIFFRACTION) and powder x-ray (X-RAY POWDER DIFFRACTION) diffraction are also powerful techniques for determining crystal structure. For many important classes of materials single crystals may be difficult or even impossible to obtain. Powder diffraction methods have the advantage of wide applicability and generally allow for much more rapid data collection than do single-crystal methods. This makes powder methods ideal for studies of materials over ranges of temperature and pressure, for example, in studies investigating phase transitions, and for in situ real-time studies. The chief disadvantage of powder methods stems from the more limited amount of information obtained in the diffraction pattern, which generally places an upper limit on the complexity of structures that can be satisfactorily treated. Progress in this regard continues at a rapid

Table 1. Neutron SCD Instruments and Facility Website Contacts

Facility	Source Type	Location	Instrument Name	Instrument Characteristics[a]	Proposal Cycle	Website URL
BENSC	Reactor	Hahn-Meitner Institute, Berlin, Germany	E5	Monochromatic thermal beam	Semiannual	http://www.hmi.de/bensc
FRJ-2	Reactor	Jülich Research Center, Jülich, Germany	SV28	Monochromatic thermal beam	Consult web site	http://www.kfa-juelich.de/iff/Institute/ins/Broschuere_NSE/
HIFAR	Reactor	ANSTO, Lucas Heights, NSW, Australia	2TANA	Monochromatic thermal beam	Consult web site	http://www.ansto.gov.au/natfac/hifar.html
HFIR	Reactor	Oak Ridge National Laboratory, Oak Ridge, Tenn., U.S.A.	HB-2A	Monochromatic thermal beam	Continuous	http://www.ornl.gov/hfir
HFR	Reactor	Institut Laue-Langevin, Grenoble, France	D9	Monochromatic hot beam	Semiannual	http://www.ill.fr
			D10	Monochromatic 4-circle and triple axis		
			D19	Monochromatic thermal beam		
			LADI	Laue with imaging plate Weissenberg geometry		
IBR-2	Reactor	Frank Laboratory of Nuclear Physics, Dubna, Russia	DN2	Monochromatic thermal beam	Consult web site	http://nfdfn.jinr.dubna.su
IPNS	Spallation	Argonne National Laboratory, Argonne, IL, U.S.A.	SCD	Time-of-flight Laue	Semiannual	http://www.pns.anl.gov
ISIS	Spallation	Rutherford Appleton Laboratory, Chilton, Oxon, U.K.	SXD	Time-of-flight Laue	Semiannual	http://www.isis.rl.ac.uk
JRR-3	Reactor	Japan Atomic Energy Research Institute, Tokai-mura, Japan	BIX-III	Monochromatic thermal beam with imaging plate. Oscillation geometry	Consult web site	http://www.jaeri.go.jp/english/
KENS	Spallation	KEK Laboratory, Tsukuba, Japan	FOX	Time-of-flight Laue	Annual	http://neutron-www.kek.jp/
LANSCE	Spallation	Los Alamos National Laboratory, Los Alamos, N.M., U.S.A.	SCD	Time-of-flight Laue	Consult web site	http://www.lansce.lanl.gov
			PCS	Time-of-flight Laue protein crystallography station	—	http://www.lansce.lanl.gov
NRU	Reactor	Chalk River Laboratories, Chalk River, Ontario, Canada	E3	Monochromatic thermal beam	Continuous	http://neutron.nrc.ca
Orphée	Reactor	Laboratoire Léon Brillouin, Saclay, France	5C2	Monochromatic hot beam	Semiannual	http://www-llb.cea.fr
R2	Reactor	Studsvik Neutron Research Laboratory, Nyköping, Sweden	SXD	Monochromatic thermal beam	Every 4 months	http://www.studsvik.uu.se
SINQ	CW Spallation	Paul Scherer Institute, Villigen, Switzerland	TriCS	Monochromatic thermal beam	Semiannual	http://www.psi.ch

[a]The instruments are 4-circle diffractometers unless otherwise noted.

pace, however, so that structures with on the order of 100 independent atoms in the crystallographic asymmetric unit may now be considered as suitable candidates for powder studies. Recently, powder-diffraction studies of proteins have even been undertaken, fitting highly constrained structure models to the data (Von Dreele, 1999). Both powder and single-crystal x-ray techniques have been enormously enhanced by the availability of synchrotron x-ray sources, with their high-intensity beams. At the third-generation synchrotron sources, x-ray intensities may be as much as nine or ten orders of magnitude larger than for the most currently available intense neutron beams.

Electron diffraction (Section 11c) is another complementary technique for crystal-structure determination—particularly when studying surfaces and interfaces, where low-energy electron diffraction (LOW-ENERGY ELECTRON DIFFRACTION) is especially powerful, or when only microcrystals are available. Progress in dynamical computations that take multiple scattering into account (n-beam dynamical diffraction computations) is extending the power of electron diffraction for the solution and refinement of crystal structures.

Nuclear magnetic resonance (NUCLEAR MAGNETIC RESONANCE IMAGING) is also a valuable technique for investigating the structure of materials. To cite just one application, high-field solid-state NMR can yield accurate distances between hydrogen atoms and—particularly in materials with high thermal motion or disorder—results may sometimes be more reliable than for comparable interatomic distances obtained from neutron diffraction.

Scope of This Unit

This unit outlines the basic principles of single-crystal neutron diffraction and provides examples to illustrate the range of materials that can be studied. Sufficient practical details are included, it is hoped, to help a non-expert get started with the technique and locate an appropriate neutron diffraction facility. A listing of the instruments that are available at various neutron sources is included in Table 1.

For readers desiring additional background, the classic monograph by Bacon (1975) provides an excellent starting point. Wilson (1999) has recently published an overview that includes a comprehensive survey of neutron diffraction results on molecular materials. For those wishing to investigate the subject in more detail, Marshall and Lovesey (1971) provide an exhaustive theoretical treatment of all aspects of neutron scattering.

PRINCIPLES OF THE METHOD

Single-crystal neutron diffraction measures elastic, coherent scattering, i.e., the Bragg reflection intensities (see X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS). Although the measurements and their interpretation are generally quite straightforward, and amount to a standard crystal-structure analysis analogous to what is done with x rays, neutron structure determinations are still relatively uncommon. This is so because of the limited number of neutron crystallography facilities and the need to obtain large samples in order to observe adequate reflection intensities (see Complementary and Related Techniques). Thus, for example, the contents of a recent release of the Cambridge Structural Database of organic and organometallic crystal structures (Allen and Kennard, 1993) consist of less than 1% neutron-determined structures (the overwhelming majority of the remaining 99% are x-ray-determined structures).

Basic Theory of Single-Crystal Neutron Diffraction

The nuclear scattering density in a crystal may be expressed in terms of the neutron structure factors, $F(h,k,l)$, by the familiar relationship

$$\rho(x,y,z) = \Sigma F(h,k,l)\exp\{-2\pi i(hx+ky+lz)\} \qquad (1)$$

where the Fourier summation (Equation 1) runs over the Miller indices h,k,l, and the coordinates x,y,z are expressed in terms of fractions of the unit-cell translations **a**, **b**, **c**. The nuclear scattering density is simply the atomic probability density at any point multiplied by the corresponding neutron scattering amplitude.

As noted above, neutrons are scattered by the atomic nuclei in a crystal, while x-rays are scattered by the electrons. Because the nuclear radius is $\sim 10^4$ times smaller than the neutron wavelengths that are used in diffraction, the neutron scattering amplitude is a constant independent of the momentum transfer, $Q = 4\pi(\sin\theta/\lambda)$. For x rays, on the other hand, the scattering amplitude, also called the scattering factor or form factor, falls off with increasing Q due to destructive interference of the scattering from different regions of the electron-density cloud. Each isotope has a unique neutron scattering amplitude, but these values lie within the same order of magnitude across the entire periodic table, as illustrated in Figure 1. For x-rays, the scattering amplitude increases with the atomic number of atoms, so that diffraction is dominated by contributions from the high-Z atoms. For this reason, neutron diffraction has important advantages in studies where hydrogen or other light atoms must be located, or

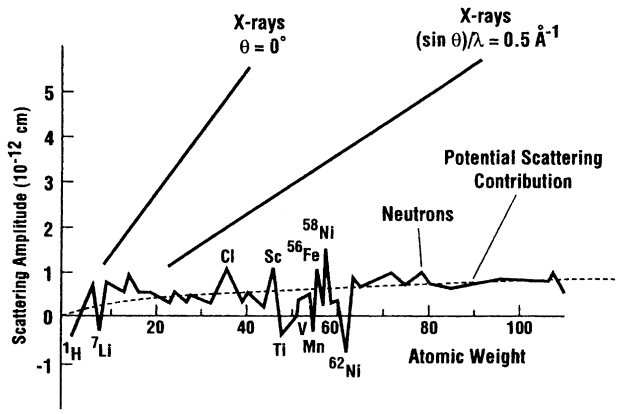


Figure 1. Neutron scattering amplitudes as a function of atomic weight. Adapted from Bacon (1975).

where isotopic substitutions may be of interest, as has been mentioned earlier. Further, because neutron diffraction images the nuclear scattering density in the crystal, it provides an unbiased estimate of the mean nuclear position. X-ray diffraction studies generally assume that the nucleus is located at the centroid of its electron-density cloud. This can be a poor assumption, e.g., where hydrogen atoms form covalent bonds. For example, C-H bond distances determined by x-ray diffraction are systematically shortened by ~0.1 Å due to the effects of bonding on the hydrogen 1*s* electron-density distribution.

It is, of course, not possible to obtain the scattering-density function $\rho(x,y,z)$ directly from the observed diffraction intensities, because of the crystallographic phase problem. In practice, an initial estimate of the neutron structure factor phases is usually obtained from a prior x-ray diffraction study. Probabilistic direct methods of phase determination developed for x-ray structure determination have, however, also been shown to work in the neutron case, e.g., for organic and organometallic crystals (Broach et al., 1979). Direct methods seem to work in practice in spite of the existence of negative scattering density whenever hydrogen is present, which actually contradicts one of the underlying assumptions of direct methods. Phase-determination methods based on anomalous dispersion can also be applied in neutron diffraction for special cases involving crystals that contain one of a handful of highly absorbing elements, including Li, B, Cd, Sm, and Gd (see, e.g., Koetzle and Hamilton, 1975, and references cited therein).

Neutron structure analysis and refinement generally utilizes standard kinematical diffraction theory. According to kinematical theory (see KINEMATIC DIFFRACTION OF X RAYS), the Bragg reflection intensity $I(h,k,l)$ is given by the relationship

$$I(h,k,l) \propto (i_o k\lambda^2 V_c/V^2)|F(h,k,l)|^2 \qquad (2)$$

where i_o is the incident neutron flux, V_c is the crystal volume, V is the unit-cell volume, and $|F(h,k,l)|^2$ is the squared structure factor. The relationship in Equation 2 holds for the monochromatic beam case. For the Laue case, i_o is a function of λ reflecting the neutron flux spectrum incident on the sample. Equation 2 omits the Lorenz term, which must be introduced to correct for the geometrical effect caused by scanning reflections to obtain integrated Bragg intensities.

The structure factor is

$$F(h,k,l) = \Sigma a_i b_i \exp(-W_i)\exp\{2\pi i(hx+ky+lz)\} \qquad (3)$$

where a_i is the atomic site occupancy, b_i is the neutron scattering amplitude, and W_i is the Debye-Waller factor, and the summation in Equation 3 runs over n atoms in the crystallographic unit cell.

From Equations 2 and 3, it is apparent that the diffracted intensity will fall off as the unit-cell volume increases. In order to achieve sufficient diffracted intensity, it is therefore necessary to grow larger single crystals for materials with larger unit cells. The use of longer neutron wavelengths is also highly desirable in this case, e.g., by performing the experiment at an instrument located on a cold moderator beamline.

The Debye-Waller factor W_i takes into account the reduction in diffracted intensity from interference produced by smearing of the scattering density due to thermal motion in the sample. Equation 4 is the isotropic approximation

$$W_i = \exp\{-U_i Q^2/2\} = \exp\{-8\pi^2 U_i(\sin^2\theta/\lambda^2)\} \qquad (4)$$

where U_i is the atomic mean-square displacement (adp). It is often the practice to introduce an anisotropic harmonic model in which second-order tensors are employed to describe the adps. In that case, the Debye-Waller factor is given by

$$\begin{aligned} W_i = \exp\{-2\pi^2(&U_{11}a^{*2}h^2 + U_{22}b^{*2}k^2 \\ &+ U_{33}c^{*2}l^2 + 2U_{12}\boldsymbol{a}^*\bullet\boldsymbol{b}^*hk + 2U_{13}\boldsymbol{a}^*\bullet\boldsymbol{c}^*hl \\ &+ 2U_{23}\boldsymbol{b}^*\bullet\boldsymbol{c}^*kl \end{aligned} \qquad (5)$$

where $\boldsymbol{a}^*$, $\boldsymbol{b}^*$, and $\boldsymbol{c}^*$ are reciprocal lattice vectors. Higher-order tensor descriptions may be used to model anharmonic effects.

Neutron diffraction is often the method of choice in studies aimed at a precise description of thermal motion and/or disorder because it is easier to determine the Debye-Waller factor in the neutron case where the atomic scattering amplitude is a constant independent of Q (see above), and there is therefore minimal correlation between the Debye-Waller factor and the site occupancy. In contrast, the x-ray atomic form factor may assume a Q dependence quite similar to that of the Debye-Waller factor, leading to high correlations in the least-squares minimization approach to structure refinement.

That the neutron scattering amplitude is independent of Q confers an additional advantage not always well appreciated. When diffraction measurements are made at cryogenic temperatures, thereby lowering the adp parameters, neutron Bragg intensities will exhibit a much reduced fall-off with Q compared to the x-ray case. This results in substantially improved precision in the atomic positions.

The standard kinematical theory outlined above ignores the effects of absorption and extinction in real crystals. Absorption is often quite minimal except in the case of crystals that contain hydrogen, where the effective absorption due to hydrogen incoherent scattering can be quite appreciable. The measured neutron intensities can be corrected for absorption using the method of Gaussian integration (Busing and Levy, 1957) or by employing an analytical technique in which the crystal is described in terms of Vornoi polyhedra (Templeton and Templeton, 1973).

The use of large crystals often requires that extinction corrections be included during neutron structure refinement to obtain a satisfactory fit to the strong reflections. A number of formalisms have been developed for this purpose. Perhaps the most widely used is that of Becker and Coppens (1975), which in its most general implementation treats both primary and secondary extinction.

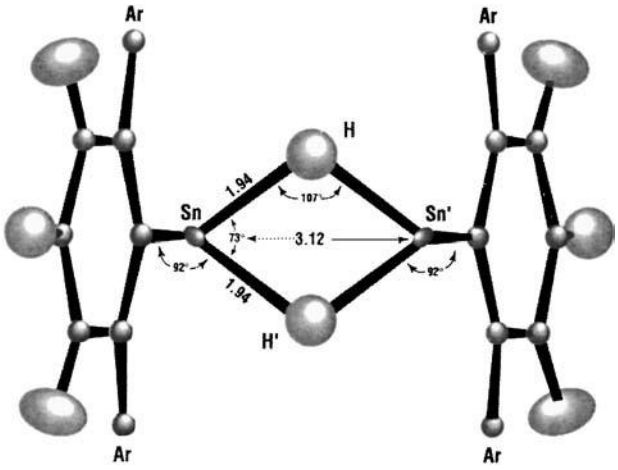


Figure 2. The molecular structure of a tin(II) hydride complex determined by a neutron diffraction study on IPNS SCD at 20 K (Koetzle et al., 2001). Estimated standard deviations in bond distances and angles are less than 0.01 Å and 0.1°, respectively.

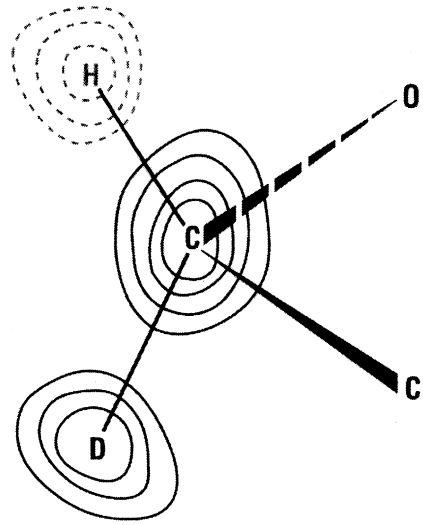


Figure 3. Contoured neutron Fourier map for (+)-neopentyl-1-d-alcohol (S) with positive nuclear scattering density at C and D shown as solid contours and negative nuclear scattering density at H shown as dashed contours (Yuan et al., 1994).

PRACTICAL ASPECTS OF THE METHOD

The initial applications of single-crystal neutron diffraction to the study of molecular structure were carried out at first generation research reactors around 1950. For additional information in this regard, see, e.g., the studies of KHF_2 (Peterson and Levy, 1952) and ice (Peterson and Levy, 1953, 1957) at Oak Ridge National Laboratory and the study of KH_2PO_4 (KDP) at Harwell (Bacon and Pease, 1953, 1955). KDP was also studied independently at about the same time at Oak Ridge (Peterson et al., 1953, 1954). These studies provided an important demonstration of the advantages of neutron diffraction for determining the positions of hydrogen atoms and for studying hydrogen bonding.

Single-crystal methods are being employed at modern reactor and spallation neutron sources to study the structures of a broad range of materials. For example, the tin hydride complex shown in Figure 2 was investigated (Koetzle et al., 2001) using the SCD spectrometer at the Intense Pulsed Neutron Source (IPNS) at Argonne National Laboratory. Collection of Bragg intensity data at a temperature of 20 K for this triclinic crystal with 111 independent atoms, unit-cell volume $V = 1986$ Å^3, and crystal volume $V_c = 15$ mm^3 required 11 days on SCD. Low-temperature studies of this quality allow full, unconstrained refinement of the structure model incorporating anisotropic Debye-Waller factors and readily yield metal-hydrogen bond distances with a precision of better than 0.01 Å. The amount of beam time used by this experiment is quite representative at present at IPNS for this type of organometallic crystal.

Figure 3 shows an application of Fourier methods to single-crystal neutron diffraction data obtained at the Brookhaven National Laboratory High Flux Beam Reactor (HFBR; note that the HFBR neutron source ceased operation in 1999). In this site-specific isotope labeling study of a chiral organic alcohol (Yuan et al., 1994), the neutron Fourier map readily gives the location of the deuterium label that was introduced in an enzymatic hydrogenation reaction.

Figure 4 shows the results of a study of the variation with temperature of the anharmonic thermal motion in the cubic form of ZnS (zincblende; Moss et al., 1980,

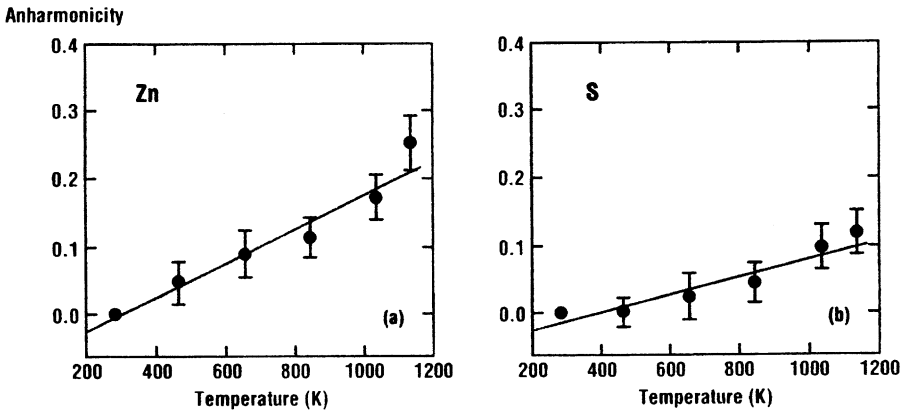


Figure 4. Anharmonicity of atomic displacements in zincblende (ZnS; m.p. 1973 K; cubic phase) as a function of temperature (Moss et al., 1980, 1983).

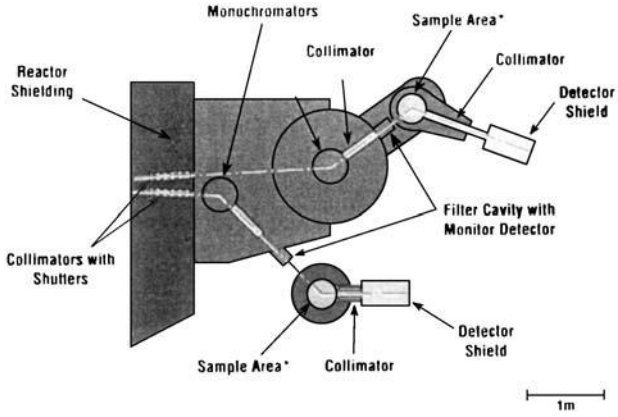


Figure 5. Schematic plan of the SCD spectrometers formerly installed at the dual thermal beam port H6 at the Brookhaven High-Flux Beam Reactor (HFBR).

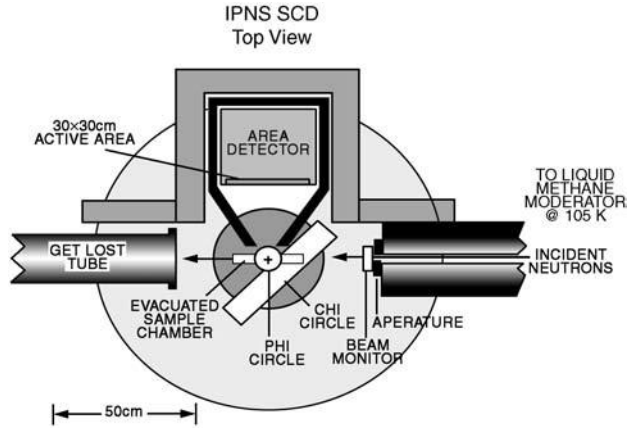


Figure 6. Schematic plan of the SCD spectrometer installed on a liquid methane moderator at the Argonne Intense Pulsed Neutron Source (IPNS).

1983). This high-resolution study, also carried out on one of the single-crystal diffractometers at the HFBR, used a relatively short neutron wavelength of 0.83 Å and successfully modeled the higher-order contributions to the Debye-Waller factor for a zincblende sample drawn from the collections of the Smithsonian Institution, Washington, D.C.

Neutron Sources and Instrumentation

As mentioned earlier, neutron sources fall into two categories: nuclear reactors, which normally operate as CW neutron sources, and spallation sources, which normally are pulsed. A pulsed nuclear reactor (IBR-30) operates at the Frank Laboratory of Nuclear physics in Dubna, Russia, and a CW spallation source (SINQ) operates at the Paul Scherer Institute in Villigen, Switzerland. A listing of neutron scattering centers offering single-crystal diffraction facilities that will host users from other organizations is given in Table 1.

Figure 5 schematically illustrates a representative SCD set-up at a reactor. Using this facility, which was formerly installed on a dual thermal beam port at the HFBR at Brookhaven, neutron beams in the wavelength range 0.8 to 1.7 Å could be extracted using a variety of monochromator crystals. Common choices for monochromators include Be, Si, Ge, Cu, and pyrolitic graphite. The incident beam horizontal divergence could be selected to match the mosaic spread of the sample using a Soller slit collimator, and a circular aperture was employed to limit the beam diameter to approximately twice the dimensions of the sample to minimize the background. Samples were mounted on one of the two four-circle diffractometers with Eulerian cradles that provided for full orientation control. The samples were generally centered in the incident beam using the observed setting angles for an azimuthal reflection. This enabled a precise centering to be carried out with the sample inside a cryostat or other specialized environment. Integrated Bragg intensities were recorded by scanning the samples in steps about the vertical (ω) axis, counting for a preset incident-beam monitor count to correct for any variation in the incident neutron flux from the reactor over time. The addition of position-sensitive area detectors has significantly enhanced the performance of many of the SCDs currently in operation at reactors (see discussion of Detectors, below). In particular, the recent emergence of neutron imaging-plate detectors and the development of improved glass scintillation detectors have been important advances.

Figure 6 is a schematic view of the SCD at Argonne's IPNS, which serves as an illustration of an instrument at a spallation neutron source. This diffractometer (Schultz, 1987), which is installed at a beam port on a liquid methane cryogenic moderator, operates by the time-of-flight Laue method (see X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS). The full spectrum of neutrons from the moderator is allowed to impinge on the sample, and the wavelength is determined by the time-of-flight required for neutrons to reach the detector, i.e., $\lambda = ht/ml$, where h is Planck's constant, t is the neutron time-of-flight, m is the neutron mass, and l is the source-to-detector distance.

The Eulerian cradle mount is used to orient the sample for a series of intensity data collection histograms, recording t and the position for each neutron arriving at the area detector. Counts are accumulated for a preset number of accelerator pulses. The IPNS proton accelerator operates at 30 Hz and a proton beam power of 7 kW. A depleted uranium spallation target is employed at present. With the geometry shown in Figure 6, a complete hemisphere of reciprocal space can be explored while collecting a total of 45 data histograms. In contrast to monochromatic beam methods, where the sample must be rotated to scan through Bragg reflections in order to obtain integrated intensities, in the Laue method the sample is held stationary while counting is in progress. This significantly simplifies the experiment and may be of particular advantage, for example, when a special sample environment is required or when carrying out rapid reciprocal-lattice surveys, e.g., while searching for phase transitions over a range of temperature and/or pressure.

In general, each neutron scattering center has developed its own instrumentation, and the design of each SCD spectrometer is unique in its details. From the prospective experimenter's point of view, however, most SCDs probably will look quite similar in the sense that they will feature an Eulerian cradle sample mount that offers the possibility of deploying cryostats, furnaces, pressure cells, etc., to obtain a wide range of sample environments. As was mentioned earlier, a successful experiment requires production of a large, well-formed single crystal. Generally, the minimum satisfactory sample size falls in the range 1 to 10 mm^3, with larger samples required for larger unit cells. These large samples are required because even at the current most powerful spallation neutron source, the 180 kW ISIS at the Rutherford Appleton Laboratory in the U.K., or at the most powerful high-flux reactor sources, the 58 MW High Flux Reactor (HFR) at the Institut Laue-Langevin in Grenoble, France, and the 85 MW High Flux Isotope Reactor (HFIR) at Oak Ridge National Laboratory in the U.S.A., the neutron flux on sample at the SCDs is several orders of magnitude reduced compared to the flux from a conventional, sealed-tube laboratory x-ray source.

Bent crystal focusing monochromators (see, e.g., Tanaka et al., 1999b, 2000) have been installed at a number of reactor SCD instruments. Significant intensity enhancement has been achieved in this way, in some cases approaching one order of magnitude.

Detectors

The first generation of SCD instruments at reactors utilized single-channel gas proportional counters filled with BF_3 or ^3He. These detectors capture neutrons by the reactions $^{10}B + {}^1n = {}^7Li + {}^4He + 2.8$ MeV, and $^3He + {}^1n = {}^1H + {}^3H + 0.76$ MeV, respectively. The thermal neutron capture cross-section of ^3He is $\sim$1.5 times that of ^{10}B, and for this reason ^3He counters are usually lighter and smaller than BF_3 counters.

Modern neutron SCD instruments generally include two-dimensional position-sensitive detectors, allowing for parallel data collection. Area detectors are, in any event, required for the time-of-flight Laue instruments at pulsed spallation sources. Neutron position-sensitive detectors are of three types: gas-filled proportional counters, generally using ^3He; scintillation counters, such as the ^6Li-glass detector on the SCD at IPNS (see Fig. 6); and imaging-plate detectors, which utilize a Gd converter. Neutron imaging plates (Niimura et al., 1994, Tazaki et al., 1999) have the important advantage of being flexible, which facilitates covering a large solid angle. Powerful new SCDs utilizing imaging plates and oscillation/Weissenberg geometry have been developed since 1990 (LADI, Wilkinson and Lehmann, 1991; Cipriani et al., 1995; BIX, Tanaka et al, 1999a). Imaging plates are integrating detectors and are therefore, unfortunately, not suitable for use on time-of-flight Laue instruments.

Future Outlook

The application of neutron scattering in general, including single-crystal diffraction, has always been constrained by the limited number of research facilities available. With the advent of the 2 MW Spallation Neutron Source (SNS), now under construction at Oak Ridge and scheduled to begin operation in 2006, the neutron scattering community will have at its disposal a state-of-the-art facility of unprecedented power and intensity, with approximately an order of magnitude increase over the power currently delivered by ISIS. Planning for major new spallation facilities is at an advanced stage in Europe, with the European Spallation Source (ESS) project, and also in Japan (JHF project). The SNS is expected to make possible a broad range of new SCD applications. For example, it should be possible to do the following.

1. Use much smaller single crystals, perhaps approaching normal x-ray sized samples ($\sim$0.01 mm^3 in volume), which will greatly expand the range of materials that can be investigated.
2. Obtain complete SCD data sets in minutes, as opposed to the days or weeks that are presently required. This should enable the study of many interesting processes in real time. To cite just one example of interest, it should be possible to investigate solid-state photochemical reactions where the mechanisms can be probed by following site-specific isotope labels (Ohgo et al., 1996, 1997).
3. Carry out extensive parametric studies, including investigating materials over a broad range of temperatures and pressures, which, in practice, at existing facilities, can only be done conveniently with powders. This would be extremely important for many materials applications, e.g., to understand the response of hydrogen-bonded ferroelectric, ferroelastic, and nonlinear optical (NLO) materials to temperature and pressure changes.

A critical challenge facing the neutron scattering community is how to replace the facilities at the current generation of research reactors. The majority of these facilities were commissioned in the 1950s and 1960s, and some are nearing the end of their projected lifetimes. Two new research reactor facilities are currently under development, at Garching, Germany, where the 20 MW FRM-II is scheduled to be operational in 2002, and at ANSTO, Lucas Heights, Australia, where a 20-MW replacement for the HIFAR reactor is scheduled to be completed in 2005.

DATA ANALYSIS AND INITIAL INTERPRETATION

As discussed above (see Principles of the Method), initial values of the neutron structure factor phases are usually calculated based on a structure model obtained from a prior x-ray diffraction study. Although the neutron phases can be obtained by direct methods (see above), the x-ray study provides an efficient structure solution and serves the important dual purpose of allowing a check for problems, such as twinning, that might make it difficult to obtain a high-quality neutron structure. Most neutron

scattering facilities strongly recommend that prospective SCD users have an x-ray structure of their material in hand before embarking on their neutron measurements.

Each SCD instrument generally provides its users with its own specially adapted suite of programs to handle data reduction, i.e., for Bragg peak indexing, integration of peak intensities, absorption correction, and merging of equivalent reflections. For Laue data, equivalent reflections are generally not merged, since they may be collected at different neutron wavelengths and their extinction corrections may therefore differ one from another. For information on data reduction software, the reader should consult the facility websites listed in Table 1. These websites generally identify the responsible instrument scientists who can advise prospective users on what to expect.

Refinement of neutron structures may be carried out with a standard crystallographic program package. Program systems in common use at neutron scattering centers include GSAS (*http://www.ncnr.nist.gov/programs/crystallography/software/gsas.html*; Larson and Von Dreele, 2000) and SHELX (*http://shelx.uni-ac.gwdg.de/SHELX/*; Sheldrick, 1997).

SAMPLE PREPARATION

As has been noted above, single-crystal neutron diffraction requires a large, high-quality crystal, generally at least 1 mm^3 in volume. There usually is no need for the crystal to be deuterated. This contrasts with the situation for neutron powder diffraction (see NEEUTRON POWDER DIFFRACTION), where gram-quantities of sample generally are required, and the material must be completely deuterated to avoid prohibitively high background from incoherent scattering of hydrogen. Single-crystal samples are normally mounted on an aluminum or vanadium pin with a suitable adhesive. Vanadium is sometimes used to reduce the background because of its extremely small neutron scattering amplitude. Aluminum will usually suffice, however, so long as the diameter of the pin is substantially smaller than that of the crystal. The personnel responsible for the facility's SCD instrument can advise as to the proper type of pin to use. Often the most suitable course is to bring unmounted samples to the neutron scattering center and to use the local laboratory facilities for final sample preparation.

Provision for on-site sample-handling facilities is frequently an important factor to consider when choosing a neutron scattering center. For example, prospective users with air-sensitive samples will want to be certain that facilities exist for handling and mounting their crystals under an inert atmosphere. The availability of well-equipped conventional laboratory facilities adjacent to the neutron source can critically affect the outcome of an experiment.

PROBLEMS

The most common reason for failure of a single-crystal neutron diffraction experiment is poor crystal quality. Ideally, smaller crystals from the same crystallization batch as the neutron samples will be examined first by x-ray diffraction, to assure their quality and to ascertain that the crystals scatter to adequate resolution (in Q). Optical examination of the samples under a polarizing microscope frequently may help to reject twinned specimens and agglomerates.

Once a neutron sample has been chosen, a careful assessment should be made of its mosaic spread along the principal crystal axes. At reactor instruments, this is usually done using ω scans, while at spallation-source instruments it is accomplished by examining preliminary data histograms that are collected to determine the sample orientation. A crystal with noticeably split diffraction profiles often will give unsatisfactory results. It is also important to avoid overlap in neighboring reflections that may result in high mosaic spreads, i.e., greater than about 2° in ω, particularly for crystals with large unit cells.

For proper estimation of intensities, it is critical that position-sensitive detectors be calibrated to remove effects of nonuniform response and to estimate detector dead-time losses, if any. At spallation-source instruments, the detector calibration must include allowance for the wavelength dependence of detector efficiency. Detector calibrations generally will have been carried out by the instrument scientist, placing an incoherent scatterer such as a vanadium rod in the sample position, and the resulting corrections will be applied automatically during data reduction.

Background corrections are also important, especially when data are acquired using a position-sensitive detector. A very successful and widely used technique that employs a variable background mask adjusted to maximize $I/\sigma(I)$, where I is the integrated intensity and $\sigma(I)$ its estimated standard deviation (Lehmann and Larsen, 1974), has been adapted to treat such data.

As noted above, the ideal kinematical expression for the Bragg reflection intensity given in Equation 2 ignores the effects of absorption and extinction that are present in real crystals. In particular, extinction is often important in neutron diffraction because of the requirement for large, well-formed crystals. Zachariasen (1967) gave an early treatment of extinction. Coppens and Hamilton (1970) extended the Zachariasen formalism to treat crystals with anisotropic extinction. Becker and Coppens (1974a,b, 1975) made further improvements, while Thornley and Nelmes (1974) treated the case of highly anisotropic extinction. For single-crystal neutron studies, it is important to have access to a crystallographic least-squares structure refinement program that allows for extinction. This should not pose a serious problem, however, since many modern refinement programs include the popular Becker-Coppens extinction formalism. When extinction is particularly severe, high correlations may occur in the least-squares refinement between extinction parameters and some of the adps, or more uncommonly even with the atomic positions. In these cases, special care should be taken, e.g., by employing restraints or damping factors, and by carefully testing the various extinction models to determine which one best fits the observed intensity data.

An absorption correction should always be applied to the observed Bragg intensities before attempting to treat

extinction in the least-squares refinement. It is preferable to measure the crystal, and index its faces, before beginning data collection. In practice, however, many samples are approximated as spheres, e.g., when the material is air-sensitive and therefore difficult to measure. Fortunately, absorption is usually quite modest for neutrons and actually rarely exceeds around 50%. An exception might be for crystals containing highly absorbing elements such as Li, B, Cd, Sm, and Gd, where it also may be important to include anomalous dispersion corrections. For many samples, however, the majority of the absorption will result from the reduction in the effective incident-beam intensity due to incoherent scattering of hydrogen. Howard et al. (1987) have determined the wavelength dependence of the hydrogen incoherent scattering cross-section, which is needed for proper absorption correction of time-of-flight Laue data from spallation sources.

ACKNOWLEDGMENTS

The author would like to thank Dr. Arthur Schultz for reading a draft of this manuscript and providing comments. Work at Argonne and Brookhaven National Laboratories was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under Contracts W-31-109-ENG-38 and DE-AC2-98CH10886, respectively. Additional financial support was provided to the author by NATO under grant PST.CLG.976225.

LITERATURE CITED

Allen, F. H. and Kennard, O. 1993. 3D search and research using the Cambridge structural database. *Chem. Des. Autom. News* 8:31–37.

Bacon, G. E. 1975. Neutron Diffraction (Third Edition). Oxford University Press, London.

Bacon, G. E. and Pease, R. S. 1953. A neutron diffraction study of potassium dihydrogen phosphate by Fourier synthesis. *Proc. Roy. Soc.* A220:397–421.

Bacon, G. E. and Pease, R. S. 1955. A neutron-diffraction study of the ferroelectric transition of potassium dihydrogen phosphate. *Proc. Roy. Soc.* A230:359–381.

Becker, P. J. and Coppens, P. 1974a. Extinction within the limit of validity of the Darwin transfer equations. I. General formalisms for primary and secondary extinction and their application to spherical crystals. *Acta Cryst.* A30:148–153.

Becker, P. J. and Coppens, P. 1974b. Extinction within the limit of validity of the Darwin transfer equations. II. Refinement of extinction in spherical crystals of SrF_2 and LiF. *Acta Cryst.* A30:148–153.

Becker, P. J. and Coppens, P. 1975. Extinction within the limit of validity of the Darwin transfer equations. III. Non-spherical crystals and anisotropy of extinction. *Acta Crystallogr.* A31:417–425.

Broach, R. W., Schultz, A. J., Williams, J. M., Brown, G. M., Manriquez, J. M., Fagan, P. J., and Marks, T. J. 1979. Molecular structure of an unusual organoactinide hydride complex determined solely by neutron diffraction. *Science* 203:172–174.

Busing, W. R. and Levy, H. A. 1957. High speed computation of the absorption correction for single crystal diffraction measurements. *Acta Crystallogr.* 10:180–187.

Cipriani, F., Castagna, J.-C., Lehmann, M. S., and Wilkinson, C. 1995. A large image-plate detector for neutrons. *Physica B* 213–214:975–977.

Coppens, P. 1997. X-ray Charge Densities and Chemical Bonding. Oxford University Press, London.

Coppens, P. and Hamilton, W. C. 1970. Anisotropic extinction corrections in the Zachariasen approximation. *Acta Cryst.* A26:71–83.

Howard, J. A. K., Johnson, O., Schultz, A. J., and Stringer, A. M. 1987. Determination of the neutron absorption cross-section for hydrogen as a function of wavelength with a pulsed neutron source. *J. Appl.Cryst.* 20:120–22.

Koetzle, T. F. and Hamilton, W. C. 1975. Neutron diffraction study of $NaSmEDTA.8H_2O$: An evaluation of methods of phase determination based on three-wavelength anomalous dispersion data. *In* Anomalous Scattering (S. Ramaseshan and S.C. Abrahams, eds.). pp. 489–502. Munksgaard, Copenhagen.

Koetzle, T. F., Henning, R., Schultz, A. J., Power, P. P., Eichler, B. E., Albinati, A., and Klooster, W. T. 2001. Neutron Diffraction Study of the First Known Tin(II) Hydride Complex, $[2,6\text{-}Trip_2C_6H_3Sn(\mu\text{-}H)]_2$, Trip = 2,4,6-tri-isopropylphenyl. XX European Crystallographic Meeting, Krakow, Poland, 2001, Abstract O.M2.05.

Larson, A. C. and Von Dreele, R. B. 2000. General Structure and Analysis System (GSAS). Los Alamos National Laboratory Report LAUR 86-7481, Los Alamos, New Mexico.

Lehmann, M. S. and Larsen, F. K. 1974. A method for location of the peaks in step-scan-measured Bragg reflections. *Acta Cryst.* A31:580–584.

Marshall, G. W., and Lovesey, S. W. 1971. Theory of Thermal Neutron Scattering. Oxford University Press, London.

Moss, B., McMullan, R. K., and Koetzle, T. F. 1980. Temperature dependence of thermal vibrations in cubic ZnS: A comparison of anharmonic models. *J. Chem. Phys.* 73:495–508.

Moss, B., Roberts, R. B., McMullan, R. K., and Koetzle, T. F. 1983. Comment on "temperature dependence of thermal vibrations in cubic ZnS: A comparison of anharmonic models." *J. Chem. Phys.* 78:7503–7505.

Niimura, N., Karasawa, Y., Tanaka, I., Miyahara, J., Takahashi, K., Saito, H., Koizumi, S., and Hidaku, M. 1994. An imaging plate neutron detector. *Nucl. Instr. Meth. Phys. Res.* A349:521–525.

Ohgo, Y., Ohashi, Y., Klooster, W. T., and Koetzle, T. F. 1996. Direct observation of hydrogen-deuterium exchange reaction in a cobaloxime crystal by neutron diffraction. *Chem. Lett.* 445-446, 579.

Ohgo, Y., Ohashi, Y., Klooster, W. T., and Koetzle, T. F. 1997. Analysis of hydrogen-deuterium exchange reaction in a crystal by neutron diffraction. *Enantiomer* 2:241–248.

Peterson, S. W. and Levy, H. A. 1952. A single crystal neutron diffraction determination of the hydrogen position in potassium bifluoride. *J. Chem. Phys.* 20:704–707.

Peterson, S. W. and Levy, H. A. 1953. A single-crystal neutron diffraction study of heavy ice. *Phys. Rev.* 92:1082.

Peterson, S. W. and Levy, H. A. 1957. A single-crystal neutron diffraction study of heavy ice. *Acta Crystallogr.* 10:70–76.

Peterson, S. W., Levy, H. A., and Simonsen, S. H. 1953. Neutron diffraction study of tetragonal potassium dihydrogen phosphate. *J. Chem. Phys.* 21:2084–2085.

Peterson, S. W., Levy, H. A., and Simonsen, S. H. 1954. Neutron diffraction study of the ferroelectric modification of potassium dihydrogen phosphate. *Phys. Rev.* 93: 1120-1121.

Schultz, A. J. 1987. Pulsed neutron single-crystal diffraction. *Trans. Am. Crystallogr. Assoc.* 23:61–69.

Sheldrick, G. M. 1997. SHELX-97. Program for the Refinement of Crystal Structures using Single Crystal Diffraction Data. University of Göttingen, Germany.

Tanaka, I., Ahmed, F. U., and Niimura, N. 2000. Application of a stacked elastically bent perfect Si monochromator with identical and different crystallographic planes for single crystal and powder neutron diffractometry. *Physica B* 283:195–298.

Tanaka, I., Kurihara, K., Haga, Y., Minezaki, Y., Fujiwara, S., Kumazawa, S., and Niimura, N. 1999a. An upgraded neutron diffractometer (BIX-I_M) for macromolecules with a neutron imaging plate. *J. Phys. Chem. Solids* 60:1623–1626.

Tanaka, I., Niimura, N., and Mikula, P. 1999b. An elastically bent silicon monochromator for a neutron diffractometer. *J. Appl. Crystallogr.* 32:525–529.

Tazaki, S., Neriishi, K., Takahashi, K., Etoh, M., Karasawa, Y., Kumazawa, S., and Niimura, N. 1999. Development of a new type of imaging plate for neutron detection. *Nucl. Instr. Meth. Phys. Res.* A424:20–25.

Templeton, D. H. and Templeton, L. K. 1973. Am. Crystallogr. Assoc. Meeting Abstracts, Abstract E10, Storrs, Connecticut.

Thornley, F. R. and Nelmes, R. J. 1974. Highly anisotropic extinction. *Acta Cryst.* A30:748–757.

Von Dreele, R. B. 1999. Combined Rietveld and stereochemical restraint refinement of a protein crystal structure. *J. Appl. Crystallogr.* 32:1084–1089.

Wilkinson, C. and Lehmann, M. S. 1991. Quasi-Laue neutron diffractometer. *Nucl. Instr. Meth. Phys. Res.* A310:411–415.

Wilson, C. C. 1999. Single Crystal Neutron Diffraction from Molecular Materials. World Scientific, Singapore.

Yuan, H. S. H., Stevens, R. C., Bau, R., Mosher, H. S., and Koetzle, T. F. 1994. Determination of the molecular configuration of (+)-Neopentyl-1-d alcohol by neutron and x-ray diffraction analysis. *Proc. Natl. Acad. Sci. U.S.A.* 91:12872–12876.

Zachariasen, W. H. 1967. A general theory of x-ray diffraction in crystals. *Acta Cryst.* 23:558–564.

KEY REFERENCES

Bacon, 1975. See above.

A classic treatment of the principles and practice of neutron diffraction. Part 4 is devoted to a description of experimental techniques.

Marshall and Lovesey, 1971. See above.

Exhaustive treatment of all aspects of neutron scattering, including inelastic and quasielastic scattering, as well as diffraction. Sometimes referred to as "the bible" of neutron scattering.

Stout, G. H. and Jensen, L. H. 1989. X-ray Structure Determination: A Practical Guide, 2nd ed. John Wiley & Sons, New York.

An excellent basic crystallography text outlining the techniques used in crystal structure analysis.

Willis, B. T. M. and Pryor, A. W. 1975. Thermal Vibrations in Crystallography. Cambridge University Press.

Provides a comprehensive overview of the use of crystallographic techniques to analyze thermal motion and disorder in solids.

Wilson, 1999. See above.

A compendium describing the current state-of-the-art and giving a review of the literature including many recent results in neutron crystal structure analysis.

THOMAS F. KOETZLE
Brookhaven National Laboratory
Upton, New York and
Argonne National Laboratory
Argonne, Illinois

PHONON STUDIES

INTRODUCTION

Until the beginning of this century, the models used to understand the properties of solids were based on the assumption that atomic nuclei are fixed at their equilibrium positions. These models have been very successful in explaining many of the low-temperature properties of solids, and especially those of metals, whose properties are determined to a large extent by the behavior of the conduction electrons. Many well-known properties of solids (such as sound propagation, specific heat, thermal expansion, conductivity, melting, and superconductivity), however, cannot be explained without considering the vibrational motion of the nuclei around their equilibrium positions. The study of lattice dynamics—or of phonons, which are the quanta of the vibrational field in a solid—was initiated by the seminal papers of Einstein (1907, 1911), Debye (1912), and Born and von Kármán (1912, 1913). Initially, most of the theoretical and experimental studies were devoted to the study of phonons in crystalline solids, as the name lattice dynamics implies. At present, the field encompasses the study of the dynamical properties of solids with defects, surfaces, and amorphous solids. It is impossible in a single unit to present even a brief overview of this field. The discussion that follows will therefore be limited to a brief outline of the neutron scattering techniques used for the study of phonons in crystalline solids. Additional information can be found in many books devoted to this subject (see Bibliography). Particularly useful is the four-volume book *Dynamical Properties of Solids* edited by Horton and Maradudin (1974), and several parts in *Neutron Scattering*, Vol. 23 of *Methods of Experimental Physics*, edited by Sköld and Price (1986).

Indirect information about the lattice dynamical properties of solids can be obtained from a variety of macroscopic measurements. For instance, in some cases, the phonon contribution to the specific heat or resistivity can be easily separated from other contributions and can provide important information about the lattice dynamical properties of a solid. In addition, measurements of sound velocities are particularly useful, since they provide the slopes of the acoustic phonon branches. Since such macroscopic measurements are relatively easy and inexpensive to make in a modern materials science laboratory, they should, in principle, be performed before one undertakes a detailed phonon study by spectroscopic techniques.

Direct measurement of the frequencies of long-wavelength optical phonons can be obtained by Raman scattering and infrared spectroscopy (Cardona, 1975–1991; Burstein, 1967). These experiments provide essential information for solids containing several atoms per unit cell and should be performed, if possible, before one undertakes a time-consuming and expensive detailed neutron scattering study. Also, for simple solids, information about the phonon dispersion curves can be obtained from x-ray diffuse scattering experiments.

Presently, the most powerful technique for the study of phonons is the inelastic scattering of thermal neutrons. The technique directly determines the dispersion relation, that is, the relationship between the frequency and propagation vector of the phonons. Its power was demonstrated by Brockhouse and Stewart (1955), who, following a suggestion put forward by Placzek and Van Hove (1954), measured the dispersion relation of aluminum.

PRINCIPLES OF THE METHOD

Given the potential seen by the nuclei of a solid, the description of their motion is reduced to the study of the oscillations of a system of particles around their equilibrium positions, which is a well-known problem in mechanics. The basic problem here is to determine the potential seen by the nuclei. If we assume that the electrons follow nuclear motion adiabatically, the effective potential $\phi(R)$ seen by the nuclei is written as,

$$\phi(R) = V(R) + V_e(R) \tag{1}$$

where R collectively denotes the nuclear positions, $V(R)$ is the nuclear interaction energy, and $V_e(R)$ is the average electronic energy determined with the nuclei fixed at R.

For a given potential, the motion of the nuclei can be immediately determined if we assume that the displacements of the nuclei from their equilibrium positions are small compared to their separation (the harmonic approximation). This approximation is very good, except for solids consisting of light nuclei, such as solid ^{4}He. If the harmonic approximation is valid, it is well known from mechanics that any nuclear motion can be considered as a superposition of a number of monochromatic waves, the normal vibrational modes or characteristic vibrations of the system. For a system containing N nuclei, the number of normal modes is $3N - 6$, the number of vibrational degrees of freedom of the system; for a solid $3N - 6 \cong 3N$, since the number of nuclei in this case is very large. The frequencies of the normal modes are obtained by diagonalizing the force constant matrix, a relatively simple problem in the case of a molecule. In an amorphous solid, one is faced with the formidable problem of diagonalizing a $3N \times 3N$ matrix. With present advances in computer technology, such direct diagonalizations are becoming possible and play an increasingly important role in the study of the vibrational properties of amorphous solids.

The translational symmetry of crystalline solids considerably simplifies the determination of the frequencies of their normal modes. In this case, the problem is reduced to the determination of the motion of the atoms contained in a unit cell. The normal modes are propagating vibrational waves with the propagation wave vector $\mathbf{q}$ determined by the crystal geometry. If the unit cell contains r nuclei, their frequencies $\omega_j(\mathbf{q})$ are obtained by diagonalizing a $3r \times 3r$ force constant matrix, which is a more tractable problem than the one encountered in amorphous solids. The relation

$$\omega = \omega_j(\mathbf{q}) \qquad (j = 1, 2, \ldots, 3r) \tag{2}$$

between the frequency and wave vector is called the dispersion relation. It consists of $3r$ branches labeled by the index j. Among these branches, there are three acoustic branches whose frequencies approach zero at the long-wavelength limit ($\mathbf{q} \to 0$). In this limit, the dispersion of the acoustic branches is linear, $\omega = \nu_j \mathbf{q}$, where ν_j is the appropriate sound velocity in the solid and the propagating vibrational waves in the solid are simply sound waves. The remaining $3r - 3$ branches tend to a finite frequency as $\mathbf{q} \to 0$ and are called optic branches, since some of these vibrational modes interact with light. For each wave vector of a branch, the motion of the nuclei in the cell is determined by solving the equations of motion. The pattern of motion in the cell is specified by the polarization vectors $\mathbf{e}_d^j(\mathbf{q})(d = 1, 2, \ldots, r)$, which provide the direction of motion of the nuclei in the cell. A vibrational mode is called longitudinal or transverse if the polarization vectors are parallel and perpendicular, respectively, to the propagation vector $\mathbf{q}$. The allowed values of the propagation vector $\mathbf{q}$ are determined by the boundary conditions on the surface of the crystal. The bulk properties of the crystal are not influenced by the specific form of the boundary conditions. For convenience, periodic boundary conditions are usually adopted. Because of the periodicity of the crystal, all physically distinct values of $\mathbf{q}$ can be obtained by restricting the allowed values of $\mathbf{q}$ in one of the primitive cells of the reciprocal lattice. Actually, instead of the primitive cell of the reciprocal lattice, it is more convenient to use the first Brillouin zone, which is the volume enclosed by planes that are the perpendicular bisectors of the lines joining a reciprocal lattice point to its neighboring points.

In the classical description outlined above, any nuclear motion can be considered as the superposition of propagating waves with various propagation vectors and polarizations. The analogy with the classical description of the electromagnetic field should be noted. As in the case of the electromagnetic field, the quantum description is easily obtained. In the harmonic approximation, the motions of the nuclei can be decoupled by a canonical transformation to normal coordinates. By this transformation, the nuclear Hamiltonian is reduced to a sum of Hamiltonians corresponding to independent harmonic oscillators, and the quantization of the vibrational field is reduced to that of the harmonic oscillator. Thus, instead of the classical wave description, we have the quantum description in terms of quanta, called phonons, that propagate through the lattice with definite energy and momentum. The energy $E_j(\mathbf{q})$ and momentum $\mathbf{p}$ of the phonon are

related to the frequency and wave vector of the corresponding vibrational wave by the well-known relations

$$E_j(\mathbf{q}) = \hbar\omega_j(\mathbf{q}) \qquad \mathbf{p} = \hbar\mathbf{q} \tag{3}$$

where $\hbar = h/2\pi$ is Planck's constant.

The free motion of noninteracting phonons corresponds to the free propagation of the monochromatic waves in the harmonic approximation. The phonons have finite lifetimes, however, since anharmonic interaction as well as interaction with other elementary excitations in the solid are always present. At thermal equilibrium, brought about by the above mentioned interactions, the average number of phonons, n_j, is given by the Bose-Einstein relation

$$n_j = \frac{1}{\exp(E_j/k_{\mathrm{B}}T) - 1} \tag{4}$$

where k_{B} is Boltzmann's constant.

Notice that at high temperatures ($E_j/k_{\mathrm{B}}T \ll 1$) the number of phonons is proportional to the temperature and inversely proportional to their energy.

From the measured phonon frequencies, one can evaluate the frequency distribution or phonon spectrum, which can be used to calculate the thermodynamic properties of the solid, such as the lattice specific heat. The frequency distribution, $Z(\omega)$, is defined so that $Z(\omega)\Delta\,\omega$ is the fraction of vibrational frequencies in the interval between ω and $\omega + \Delta\omega$.

$$Z(\omega) = \frac{1}{3N}\sum_{j,\mathbf{q}} \delta[\omega - \omega_j(\mathbf{q})] \tag{5}$$

where δ is Dirac's delta function.

As mentioned earlier, the most detailed information about the lattice dynamical properties of solids is obtained by studying the scattering of thermal neutrons by a crystalline specimen of solid. This is because the energy and wave vector of thermal neutrons are of the same order as those of the normal vibrations in a solid.

The scattering cross-section consists of an incoherent and a coherent part. The most detailed information about the lattice dynamical properties of the solid is obtained by measuring the coherent one-phonon inelastic neutron scattering cross-section from a single crystal. In such studies, incoherent and multiphonon processes are of importance only because their contribution to the background scattering may complicate observation of the one-phonon coherent processes.

The coherent cross-section for scattering of the neutron to a final state characterized by a wave vector $\mathbf{k}_1$ (and energy $E_1 = \hbar k_1/2m$) due to the creation or annihilation of a single phonon of frequency $\omega_j(\mathbf{q})$ is essentially (see, e.g., Chapter 1 of Sköld and Price, 1986) given by

$$\frac{k_1}{k_0}\frac{\left[n_j(\mathbf{q}) + \frac{1}{2} \pm \frac{1}{2}\right]}{\omega_j(\mathbf{q})}|F(\mathbf{Q})|^2\delta[E_1 - E_0 \pm \hbar\omega_j(\mathbf{q})]\delta(\mathbf{Q} \mp \mathbf{q} - \tau) \tag{6}$$

In this equation,

$$\mathbf{Q} = \mathbf{k}_0 - \mathbf{k}_1 \tag{7}$$

is the scattering vector; $[n_j(\mathbf{q}) + \frac{1}{2} \pm \frac{1}{2}]$ is the population factor with $n_j(\mathbf{q})$ given by Equation 4, the upper and lower signs correspond to phonon creation and phonon annihilation, respectively; the delta functions assure conservation of energy and momentum in the scattering process

$$\mathbf{Q} = \mathbf{q} + \tau \qquad E_1 - E_0 = \mp\hbar\omega_j(\mathbf{q}) \tag{8}$$

and the inelastic structure factor $F(\mathbf{Q})$ is defined by

$$F(\mathbf{Q})\sum_d M_d^{-1/2} b_d \mathbf{e}_d^j(\mathbf{q}) \cdot \mathbf{Q} e^{i\mathbf{Q}\cdot\mathbf{d}} e^{-W_d} \tag{9}$$

with the sum extending over the atoms of the unit cell. In this equation, M_d is the mass of the dth atom in the unit cell, b_d its coherent neutron scattering length, $\mathbf{d}$ its position vector, and e^{-W_d} the Debye-Waller factor.

PRACTICAL ASPECTS OF THE METHOD

The study of the coherent one-phonon inelastic neutron scattering from a single crystal is the most powerful technique for detailed investigation of the lattice dynamics of crystalline solids. Simple inspection of the coherent one-phonon neutron scattering cross-section (Equation 6) provides valuable information for the practical implementation of the technique.

The cross-section is inversely proportional to the phonon frequency and proportional to the population factor ($n_j + \frac{1}{2} \pm \frac{1}{2}$). Thus, for the low-lying modes ($\hbar\omega/kT < 1$), the intensity is inversely proportional to the square of the phonon frequency. The population factor is n_j for processes with neutron energy gain (phonon annihilation) and $(n_j + 1)$ for processes with neutron energy loss (phonon creation). Since $n_j(\mathbf{q})$ rapidly decreases with increasing $\hbar\omega/kT$, experiments with neutron energy loss are generally preferred. The k_1/k_0 factor in the cross-section, which favors energy-gain processes, cannot outweigh the gain in intensity due to the population factor for the energy-loss processes. Measurements with neutron energy gain, however, can be very useful in checking, whenever necessary, the results obtained by energy-loss measurements. Also, in experiments using neutrons of very low-incident energies (e.g., 5 meV), one is restricted in most cases to the observation of energy-gain processes. It should also be noted that because of the population factor, the intensity of the observed neutron groups usually increases with increasing sample temperature. Actually, for the low-lying modes [$(\hbar\omega/kT) \ll 1$], the intensity of the neutron groups is proportional to the sample temperature. Because of this gain in intensity with increasing temperature, most detailed studies of the dispersion curves have been performed at room or higher temperatures. The decrease in the Debye-Waller factor e^{-2W} with increasing temperature, with few exceptions, only moderates the increase of the intensity with increasing temperature. At high temperatures,

however, the increase in background scattering and the changes in the shape and width of the observed neutron groups, resulting from anharmonic effects, may considerably complicate the measurements.

The intensity of a neutron group corresponding to a certain mode depends on the crystal structure of the material through the square of the inelastic structure factor. It can be seen that there are definite advantages to performing the measurements, whenever possible, in mirror symmetry planes—such as the (110) and (001) planes in the case of cubic crystals. In fact, if the neutron scattering vector **Q** lies in a mirror plane, in principle no one-phonon scattering can be observed from modes polarized perpendicularly to the mirror plane. This considerably simplifies the task of identifying the modes with polarization vectors lying in the plane. In addition, the frequencies and polarization vectors of these modes vary slowly in the vicinity of and perpendicular to the mirror plane. As a result, the vertical collimation of the incident and scattered neutron beams can usually be much more relaxed than the horizontal collimation, thus obtaining more intensity without any significant loss in the accuracy of the measurements. The polarization dependence of the inelastic structure factor is exploited in practice to differentiate between various normal modes and to adopt the best experimental arrangement for their detection. For instance, if in a certain direction the modes are by symmetry purely longitudinal or transverse, the experimental configuration can be arranged so that the intensity is maximized for either the longitudinal or the transverse modes.

To illustrate these elementary considerations, let us consider the simple case of a monatomic face-centered cubic (fcc) structure like Cu. In this case, the modes along the [100], [110], and [111] directions are, by symmetry, purely longitudinal or transverse. Also, by symmetry, the transverse branches along the [100] and [111] directions are degenerate. The transverse branches along the [110] direction, on the other hand, are not degenerate and usually are denoted by T_2 (polarization parallel to the cube edge) and T_1 (polarization parallel to a face diagonal). Most of the measurements of the dispersion curves in these directions can be obtained by studying the scattering in the (110) mirror plane of the crystal. With the crystal in this orientation, all the longitudinal branches, the transverse branches along [111] and [100], and the T_2 [110] branch can be obtained. Typical scattering geometries for the detection of longitudinal and transverse [100] modes in the $(1\bar{1}0)$ scattering plane are indicated in Figure 1. The T_1 (110) branch can be observed by measurements in the (100) plane of the crystal. With the crystal in this orientation, measurements can be made along the [1ξ0] direction as well. These modes, usually designated Λ and Π, are not purely longitudinal or transverse. The dispersion curves of Cu are given in Figure 2. Notice that the measurements along the [110] direction were extended beyond the zone boundary to point X, where the L[110] branch is degenerate with T[100] and T_2[110] is degenerate with the L[100].

In simple cases, such as that of Cu, the symmetry of the structure determines the polarization vectors in high-symmetry directions and allows selection of the proper experimental arrangements. In more complicated structures with unit cells containing several atoms, planning of experiments is difficult, if not impossible, without some prior knowledge regarding the inelastic structure factor. In these cases, one proceeds by adopting a Born–von Kármán model with forces extending to a couple of nearest neighbors. By assigning reasonable values to the force constants (using, e.g., information available about the lattice dynamics of the system, such as measured values of the elastic constants), the structure factors can be evaluated. These calculated inelastic structure factors can then be used to plan the experimental measurements. As soon as some data have been obtained, the force constants can be readjusted, and a more realistic set of intensities can be calculated to serve as a guide for the selection of experimental conditions.

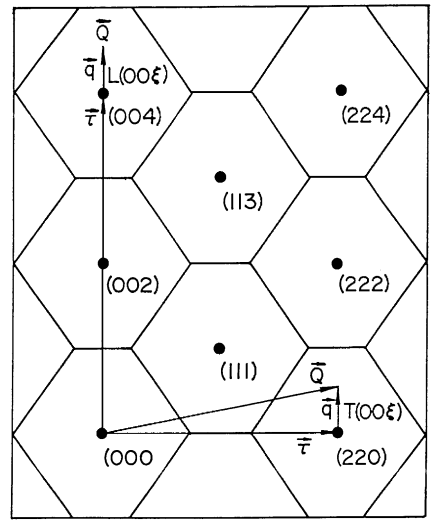


Figure 1. Neutron scattering amplitudes as a function of atomic weight. Adapted from Bacon (1975).

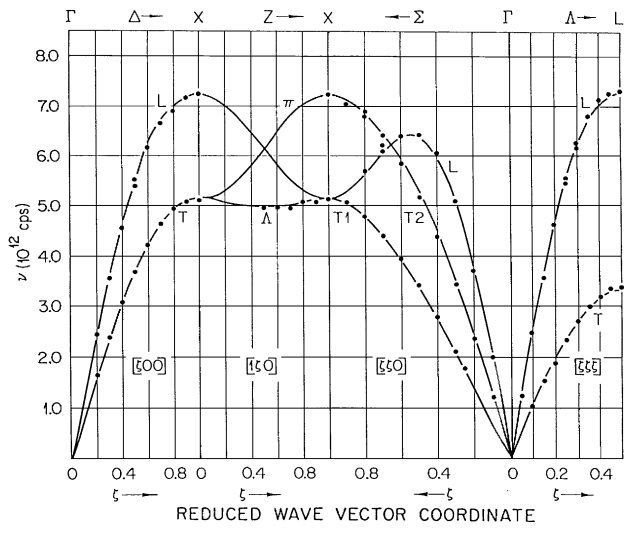


Figure 2. The molecular structure of a tin(II) hydride complex determined by a neutron diffraction study on IPNS SCD at 20 K (Koetzle et al., 2001). Estimated standard deviations in bond distances and angles are less than 0.01 Å and 0.1°, respectively.

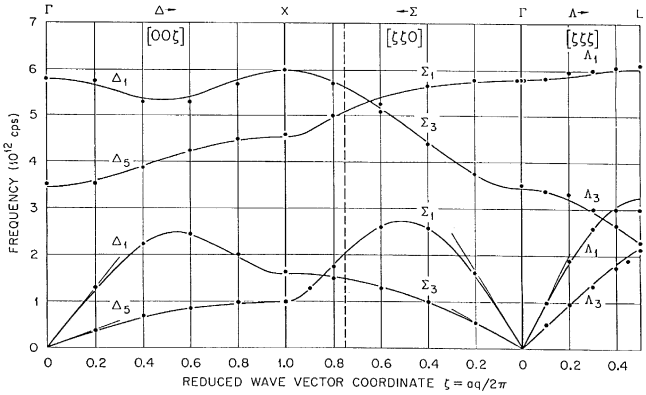


Figure 3. Contoured neutron Fourier map for (+)-neopentyl-1-d-alcohol (S) with positive nuclear scattering density at C and D shown as solid contours and negative nuclear scattering density at H shown as dashed contours (Yuan et al., 1994).

To illustrate this point, let us consider AgCl, which is a slightly more complicated case than that of Cu. This ionic crystal has the sodium chloride structure, and the Bragg intensities with even and odd Miller indices are proportional, respectively, to $(b_{Ag} + b_{Cl})^2$ (strong reflections) and $(b_{Ag} - b_{Cl})^2$ (weak reflections). By symmetry, the high-symmetry-direction branches are purely longitudinal or transverse. At Γ, the two atoms in the unit cell move in phaseand out of phase, respectively, for the acoustic and optical branches. It is easily seen by using Equation 9 that, close to Γ, measurements for the acoustic and optical branches should be performed around strong and weak reflections, respectively—a rule of thumb applicable to systems with two atoms per unit cell. This simple rule is only valid close to the zone center, however. In this particular experiment, the measurements away from Γ were planned using shell-model calculations of the inelastic structure factor. The measured dispersion curves of AgCl are given in Figure 3. Notice the similarity of the acoustic branches to those of the fcc structure, in particular the degeneracies at the point *X*.

The inelastic structure factor, and therefore the intensity, increases as Q^2, although this increase is moderated at large Q values by the Debye-Waller factor. This decrease of the Debye-Waller factor is not usually significant. The case of quantum crystals, such as ^{4}He, is a notable exception. In these crystals, the displacements of the atoms from their equilibrium positions are so large that the decrease with increasing Q of the Debye-Waller factor makes the observation of phonons impossible except around a few reflections with relatively small Miller indices. Thus, with some exceptions, such as the case of ^{4}He mentioned above, the measurements must be performed around reflections with large Miller indices.

Once the experimental conditions have been defined, following the general principles just outlined, the intensity of the observed neutron groups is determined by the intensity of the neutron source and the resolution of the instrument. Since the intensities vary inversely as a relatively high power (~4) of the overall resolution of the instrument, one must adopt a compromise between high resolution and reasonable intensity. With the neutron fluxes available in present-day sources, the energy resolution for a typical experiment is of the order of 1%, a relatively poor resolution compared to those ordinarily obtained in light-scattering experiments. There are essentially two methods allowing the determination of the neutron energy with such precision: (1) Bragg scattering from a single crystal (used as monochromator or analyzer) and (2) measuring the time of flight of the neutrons, chopped into pulses, over a known distance. A large number of instruments have been designed utilizing the first, the second, or a combination of these methods.

The choice of a particular instrument depends on the type of problem to be investigated. For single crystals, the region of interest in $(\mathbf{Q}\omega)$ space is considerably reduced by the symmetry of the crystal. Most of the relevant information regarding the lattice dynamics can be obtained by measuring a relatively small number of phonon frequencies at points or along lines of high symmetry of reciprocal space. For this type of experiment, one needs to perform a variety of rather specific scans in $(\mathbf{Q}\omega)$ space. It is generally accepted that, for these types of measurements, the triple-axis spectrometer has a definite advantage over other instruments. For incoherent scattering experiments, on the other hand, time-of-flight instruments are more convenient. These instruments are also preferred in coherent scattering experiments involving the measurement of a large number of frequencies in a relatively extended region of $(\mathbf{Q}\omega)$ space (as, e.g., in amorphous solids and complex compounds with a very large primitive unit cell).

The arguments presented here are mainly valid for instruments installed at nuclear reactors. Time-of-flight instruments are, of course, the natural machines to use with a pulsed source. Several time-of-flight machines, such as the multianalyzer crystal instrument and the constant-Q spectrometer, are presently being used at pulsed sources for coherent inelastic scattering experiments. Of these instruments, only the multianalyzer crystal instrument can reach the versatility of the triple-axis spectrometers (operating at continuous neutron sources) for coherent neutron scattering studies. Since the various instruments and their characteristics are described in detail in recent reviews (Dolling, 1974; Windsor, 1981, 1986), we present here only a few remarks regarding the use of the triple-axis spectrometer.

Triple-Axis Spectrometry

The design of present-day triple-axis spectrometers has changed little from that originally proposed by Brockhouse (1961). A schematic diagram of such an instrument is shown in Figure 4. A monochromatic beam of neutrons, obtained by Bragg reflection (through an angle $2\theta_M$) from a monochromator crystal, is scattered by the sample (through an angle ϕ) and the energy of this scattered beam is determined by Bragg scattering (through an angle $2\theta_A$) from an analyzer crystal. The orientation of the sample is defined by the angle χ between a certain crystal axis and the incident beam. The overall resolution of the instruments is governed by the collimators C_1, C_2, C_3, and C_4

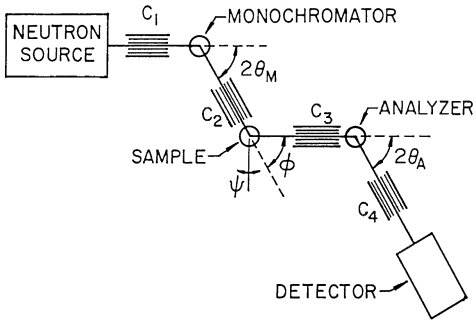


Figure 4. Anharmonicity of atomic displacements in zincblende (ZnS; m.p. 1973 K; cubic phase) as a function of temperature (Moss et al., 1980, 1983).

and the mosaic spreads of the crystals (monochromator, analyzer, and sample).

In some designs, no collimator (C_1) between the source and the monochromating crystal is used, since the natural collimation of the incident beam due to the relatively large distance between source and monochromator may be sufficient for most practical applications. Some horizontal collimation, however, is desirable, especially if it is easily adjustable, since it provides additional flexibility and reduces the overall background (for most experiments, a horizontal collimation before the monochromator between 40′ and 1° is sufficient). Since the distances between the sample and, respectively, the monochromator and the analyzer crystal are relatively short, it is necessary to introduce collimators (C_2, C_3) in order to restrict the horizontal divergence of the neutrons incident on and scattered from the specimen. In typical experiments, the horizontal collimation of the incident and scattered beam is 10′ to 40′. In most experiments no collimator (C_4) is used between the analyzer and the counter; a coarse (40′ to 2°) horizontal collimation can, however, reduce background if this is a serious problem in the measurements. Since most of the measurements in coherent neutron scattering experiments are performed in mirror planes, usually no vertical collimation is necessary. If, however, it is suspected that in a particular experiment scattering by modes off the horizontal scattering plane has been observed, the experiment should be repeated with some vertical collimation.

A low-sensitivity counter (usually a ^{235}U fission chamber) is placed before the sample to monitor the flux of neutrons incident on the sample. This monitor counter is normally used to control the counting times, so that no corrections are necessary for any changes in the incident neutron flux during the experiment. A second monitor counter is usually placed after the sample to detect Bragg scattering from the specimen, which may give rise to a spurious neutron group by incoherent scattering from the analyzer crystal. The signal counter is usually a high-efficiency (80% to 90%) $^{10}BF_3$ gas detector. Recently, the more compact ^3He gas detectors have replaced the $^{10}BF_3$ counters in several installations.

The intensity obtained in an experiment depends critically on the monochromator and analyzer crystals used in the spectrometer. In principle, the nuclei of the material used for a good monochromator (or analyzer) should have negligible absorption and incoherent scattering and fairly large coherent scattering cross-sections. Unfortunately, it is not always possible to obtain large single crystals of sufficient mosaic spread to obtain high reflectivity for all materials with favorable neutron characteristics. There are many methods of improving the reflectivities and other characteristics of the crystals used as monochromators (or analyzers), the most common being the mechanical distortion of the crystal at room or high temperatures. For details on this subject, we refer to the review by Freund (1976). The most useful monochromator (and analyzer) at present is pyrolytic graphite, which is ordered only along the c axis. In typical experiments, the (002) reflection of pyrolytic graphite is used at energies below ~50 meV. At higher energies the Bragg scattering angles become quite small [because of the relatively large lattice constant ($c = 6.71$ Å) of pyrolytic graphite]. This results in relatively poor energy resolution (since from Bragg's law $\Delta E = 2E \cot\theta_M \Delta\theta_M$). In addition, at these higher energies, the reflectivity decreases because of the scattering by parasitic reflections. Thus, at higher energies Be [usually reflecting from the (002) planes] is often used as a monochromator and/or analyzer in typical experiments. In practice, it is highly desirable to have many good monochromator crystals available in order to optimize intensity and resolution in a particular experiment. In most neutron scattering research centers, pyrolytic graphite as well as Be and (distorted) Ge crystals are available to be used as monochromators (or analyzers).

One of the most important advantages of the triple-axis spectrometer is its versatility. Highly specific scans along selected lines of the ($\mathbf{Q}\omega$) space can be performed under various experimental conditions, since both the neutron energy and the instrumental resolution can be easily adjusted during the experiments. There are several modes of operation. Irrespective of the mode of operation, however, the experiment determines ω as a function of $\mathbf{Q}$ (or $\mathbf{q}$). In the experimental plane, the energy and momentum conservation conditions determine the three unknowns, ω and the two components of $\mathbf{Q}$ along two axes in the scattering plane. Since in a crystal spectrometer there are four independently variable parameters (E_0, E_1, ϕ, and ψ), there are evidently a large number of ways in which to perform an experimental scan. Usually E_1 or E_0 is kept fixed during a particular scan.

The triple-axis spectrometer can, like a time-of-flight machine, be operated in the so-called conventional mode. In this mode $\mathbf{k}_0$ (and E_0) remain fixed, and the energies of the neutrons scattered at a fixed angle are analyzed by changing the setting of the analyzer crystal. As the magnitude (but not the direction) of $\mathbf{k}_1$ (and E_1) is varied, the extremity of the scattering vector moves through reciprocal space causing $\mathbf{q}$ to change, and neutron groups are observed if the energy and momentum conservation conditions are fulfilled. The disadvantage of this method is that it cannot be used to determine phonon frequencies at specified values of the wave vector $\mathbf{q}$, except by interpolation between values obtained in a series of experiments.

To determine the phonon frequencies at preselected values of $\mathbf{q}$, the constant-$\mathbf{Q}$ method is used. The ease

with which measurements with constant $\mathbf{Q}$ (and $\mathbf{q}$) can be performed with a triple-axis spectrometer is one of its main advantages over time-of-flight methods. The principle of this method, first introduced by Brockhouse (1960, 1961), is very simple. The choice of $\mathbf{q}$ and the reciprocal lattice point around which the scan is to be performed fix the components, say Q_x, Q_y, of the vector $\mathbf{Q}$, which is to remain constant during the scan. For instance, by assuming that E_1 is to be kept constant during the scan, a range of values for E_0 (or equivalently, $2\ \theta_M$) is chosen using a guess for the energy of the phonon. For each of these values of E_0, the appropriate values of ϕ and ψ are obtained from the momentum conservation conditions. If a neutron group is observed in the scan, its center will determine the frequency of a phonon with propagation vector $\mathbf{q}$. Similarly, the constant-$\mathbf{Q}$ scan can be performed keeping the incident neutron energy E_0 fixed. In most cases it is preferable to keep E_1 fixed, since in a constant-E_0 scan the variation in the sensitivity of the analyzer spectrometer as the scattered neutron energy varies during the scan can distort the observed neutron groups. In a constant-E_1 scan, no correction is needed for the k_1/k_0 factor in the cross- section since k_1 is fixed and k_0 is canceled by the $1/\nu_0$ sensitivity of the monitor counter. Another important advantage of the constant-$\mathbf{Q}$ method is that integrated intensities can be obtained and interpreted more easily than with other methods. This can be easily seen by momentarily neglecting the finite resolution of the instrument. The theoretical intensity is then obtained by simply integrating the theoretical cross-section over the phonon modes observed during an experimental scan. This matter is not trivial (Waller and Fröman, 1952), since E_0, E_1, k_0, and k_1, which appear in the delta functions, are related. For a general scan, the result of the integration is to multiply the expression preceding the delta functions by $|\mathrm{J}_j|^{-1}$, where the Jacobian J_j depends on the slope of the dispersion curve under study. This dependence on the slope of the branch being studied complicates the analysis of the measured intensities. For a constant-$\mathbf{Q}$ scan, on the other hand, $\omega_j(\mathbf{q})$ is independent of the energy transfer, since $\mathbf{q}$ is constant, and the integration is trivial, $\mathrm{J}_j = 1$.

In addition to the conventional and constant-$\mathbf{Q}$ scans, a variety of other scans can be programmed with a triple-axis spectrometer. One of the most useful is the constant-energy-transfer scan. In this type of scan, E_0 and E_1 are kept fixed and the angles ϕ and ψ are varied so that the measurement is performed with constant energy transfer along a predetermined line in reciprocal space. This type of scan is preferred over a constant-$\mathbf{Q}$ scan for very steep dispersion curves or whenever there is a sharp dip in a dispersion curve—as, e.g., in the study of the longitudinal phonon frequencies in the vicinity of $\mathbf{q} = \frac{2}{3}[111]$ of some body-centered cubic (bcc) metals.

Finally, a few comments should be made regarding the resolution of a triple-axis spectrometer and its focusing properties. Assume that the instrument has been set to detect neutrons scattered by the sample with momentum transfer $\mathbf{Q}_0$ and energy transfer $\hbar\omega_0$. Because of the finite collimation of the neutron beams and the finite mosaic spreads of the crystals (monochromator, analyzer, and sample), there is also a finite probability for the instrument to detect neutrons experiencing momentum transfer $\mathbf{Q}$ and energy transfer $\hbar\omega$ differing from $\mathbf{Q}_0$ and $\hbar\omega_0$ by $\Delta\mathbf{Q}$ and $\hbar\Delta\omega$, respectively. This probability is usually denoted by $R(\mathbf{Q} - \mathbf{Q}_0, \omega - \omega_0)$ and defines the resolution or transmission function of the instrument. The intensity observed for a particular setting of the spectrometer $(\mathbf{Q}_0, \omega_0)$ is then simply the convolution of the one-phonon coherent scattering cross-section with the resolution function around $\mathbf{Q}_0, \omega_0$:

$$I(\mathbf{Q}_0, \omega_0) = \int R(\mathbf{Q} - \mathbf{Q}_0, \omega - \omega_0)\sigma(\mathbf{Q}, \omega)d\mathbf{Q}\,d\omega \qquad (10)$$

The course of any particular scan can then be visualized as the stepwise motion of the resolution function through the dispersion surface. If one assumes that the transmission functions of the collimators and single crystals (monochromator, analyzer, and sample) are Gaussian distributions, a relatively simple analytical expression for the resolution of the instrument can be obtained (Cooper and Nathans, 1967; Mller et al., 1970),

$$R(\mathbf{Q} - \mathbf{Q}_0, \omega - \omega_0) = R_0 \exp\left\{-\frac{1}{2}\sum_{k,l=1}^{4} M_{kl}x_k x_l\right\} \qquad (11)$$

where $x_i = \Delta Q_i (i = 1, 2, 3)$, $x_4 = \Delta\omega$, and R_0 and M_{kl} are rather complicated expressions specified by the collimations and the mosaic spreads of the crystals. In writing this expression, x_1 was taken along $\mathbf{Q}_0$, and x_3 was vertical. The surface over which $R = R_0/2$ is represented by

$$\sum_{k,l=1}^{4} M_{kl}x_k x_l = 1.386 \qquad (12)$$

and is referred to as the resolution ellipsoid. Since the experiment is performed in a plane that is usually horizontal, in practically all cases, one is only concerned with the three-dimensional ellipsoid in the $\Delta Q_| = \Delta Q_\parallel, \Delta Q_2 = \Delta Q_\perp, \Delta\omega$ space (where $\Delta Q_\parallel$ and $\Delta Q_\perp$ denote the components of $\Delta\mathbf{Q}$ parallel and perpendicular to $\mathbf{Q}_0$, respectively). Typically, the resolution ellipsoid is quite elongated along $\Delta\omega$, and it is more elongated along $\Delta Q_\parallel$ than along $\Delta Q_\perp$. The elongated nature of the resolution ellipsoid is responsible for the focusing characteristics of the spectrometer: the width of the observed neutron groups will depend on the orientation of the long axes of the ellipsoid with respect to the dispersion surface. Clearly (see Fig. 5), the sharpest neutron group will be obtained if the experimental configuration is such that the long axis of the resolution ellipsoid is parallel to the dispersion surface. Because of the shape of the ellipsoid, focusing effects are more pronounced for transverse acoustic phonons (with $\Delta_q\omega$ perpendicular to $\mathbf{Q}_0$) than for longitudinal phonons (with $\Delta_q\omega$ parallel to $\mathbf{Q}_0$). In the experiments, the transverse phonons are always measured with a focused configuration. Sometimes, it may also be useful to measure longitudinal phonons not in a purely longitudinal configuration ($\mathbf{Q} \| \mathbf{q}$) but in a configuration favorable for focusing. To determine the focused configuration for a given

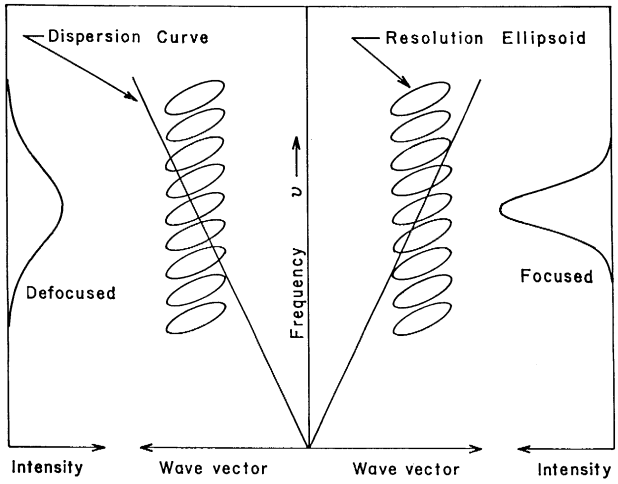


Figure 5. Schematic plan of the SCD spectrometers formerly installed at the dual thermal beam port H6 at the Brookhaven High-Flux Beam Reactor (HFBR).

instrument, one usually measures the neutron groups corresponding to a low-lying transverse acoustic phonon with **Q** taken, in the clockwise and counterclockwise sense, respectively, from a selected reciprocal lattice point. The sharpest of the two neutron groups (see Fig. 6) indicates which of the two configurations is focused, and thus establishes in which direction (clockwise or counterclockwise) from a reciprocal point the propagation vector of a transverse phonon (with ω increasing with increasing $\mathbf{q}$) should be taken to exploit focusing.

Computer programs for the evaluation of the resolution function are available in all neutron scattering centers. In many studies (e.g., the determination of intrinsic phonon widths), such detailed calculations of the resolution function are essential. For the planning of most experiments, however, simple estimates of the resolution are sufficient since they provide the information needed for the proper choice of the instrumental parameters. The contribution to the energy resolution of the spectrometer of each collimator and crystal involved can be evaluated, and an estimate of the overall energy resolution can be obtained by taking the square root of the sum of the squares of these contributions. Each contribution, $\Delta\omega$, to the energy resolution is obtained (Stedman and Nilsson, 1966), from

$$\Delta\omega = \Delta\mathbf{k} \cdot \left[\frac{\hbar}{m}\mathbf{k} - \nabla_q \omega(\mathbf{q})\right] \tag{13}$$

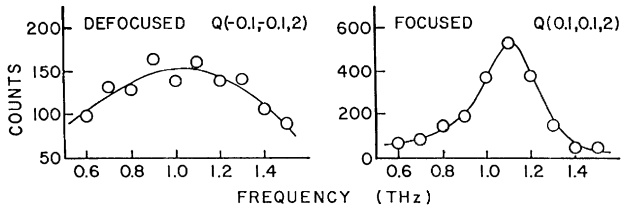


Figure 6. Schematic plan of the SCD spectrometer installed on a liquid methane moderator at the Argonne Intense Pulsed Neutron Source (IPNS).

In this equation, $\mathbf{k}$ stands for the average neutron wave vectors ($\mathbf{k}_0$ for the incident or $\mathbf{k}$, or the scattered beam) and $\Delta\mathbf{k}$ for the spread around these average values introduced by the collimators and finite mosaic spreads of the monochromator and analyzer crystal. To be more specific, let us assume that the monochromator is set to reflect neutrons of wave vector $\mathbf{k}_0$. As a result of the finite collimation α of the beam and the mosaic width β of the monochromator, the distribution of incident neutron wave vectors around $\mathbf{k}_0$ has a width $\Delta\mathbf{k}_0 = \Delta\mathbf{k}_{0c} + \Delta\mathbf{k}_{0\eta}$ (where $\Delta\mathbf{k}_{0c}$ from the collimation and $\Delta\mathbf{k}_{0\eta}$ from the mosaic width of the monochromator). It is easily shown, using Bragg's law, that $\Delta\mathbf{k}_{0c}(|\Delta\mathbf{k}_{0c}| = \alpha k_0 \theta_M)$ is parallel to the reflecting planes of the monochromator, whereas $\Delta\mathbf{k}_{0\eta}(|\Delta\mathbf{k}_{0\eta}| = \beta k_0 \cot\theta_M)$ is along $\mathbf{k}_0$. Notice that the distribution has a pronounced elongation along $\Delta\mathbf{k}_{0c}$, since $|\Delta\mathbf{k}_{0c}|/|\Delta\mathbf{k}_{0\eta}|$ is typically between 2.5 and 5, a feature alluded to earlier in this section. It can be seen from Equation 13 that to minimize this contribution to the energy resolution, the experimental arrangement should be such that the vector $[(\hbar\mathbf{k}_0/m) - \nabla_q\omega]$ is approximately normal to the monochromator reflecting planes, and this is the focusing condition at the monochromator. When this condition is approximately satisfied, the spread of incident energy around E_0 due to the distribution of incident wave vectors (first term in the right-hand side of Equation 13) approximately matches the spread in ω (second term of Equation 13) brought about by the spread in $\mathbf{q}$ due to Δk_0. Similar considerations apply to the scattered beam. In particular, the configuration of the analyzer spectrometer can be chosen to satisfy the focusing condition: $[(\hbar\mathbf{k}_0/m) - \nabla_q\omega]$ parallel to the normal to the analyzer reflecting planes. By using such simple considerations, the optimum focusing configuration can be adopted and a reasonable estimate of the resolution of the instrument can be obtained. The resolution widths calculated by this simple method are accurate to within $\sim$15%, which is sufficient for typical experiments.

METHOD AUTOMATION

Presently, all neutron scattering spectrometers used for the study of phonons are computer operated. By simple commands, the user can direct the instrument to perform any given scan or a series of scans with different parameters. Of course, software limits must be set so that the spectrometer does not assume a configuration incompatible with the experimental setup. In a triple-axis spectrometer, it is recommended to use limit-switches, in addition to the software limits, to restrict the motion of the monochromator drum, counter, and sample table within the limits set by a particular experimental setup. This is particularly important if the sample is in a cryostat, dilution refrigerator, or superconducting magnet.

In practically all neutron scattering spectrometers, the data can be printed and/or displayed graphically. In addition, in most spectrometers the user can perform a preliminary analysis of the collected data by using the computer programs available at the instrument site.

DATA ANALYSIS

Over the last 35 years, considerable progress has been made toward understanding of the connection between phonon measurements and the crystalline and electronic structure of solids. Until the pioneering work of Toya (1958), on the dispersion curves of Na, the measured dispersion curves of solids were analyzed in terms of simple phenomenological models based on the Born–von Kármán formalism.

As mentioned earlier, the vibration frequencies of a crystalline solid, in the harmonic approximation, are determined by a $3r \times 3r$ force constant (or dynamical) matrix, which depends on the forces between the nuclei. In a Born–von Kármán model, only forces extending to a few nearest neighbors are introduced. The number of independent force constants one needs to introduce is considerably reduced by the crystal symmetry and the requirement of translational and rotational invariance of the solid as a whole. The force constants introduced are considered as adjustable parameters to be determined by fitting to the experimental data. With few exceptions, however, Born–von Kármán models with a few adjustable parameters cannot adequately describe the measured dispersion curves. This is because of the different ways with which the electrons screen and modify the internuclear Coulomb forces in the various classes of materials. By making reasonable assumptions about the response of the electronic medium, however, a large number of phenomenological models have been advanced to explain the lattice dynamical properties of various classes of solids. A detailed description of these phenomenological models can be found in many books (Born and Huang, 1954; Bilz and Kress, 1979; Hardy and Karo, 1979; Brüesch, 1982; Venkataraman et al., 1975; Horton and Maradudin, 1974) and review articles (Hardy, 1974; Cochran, 1971; Sinha, 1973).

Phenomenological models are extremely useful since they can be used to calculate the phonon density of states and the lattice specific heat. Also, they are used to calculate the inelastic structure factors that are invaluable in planning experimental measurement of the dispersion curves of solids containing several atoms per unit cell. The main drawback of phenomenological models is that in most cases the parameters introduced to characterize the electronic response to the nuclear motion have no clear physical interpretation. In many cases, it is difficult, if not impossible, to understand how the results are related to the detailed electronic structure of materials, as provided by energy-band calculations.

Following the work of Toya (1958) on the dispersion curves of Na, the lattice dynamics of simple metals was developed within the framework of the pseudopotential theory of metals (Harrison, 1966; Heine, 1968; Cohen and Heine, 1968; Heine and Weaire, 1968). For most of the other classes of materials, however, it was impossible until the mid-1970s to relate the measured dispersion curves to the electronic properties of the materials. Progress in computational techniques has presently made possible first-principles calculations. The most direct approach describes the electronic response to the nuclear motion by a dielectric matrix (or electronic susceptibility). In practical applications, however, one is faced with the problem of inverting this dielectric matrix, a rather formidable problem. Because of this problem, only few calculations have been performed to date with this direct (or inversion) method.

Various methods have been proposed to overcome the computational problems of the dielectric function formalism of the general theory. The basic idea is to separate the band structure contribution to the dynamical matrix. One of the most commonly used methods is that put forward by Varma and Weber (1977, 1979). In this approach, the dynamical matrix is reduced to

$$D = D_{\mathrm{sr}} + D_2 \tag{14}$$

where D_{sr} is determined by relatively short-range forces and D_2 is the main band structure contribution. In practical applications the only calculation needed is that of D_2, since D_{sr} is parametrized by using short-range force constants. Many phonon anomalies observed in the dispersion curves of metals have been explained by this method.

Finally, it should be emphasized that presently it is possible to evaluate phonon frequencies directly and to compare them with the experimental results. These *ab initio* or frozen phonon calculations (Chadi and Martin, 1976; Wendel and Martin, 1978, 1979; Yin and Cohen, 1980; Kunc and Martin, 1981; Harmon et al., 1981; Yin and Cohen, 1982; Kunc and Martin, 1982; Ho et al., 1982) are particularly powerful in relating observed phonon anomalies in the dispersion curves to the electronic properties of the solid. To illustrate the principle of the method, consider a perfect (monatomic) crystal and another obtained from the first by displacing the atoms to the positions they assume instantaneously in the presence of a lattice wave of wave vector $\mathbf{q}$ and frequency $\omega(\mathbf{q})$. Clearly, the second lattice is obtained from the perfect lattice by displacing the atoms, at the equilibrium positions $\mathbf{l}$ by

$$\Delta\mathbf{R}_l = \mathbf{u}_0 \cos(\mathbf{q} \cdot \mathbf{l}) \tag{15}$$

where $\mathbf{u}_0$ defines the amplitude and direction of the displacement. The energy difference per atom, ΔE, between the two lattices is simply the potential energy per atom for the "frozen" phonon mode,

$$\Delta E = \frac{1}{N}\sum_l \frac{1}{2} M\omega^2(\mathbf{q})|\delta\mathbf{R}_l|^2 \tag{16}$$

which can be written as

$$\Delta E = \frac{1}{2} M\omega^2(\mathbf{q})\left[\frac{1}{N}|\mathbf{u}_0|^2 \sum_l \cos^2(\mathbf{q} \cdot \mathbf{l})\right] \tag{17}$$

For a zone boundary phonon [$\mathbf{q} = (\tau/2)$], the expression in brackets is simply $|\mathbf{u}_0|^2$, whereas for arbitrary $\mathbf{q}$, it is $\frac{1}{2}|\mathbf{u}_0|^2$. To obtain the phonon frequency, the energy difference ΔE is evaluated as a function of the displacement $|\mathbf{u}_0|$ by calculating the total energies of the distorted and

perfect crystal. The phonon frequency is then obtained from the curvature of the energy-difference versus displacement curve.

The method is used to obtain frequencies of phonon modes with wave vector $\mathbf{q}$ such that $n\mathbf{q} = \tau$ (where n is an integer), since only in this case is the distorted crystal also periodic (with a larger real-space cell), and its total energy can be calculated by the same techniques as those used to obtain the total energy of the perfect crystal. The computational times increase, of course, as the number of atoms in the larger cell of the distorted crystal increases. Accurate calculations of total energies for crystal structures with 10 to 15 atoms per cell are presently feasible.

Frozen phonon calculations have been performed for simple metals and semiconductors, as well as transition metals. In general, the agreement between calculated and experimentally determined frequencies is excellent. In a slightly different approach (Kunc and Martin, 1982; Yin and Cohen, 1982), the phonon dispersion curves along high symmetry directions can be obtained by calculating the force constants between atomic layers perpendicular to the direction of wave propagation. In these calculations, a supercell containing a sufficient number of layers is adopted and the interlayer force constants are obtained from the Hellmann-Feynman forces acting on individual atomic layers when a certain layer is displaced along the direction of wave propagation.

Probably the most appealing aspect of this method is that it provides not only phonon frequencies, but also additional information for comparison with the experimental results: lattice constants, bulk moduli, and elastic constants (obtained by subjecting the perfect crystal to a homogeneous deformation). Also, by studying the departure from a parabola of the energy versus displacement curves, one can obtain information about anharmonic effects. In addition, extra minima in the energy versus displacement curves may indicate instability of the lattice toward a certain phase transformation. It should be pointed out, however, that the calculations are valid essentially at $T = 0$, a fact that is sometimes overlooked in comparing the calculations to the experimental results.

In practice, the analysis and interpretation of the results is facilitated considerably if good quality data are collected. Given a sample of good quality this can be achieved by minimizing the background, optimizing the resolution, and adopting sufficient counting times for satisfactory counting statistics. The phonon frequencies, obtained by fitting the phonon peaks, are assigned to the various branches using Born–von Kármán model calculations. The dispersion curves are then fitted to those given by various theoretical or phenomenological models. A preliminary analysis of the results can be made as the experiment is in progress by making use of the programs available in practically all neutron scattering facilities.

SAMPLE PREPARATION

A few remarks should be made regarding the samples used for phonon measurements. From the expression of the inelastic structure factor, it is evident that light elements with relatively large coherent cross-sections are favored for coherent inelastic scattering experiments. Actually, the small mass of ^{4}He made possible the measurements of the dispersion curves of this quantum crystal. If the incoherent cross-section that contributes to the background is not sufficiently small, it may also complicate the measurements. In such cases the use of single isotopes, as in materials containing Ni, can be very helpful. The use of appropriate isotopes may actually be essential if the material to be studied has a significant neutron absorption cross-section (much larger than $\sim$100 b). For instance, the use of nonabsorbing isotopes made possible the study of the dispersion curves of Cd (Chernyshov et al., 1979; Dorner et al., 1981). In such extreme cases, however, one must seriously examine the possibility of obtaining the desired information by an x-ray diffuse scattering experiment instead (see X-RAY AND NEUTRON DIFFUSE SCATTERING MEASUREMENTS). For instance, such measurements have been performed on V, which, because of its low coherent scattering length (0.05×10^{-12} cm), is difficult to study with standard coherent inelastic scattering techniques.

For coherent neutron scattering experiments, it is almost essential to have a single-crystal specimen whose size is dictated by the material properties and the intensity of the neutron source. For most materials, information about lattice dynamics can be obtained in a high-flux source on samples as small as 0.05 cm^3. In general, for detailed studies, one needs specimens of the order of 1 cm^3 or larger. If it is available, a single crystal of high perfection (mosaic spread 0.2° should be used. In high-temperature or -pressure experiments, as well as in studies where the sample is brought through a phase transformation, however, the experimentalist may have little, if any, control of the mosaic spread of the sample. In such cases, measurements may have to be performed on crystals with mosaic spreads as large as 5°. Difficulties arising from loss in experimental resolution and Bragg-phonon scattering processes in such cases are discussed in the Problems section.

When possible, high-purity samples should be used, since impurities as well as defects increase the background scattering. It should be noted, however, that impurity contents as high as 1000 ppm have, in most cases, very little effect on the dispersion curves of the material being studied. Significant hydrogen content, on the other hand, can complicate the experiments considerably. In addition to its large incoherent cross-section, which considerably increases the background, hydrogen can precipitate as a hydride at a certain temperature and cause a large increase in the mosaic of the sample. This is particularly troublesome in studies of lattice dynamics as a function of temperature.

As mentioned earlier, most of the detailed coherent neutron scattering studies have been performed on single-crystal specimens. Recently, however, it has been demonstrated (Buchenau et al., 1981) that considerable information about the lattice dynamics of materials can be obtained by studies of the coherent neutron scattering by polycrystalline samples. In the case of Ca, the dispersion curves obtained by this technique were found to be in good agreement with measurements performed on

single-crystal specimens (Stassis et al., 1983). For materials for which relatively large single crystals are not available, study of the coherent scattering by a polycrystalline sample may prove to be extremely valuable.

SPECIMEN MODIFICATION

In general, the scattering of thermal neutrons is a nondestructive experimental technique. However, after irradiation practically all samples become to some extent radioactive. It is therefore important before performing an experiment to estimate the activity of the samples after irradiation and to avoid, if possible, the use of elements that upon irradiation become long-lived radioactive isotopes. After the experiment is completed, users are required to have the samples checked by the institutional radiation safety office (Health Physics). In many cases the sample is stored for several days or weeks to allow its activity to decay to admissible levels before it is shipped to the user.

PROBLEMS

Collection of data on a triple-axis spectrometer is complicated by the observation of peaks arising from higher-order contamination. If a monochromator is set to reflect neutrons of energy E_0 from the (hkl) plane, it is also set to reflect in the same direction neutrons of energy $n^2 E_0$ from the $(nh\ nk\ nl)$ plane. Similarly, if the analyzer is set to reflect neutrons of energy E_1, it will also reflect in the direction of the counter neutrons of energy $m^2 E_1$. It is easily seen that, as a result, various spurious neutron groups may be observed. One of the most common is due to the elastic incoherent scattering by the sample, in the direction of the analyzer, of the nth-order component of the incident beam. Clearly, if n and m are such that $n^2 E_0 = m^2 E_1 = m^2(E_0 \pm \hbar\omega)$, a spurious neutron group will be observed that could be attributed to the creation (or annihilation) of a phonon of energy $\hbar\omega$. If such a problem is suspected, one can repeat the measurement using a sufficiently high E_1 (or E_0) so that this process is not energetically possible for any values of n and m. Quite generally, by measurements at different energies one can avoid ambiguities arising from spurious neutron groups due to higher-order contamination. This problem is serious, however, in studies of complex systems, and various methods are used to minimize it.

The simplest method to deal with the problem of order contamination is to choose the incident energy close to the maximum of the reactor spectrum so that the higher-order contamination in the incident beam arises from the low-flux region of the spectrum. In most experiments, however, one needs lower energies for better energy resolution. It is, therefore, highly desirable to decrease the flux of higher-energy neutrons incident on the monochromator crystal. This can be achieved by placing a single crystal of high perfection before the monochromator. The choice of the best filter for this purpose is restricted by the difficulty of obtaining large, highly perfect, single crystals of the desired material. Quartz, silicon, and sapphire have been used with considerable success. Of these, silicon and quartz are used at liquid nitrogen temperatures and have similar characteristics. Sapphire seems to be more effective and can be used at room temperature without significant loss of efficiency. Note that mechanical velocity selectors and curved neutron guides can also be used for the same purpose. For instance, an S-shaped neutron guide is installed between the cold source and the IN-12 triple-axis spectrometer of the high flux reactor at the Institut Laue-Langevin.

Even when the higher-energy neutron flux incident on the monochromator has been minimized as described above, additional discrimination of the higher-order contamination is desirable in a large number of experiments. This is usually achieved by using appropriate filters. Pyrolitic graphite is probably the most useful, and it is used routinely in typical experiments. It is particularly efficient in a narrow energy range around 13.7 and 14.7 meV. Pyrolitic graphite filters are also quite efficient at 30.5 and 42 meV. The filter, typically several inches thick, is oriented so that the c axis is along the beam and is positioned in front of either the analyzer or the sample (depending on whether the experiment is performed at fixed scattered or incident neutron energy). For experiments at low energies ($\leq$5 meV), polycrystalline filters such as Be are frequently used. In this case, one exploits the fact that no coherent elastic scattering occurs if the neutron wavelength exceeds twice the largest spacing between planes in the material. Finally, it should be mentioned that one can eliminate a specific higher-order contamination by appropriate choice of the analyzer or monochromator. For instance, one of the most troublesome processes is the incoherent elastic scattering by the sample of the primary beam, in the direction of the analyzer, followed by second-order scattering from the analyzer $(n = 1, m = 2)$. This process can be eliminated by using Ge (or Si) reflecting from the (111) planes as analyzer, since for all practical purposes the (222) reflection is forbidden for diamond-structure crystals.

Since we discussed in some detail how to deal with the spurious neutron groups arising from higher-order contamination, it is appropriate now to mention three other processes that may introduce similar problems. The first process, alluded to earlier in this section, involves Bragg scattering of the primary incident beam (of energy E_0) by the sample in the direction of the scattered beam followed by incoherent scattering by the analyzer in the direction of the counter. In the second process, neutrons with energy E_1 in the incident beam may be Bragg reflected from the sample and then from the analyzer into the detector. In both cases, a relatively high count rate will be observed in the second monitor, which is placed behind the sample. The problem is then easily eliminated by a slight change of experimental conditions. The third case involves a double-scattering process in the sample: coherent one-phonon scattering followed or preceded by Bragg scattering. This process is particularly troublesome since it gives rise to neutron groups corresponding to modes that in principle should not be observed under the experimental conditions of the measurement. For instance, quite often one observes in a longitudinal scan, neutron groups corresponding

to transverse modes. In principle, this problem can be practically eliminated by using a single crystal of small mosaic spread. In experiments with crystals of large mosaic spreads (as is often the case in high-pressure or -temperature studies), this double-scattering process can complicate the measurements considerably.

Finally, it should be mentioned that radiation damage of the sample is not usually a problem because the sample is exposed to the relatively low intensity and low energy beam reflected by the monochrometer. A notable exception is the cast of some classes of biological materials.

LITERATURE CITED

Bilz, H. and Kress, W. 1979. Phonon Dispersion Relations in Insulators. Springer-Verlag, Berlin and New York.

Born, M. and Huang, K. 1954. Dynamical Theory of Crystal Lattices. Oxford University Press, London.

Born, M. and von Kàrmàn, T. 1912. *Z. Phys.* 14:65.

Born, M. and von Kàrmàn, T. 1913. *Z. Phys.* 13:297.

Brockhouse, B. N. 1960. *Bull. Am. Phys. Soc.* 5:462.

Brockhouse, B. N. 1961. *In* Inelastic Scattering of Neutrons in Solids and Liquids, p. 113. IAEA, Vienna.

Brockhouse, B. N. and Stewart, A. T. 1955. *Phys. Rev.* 100:756.

Brüesch, P. 1982. Phonons: Theory and Experiments. Springer-Verlag, Berlin and New York.

Buchenau, U., Schober, H. R., and Wagner, R. 1981. *J. Phys.* (*Orsay, France*) 42:C6–365.

Burstein, E. 1967. *In* Phonons and Phonon Interactions (T. A. Bak, ed.) p. 276. W. A. Benjamin, New York.

Cardona, M. and Güntherodt, G. 1975–1991. Light Scattering in Solids. Vols. 1–6, Springer-Verlag, Berlin.

Chadi, J. D. and Martin, R. M. 1976. Solid State Commun. 19:643.

Chernyshov, A. A., Pushkarev, V. V., Rumpantsev, A. Y., Dorner, B., and Pynn, R. 1979. *J. Phys. F* 9:1983.

Cochran, W. 1971. *Crit. Rev Solid State Sci.* 2:1.

Cohen, M. L. and Heine, V. 1968. *Solid State Phys.* 24:38 [and references cited therein].

Cooper, M. J. and Nathans, R. 1967. *Acta Crystallogr.* 23:357.

Debye, P. 1912. *Ann. Phys.* (*Leipzig)* 39:789.

Dolling, G. 1974. *In* Dynamical Properties of Solids (G. K. Horton and A. A. Maradudin, eds.) Vol. 1, p. 541. North-Holland, Amsterdam.

Dorner, B., Chernysov, A.A., Pushkarev, V. V., Rumyantsev, A. Y., and Pynn, R. 1981. J. Phys. F 11:365.

Einstein, A. 1907. *Ann. Phys.* (*Leipzig*) 22:180.

Einstein, A. 1911. *Ann. Phys.* (*Leipzig*) 35:679.

Freund, A. 1976. *In* Proceedings of the Conference on Neutron Scattering. Gatlinburg, Tenn. p. 1143.

Hardy, J. R. 1974. *In* Dynamical Properties of Solids (G. K. Horton and A. A. Maradudin, eds.) Vol. 7, p. 157. North-Holland, Amsterdam.

Hardy, J. R. and Karo, A. M. 1979. The Lattice Dynamics and Statics of Alkali Halide Crystals. Plenum, New York.

Harmon, B. N., Weber, W., and Hamann, D. R. 1981. *J. Phys.* (*Orsay, France*) 42:C6–628.

Harrison, W. A. 1966. Pseudopotentials in the Theory of Metals. Benjamin, New York.

Heine, V. 1968. *Solid State Phys.* 24:1 [and references cited therein].

Heine, V. and Weaire, D. 1968. *Solid State Phys.* 24:250 [and references cited therein].

Ho, K.-M., Fu, C.-L, Harmon, B. N., Weber, W., and Hamann, D. R. 1982. *Phys. Rev. Lett.* 49:673.

Horton, G. K. and Maradudin, A. A. (eds.) 1974. Dynamical Properties of Solids, Vols. 1 to 4. North-Holland, Amsterdam.

Kunc, K. and Martin, R. M. 1981. *J. Phys.* (*Orsay, France)* 42:C6–649.

Kunc, K. and Martin, R. M. 1982. *Phys. Rev. Lett.* 48:406.

Mller, H. Bjerrum, and Nielsen, M. 1970. *In* Instrumentation for Neutron Inelastic Scattering Research. p. 49. IAEA, Vienna.

Nicklow, R. M. Gilat, G., Smith, H. G., Raubenheimer, L. J., and Wilkinson, M. K. 1967. Phys. Rev. 164:922.

Placzek, G. and Van Hove, L. 1954. *Phys. Rev.* 93:1207.

Sinha, S. K. 1973. *Crit. Rev. Solid State Sci.* 4:273.

Sköld, K. and Price, D. L. (eds.) 1986. Neutron Scattering (Methods of Experimental Physics, Vol. 23). Academic, New York.

Stassis, C., Zarestky, J., Misemer, D. K., Skriver, H. L., and Harmon, B. N. 1983. *Phys. Rev. B* 27:3303.

Stedman, R. and Nilsson, G. 1966. *Phys. Rev.* 145:492.

Toya, T. 1958. *J. Res. Inst. Catal. Hokkaido Univ.* 6:183.

Varma, C. M. and Weber, W. 1977. *Phys. Rev. Lett.* 39:1094.

Varma, C. M. and Weber, W. 1979. *Phys. Rev B* 19:6142.

Venkataraman, G., Feldkamp, L. A., and Shani, V. C. 1975. Dynamics of Perfect Crystal. MIT Press, Cambridge, Mass.

Vijayaraghavan, P. R., Nicklow, R. M., Smith, H. G., and Wilkinson, M. K. 1970. *Phys. Rev. B* 1:4819.

Waller, I. and Fröman, P. O. 1952. *Ark. Fys.* 4:183.

Wendel, H. and Martin, R. M. 1978. *Phy. Rev. Lett.* 40:950.

Wendel, H. and Martin, R. M. 1979. *Phys. Rev. B* 19:5251.

Windsor, C. G. 1981. Pulsed Neutron Scattering. Taylor & Francis, Halsted Press, New York.

Windsor, C. G. 1986. *In* Methods of Experimental Physics (K. Sköld and D. L. Price, eds.) Vol. 23, p. 197. Academic, New York.

Yin, M. T. and Cohen, M. L. 1980. *Phys. Rev. Lett.* 45:1004.

Yin, M. T. and Cohen, M. L. 1982. *Phys. Rev. D* 25:4317.

Yin, M. T. and Cohen, M. L. 1982. *Phys. Rev. B* 26:3259 [and references cited therein].

KEY REFERENCES

Born and Huang, 1954. See above.

Although quite old, still the best introduction to the subject of lattice dynamics.

Venkataraman, Feldkamp, and Sahni, 1975. See above.

A relatively up-to-date treatment of lattice dynamics. It contains a detailed outline of the Born-von Kármán formalism.

Sköld and Price (eds.), 1986. See above.

An excellent source for information regarding phonons in various materials. The first part is a concise introduction to neutron scattering.

BIBLIOGRAPHY

Bilz, H. and Kress, W. 1979. Phonon Dispersion Relations in Insulators. Springer-Verlag, Berlin and New York.

Born, M. and Huang, K. 1954. Dynamical Theory of Crystal Lattices. Oxford University Press, London.

Brillouin, L. 1953. Wave Propagation in Periodic Structures. Dover, New York.

Brüesch, P. 1982. Phonons: Theory and Experiments. Springer-Verlag, Berlin and New York.

Choquard, Ph. 1967. The Anharmonic Crystal. Benjamin, New York.

Cochran, W. 1973. The Dynamics of Atoms in Crystals. Arnold, London.

Decius, J. C. and Hexter, R. M. 1977. Molecular Vibrations in Crystals. McGraw-Hill, New York.

Donovan, B. and Angress, J. F. 1971. Lattice Vibrations. Chapman & Hall, London.

Dorner, B. 1982. Coherent Inelastic Neutron Scattering in Lattice Dynamics. Springer-Verlag, Berlin and New York.

Hardy, J. R. and Karo, A. M. 1979. The Lattice Dynamics and Statics of Alkali Halide Crystals. Plenum, New York.

Horton, G. K. and Maradudin, A., eds. 1974. Dynamical Properties of Solids, Vols. 1 to 4. North-Holland, Amsterdam.

Ludwig, W. 1967. Recent Developments in Lattice Theory. Springer-Verlag, Berlin and New York.

Maradudin, A. A., Ipatova, I. P., Montroll, E. W., and Weiss, G.H. 1971. Theory of Lattice Dynamics in the Harmonic Approximation. *In* Solid States Physics, Suppl. 3, 2nd Ed. Academic, New York.

Reissland, J. A. 1973. The Physics of Phonons. Wiley, New York.

Venkataraman, G., Feldkamp, L. A., and Sahni, V. C. 1975. Dynamics of Perfect Crystals. MIT Press, Cambridge, Mass.

C. Stassis
Iowa State University
Ames, Iowa

MAGNETIC NEUTRON SCATTERING

INTRODUCTION

The neutron is a spin-$\frac{1}{2}$ particle that carries a magnetic dipole moment of -1.913 nuclear magnetons. Magnetic neutron scattering then originates from the interaction of the neutron's spin with the unpaired electrons in the sample, either through the dipole moment associated with an electron's spin or via the orbital motion of the electron. The strength of this magnetic dipole-dipole interaction is comparable to the neutron-nuclear interaction, and thus there are magnetic cross-sections that are analogous to the nuclear ones, which reveal the structure and dynamics of materials over wide ranges of length scale and energy. Magnetic neutron scattering plays a central role in determining and understanding the microscopic properties of a vast variety of magnetic systems—from the fundamental nature, symmetry, and dynamics of magnetically ordered materials—to the elucidation of the magnetic characteristics essential in technological applications.

One traditional role of magnetic neutron scattering has been the measurement of magnetic Bragg intensities in the magnetically ordered regime. Such measurements can be used to determine the spatial arrangement and directions of the atomic magnetic moments, the atomic magnetization density of the individual atoms in the material, and the value of the ordered moments as a function of thermodynamic parameters such as temperature, pressure, and applied magnetic field. These types of measurements can be carried out on single crystals, powders, thin films, and artificially grown multilayers, and often the information collected can be obtained by no other experimental technique. For magnetic phenomena that occur over length scales that are large compared to atomic distances, the technique of magnetic small angle neutron scattering (SANS) can be applied, in analogy to structural SANS. This is an ideal technique to explore domain structures, long wavelength oscillatory magnetic states, vortex structures in superconductors, and other spatial variations of the magnetization density on length scales from 1 to 1000 nm. Another specialized technique is neutron reflectometry, which can be used to investigate the magnetization profile in the near-surface regime of single crystals, as well as the magnetization density of thin films and multilayers, in analogy with structural reflectometry techniques. This particular technique has enjoyed dramatic growth during the last decade due to the rapid advancement of atomic deposition capabilities.

Neutrons can also scatter inelastically, to reveal the magnetic fluctuation spectrum of a material over wide ranges of energy ($\sim 10^{-8} - 1$ eV) and over the entire Brillouin zone. Neutron scattering plays a truly unique role in that it is the only technique that can directly determine the complete magnetic excitation spectrum, whether it is in the form of the dispersion relations for spin wave excitations, wave vector and energy dependence of critical fluctuations, crystal field excitations, magnetic excitons, or moment fluctuations. In the present overview we will discuss some of these possibilities.

Competitive and Related Techniques

Much of the information that can be obtained from neutron scattering, such as ordering temperatures and ferromagnetic moments, can and should be compared with magnetic measurements of quantities such as specific heat, susceptibility, and magnetization (Magnetism and magnetic measurements). Indeed it is often highly desirable to have measurements available from these traditional techniques before undertaking the neutron measurements, as a way of both characterizing the particular specimen being used and to be able to focus the neutron data collection to the range of temperature, applied field, etc., of central interest. Thus these complementary measurements are highly desirable.

For magnetic diffraction, the technique that is directly competitive is magnetic x-ray scattering. The x-ray magnetic cross-sections are typically $\sim 10^{-6}$ of those of charge scattering, but with the high fluxes available from synchrotron sources x-ray scattering from these cross-sections is readily observable. The information that can be obtained is also complementary to neutrons in that neutrons scatter from the total magnetization in the system, whether it originates from spin or orbital motions, while for x rays

the spin and orbital contributions can be distinguished. X rays also have the advantage that the wavelength can be tuned to an electronic resonance, which can increase the cross-section by as much as $\sim 10^4$, and thus it can be an element-specific technique. However, the limited number of resonances available restricts the materials that can be investigated, while the high absorption cross-sections and the generally more stringent experimental needs and specimen requirements are disadvantages. Each x-ray photon also carries $\sim 10^6$ times the energy that a neutron carries, and the combination of this with the brightness of a synchrotron source can cause rather dramatic changes in the physical or chemical properties of the sample during the investigation. Neutrons are a much gentler probe, but the low flux relative to synchrotron sources also means that larger samples are generally needed. Both techniques must be carried out at centralized facilities. Magnetic x-ray diffraction is a relatively young technique, though, and its use is expected to grow as the technique develops and matures. However, neutron scattering generally will continue to be the technique of choice when investigating new materials.

For the investigation of spin dynamics, there are a number of experimental techniques that can provide important information. Mössbauer spectroscopy (MÖSSBAUER SPECTROMETRY), muon spin precession, and perturbed angular correlation of nuclear radiation measurements can reveal characteristic time scales, as well as information about the site symmetry of the magnetism and the ordered moment as a function of temperature. Raman and infrared (IR) spectroscopy can reveal information about the long-wavelength (Brillouin zone–center) excitations (RAMAN SPECTROSCOPY OF SOLIDS), and the magnon density of states via two-magnon processes. Microwave absorption can determine the spin wave excitation spectrum at very long wavelengths (small wave vectors $\mathbf{q}$), while nuclear magnetic resonance (NUCLEAR MAGNETIC RESONANCE IMAGING) can give important information about relaxation rates. However, none of these techniques can explore the full range of dynamics that neutrons can, and in this regard there is no competitive experimental technique for measurements of the spin dynamics.

PRINCIPLES OF THE METHOD

Magnetic Diffraction

The cross-section for magnetic Bragg scattering (Bacon, 1975) is given by

$$I_{\mathrm{M}}(\mathbf{g}) = C\left(\frac{\gamma e^2}{2mc^2}\right)^2 \mathbf{M}_{\mathrm{g}} A(\theta_{\mathrm{B}}) |F_{\mathrm{M}}(\mathbf{g})|^2 \qquad (1)$$

where I_{M} is the integrated intensity for the magnetic Bragg reflection located at the reciprocal lattice vector, $\mathbf{g}$, the neutron-electron coupling constant enclosed in large parentheses evaluates as -0.27×10^{-12} cm, C is an instrumental constant that includes the resolution of the measurement, $A(\theta_{\mathrm{B}})$ is an angular factor that depends on the method of measurement (e.g., sample angular rotation, or $\theta : 2\theta$ scan), and $\mathbf{M}_{\mathrm{g}}$ is the multiplicity of the reflection (for a powder sample). The magnetic structure factor $F_{\mathrm{M}}(\mathbf{g})$ is given in the general case (Blume, 1961) by

$$F_{\mathrm{M}}(\mathbf{g}) = \sum_{j=1}^{N} e^{ig\cdot\mathbf{r}_j} \hat{\mathbf{g}} \times [\mathbf{M}_j(\mathbf{g}) \times \hat{\mathbf{g}}] e^{-W_j} \qquad (2)$$

where $\hat{\mathbf{g}}$ is a unit vector in the direction of the reciprocal lattice vector $\mathbf{g}$, $M_j(\mathbf{g})$ is the vector form factor of the jth ion located at $\mathbf{r}_j$ in the unit cell, W_j is the Debye-Waller factor (see MÖSSBAUER SPECTROMETRY) that accounts for the thermal vibrations of the jth ion, and the sum is over all (magnetic) atoms in the unit cell. The triple cross product originates from the vector nature of the dipole-dipole interaction of the neutron with the electron. A quantitative calculation of $M_j(\mathbf{g})$ in the general case involves evaluating matrix elements of the form $\langle \pm | 2\boldsymbol{S} e^{i\mathbf{g}\bullet\mathbf{R}} + \boldsymbol{O} | \pm \rangle$, where $\boldsymbol{S}$ is the (magnetic) spin operator, $\boldsymbol{O}$ is the symmetrized orbital operator introduced by Trammell (1953), and $|\pm\rangle$ represents the angular momentum state. This can be quite a complicated angular-momentum computation involving all the electron orbitals in the unit cell. However, usually the atomic spin density is *collinear*, by which we mean that at each point in the spatial extent of the electron's probability distribution, the *direction* of the atomic magnetization density is the same. In this case, the direction of $M_j(\mathbf{g})$ does not depend on $\mathbf{g}$, and the form factor is just a scalar function, $f(\mathbf{g})$, which is simply related to the Fourier transform of the magnetization density. The free-ion form factors have been tabulated for essentially all the elements. If one is familiar with x-ray diffraction, then it is helpful to note that the magnetic form factor for neutron scattering is analogous to the form factor for charge scattering of x rays, except that for x rays it corresponds to the Fourier transform of the total charge density of all the electrons, while in the magnetic neutron case it is the transform of the "magnetic" electrons, which are the electrons whose spins are unpaired. Recalling that a Fourier transform inverts the relative size of objects, the magnetic form factor typically decreases much more rapidly with $\mathbf{g}$ than for the case of x-ray charge scattering, since the unpaired electrons are usually the outermost ones of the ion. This dependence of the scattering intensity on $f(\mathbf{g})$ is a convenient way to distinguish magnetic cross-sections from nuclear cross-sections, where the equivalent of the form factor is just a constant (the nuclear coherent scattering amplitude b).

If, in addition to the magnetization density being collinear, the magnetic moments in the ordered state point along a unique direction (i.e., the magnetic structure is a ferromagnet, or a simple $+-+-$ type antiferromagnet), then the square of the magnetic structure factor simplifies to

$$|F_{\mathrm{M}}(\mathbf{g})|^2 = \langle 1 - (\hat{\mathbf{g}} \bullet \hat{\boldsymbol{\eta}})^2 \rangle \left| \sum_j \eta_j \langle \mu_j^z \rangle f_j(\mathbf{g}) e^{-W_j} e^{i\mathbf{g}\bullet\mathbf{r}_j} \right|^2 \qquad (3)$$

where $\hat{\boldsymbol{\eta}}$ denotes the (common) direction of the ordered moments and η_j the sign of the moment (± 1), $\langle \mu_j^2 \rangle$ is the

average value of the ordered moment in thermodynamic equilibrium at (T, H, P, ...), and the orientation factor $\langle 1-(\hat{\mathbf{g}}\bullet\hat{\boldsymbol{\eta}})^2\rangle$ represents an average over all possible domains. If the magnetic moments are the same type, then this expression further simplifies to

$$|F_{\mathrm{M}}(\mathbf{g})|^2 = \langle 1-(\hat{\mathbf{g}}\bullet\hat{\boldsymbol{\eta}})^2\rangle\langle\mu^z\rangle^2 f^2(\mathbf{g})e^{-2W_j}\left|\sum_j \eta_j e^{i\mathbf{g}\bullet\mathbf{r}_j}\right|^2 \quad (4)$$

We see from these expressions that neutrons can be used to determine several important quantities: the location of magnetic atoms in the unit cell and the spatial distribution of their magnetic electrons, as well as the dependence of $\langle\mu^z\rangle$ on temperature, field, pressure, or other thermodynamic variables, which is directly related to the order parameter for the phase transition (e.g., the sublattice magnetization). Often the preferred magnetic axis, $\hat{\boldsymbol{\eta}}$, can also be determined from the relative intensities. Finally, the scattering can be put on an absolute scale by internal comparison with the nuclear Bragg intensities I_{N} from the same sample, given by

$$I_{\mathrm{N}}(\mathbf{g}) = CM_{\mathbf{g}}A(\theta_{\mathrm{B}})|F_{\mathrm{N}}(\mathbf{g})|^2 \quad (5)$$

with

$$|F_{\mathrm{N}}(\mathbf{g})|^2 = \left|\sum_j b_j e^{-W_j}e^{i\mathbf{g}\bullet\mathbf{r}_j}\right|^2 \quad (6)$$

Here b_j is the coherent nuclear scattering amplitude for the jth atom in the unit cell, and the sum is over all atoms in the unit cell. Often the nuclear structure is known accurately and F_{N} can be calculated, whereby the saturated value of the magnetic moment can be obtained.

Subtraction Technique

There are several ways that magnetic Bragg scattering can be distinguished from the nuclear scattering from the structure. Above the magnetic ordering temperature, all the Bragg peaks are nuclear in origin, while as the temperature drops below the ordering temperature the intensities of the magnetic Bragg peaks rapidly develop, and for unpolarized neutrons the nuclear and magnetic intensities simply add. If these new Bragg peaks occur at positions that are distinct from the nuclear reflections, then it is straightforward to distinguish magnetic from nuclear scattering. In the case of a ferromagnet, however, or for some antiferromagnets that contain two or more magnetic atoms in the chemical unit cell, these Bragg peaks can occur at the same position. One standard technique for identifying the magnetic Bragg scattering is to make one diffraction measurement in the paramagnetic state well above the ordering temperature, and another in the ordered state at the lowest temperature possible, and then subtract the two sets of data. In the paramagnetic state, the (free ion) diffuse magnetic scattering is given (Bacon, 1975) by

$$I_{\mathrm{Para}} = \frac{2}{3}C\left(\frac{\gamma e^2}{2mc^2}\right)^2 \mu_{\mathrm{eff}}^2 f(\mathbf{Q})^2 \quad (7)$$

where μ_{eff} is the effective magnetic moment $\{= g[J(J+1)]^{1/2}$ for a free ion$\}$. This is a magnetic incoherent cross-section, and the only angular dependence is through the magnetic form factor $f(\mathbf{Q})$. Hence this scattering looks like "back-ground." There is a sum rule on the magnetic scattering in the system, though, and in the ordered state most of this diffuse scattering shifts into the coherent magnetic Bragg peaks. A subtraction of the high-temperature data (Equation 7) from the data obtained at low temperature (Equation 1) will then yield the magnetic Bragg peaks, on top of a deficit (negative) of scattering away from the Bragg peaks due to the disappearance of the diffuse paramagnetic scattering in the ordered state. On the other hand, usually none of the nuclear cross-sections change significantly with temperature, and hence drop out in the subtraction. A related subtraction technique is to apply a large magnetic field in the paramagnetic state, to induce a net (ferromagnetic-like) moment. The zero-field (nuclear) diffraction pattern can then be subtracted from the high-field pattern to obtain the induced-moment diffraction pattern.

Polarized Beam Technique

When the neutron spins that impinges on a sample has a well-defined polarization state, then the nuclear and magnetic scatterin that originates from the sample interferes *coherently*, in contrast to being separate cross-sections like Equations 1 and 5, where magnetic and nuclear intensities just add. A simple example is provided by the magnetic and nuclear scattering from elemental iron, which has the body-centered cubic (bcc) structure and a (saturated) ferromagnetic moment of $2.2\mu_{\mathrm{B}}$ at low temperature. It is convenient to define the magnetic scattering amplitude p (Bacon, 1975) as

$$p = \left(\frac{\gamma e^2}{2mc^2}\right)\langle\mu\rangle f(\mathbf{g}) \quad (8)$$

and then the structure factor for the scattering of polarized neutrons is given by

$$F(\mathbf{g}) = (b\pm p)e^{-W_j}(1+e^{i\pi(h+k+l)}) \quad (9)$$

where the $\pm$ indicates the two polarization states of the incident neutrons. The phase factor in parentheses originates from the sum over the unit cell, which in this case is just two identical atoms located at the origin (0,0,0) and the body-centered position (1/2,1/2,1/2), and (hkl) are Miller indices (SYMMETRY IN CRYSTALLOGRAPHY). This gives the familiar selection rule that for a bcc lattice $h+k+l$ must be an even integer for Bragg reflection to occur. The smallest nonzero reciprocal lattice vector is then the (1,1,0), and for this reflection ($\mathbf{g} = 3.10$ Å^{-1}) the magnetic scattering amplitude is as follows:

$$p = (0.27 \times 10^{-12}\ \text{cm})(2.2)(0.59) = 0.35 \times 10^{-12}\ \text{cm} \quad (10)$$

The coherent nuclear scattering amplitude for iron is $b = 0.945 \times 10^{-12}$ cm and is independent of the Bragg reflection. The flipping ratio R is defined as the ratio of the intensities for the two neutron polarizations; for this reflection we have

$$R = \left(\frac{0.945 + 0.35}{0.945 - 0.35}\right)^2 = 4.74 \quad (11)$$

Note that, for a different Bragg reflection, the only quantity that changes is $f(\mathbf{g})$, so that a measurement of the flipping ratio at a series of Bragg peaks (hkl) can be used to make precision determinations of the magnetic form factor, and thus by Fourier transform the spatial distribution of the atomic magnetization density. It also should be noted that if $b \approx p$, then the flipping ratio will be very large, and this is in fact one of the standard methods employed to produce a polarized neutron beam in the first place.

In the more general situation when the magnetic structure is not a simple ferromagnet, polarized beam diffraction measurements with polarization analysis of the scattered neutrons must be used to establish unambiguously which peaks are magnetic, which are nuclear, and more generally to separate the magnetic and nuclear scattering at Bragg positions where there are both nuclear and magnetic contributions. The polarization analysis technique as applied to this problem is in principle straightforward; complete details can be found elsewhere (Moon et al., 1969; Williams, 1988). Nuclear coherent Bragg scattering never causes a reversal, or spin-flip, of the neutron spin direction upon scattering. Thus the nuclear peaks will only be observed in the non-spin-flip scattering geometry. We denote this configuration as $(+\ +)$, where the neutron is incident with up spin, and remains in the up state after scattering. Non-spin-flip scattering also occurs if the incident neutron is in the down state, and remains in the down state after scattering [denoted $(-\ -)$]. The magnetic cross-sections, on the other hand, depend on the relative orientation of the neutron polarization $\mathbf{P}$ and the reciprocal lattice vector $\mathbf{g}$. In the configuration where $\mathbf{P} \perp \mathbf{g}$, half the magnetic Bragg scattering involves a reversal of the neutron spin [denoted by $(-\ +)$ or $(+\ -)$], and half does not. Thus, for the case of a purely magnetic reflection the spin-flip $(-\ +)$ and non-spin-flip $(+\ +)$ intensities should be equal in intensity. For the case where $\mathbf{P} \parallel \mathbf{g}$, all the magnetic scattering is spin-flip. Hence for a pure magnetic Bragg reflection the spin-flip scattering should be twice as strong as for the $\mathbf{P} \perp \mathbf{g}$ configuration, while ideally no non-spin-flip scattering will be observed. The analysis of these cross-sections can be used to unambiguously identify nuclear from magnetic Bragg scattering.

Inelastic Scattering

Magnetic inelastic scattering plays a unique role in determining the spin dynamics in magnetic systems, as it is the only probe that can directly measure the complete magnetic excitation spectrum. Typical examples are spin wave dispersion relations, critical fluctuations, crystal field excitations, and moment/valence fluctuations.

As an example, consider identical spins S on a simple cubic lattice, with a coupling given by $-JS_i \cdot S_j$ where J is the (Heisenberg) exchange interaction between neighbors separated by the distance a. The collective excitations for such an ensemble of spins are magnons (or loosely termed "spin waves"). If we have $J > 0$ so that the lowest energy configuration is where the spins are parallel (a ferromagnet), then the magnon dispersion along the edge of the cube (the [100] direction) is given by $E(q) = 8JS[\sin^2(qa/2)]$. At each wave vector $\mathbf{q}$, the magnon energy is different, and a neutron can interact with the system of spins and either create a magnon at $(\mathbf{q}, E)$, with a concomitant change of momentum and loss of energy of the neutron, or conversely destroy a magnon with a gain in energy. The observed change in momentum and energy for the neutron can then be used to map out the magnon dispersion relation. Neutron scattering is particularly well suited for such inelastic scattering studies since neutrons typically have energies that are comparable to the energies of excitations in the solid, and therefore the neutron energy changes are large and easily measured. The dispersion relations can then be measured over the entire Brillouin zone (see, e.g., Lovesey, 1984).

Additional information about the nature of the excitations can be obtained by polarized inelastic neutron scattering techniques, which are finding increasing use. Spin wave scattering is represented by the raising and lowering operators $S\pm = S_x \pm iS_y$, which cause a reversal of the neutron spin when the magnon is created or destroyed. These "spin-flip" cross-sections are denoted by $(+\ -)$ and $(-\ +)$. If the neutron polarization $\mathbf{P}$ is parallel to the momentum transfer $\mathbf{Q}$, $\mathbf{P} \parallel \mathbf{Q}$, then spin angular momentum is conserved (as there is no orbital contribution in this case). In this experimental geometry, we can only create a spin wave in the $(-\ +)$ configuration, which at the same time causes the total magnetization of the sample to decrease by one unit ($1\mu_B$). Alternatively, we can destroy a spin wave only in the $(+\ -)$ configuration, while increasing the magnetization by one unit. This gives us a unique way to unambiguously identify the spin wave scattering, and polarized beam techniques in general can be used to distinguish magnetic from nuclear scattering in a manner similar to the case of Bragg scattering.

Finally, we note that the magnetic Bragg scattering is comparable in strength to overall magnetic inelastic scattering. However, all the Bragg scattering is located at a single point in reciprocal space, while the inelastic scattering is distributed throughout the (three-dimensional) Brillouin zone. Hence, when actually making inelastic measurements to determine the dispersion of the excitations, one can only observe a small portion of the dispersion surface at any one time, and thus the observed inelastic scattering is typically two to three orders of magnitude less intense than the Bragg peaks. Consequently these are much more time-consuming measurements, and larger samples are needed to offset the reduction in intensity. Of course, a successful determination of the dispersion relations yields a complete determination of the fundamental atomic interactions in the solid.

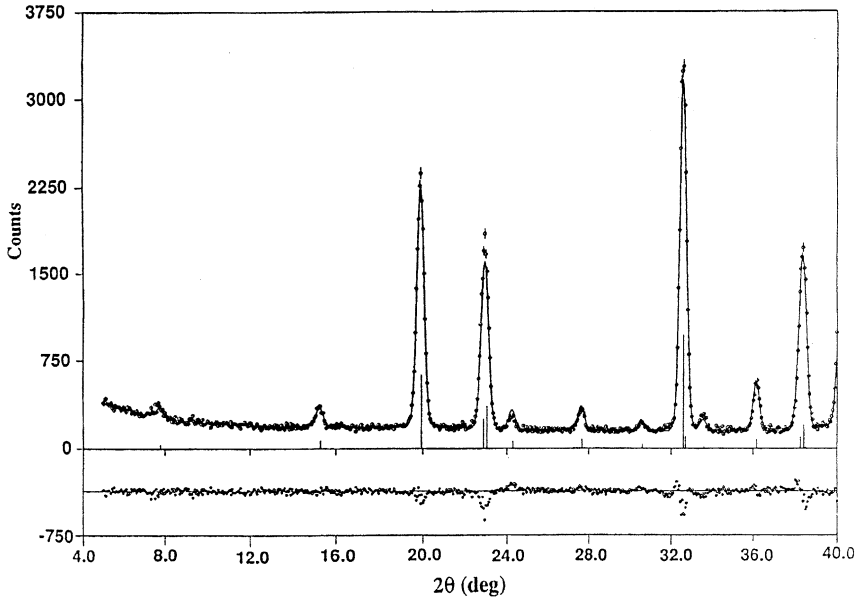


Figure 1. Calculated (solid curve) and observed intensities at room temperature for a powder of antiferromagnetically ordered $YBa_2Fe_3O_8$. The differences between calculated and observed are shown at the bottom. (Huang et al., 1992.)

PRACTICAL ASPECTS OF THE METHOD

Diffraction

As an example of Bragg scattering, a portion of the powder diffraction pattern from a sample of $YBa_2Fe_3O_8$ is shown in Figure 1 (Huang et al., 1992). The solid curve is a Rietveld refinement (Young, 1993) of both the antiferromagnetic and crystallographic structure for the sample. From this type of data, we can determine the full crystal structure; lattice parameters, atomic positions in the unit cell, site occupancies, etc. We can also determine the magnetic structure and value of the ordered moment. The results of the analysis are shown in Figure 2; the crystal structure is identical to the structure for the $YBa_2Cu_3O_7$ high-T_C cuprate superconductor, with the Fe replacing the Cu, and the magnetic structure is also the same as has been observed for the Cu spins in the oxygen-reduced ($YBa_2Cu_3O_6$) semiconducting material.

Experimentally, we can recognize the magnetic scattering by several characteristics. First, it should be temperature-dependent, and the Bragg peaks will vanish above the ordering temperature. Figure 3 shows the temperature dependence of the intensity of the peak at a scattering angle of 19.5° in Figure 1. The data clearly reveal a phase transition at the Néel temperature for $YBa_2Fe_3O_8$ of 650 K (Natali Sora et al., 1994); the Néel temperature is where long-range antiparallel order of the spins first occurs. Above the antiferromagnetic phase transition, this peak completely disappears, indicating that it is a purely magnetic Bragg peak. A second characteristic is that the magnetic intensities become weak at high scattering angles (not shown), as $f(\mathbf{g})$ typically falls off strongly with increasing angle. A third, more elegant, technique is to use polarized neutrons. The polarization technique can be used at any temperature, and for any material, regardless of whether or not it has a crystallographic distortion (e.g., via magnetoelastic interactions) associated with the magnetic transition. It is more involved and time-consuming experimentally, but yields an unambiguous identification and separation of magnetic and nuclear Bragg peaks.

First consider the case where $\mathbf{P} \parallel \mathbf{g}$, which is generally achieved by having a horizontal magnetic field, which must also be oriented along the scattering vector. In this geometry, all the magnetic scattering is spin-flip, while the nuclear scattering is always non-spin-flip. Hence for a magnetic Bragg peak the spin-flip scattering should be twice as strong as for the $\mathbf{P} \perp \mathbf{g}$ configuration (vertical

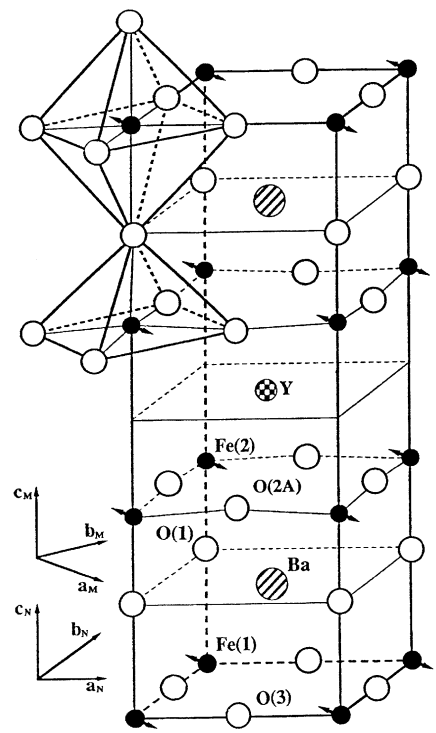


Figure 2. Crystal and magnetic structure for $YBa_2Fe_3O_8$ deduced from the data of Figure 1. (Huang et al., 1992.)

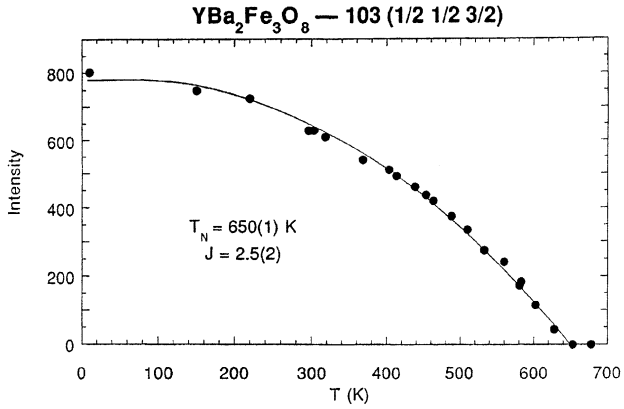


Figure 3. Temperature dependence of the intensity of the magnetic reflection found at a scattering angle of 19.5° in Figure 1. The Néel temper ature is 650 K. (Natali et al., 1994.)

field), while the nuclear scattering is non-spin-flip scattering and independent of the orientation of **P** and **g**. Figure 4 shows the polarized beam results for the same two peaks, at scattering angles (for this wavelength) of 30° and 35°; these correspond to the peaks at 19.5° and 23° in Figure 1. The top section of the figure shows the data for the $\mathbf{P} \perp \mathbf{g}$ configuration. The peak at 30° has the identical intensity for both spin-flip and non-spin-flip scattering, and hence we conclude that this scattering is purely magnetic in origin, as inferred from Figure 3. The peak at 35°, on the other hand, has strong intensity for (+ +), while the intensity for (− +) is smaller by a factor of 1/11, the instrumental flipping ratio in this measurement. Hence, ideally there would be no spin-flip scattering, and this peak is identified as a pure nuclear reflection. The center row shows the same peaks for the $\mathbf{P} \parallel \mathbf{g}$ configuration, while the bottom row shows the subtraction of the $\mathbf{P} \perp \mathbf{g}$ spin-flip scattering from the $\mathbf{P} \parallel \mathbf{g}$ spin-flip scattering. In this subtraction procedure, instrumental background, as well as all nuclear scattering cross-sections, cancel, isolating the magnetic scattering. We see that there is magnetic intensity only for the low angle position, while no intensity survives the subtraction at the 35° peak position. These data unambiguously establish that the 30° peak is purely magnetic, while the 35° peak is purely nuclear. This simple example demonstrates how the technique works; obviously it would play a much more critical role in cases where it is not clear from other means what is the origin of the peaks, such as in regimes where the magnetic and nuclear peaks overlap, or in situations where the magnetic transition is accompanied by a structural distortion. If needed, a complete "magnetic diffraction pattern" can be obtained and analyzed with these polarization techniques.

In cases where there is no significant coupling of the magnetic and lattice systems, on the other hand, the subtraction technique can also be used to obtain the magnetic diffraction pattern (see, e.g., Zhang et al., 1990). This technique is especially useful for low-temperature magnetic phase transitions where the Debye-Waller effects can be

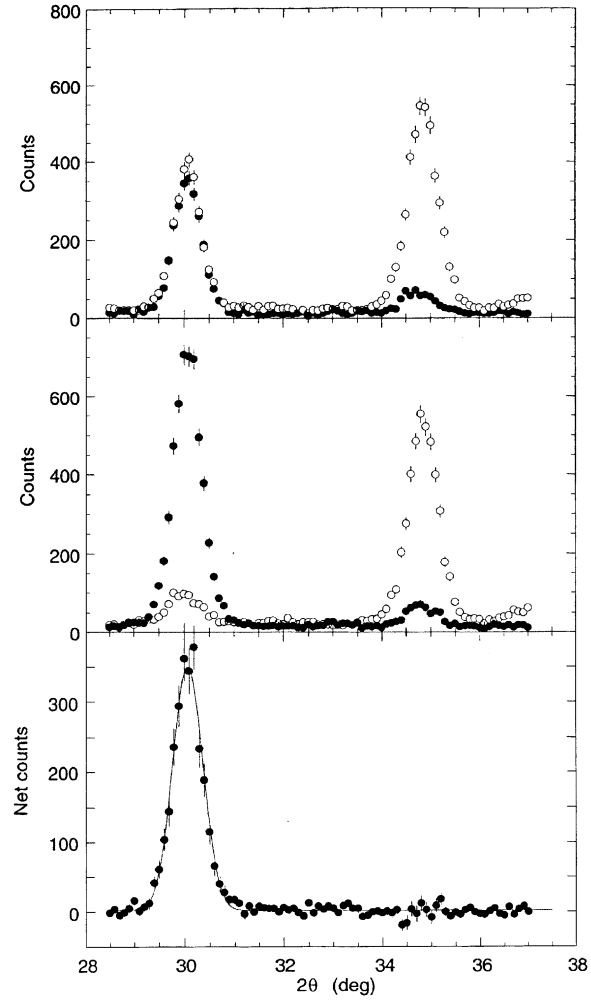


Figure 4. Polarized neutron scattering. The top portion of the figure is for $\mathbf{P} \perp \mathbf{g}$, where the open circles show the non-spin-flip scattering and the filled circles are the observed scattering in the spin-flip configuration. The low angle peak has equal intensity for both cross-sections, and thus is identified as a pure magnetic reflection, while the ratio of the (++) to (−+) scattering for the high-angle peak is 11, the instrumental flipping ratio. Hence this is a pure nuclear reflection. The center portion of the figure is for $\mathbf{P} \perp \mathbf{g}$ and the bottom portion is the subtraction of the $\mathbf{P} \perp \mathbf{g}$ spin-flip scattering from the data for $\mathbf{P} \parallel \mathbf{g}$. Note that in the subtraction procedure all background and nuclear cross-sections cancel, isolating the magnetic scattering. (Huang et al., 1992.)

safely neglected. Figure 5 shows the diffraction patterns for $Tm_2Fe_3Si_5$, which is an antiferromagnetic material that becomes superconducting under pressure (Gotaas et al., 1987). The top part of the figure shows the diffraction pattern obtained above the antiferromagnetic ordering temperature of 1.1 K, where just the nuclear Bragg peaks are observed. The middle portion of the figure shows the low-temperature diffraction pattern in the ordered state, which contains both magnetic and nuclear Bragg peaks, and the bottom portion shows the subtraction, which gives the magnetic diffraction pattern.

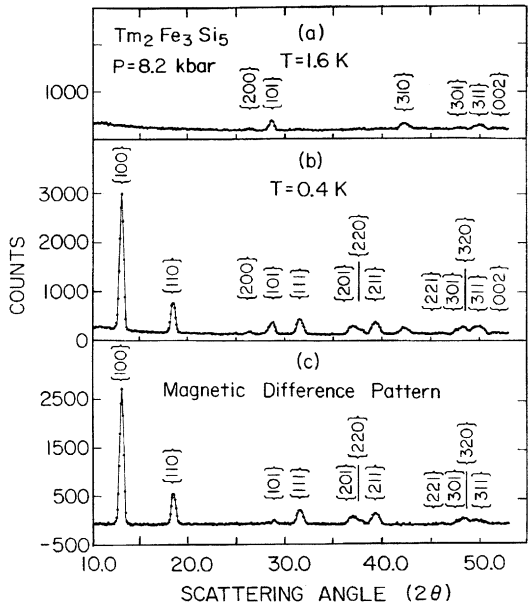


Figure 5. Diffraction patterns for the antiferromagnetic superconductor $Tm_2Fe_3Si_5$. The top part of the figure shows the nuclear diffraction pattern obtained above the antiferromagnetic ordering temperature of 1.1 K, the middle portion of the figure shows the low-temperature diffraction pattern in the ordered state, and the bottom portion shows the subtraction of the two, which gives the magnetic diffraction pattern. (Gotaas et al., 1987.)

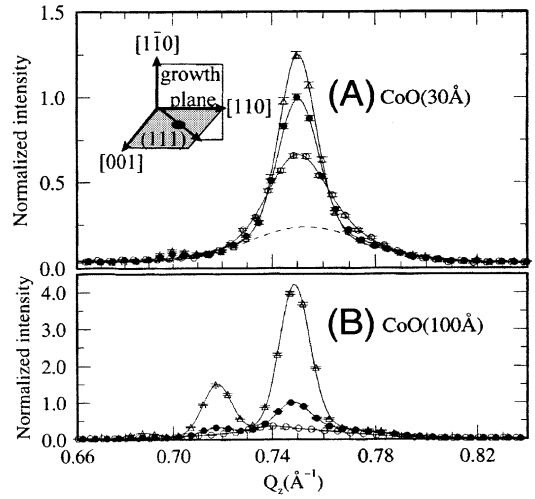


Figure 7. Neutron diffraction scans of the (111) reflection along the [001] growth axis direction (scattering vector $\mathbf{Q}_z$) for (**A**) $[Fe_3O_4$/(100 Å)CoO(30 Å)$]_{50}$ and (**B**) $[Fe_3O_4$/(100 Å)CoO(100 Å)$]_{50}$ superlattices, taken at 78 K in zero applied field. The closed circles, open circles, and triangles indicate data taken after zero-field cooling (initial state), field cooling with an applied field ($H = 14$ kOe) in the [110] direction, and field cooling with an applied field in the [110] direction, respectively. The inset illustrates the scattering geometry. The dashed lines indicate the temperature-and field-independent Fe_3O_4 component (Ijiri et al., 1998).

Another example of magnetic Bragg diffraction is shown in Figure 6. Here we show the temperature dependence of the intensity of an antiferromagnetic Bragg peak on a single crystal of the high-T_C superconductor ($T_C \approx 92$ K) $ErBa_2Cu_3O_7$ (Lynn et al., 1989). The magnetic interactions of the Er moments in this material are highly anisotropic, and this system turns out to be an ideal two-dimensional (2D) (planar) Ising antiferromagnet; the solid curve is Onsager's exact solution to the $S = \frac{1}{2}$, 2D Ising model (Onsager, 1944), and we see that it provides an excellent representation of the experimental data.

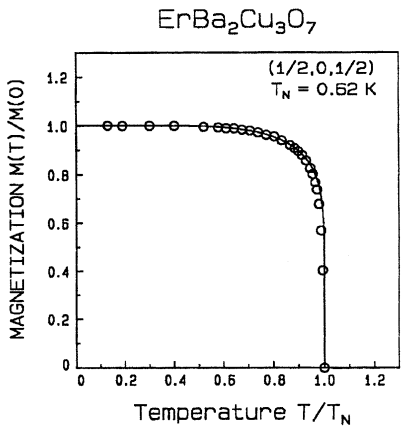


Figure 6. Temperature dependence of the sublattice magnetization for the Er spins in superconducting $ErBa_2Cu_3O_7$, measured on a single crystal weighing 31 mg. The solid curve is Onsager's exact theory for the 2D, $S = \frac{1}{2}$, Ising model (Lynn et al., 1989).

A final diffraction example is shown in Figure 7, where the data for two Fe_3O_4/CoO superlattices are shown (Ijiri et al., 1998). The superlattices consist of 50 repeats of 100-Å thick layers of magnetite, which is ferrimagnetic, and either 30- or 100-Å thick layers of the antiferromagnet CoO. The superlattices were grown epitaxially on single-crystal MgO substrates, and thus these may be regarded as single-crystal samples. These scans are along the growth direction (Q_z), and show the changes in the magnetic scattering that occur when the sample is cooled in an applied field, versus zero-field cooling. These data, together with polarized beam data taken on the same samples, have elucidated the origin of the technologically important exchange biasing effect that occurs in these magnetic superlattices.

It is interesting to compare the type and quality of data that are represented by these three examples. The powder diffraction technique is quite straightforward, both to obtain and analyze the data. In this case, typical sample sizes are ~1–20 g, and important and detailed information can be readily obtained with such sample sizes in a few hours of spectrometer time, depending on the particulars of the problem. The temperature dependence of the order parameter in $ErBa_2Cu_3O_7$, on the other hand, was obtained on a single crystal weighing only 31 mg. Note that the statistical quality of the data is much better than for the powder sample, even though the sample is more than 2 orders of magnitude smaller; this is because it is a single crystal and all the scattering is directed into a single peak, rather than scattering into a powder-diffraction ring. The final example was for Fe_3O_4/CoO

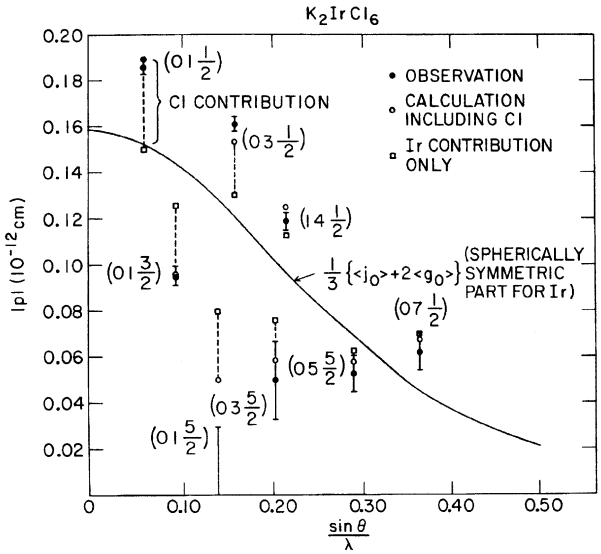


Figure 8. Comparison of the observed and calculated magnetic form factor for the cubic K_2IrCl_6 system, showing the enormous directional anisotropy along the spin direction ([001] direction) compared with perpendicular to the spin direction. The contribution of the Cl moments along the spin direction, arising from the covalent bonding, is also indicated; no net moment is transferred to the Cl ions that are perpendicular to the spin direction. (Lynn et al., 1976.)

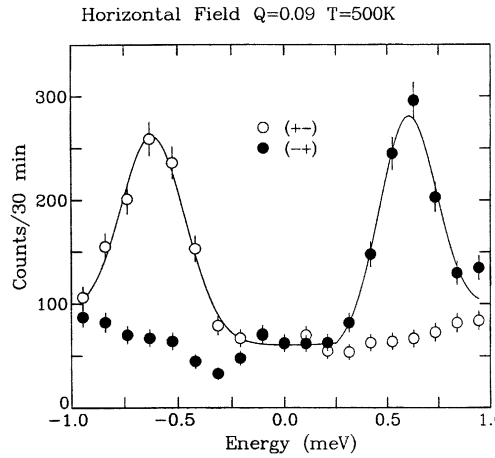


Figure 9. Spin-flip scattering observed for the amorphous Invar $Fe_{86}B_{14}$ isotropic ferromagnetic system in the $\mathbf{P} \parallel \mathbf{Q}$ configuration. Spin waves are observed for neutron energy gain ($E < 0$) in the $(+-)$ cross-section, and for neutron energy loss ($E > 0$) in the $(-+)$ configuration. (Lynn et al., 1993.)

superlattices, where the weight of the superlattices that contribute to the scattering is ~1 mg. Thus it is clear that interesting and successful diffraction experiments can be carried out on quite small samples.

The final example in this section addresses the measurement of a magnetic form factor, in the cubic antiferromagnetic K_2IrCl_6. The Ir ions occupy a face-centered cubic (fcc) lattice, and the *5d* electrons carry a magnetic moment that is covalently bonded to the six Cl ions located octahedrally around each Ir ion. One of the interesting properties of this system is that there is charge transfer from the Ir equally onto the six Cl ions. A net spin transfer, however, only occurs for the two Cl ions along the direction in which the Ir spin points. This separation of spin and charge degrees of freedom leads to a very unusual form factor that is highly anisotropic, as shown in Figure 8 (Lynn et al., 1976). Note the rapid decrease in the form factor as one proceeds along the (*c* axis) spin direction, from the (0,1,1/2), to the (0,1,3/2), then to the (0,1,5/2) Bragg peak, which in fact has an unobservable intensity. In this example, ~30% of the total moment is transferred onto the Cl ions, which is why these covalency effects are so large in this compound. It should be noted that, in principle, x-ray scattering should be able to detect the charge transferred onto the (six) Cl ions as well, but this is much more difficult to observe because it is a very small fraction of the total charge. In the magnetic case, in contrast, it is a large percentage effect and has a different symmetry than the lattice, which makes the covalent effects much easier to observe and interpret.

Inelastic Scattering

There are many types of magnetic excitations and fluctuations that can be measured with neutron scattering techniques, such as magnons, critical fluctuations, crystal field excitations, magnetic excitons, and moment/valence fluctuations. To illustrate the basic technique, consider an isotropic ferromagnet at sufficiently long wavelengths (small q). The spin wave dispersion relation is given by $E_{sw} = D(T)q^2$, where D is the spin wave "stiffness" constant. The general form of the spin wave dispersion relation is the same for all isotropic ferromagnets, a requirement of the (assumed) perfect rotational symmetry of the magnetic system, while the numerical value of D depends on the details of the magnetic interactions and the nature of the magnetism. One example of a prototypical isotropic ferromagnet is amorphous $Fe_{86}B_{14}$. Figure 9 shows an example of polarized beam inelastic neutron scattering data taken on this system (Lynn et al., 1993). These data were taken with the neutron polarization $\mathbf{P}$ parallel to the momentum transfer $\mathbf{Q}(\mathbf{P} \parallel \mathbf{Q})$, where we should be able to create a spin wave only in the $(-+)$ configuration, or destroy a spin wave only in the $(+-)$ configuration. This is precisely what we see in the data—for the $(-+)$ configuration the spin waves can only be observed for neutron energy loss scattering ($E > 0$), while for the $(+-)$ configuration, spin waves can only be observed in neutron energy gain ($E < 0$). This behavior of the scattering uniquely identifies these excitations as spin waves.

Data like these can be used to measure the renormalization of the spin waves as a function of temperature, as well as to determine the lifetimes as a function of wave vector and temperature. An example of the renormalization of the "stiffness" constant D for the magnetoresistive oxide $Tl_2Mn_2O_7$ is shown in Figure 10 (Lynn et al., 1998). Here, the wave vector dependence of the dispersion relation has been determined at a series of q's, and the stiffness parameter is extracted from the data. The variation in the stiffness parameter is then plotted, and indicates a

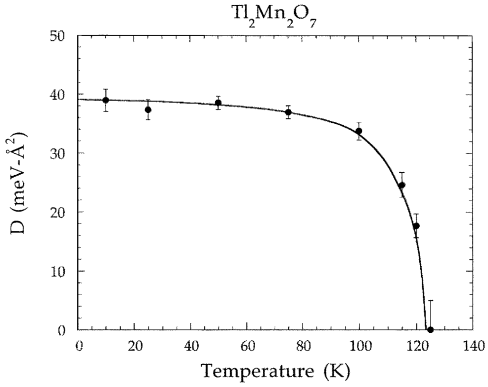


Figure 10. Temperature dependence of the spin wave stiffness $D(T)$ in the magnetoresistive $Tl_2Mn_2O_7$ pyrochlore, showing that the spin waves renormalize as the ferromagnetic transition is approached. (Lynn et al., 1998.) Below T_C the material is a metal, while above T_C it exhibits insulator behavior.

smooth variation with a ferromagnetic transition temperature of 123 K. These measurements can then be compared directly with theoretical calculations. They can also be compared with other experimental observations, such as magnetization measurements, whose variation with temperature originates from spin wave excitations.

Finally, these types of measurements can be extended all the way to the zone boundary on single crystals. Figure 11 shows an example of the spin wave dispersion relation for the ferromagnet $La_{0.85}Sr_{0.15}MnO_3$, a system that undergoes a metal-insulator transition that is associated with the transition to ferromagnetic order (Vasiliu-Doloc et al., 1998). The top part of the figure shows the dispersion relation for the spin wave energy along two high-symmetry directions, and the solid curves are fits to a simple nearest-neighbor spin-coupling model. The overall trend of the data is in reasonable agreement with the model, although there are some clear discrepancies as well, indicating that a more sophisticated model will be needed in order to obtain quantitative agreement. In addition to the spin wave energies, though, information about the intrinsic lifetimes can also be determined, and these linewidths are shown in the bottom part of the figure for both symmetry directions. In the simplest type of model, no intrinsic spin wave linewidths at all would be expected at low temperatures, while we see here that the observed linewidths are very large and highly anisotropic, indicating that an itinerant-electron type of model is more appropriate for this system.

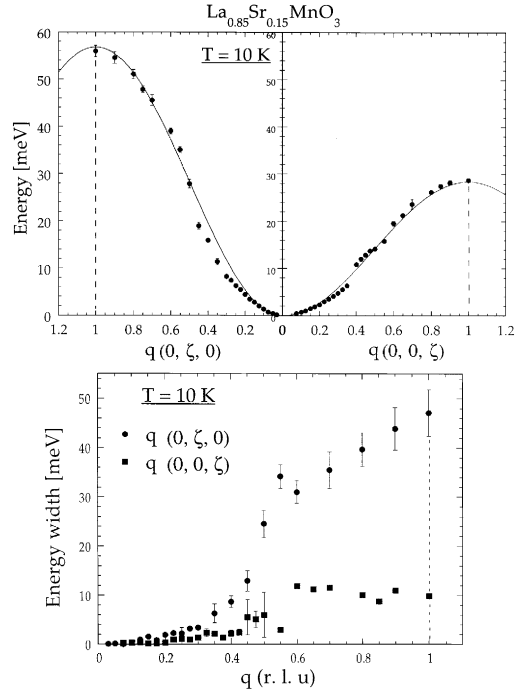


Figure 11. Ground state spin wave dispersion along the (0,1,0) and (0,0,1) directions measured to the zone boundary for the magnetoresistive manganite $La_{0.85}Sr_{0.15}MnO_3$. The solid curves are a fit to the dispersion relation for a simple Heisenberg exchange model. The bottom part of the figure shows the intrinsic linewidths of the excitations. In the standard models, intrinsic linewidths are expected only at elevated temperatures. The large observed linewidths demonstrate the qualitative inadequacies of these models. (Vasiliu-Doloc et al., 1998.)

METHOD AUTOMATION

Neutrons for materials research are produced by one of two processes; fission of U^{235} in a nuclear reactor, or by the spallation process where high energy protons from an accelerator impact on a heavy-metal target like W and explode the nuclei. Both techniques produce high-energy (MeV) neutrons, which are then thermalized or sub-thermalized in a moderator (such as heavy water) to produce a Maxwellian spectrum of thermal or cold neutrons, with energies in the meV range. Thus all neutron scattering instrumentation is located at centralized national facilities, and each facility generally has a wide variety of instrumentation that has been designed and constructed for the specific facility, and is maintained and scheduled by facility scientists. Generally, any of these instruments can be used to observe magnetic scattering, and with the number and diversity of spectrometers in operation it is not practical to review the instrumentation here; the interested user should contact one of the facilities directly. The facilities themselves operate for periods of weeks at a time continuously, and hence, since the early days of neutron scattering all the data collection has been accomplished by automated computer control.

SAMPLE PREPARATION

As a general rule, there is no particular preparation that is required for the sample, in the sense that there is no need, for example, to polish the surface, cleave it under high vacuum, or perform similar procedures. Neutrons are a deeply penetrating bulk probe, and hence generally do not interact with the surface. There are two exceptions to this rule. One is when utilizing polarized neutrons in the investigation of systems with a net magnetization, where a rough surface can cause the variations in the local

magnetic field that depolarize the neutrons. In such cases, the surface may need to be polished. The second case is for neutron reflectometry, where small-angle mirror reflection is used to directly explore the magnetization profile of the surface and interfacial layers. In this case, the samples need to be optically flat over the full size of the sample (typically 1 to 10 cm^2), and of course the quality of the film in terms of, e.g., surface roughness and epitaxial quality becomes part of the investigation.

For magnetic single crystal diffraction, typical sample sizes should be no more than a few mm^3 in order to avoid primary and secondary extinction effects; if the sample is too large, extinction will always limit the accuracy of the data obtained. Since the magnetic Bragg intensity is proportional to the square of the ordered moment, any extinction effects will depend on the value of $\langle \mu_z \rangle$ at a particular temperature and field, while the ideal size of the sample can vary by three to four orders of magnitude, depending on the value of the saturation moment.

Powder diffraction requires gram-sized samples, generally in the range of 1 to 20 g. The statistical quality of the data that can be obtained is directly related to the size, so in principle there is a direct trade-off between sample size and the time required to collect a set of diffraction data at a particular temperature or magnetic field. Sample sizes smaller than 1 g can also be measured, but since the neutron facilities are heavily oversubscribed, in practice a smaller sample size translates into fewer data sets, so an adequate-sized sample is highly desirable.

The cross-sections for inelastic scattering are typically two to three orders of magnitude smaller than elastic cross-sections, and therefore crystals for inelastic scattering typically must be correspondingly larger in comparison with single-crystal diffraction. Consequently, for both powder diffraction and inelastic scattering, the general rule is the bigger the sample the better the data. The exception is if one or more of the elements in a material has a substantial nuclear absorption cross-section, in which case the optimal size of the sample is then determined by the inverse of this absorption length. The absorption cross-sections are generally directly proportional to the wavelength of the neutrons being employed for a particular measurement, and the optimal sample size then depends on the details of the spectrometer and the neutron wavelength(s). For the absorption cross-sections of all the elements, see Internet Resources. In some cases, particular isotopes can be substituted to avoid some large absorption cross-sections, but generally these are expensive and/or are only available in limited quantity. Therefore isotopic substitution can be employed only for a few isotopes that are relatively inexpensive, such as deuterium or boron, or in scientifically compelling cases where the cost can be justified.

DATA ANALYSIS AND INITIAL INTERPRETATION

One of the powers of neutron scattering is that it is a very versatile tool and can be used to probe a wide variety of magnetic systems over enormous ranges of length scale ($\sim$0.1 to 10000 Å) and dynamical energy ($\sim$10 neV to 1 eV). The instrumentation to provide these measurement capabilities is equally diverse, but data analysis software is generally provided for each instrument to perform the initial interpretation of the data. Moreover, the technique is usually sufficiently fundamental and the interpretation sufficiently straightforward that the initial data analysis is often all that is needed.

PROBLEMS

Neutron scattering is a very powerful technique, but in general it is a flux-limited, and this usually requires a careful tradeoff between instrumental resolution and signal. There can also be unwanted cross-sections that appear and contaminate the data, and so one must of course exercise care in collecting the data and be vigilant in its interpretation. If a time-of-flight spectrometer is being employed, for example, there may be frame overlap problems that can mix the fastest neutrons with the slowest, and give potentially spurious results. Similarly, if the spectrometer employs one or more single crystals to monochromate or analyze the neutron energies, then the crystal can reflect higher-order wavelengths [since Bragg's law is $n\lambda = 2d\sin(\theta)$] which can give spurious peaks. Sometimes these can be suppressed with the use of filters, but more generally it is necessary to vary the spectrometer conditions and retake the data in order to identify genuine cross-sections from spurious ones. This identification process can consume considerable beam time.

Another problem can be encountered in identifying the magnetic Bragg scattering in powders using the subtraction technique. If the temperature is sufficiently low, the nuclear spins may begin to align with the electronic spins, particularly if there is a large hyperfine field. Significant polarization can occur at temperatures of a few degrees kelvin and lower, and the nuclear polarization can be confused with electronic magnetic ordering. Another problem can be encountered if the temperature dependence of the Debye-Waller factor is significant, or if there is a structural distortion associated with the magnetic transition, such as through a magnetoelastic interaction. In this case, the nuclear intensities will not be identical, and thus will not cancel correctly in the subtraction. It is up to the experimenter to decide if this is a problem, within the statistical precision of the data. A problem can also occur if there is a significant thermal expansion, where the diffraction peaks shift position with temperature. By significant, we mean that the shift is noticeable in comparison with the instrumental resolution employed. Again it is up to the experimenter to decide if this is a problem. In both these latter cases, though, a full refinement of the combined magnetic and nuclear structures in the ordered phase can be carried out. Alternatively, polarized beam techniques can be employed to unambiguously separate the magnetic and nuclear cross-sections. Finally, if one uses the subtraction technique in the method where a field is applied, the field can cause the powder particles to reorient if there is substantial crystalline magnetic anisotropy in the sample. This preferred orientation will remain when the field is removed, and will be evident in the nuclear peak intensities, but must be taken into account.

Finally, we remark about the use of polarized neutrons. Highly polarized neutron beams can be produced by single-crystal diffraction from a few special magnetic materials, from magnetic mirrors, or from transmission of the beam through a few specific nuclei where the absorption cross-section is strongly spin dependent. The choices are limited, and consequently most spectrometers are not equipped with polarization capability. For instruments that do have a polarized beam option, generally one has to sacrifice instrumental performance in terms of both resolution and intensity. Polarization techniques are then typically not used for many problems in a routine manner, but rather are usually used to answer some specific question or make an unambiguous identification of a cross-section, most often after measurements have already been carried out with unpolarized neutrons.

ACKNOWLEDGMENTS

I would like to thank my colleagues, Julie Borchers, Qing Huang, Yumi Ijiri, Nick Rosov, Tony Santoro, and Lida Vasiliu-Doloc, for their assistance in the preparation of this unit.

There are vast numbers of studies in the literature on various aspects presented here. The specific examples used for illustrative purposes were chosen primarily for the author's convenience in obtaining the figures, and his familiarity with the work.

LITERATURE CITED

Bacon, G. E. 1975. Neutron Diffraction, 3rd ed. Oxford University Press, Oxford.

Blume, M. 1961. Orbital contribution to the magnetic form factor of Ni^{++}. *Phys. Rev.* 124:96–103.

Gotaas, J. A., Lynn, J. W., Shelton, R. N., Klavins, P., and Braun, H. F. 1987. Suppression of the superconductivity by antiferromagnetism in $Tm_2Fe_3Si_5$. *Phys. Rev. B* 36:7277–7280.

Huang, Q., Karen, P., Karen, V. L., Kjekshus, A., Lynn, J. W., Mighell, A. D., Rosov, N., and Santoro, A. 1992. Neutron powder diffraction study of the nuclear and magnetic structures of $YBa_2Cu_3O_8$, *Phys. Rev. B* 45:9611–9619.

Ijiri, Y., Borchers, J. A., Erwin, R. W., Lee, S.-H., van der Zaag, P. J., and Wolf, R. M. 1998. Perpendicular coupling in exchange-biased Fe_3O_4/CoO superlattices. *Phys. Rev. Lett.* 80:608–611.

Lovesey, S. W. 1984. Theory of Neutron Scattering from Condensed Matter, Vol. 2. Oxford University Press, New York.

Lynn, J. W., Clinton, T. W., Li, W.-H., Erwin, R. W., Liu, J. Z., Vandervoort, K., and Shelton, R. N. 1989. 2D and 3D magnetic order of Er in $ErBa_2Cu_3O_7$. *Phys. Rev. Lett.* 63:2606–2610.

Lynn, J. W., Rosov, N., and Fish, G. 1993. Polarization analysis of the magnetic excitations in Invar and non-Invar amorphous ferromagnets. *J. Appl. Phys.* 73:5369–5371.

Lynn, J. W., Shirane, G., and Blume, M. 1976. Covalency effects in the magnetic form factor of Ir in K_2IrCl_6. *Phys. Rev. Lett.* 37:154–157.

Lynn, J. W., Vasiliu-Doloc, L., and Subramanian, M. 1998. Spin dynamics of the magnetoresistive pyrochlore $Tl_2Mn_2O_7$. *Phys. Rev. Lett.* 80:4582–4586.

Moon, R. M., Riste, T., and Koehler, W. C. 1969. Polarization analysis of thermal neutron scattering. *Phys. Rev.* 181:920–931.

Natali, S. I., Huang, Q., Lynn, J. W., Rosov, N., Karen, P., Kjekshus, A., Karen, V. L., Mighell, A. D., and Santoro, A. 1994. Neutron powder diffraction study of the nuclear and magnetic structures of the substitutional compounds $(Y_{1-x}Ca_x)$ $Ba_2Fe_3O_{8+\delta}$. *Phys. Rev. B* 49:3465–3472.

Onsager, L. 1944. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Phys. Rev.* 65:117–149.

Trammell, G. T. 1953. Magnetic scattering of neutrons from rare earth ions. *Phys. Rev.* 92:1387–1393.

Vasiliu-Doloc, L., Lynn, J. W., Moudden, A. H., de Leon-Guevara, A. M., and Revcolevschi, A. 1998. Structure and spin dynamics of $La_{0.85}Sr_{0.15}MnO_3$. *Phys. Rev. B* 58:14913–14921.

Williams, G. W. 1988. Polarized Neutrons. Oxford University Press, New York.

Young, R. A. 1993. The Rietveld Method. Oxford University Press, Oxford.

Zhang, H., Lynn, J. W., Li, W-H., Clinton, T. W., and Morris, D. E. 1990. Two- and three-dimensional magnetic order of the rare earth ions in $RBa_2Cu_4O_8$. *Phys. Rev. B* 41:11229–11236.

KEY REFERENCES

Bacon, 1975. See above.

This text is more for the experimentalist, treating experimental procedures and the practicalities of taking and analyzing data. It does not contain some of the newest techniques, but has most of the fundamentals. It is also rich in the history of many of the techniques.

Balcar, E. and Lovesey, S. W. 1989. Theory of Magnetic Neutron and Photon Scattering. Oxford University Press, New York.

More recent work that specifically addresses the theory for the case of magnetic neutron as well as x-ray scattering.

Lovesey, 1984. See above.

This text treats the theory of magnetic neutron scattering in depth. Vol. 1 covers nuclear scattering.

Moon et al., 1969. See above.

This is the classic article that describes the triple-axis polarized beam technique, with examples of all the fundamental measurements that can be made with polarized neutrons. Very readable.

Price, D. L. and Sköld, K. 1987. Methods of Experimental Physics: Neutron Scattering. Academic Press, Orlando, Fla.

A recent compendium that covers a variety of topics in neutron scattering, in the form of parts by various experts.

Squires, G. L. 1978. Thermal Neutron Scattering. Cambridge University Press, New York.

This book is more of a graduate introductory text to the subject of neutron scattering.

Williams, 1988. See above.

This textbook focuses on the use of polarized neutrons, with all the details.

Young, 1993. See above.

This text details the profile refinement technique for powder diffraction.

INTERNET RESOURCES

Magnetic Form Factors

http://papillon.phy.bnl.gov/form.html

Numerical Values of the Free-Ion Magnetic Form Factors

http://www.ncnr.nist.gov/resources/n-lengths/

Values of the coherent nuclear scattering amplitudes and other nuclear cross-sections.

J. W. LYNN
University of Maryland
College Park, Maryland

This page intentionally left blank

INDEX

Ab initio theory
  computational analysis, 72–74
  electronic structure analysis
    local density approximation (LDA) + U theory, 87
    phase diagram prediction, 101–102
  metal alloy magnetism, first-principles calculations, paramagnetic Fe, Ni, and Co, 189
  molecular dynamics (MD) simulation, surface phenomena, 156, 158
  neutron powder diffraction, power data sources, 1297–1298
  phonon analysis, 1324–1325
  x-ray powder diffraction, structure determination, 845
Abrasion testing, basic principles, 317
Absolute zero, thermodynamic temperature scale, 32
Absorption coefficient, x-ray absorption fine structure (XAFS) spectroscopy, 875
Acceleration, tribological testing, 333
Accidental channeling, nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1207–1208
Accuracy calculations
  energy-dispersive spectrometry (EDS), standardless analysis, 1150
  metal alloy bonding, 139–141
"Acheson" graphite, combustion calorimetry, 372
Achromatic lenses, optical microscopy, 670
Acoustic analysis. *See also* Scanning acoustic microscopy (SAM)
  impulsive stimulated thermal scattering (ISTS) *vs.,* 744–749
Adhesive mechanisms, tribological and wear testing, 324–325
Adsorbate materials, scanning tunneling microscopy (STM) analysis, 1115
Adsorption
  cryopump operation, 9–10
  vacuum system principles, outgassing, 2–3
Aerial image, optical microscopy, 668
ALCHEMI (atom location by channeling electron microscopy), diffuse intensities, metal alloys, concentration waves, multicomponent alloys, 259–260
ALGOL code, neutron powder diffraction, Rietveld refinements, 1306–1307
Alignment protocols
  Auger electron spectroscopy (AES), specimen alignment, 1161
  ellipsometry polarizer/analyzer, 738–739
    PQSA/PCSA arrangements, 738–739
  impulsive stimulated thermal scattering (ISTS), 751–752
  liquid surface x-ray diffraction, instrumetation criteria, 1037–1038
  Optical alignment, Raman spectroscopy of solids, 706
  surface x-ray diffraction
    beamline alignment, 1019–1020
    diffractometer alignment, 1020–1021
    laser alignment, 1014
    sample crystallographic alignment, 1014–1015
  ultraviolet photoelectron spectroscopy (UPS), 729–730
Alkanes, liquid surface x-ray diffraction, surface crystallization, 1043
All-electron methods, metal alloy bonding, precision measurements, self-consistency, 141–142
Allen-Cahn equation, microstructural evolution, 122
All-metal flange seal, configuration, 18
Alloys. *See* Metal alloys
Alternating current (AC) losses, superconductors, electrical transport measurement
  alternatives to, 473
  applications, 472
Alternating-gradient magnetometer (AGM), principles and applications, 534–535
Altitude, mass, weight, and density definitions, 25–26
Aluminum
  metallographic analysis
    deformed high-purity aluminum, 69
    7075-T6 aluminum alloys, microstructural evaluation, 68–69
  vacuum system construction, 17
  weight standards, 26–27
Aluminum-lithium alloys, phase diagram prediction, 108–110
Aluminum-nickel alloys, phase diagram prediction, 107
American Society for Testing and Materials (ASTM)
  balance classification, 27–28
  fracture toughness testing
    crack extension measurement, 308
    crack tip opening displacement (CTOD) (δ), 307
    standards for, 302
  sample preparation protocols, 287
  tensile test principles, 284–285
  weight standards, 26–27
Amperometric measurements, electrochemical profiling, 579–580
Amplitude reduction, x-ray absorption fine structure (XAFS) spectroscopy, single scattering picture, 871–872
Analog-to-digital converter (ADC)
  heavy-ion backscattering spectrometry (HIBS), 1279
  ion-beam analysis (IBA) experiments, ERD/RBS techniques, 1185–1186
  polarization-modulation ellipsometer, 739
  superconductors, electrical transport measurements, voltmeter properties, 477
Analytical chemistry, combustion calorimetry, 377–378
Analytical electron microscope (AEM), energy-dispersive spectrometry (EDS), automation, 1141
Analyzers
  ellipsometry, 738
    relative phase/amplitude calculations, 739–740
  trace element accelerator mass spectrometry (TEAMS), magnetic/electrostatic analyzer calibration, 1241–1242
  x-ray photoelectron spectroscopy (XPS), 980–982
ANA software, surface x-ray diffraction
  data analysis, 1024
  lineshape analysis, 1019
Anastigmats lenses, optical microscopy, 669
Angle calculations
  liquid surface x-ray diffraction, instrumetation criteria, 1037–1038
  surface x-ray diffraction, 1021
Angle-dependent tensors, resonant scattering analysis, 909–910
Angle-dispersive constant-wavelength diffractometer, neutron powder diffraction, 1289–1290
Angle-resolved x-ray photoelectron spectroscopy (ARXPS)
  applications, 986–988
  silicon example, 997–998
Angular dependence, resonant scattering, 910–912
  tensors, 916
Angular efficiency factor, x-ray photoelectron spectroscopy (XPS), elemental composition analysis, 984
Anisotropy
  metal alloy magnetism, 191–193
    MAE calculations, 192
    magnetocrystalline Co-Pt alloys, 199–200
    pure Fe, Ni, and Co, 193
  permanent magnets, 497–499
Annular detectors, scanning transmission electron microscopy (STEM), 1091
Anomalous transmission
  resonant magnetic x-ray scattering, ferromagnets, 925
  two-beam diffraction, 231–232

Antiferromagnetism
  basic principles, 494
  magnetic neutron scattering, 1335
  magnetic x-ray scattering
    nonresonant scattering, 928–930
    resonant scattering, 930–932
  neutron powder diffraction, 1288–1289
  principles and equations, 524
Antiphase domain boundary (APB), coherent ordered precipitates, microstructural evolution, 123
Antiphase domains (APDs), coherent ordered precipitates, microstructural evolution, 122–123
Antishielding factors, nuclear quadrupole resonance (NQR), 777–778
Anti-Stokes scattering, Raman spectroscopy of solids, semiclassical physics, 701–702
Aperture diffraction
  scanning transmission electron microscopy (STEM)
    phase-contrast illumination *vs.,* 1095–1097
    probe configuration, 1097–1098
  transmission electron microscopy (TEM), 1078
Archard equation, wear testing, 334
Areal density
  heavy-ion backscattering spectrometry (HIBS), 1281
  ion-beam analysis (IBA), ERD/RBS equations, 1188–1189
Argon hangup, cryopump operation, 10
Array detector spectrometers, ultraviolet/visible absorption (UV-VIS) spectroscopy, 693
Arrhenius behavior
  chemical vapor deposition (CVD) model, gas-phase chemistry, 169
  deep level transient spectroscopy (DLTS), semiconductor materials, 423
    data analysis, 425
  thermal analysis and, 343
  thermogravimetric (TG) analysis, 346–347
    kinetic theory, 352–354
Arrott plots, Landau magnetic phase transition, 529–530
Assembly procedures, vacuum systems, 19
Associated sets, lattice alloys, x-ray diffraction, local atomic correlation, 217
Astigmatism, transmission electron microscopy (TEM), 1079
Asymmetric units
  crystallography, space groups, 47–50
  diffuse scattering techniques, 889
Asymmetry parameter
  neutron powder diffraction, axial divergence peak asymmetry, 1292–1293
  nuclear quadrupole resonance (NQR)
    nutation nuclear resonance spectroscopy, 783–784
    zero-field energy leevels, 777–778
Atmospheric characteristics, gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 394
Atomic absorption spectroscopy (AAS), particle-induced x-ray emission (PIXE) and, 1211
Atomic concentrations, x-ray photoelectron spectroscopy (XPS)
  composition analysis, 994–996
  elemental composition analysis, 984–986
Atomic emission spectroscopy (AES), particle-induced x-ray emission (PIXE) and, 1211
Atomic energy levels, ultraviolet photoelectron spectroscopy (UPS), 727–728
Atomic force microscopy (AFM)
  liquid surfaces and monomolecular layers, 1028
  low-energy electron diffraction (LEED), 1121
  magnetic domain structure measurements, magnetic force microscopy (MFM), 549–550
  scanning electrochemical microscopy (SECM) and, 637
  scanning transmission electron microscopy (STEM) *vs.,* 1093
  scanning tunneling microscopy (STM) *vs.,* 1112
  surface x-ray diffraction and, 1007–1008
Atomic layer construction, low-energy electron diffraction (LEED), 1134–1135
Atomic magnetic moments
  local moment origins, 513–515
  magnetism, general principles, 491–492
Atomic probe field ion microscopy (APFIM), transmission electron microscopy (TEM) and, 1064
Atomic resolution spectroscopy (ARS), scanning transmission electron microscopy (STEM), Z-contrast imaging, 1103–1104
Atomic short-range ordering (ASRO) principles
  diffuse intensities, metal alloys, 256
    basic definitions, 252–254
    competitive strategies, 254–255
    concentration waves
      density-functional approach, 260–262
      first-principles, electronic-structure calculations, 263–266
      multicomponent alloys, 257–260
    effective cluster interactions (ECI), high-temperature experiments, 255–256
    hybridization in NiPt alloys, charge correlation effects, 266–268
    magnetic coupling and chemical order, 268
    mean-field results, 262
    mean-field theory, improvement on, 262–267
    multicomponent alloys, Fermi-surface nesting and van Hove singularities, 269–270
    pair-correlation functions, 256–257
    sum rules and mean-field errors, 257
    temperature-dependent shifts, 273
  magnetism, metal alloys, 190–191
    gold-rich AuFe alloys, 198–199
    iron-vanadium alloys, 196–198
    Ni-Fe alloys, energetics and electronic origins, 193–196
Atomic sphere approximation (ASA)
  linear muffin tin orbital method (LMTO), electronic topological transitions, van Hove singularities in CuPt, 272–273
  metal alloy bonding, precision calculations, 143
  phase diagram prediction, 101–102
Atomic structure
  surface x-ray diffraction, 1010–1011
  ultraviolet photoelectron spectroscopy (UPS), symmetry characterization, photoemission process, 726–727
  x-ray powder diffraction, candidate atom position search, 840–841
Atomistic Monte Carlo method, microstructural evolution, 117
Atomistic simulation, microstructural evolution modeling and, 128–129
Auger electron spectroscopy (AES)
  automation, 1167
  basic principles, 1159–1160
  chemical effects, 1162–1163
  competitive and related techniques, 1158–1159
  data analysis and interpretation, 1167–1168
  depth profiling, 1165–1167
  fluorescence and diffraction analysis, 940
  instrumentation criteria, 1160–1161
  ion-beam analysis (IBA) *vs.,* 1181
  ion-excitation peaks, 1171–1173
  ionization loss peaks, 1171
  limitations, 1171–1173
  line scan properties, 1163–1164
  low-energy electron diffraction (LEED), sample preparation, 1125–1126
  mapping protocols, 1164–1165
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1201–1202
  plasmon loss peaks, 1171
  qualitative analysis, 1161–1162
  quantitative analysis, 1169
  research background, 1157–1159
  sample preparation, 1170
  scanning electron microscopy (SEM) and, 1051–1052
  scanning tunneling microscopy (STM) and
    basic principles, 1113
    sample preparation, 1117
  sensitivity and detectability, 1163
  specimen modification, 1170–1171
    alignment protocols, 1161
  spectra categories, 1161
  standards sources, 1169–1170, 1174
  x-ray absorption fine structure (XAFS), detection methods, 876–877
  x-ray photoelectron spectroscopy (XPS), 978–980
    comparisons, 971
    final-state effects, 975–976
    kinetic energy principles, 972
    sample charging, 1000–1001
Auger recombination, carrier lifetime measurement, 431
  free carrier absorption (FCA), 441–442
Automated procedures
  Auger electron spectroscopy (AES), 1167
  bulk measurements, 404
  carrier lifetime measurement
    free carrier absorption (FCA), 441
    photoconductivity (PC), 445–446
    photoluminescence (PL), 451–452
  deep level transient spectroscopy (DLTS), semiconductor materials, 423–425
  differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 370
  diffuse scattering techniques, 898–899
  electrochemical quartz crystal microbalance (EQCM), 659–660
  electron paramagnetic resonance (EPR), 798
  ellipsometry, 739
  energy-dispersive spectrometry (EDS), 1141
  gas analysis, simultaneous techniques, 399
  hardness test equipment, 319
  heavy-ion backscattering spectrometry (HIBS), 1280
  impulsive stimulated thermal scattering (ISTS), 753
  low-energy electron diffraction (LEED), 1127
  magnetic domain structure measurements
    holography, 554–555
    Lorentz transmission electron microscopy, 552
    spin polarized low-energy electron microscopy (SPLEEM), 557
    x-ray magnetic circular dichroism (XMCD), 555–556
  magnetic neutron scattering, 1336
  magnetometry, 537
    limitations, 538
  magnetotransport in metal alloys, 565–566
  medium-energy backscattering, 1269
  micro-particle-induced x-ray emission (Micro-PIXE), 1216
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1205–1206

particle-induced x-ray emission (PIXE), 1216
phonon analysis, 1323
photoluminescence (PL) spectroscopy, 686
scanning electron microscopy (SEM), 1057–1058
scanning tunneling microscopy (STM) analysis, 1115–1116
semiconductor materials, Hall effect, 414
single-crystal x-ray structure determination, 860
superconductors, electrical transport measurements, 480–481
surface magneto-optic Kerr effect (SMOKE), 573
surface measurements, 406
test machine design conditions, 286–287
thermal diffusivity, laser flash technique, 387
thermogravimetric (TG) analysis, 356
thermomagnetic analysis, 541–544
trace element accelerator mass spectrometry (TEAMS), 1247
transmission electron microscopy (TEM), 1080
tribological and wear testing, 333–334
ultraviolet photoelectron spectroscopy (UPS), 731–732
ultraviolet/visible absorption (UV-VIS) spectroscopy, 693
x-ray absorption fine structure (XAFS) spectroscopy, 878
x-ray magnetic circular dichroism (XMCD), 963
x-ray microfluorescence/microdiffraction, 949
x-ray photoelectron spectroscopy (XPS), 989
Automatic recording beam microbalance, thermogravimetric (TG) analysis, 347–350
Avalanche mechanisms, *pn* junction characterization, 469
AVE averaging software, surface x-ray diffraction, data analysis, 1025
Avrami-Erofeev equation, time-dependent neutron powder diffraction, 1299–1300
Axial dark-field imaging, transmission electron microscopy (TEM), 1071
Axial divergence, neutron powder diffraction, peak asymmetry, 1292–1293

Background properties
diffuse scattering techniques, inelastic scattering backgrounds, 890–893
fluorescence analysis, 944
heavy-ion backscattering spectrometry (HIBS), 1281
single-crystal neutron diffraction, 1314–1315
x-ray absorption fine structure (XAFS) spectroscopy, 878–879
x-ray photoelectron spectroscopy (XPS)
data analysis, 990–991
survey spectrum, 974
Background radiation
energy-dispersive spectrometry (EDS), filtering algorithgm, 1143–1144
liquid surface x-ray diffraction, error sources, 1045
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1205
Back-projection imaging sequence, magnetic resonance imaging (MRI), 768
Backscattered electrons (BSEs), scanning electron microscopy (SEM)
data analysis, 1058
detector criteria and contrast images, 1055–1056
signal generation, 1051–1052
Backscatter loss correction
Auger electron spectroscopy (AES), quantitative analysis, 1169
energy-dispersive spectrometry (EDS)
standardless analysis, 1149
stray radiation, 1155
Backscatter particle filtering
ion-beam analysis (IBA), ERD/RBS examples, 1194–1197
medium-energy backscattering, 1261–1262
data analysis protocols, 1269–1270
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), particle filtering, 1204
x-ray absorption fine structure (XAFS) spectroscopy, 871–872
Bain transformation, phase diagram prediction, static displacive interactions, 105–106
Bakeout procedure, surface x-ray diffraction, ultrahigh-vacuum (UHV) systems, 1023–1024
Balance design and condition
classification of balances, 27–28
mass measurement techniques, mass, weight, and density definitions, 25–26
Ball indenter, Rockwell hardness values, 323
"Ballistic transport reaction" model, chemical vapor deposition (CVD)
free molecular transport, 170
software tools, 174
Band energy only (BEO) theory, diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 264–266
Band gap measurements, semiconductor materials, semiconductor-liquid interfaces, 613–614
Bandpass thermometer, operating principles, 37
Band theory of solids
magnetic moments, 515–516
metal alloy magnetism, electronic structure, 184–185
transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
Band-to-band recombination, photoluminescence (PL) spectroscopy, 684
Baseline determination, superconductors, electrical transport measurement, 481–482
Basis sets, metal alloy bonding, precision calculations, 142–143
Bath cooling, superconductors, electrical transport measurements, 478
Bayard-Alpert gauge, operating principles, 14–15
Beamline components
liquid surface x-ray diffraction, alignment protocols, 1037–1038
magnetic x-ray scattering, 925–927
harmonic contamination, 935
surface x-ray diffraction, alignment protocols, 1019–1020
Bearings, turbomolecular pumps, 7–8
Beer-Lambert generation function
laser spot scanning (LSS), semiconductor-liquid interface, 626–628
ultraviolet/visible absorption (UV-VIS) spectroscopy, quantitative analysis, 690–691
Beer's law, photoconductivity (PC), carrier lifetime measurement, 444–445
Bellows-sealed feedthroughs, ultrahigh vacuum (UHV) systems, 19
Bending magnetic devices
magnetic x-ray scattering, beamline properties, 926–927
x-ray magnetic circular dichroism (XMCD), 957–959
Benzoic acid, combustion calorimetry, 372
Bessel function, small-angle scattering (SAS), 220–221

Bethe approximation
energy-dispersive spectrometry (EDS), standardless analysis, 1148–1149
ion beam analysis (IBA), 1176
multiple-beam diffraction, 239
Bethe-Slater curve, ferromagnetism, 525–526
Bias conditions, deep level transient spectroscopy (DLTS), semiconductor materials, 421–422
Bidimensional rotating frame NQRI, emergence of, 787–788
Bilayer materials, angle-resolved x-ray photoelectron spectroscopy (ARXPS), 987–988
Bimolecular reaction rates, chemical vapor deposition (CVD) model, gas-phase chemistry, 169
Binary collisions, particle scattering, 51
general formula, 52
Binary/multicomponent diffusion
applications, 155
dependent concentration variable, 151
frames of reference and diffusion coefficients, 147–150
Kirkendall effect and vacancy wind, 149
lattice-fixed frame of reference, 148
multicomponent alloys, 150–151
number-fixed frame of reference, 147–148
tracer diffusion and self-diffusion, 149–150
transformation between, 148–149
volume-fixed frame of reference, 148
linear laws, 146–147
Fick's law, 146–147
mobility, 147
multicomponent alloys
Fick-Onsager law, 150
frames of reference, 150–151
research background, 145–146
substitutional and interstitial metallic systems, 152–155
B2 intermetallics, chemical order, 154–155
frame of reference and concentration variables, 152
interdiffusion, 155
magnetic order, 153–154
mobilities and diffusivities, 152–153
temperature and concentration dependence of mobilities, 153
Binary phase information, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 368
Binding energy
Auger electron spectroscopy (AES), 1159–1160
x-ray photoelectron spectroscopy (XPS)
chemical state information, 986
final-state effects, 975–976
B2 intermetallics, binary/multicomponent diffusion, 154–155
Biological materials
scanning electrochemical microscopy (SECM), 644
scanning tunneling microscopy (STM) analysis, 1115
Biot-Savart law, electromagnet structure and properties, 499–500
Bipolar junction transistors (BJTs), characterization
basic principles, 467–469
competitive and complementary techniques, 466–467
limitations of, 471
measurement equipment sources and selection criteria, 471
protocols and procedures, 469–470
research background, 466–467
sample preparation, 470–471

Bipolar magnetic field gradient, magnetic resonance imaging (MRI), flow imaging sequence, 769–770
Bipotentiostat instrumentation, scanning electrochemical microscopy (SECM), 642–643
Birefringent polarizer, ellipsometry, 738
Bitter pattern imaging, magnetic domain structure measurements, 545–547
Black-body law, radiation thermometer, 36–37
"Bleach-out" effect, nuclear quadrupole resonance (NQR), spin relaxation, 780
Bloch wave vector
  diffuse intensities, metal alloys, first-principles calcuations, electronic structure, 265–266
  dynamical diffraction, 228
  scanning transmission electron microscopy (STEM)
    dynamical diffraction, 1101–1103
    strain contrast imaging, 1106–1108
  solid-solution alloy magnetism, 183–184
Bode plots
  corrosion quantification, electrochemical impedance spectroscopy (EIS), 599–603
  semiconductor-liquid interfaces, differential capacitance measurements, 617–619
Body-centered cubic (bcc) cells
  iron (Fe) magnetism
    local moment fluctuation, 187–188
    Mössbauer spectroscopy, 828–830
  x-ray diffraction, structure-factor calculations, 209
Bohr magnetons
  magnetic fields, 496
  magnetic moments, atomic and ionic magnetism, local moment origins, 513–515
  paramagnetism, 493–494
  transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
Bohr velocity, ion beam analysis (IBA), 1175–1176
Boltzmann constant
  chemical vapor deposition (CVD) model
    gas-phase chemistry, 169
    kinetic theory, 171
  liquid surface x-ray diffraction, simple liquids, 1040–1041
  phonon analysis, 1318
  pn junction characterization, 467–468
  thermal diffuse scattering (TDS), 213
Boltzmann distribution function
  Hall effect, semiconductor materials, 412
  quantum paramagnetiic response, 521–522
  semiconductor-liquid interface, dark current-potential characteristics, 607
Bond distances
  diffuse scattering techniques, 885
  single-crystal x-ray structure determination, 861–863
Bonding-antibonding
  Friedel's *d*-band energetics, 135
  metal alloys, transition metals, 136–137
Bonding principles
  metal alloys
    limits and pitfalls, 138–144
      accuracy limits, 139–141
      all-electron *vs.* pseudopotential methods, 141–142
      basis sets, 142–143
      first principles *vs.* tight binding, 141
      full potentials, 143
      precision issues, 141–144
      self-consistency, 141
      structural relaxations, 143–144
    phase formation, 135–138
      bonding-antibonding effects, 136–137
      charge transfer and electronegativities, 137–138
      Friedel's *d*-band energetics, 135
      size effects, 137
      topologically close-packed phases, 135–136
      transition metal crystal structures, 135
      wave function character, 138
    research background, 134–135
  surface phenomena, molecular dynamics (MD) simulation, 156–157
  x-ray photoelectron spectroscopy (XPS), initial-state effects, 976–978
Boonton 72B meter, capacitance-voltage (C-V) characterization, 460–461
Borie-Sparks (BS) technique, diffuse scattering techniques, 886–889
  data interpretation, 895–897
Born approximation
  distorted-wave Born approximation
    grazing-incidence diffraction (GID), 244–245
    surface x-ray diffraction measurements, 1011
  liquid surface x-ray diffraction
    non-specular scattering, 1033–1034
    reflectivity measurements, 1032–1033
  multiple-beam diffraction, second-order approximation, 238–240
Born-Oppenheimer approximation
  electronic structure computational analysis, 74–75
  metal alloy bonding, accuracy calculations, 140–141
Born-von Kármán model, phonon analysis, 1319–1323
  data analysis, 1324–1325
Borosilicate glass, vacuum system construction, 17
Boundary conditions
  dynamical diffraction, basic principles, 229
  impulsive stimulated thermal scattering (ISTS) analysis, 754–756
  multiple-beam diffraction, NBEAM theory, 238
  two-beam diffraction
    diffracted intensities, 233
    dispersion surface, 230–231
Bound excitons, photoluminescence (PL) spectroscopy, 682
Boxcar technique, deep level transient spectroscopy (DLTS), semiconductor materials, 424–425
Bragg-Brentano diffractometer, x-ray powder diffraction, 838–839
Bragg-Fresnel optics, x-ray microprobes, 946–947
Bragg reflection principles
  diffuse intensities, metal alloys, atomic short-range ordering (ASRO) principles, 257
  diffuse scattering techniques
    intensity computations, 901–904
    research background, 882–884
  dynamical diffraction
    applications, 225
    multiple Bragg diffraction studies, 225–226
  grazing-incidence diffraction (GID), 241–242
    inclined geometry, 243
  ion-beam analysis (IBA), ERD/RBS equations, 1188–1189
  liquid surface x-ray diffraction, instrumetation criteria, 1037–1038
  local atomic order, 214–217
  magnetic neutron scattering
    diffraction techniques, 1332–1335
    inelastic scattering, 1331
    magnetic diffraction, 1329–1330
    polarized beam technique, 1330–1331
    subtraction technque, 1330
  magnetic x-ray scattering
    data analysis and interpretation, 934–935
    resonant antiferromagnetic scattering, 931–932
    surface magnetic scattering, 932–933
  multiple-beam diffraction, basic principles, 236–237
  neutron powder diffraction
    angle-dispersive constant-wavelength diffractometer, 1290
    microstrain broadening, 1294–1296
    particle size effect, 1293–1294
    peak shape, 1292
    positional analysis, 1291–1292
    probe configuration, 1286–1289
    time-of-flight diffractomer, 1290–1291
  nonresonant magnetic x-ray scattering, 920–921
  phonon analysis, 1320
  scanning transmission electron microscopy (STEM)
    incoherent scattering, 1099–1101
    phase-contrast illumination, 1094–1097
  single-crystal neutron diffraction
    instrumentation criteria, 1312–1313
    protocols, 1311–1313
    research background, 1307–1309
  small-angle scattering (SAS), 219–222
  surface/interface x-ray diffraction, 218–219
  surface x-ray diffraction
    crystallographic alignment, 1014–1015
    error detection, 1018–1019
  thermal diffuse scattering (TDS), 211–214
  transmission electron microscopy (TEM)
    Ewald sphere construction, 1066–1067
    extinction distances, 1068
    Kikuchi line origin and scattering, 1075–1076
    selected-area diffraction (SAD), 1072–1073
    structure and shape factor analysis, 1065–1066
  two-beam diffraction
    boundary conditions and Snell's law, 230–231
    Darwin width, 232
    diffracted intensities, 234–235
    diffracted wave properties, 229–230
    dispersion equation solution, 233
    hyperboloid sheets, 230
    wave field amplitude ratios, 230
    x-ray birefringence, 232–233
    x-ray standing wave (XWS) diffraction, 232, 235–236
  x-ray absorption fine structure (XAFS) spectroscopy, energy resolution, 877
  x-ray diffraction, 209
  X-ray microdiffraction analysis, 945
  x-ray powder diffraction, 837–838
    crystal lattice and space group determination, 840
    data analysis and interpretation, 839–840
    error detection, 843
    integrated intensities, 840
Brale indenter, Rockwell hardness values, 323
Bravais lattices
  diffuse intensities, metal alloys, concentration waves, 258–260
  Raman active vibrational modes, 709–710, 720–721
  single-crystal x-ray structure determination, crystal symmetry, 854–856
  space groups, 46–50
  transmission electron microscopy (TEM), structure and shape factor analysis, 1065–1066
Breakdown behavior, *pn* junction characterization, 469

Breit-Wigner equation, neutron powder diffraction, probe configuration, 1287–1289
Bremsstrahlung ischromat spectroscopy (BIS)
  ultraviolet photoelectron spectroscopy (UPS) and, 723
  x-ray photoelectron spectroscopy (XPS), 979–980
Brewster angle microscopy (BAM)
  carrier lifetime measurement, free carrier absorption (FCA), 441
  liquid surfaces and monomolecular layers, 1028
  liquid surface x-ray diffraction
    Langmuir monolayers, 1043
    p-polarized x-ray beam, 1047
Bright-field imaging
  reflected-light optical microscopy, 676–680
  transmission electron microscopy (TEM)
    complementary dark-field and selected-area diffraction (SAD), 1082–1084
    deviation vector and parameter, 1068
    protocols and procedures, 1069–1071
    thickness modification, deviation parameter, 1081–1082
Brillouin zone dimension
  diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 265–266
  magnetic neutron scattering, inelastic scattering, 1331
  magnetotransport in metal alloys, magnetic field behavior, 563–565
  metal alloy bonding, precision calculations, 143
  quantum paramagnetiic response, 521–522
  ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 725
Brinell hardness testing
  automated methods, 319
  basic principles, 317–318
  data analysis and interpretation, 319–320
  hardness values, 317–318, 323
  limitations and errors, 322
  procedures and protocols, 318–319
  research background, 316–317
  sample preparation, 320
  specimen modification, 320–321
Brittle fracture, fracture toughness testing
  load-displacement curves, 304
  Weibull distribution, 311
Broadband radiation thermometer, operating principles, 37
Bulk analysis
  secondary ion mass spectrometry (SIMS) and, 1237
  trace element accelerator mass spectrometry (TEAMS)
    data analysis protocols, 1249–1250
    impurity measurements, 1247–1249
Bulk capacitance measurements, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 624
Bulk chemical free energy
  coherent ordered precipitates, microstructural evolution, 123
  metal alloy bonding, accuracy calculations, 140–141
  microstructural evolution, 119–120
Bulk crystals, surface x-ray diffraction, 1008–1009
Bulk measurements
  automation, 404
  basic principles, 403
  limitations, 405
  protocols and procedures, 403–404
  sample preparation, 404–405
  semiconductor materials, Hall effect, 416–417
Burger's vector
  stress-strain analysis, yield-point phenomena, 283–284
  transmission electron microscopy (TEM), defect analysis, 1085
Burial pumping, sputter-ion pump, 11
Butler-Volmer equation, corrosion quantification, Tafel technique, 593–596

Cadmium plating, metallographic analysis, measurement on 4340 steel, 67
Cagliotti coefficients, neutron powder diffraction, constant wavelength diffractometer, 1304–1305
Cahn-Hilliard (CH) diffusion equation
  continuum field method (CFM), 114
    diffusion-controlled grain growth, 127–128
    interfacial energy, 120
    numerical algorithms, 129–130
  microscopic field model (MFM), 116
  microstructural evolution, field kinetics, 122
Calibration protocols
  combustion calorimetry, 379
  differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 366–367
  diffuse scattering techniques, absolute calibration, measured intensities, 894
  electron paramagnetic resonance (EPR), 797
  energy-dispersive spectrometry (EDS), 1156
  gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 393
  hot cathode ionization gauges, 16
  magnetometry, 536
  mass measurement process assurance, 28–29
  trace element accelerator mass spectrometry (TEAMS)
    data acquisition, 1252
    magnetic/electrostatic analyzer calibration, 1241–1242
  transmission electron microscopy (TEM), 1088–1089
  x-ray absorption fine structure (XAFS) spectroscopy, 877
Calorimetric bomb apparatus, combustion calorimetry, 375–377
Calorimetry. *See* Combustion calorimetry
CALPHAD group
  diffuse intensities, metal alloys, basic principles, 254
  Phase diagram predictions
    aluminum-lithium analysis, 108–110
    aluminum-nickel analysis, 107
    basic principles, 91
    cluster approach, 91–92, 96–99
    cluster expansion free energy, 99–101
    electronic structure calculations, 101–102
    ground-state analysis, 102–104
    mean-field approach, 92–96
    nickel-platinum analysis, 107–108
    nonconfigurational thermal effects, 106–107
    research background, 90–91
    static displacive interactions, 104–106
Camera equation, transmission electron microscopy (TEM), selected-area diffraction (SAD), 1072–1073
Cantilever magnetometer, principles and applications, 535
Capacitance diaphragm manometers, applications and operating principles, 13
Capacitance meters, capacitance-voltage (C-V) characterization
  limitations of, 463–464
  selection criteria, 460–461
Capacitance-voltage (C-V) characterization
  pn junctions, 467
  research background, 401
  semiconductors
    basic principles, 457–458
    data analysis, 462–463
    electrochemical profiling, 462
    instrument limitations, 463–464
    mercury probe contacts, 461–462
    profiling equipment, 460–461
    protocols and procedures, 458–460
    research background, 456–457
    sample preparation, 463
    trapping effects, 464–465
Capillary tube electrodes, scanning electrochemical microscopy (SECM), 648–659
Carbon materials, Raman spectroscopy of solids, 712–713
Carnot efficiency, thermodynamic temperature scale, 32
Carrier decay transient extraction, carrier lifetime measurement, free carrier absorption (FCA), 441
Carrier deficit, carrier lifetime measurement, 428–429
Carrier lifetime measurement
  characterization techniques, 433–435
    device related techniques, 435
    diffusuion-length-based methods, 434–435
    optical techniques, 434
  cyclotron resonance (CR), 805–806
  free carrier absorption, 438–444
    automated methods, 441
    basic principles, 438–440
    carrier decay transient, 441
    computer interfacing, 441
    data analysis and interpretation, 441–442
    depth profiling, sample cross-sectioning, 443
    detection electronics, 440–441
    geometrical considerations, 441
    lifetime analysis, 441–442
    lifetime depth profiling, 441
    lifetime mapping, 441–442
    limitations, 443–444
    probe laser selection, 440
    processed wafers, metal and highly doped layer removal, 443
    pump laser selection, 440
    sample preparation, 442–443
    virgin wafers, surface passivation, 442–443
  generation lifetime, 431
  photoconductivity, 444–450
    basic principles, 444
    data analysis, 446–447
    high-frequency range, automation, 445–446
    limitations, 449–450
    microwave PC decay, 447–449
    radio frequency PC decay, 449
    sample preparation, 447
    standard PC decay method, 444–445
  photoluminescence, 450–453
    automated procedures, 451–452
    data analysis and interpretation, 452–453
    deep level luminescence, 450–451
    limitations, 453
    near-band-gap emission, 450
    photon recycling, 451
    shallow impurity emission, 450
  physical quantities, 433
  recombination mechanisms, 429–431
  selection criteria, characterization methods, 453–454
  steady-state, modulated, and transient methods, 435–438
    data interpretation problems, 437
    limitations, 437–438
    modulation-type method, 436
    pulsed-type methods, 437–438

Carrier lifetime measurement (*Continued*)
quasi-steady-state-type method, 436–437
surface recombination and diffusion, 432–433
theoretical background, 401, 427–429
trapping techniques, 431–432
Cartesian coordinates, Raman active vibrational modes, 710
Case hardening, microindentation hardness testing, 318
Catalyst analysis, thermogravimetric (TG) analysis, gas-solid reactions, 356
Cathode burial, sputter-ion pump, 11
Cathode-ray tubes (CRT), scanning electron microscopy (SEM), 1052–1053
Cathodoluminescence (CL) technique, *pn* junction characterization, 466–467
Cauchy's inequality, single-crystal x-ray structure determination, direct method computation, 865–868
Cavity resonators, microwave measurement techniques, 409
Cellular automata (CA) method, microstructural evolution, 114–115
Center-of-mass reference frame
forward-recoil spectrometry, data interpretation, 1270–1271
particle scattering, kinematics, 54–55
Central-field theory, particle scattering, 57–61
cross-sections, 60–61
deflection functions, 58–60
approximation, 59–60
central potentials, 59
hard spheres, 58–59
impact parameter, 57–58
interaction potentials, 57
materials analysis, 61
shadow cones, 58
Central potentials, particle scattering, central-field theory, 57
deflection function, 59
Ceramic materials
turbomolecular pumps, 7
x-ray photoelectron spectroscopy (XPS), 988–989
CFD-ACE, chemical vapor deposition (CVD) model, hydrodynamics, 174
Cgs magnetic units
general principles, 492–493
magnetic vacuum permeability, 511
Chain of elementary reactions, chemical vapor deposition (CVD) model, gas-phase chemistry, 168–169
Channeling approximation, scanning transmission electron microscopy (STEM), dynamical diffraction, 1102–1103
Channel plate arrays, x-ray photoelectron spectroscopy (XPS), 982–983
Channeltrons, x-ray photoelectron spectroscopy (XPS), 982–983
Character table, vibrational Raman spectroscopy, group theoretical analysis, 718
Charge-balance equation (CBE), semiconductor materials, Hall effect, 412, 416
Charge correlation effects, diffuse intensities, metal alloys
concentration waves, first-principles calcuations, electronic structure, 266
hybriziation in NiPt alloys and, 266–268
Charge-coupled device (CCD)
low-energy electron diffraction (LEED), instrumentation and automation, 1126–1127
photoluminescence (PL) spectroscopy, 684
Raman spectroscopy of solids, dispersed radiation measurment, 707–708
single-crystal x-ray structure determination
basic principles, 850
detector components, 859–860
transmission electron microscopy (TEM)
automation, 1080
phase-contrast imaging, 1096–1097
Charged particle micropobes, x-ray microprobes, 940–941
Charge-funneling effects, ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1228
Charge transfer
cyclic voltammetry, nonreversible charge-transfer reaactions, 583–584
metal alloy bonding, 137–138
semiconductor-liquid interface
photoelectrochemistry, thermodynamics, 606–607
time-resolved photoluminescence spectroscopy (TRPL), 630–631
Charpy V-Notch impact testing, fracture toughness testing, 302
Chemically modified electrodes (CMEs), electrochemical profiling, 579–580
Chemical order
diffuse intensities, metal alloys, magnetic effects and, 268
diffuse scattering techniques
basic principles, 884–885
research background, 882–884
substitutional and interstitial metallic systems, 154–155
Chemical polishing, ellipsometry, surface preparations, 741–742
Chemical shift imaging (CSI)
Auger electron spectroscopy (AES), 1162–1163
nuclear magnetic resonance
basic principles, 764–765
imaging sequence, 770
x-ray photoelectron spectroscopy (XPS)
applications, 986
initial-state effects, 976–978
reference spectra, 996–998
research background, 970–972
Chemical thinning, transmission electron microscopy (TEM), sample preparation, 1086
Chemical vapor deposition (CVD), simulation models
basic components, 167–173
free molecular transport, 169–170
gas-phase chemistry, 168–169
hydrodynamics, 170–172
kinetic theory, 172
plasma physics, 170
radiation, 172–173
surface chemistry, 167–168
limitations, 175–176
research background, 166–167
software applications
free molecular transport, 174
gas-phase chemistry, 173
hydrodynamics, 174
kinetic theory, 175
plasma physics, 174
radiation, 175
surface chemistry, 173
Chemisorption, sputter-ion pump, 10–11
CHEMKIN software, chemical vapor deposition (CVD) models
limitations, 175–176
surface chemistry, 173
Chromatic aberration
optical microscopy, 669–670
transmission electron microscopy (TEM), lens defects and resolution, 1078
Chromatograpy, evolved gas analysis (EGA) and, 396–397
Circular dichroism. *See* X-ray magnetic circular dichroism
Circularly polarized x-ray (CPX) sources, x-ray magnetic circular dichroism (XMCD), 957–959
figure of merit, 969–970
Classical mechanics
magnetic x-ray scattering, 919–920
Raman spectroscopy of solids, electromagnetic radiation, 699–701
resonant scattering analysis, 906–907
surface magneto-optic Kerr effect (SMOKE), 570–571
Classical phase equilibrium, phase diagram predictions, 91–92
Clausius-Mosotti equation, electric field gradient, nuclear couplings, nuclear quadrupole resonance (NQR), 777
Cleaning procedures, vacuum systems, 19
Cllip-on gage, mechanical testing, extensometers, 285
Closed/core shell diamagnetism, dipole moments, atomic origin, 512
Cluster magnetism, spin glass materials, 516–517
Cluster variational method (CVM)
diffuse intensities, metal alloys, 255
concentration waves, multicomponent alloys, 259–260
mean-field theory and, 262–263
phase diagram prediction
nickel-platinum alloys, 107–108
nonconfigurational thermal effects, 106–107
phase diagram predictions, 96–99
free energy calculations, 99–100
Cmmm space group, crystallography, 47–50
Coarse-grained approximation, continuum field method (CFM), 118
Coarse-grained free energy formulation, microstructural evolution, 119
Coaxial probes, microwave measurement techniques, 409
Cobalt alloys
magnetism, magnetocrystalline anisotropy energy (MAE), 193
magnetocrystalline anisotropy energy (MAE), Co-Pt alloys, 199–200
paramagnetism, first-principles calculations, 189
Cobalt/copper superlattices, surface magneto-optic Kerr effect (SMOKE), 573–574
Cockroft-Walton particle accelerator, nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1203–1204
Coherence analysis
low-energy electron diffraction (LEED), estimation techniques, 1134
Mössbauer spectroscopy, 824–825
phonon analysis, neutron scattering, 1325–1326
scanning transmission electron microscopy (STEM)
phase-contrast imaging *vs.*, 1093–1097
research background, 1091–1092
strain contrast imaging, 1106–1108
Coherent ordered precipitates, microstructural evolution, 122–126
Coherent potential approximation (CPA)
diffuse intensities, metal alloys
concentration waves, first-principles calcuations, electronic structure, 264–266
hybridization in NiPt alloys, charge correlation effects, 267–268
metal alloy magnetism, magnetocrystalline anisotropy energy (MAE), 192–193
solid-solution alloy magnetism, 183–184
Cohesive energies, local density approximation (LDA), 79–81

Cold cathode ionization gauge, operating principles, 16
Collective magnetism, magnetic moments, 515
Collision diameter, particle scattering, deflection functions, hard spheres, 58–59
Combination bearings systems, turbomolecular pumps, 7
Combustion calorimetry
  commercial sources, 383
  data analysis and interpretation, 380–381
  limitations, 381–382
  protocols and procedures, 375–380
    analytical chemistry and, 377–378
    calibration protocols, 379
    experimental protocols, 377
    measurement protocols, 378–379
    non-oxygen gases, 380
    standard state corrections, 379–380
  research background, 373–374
  sample preparation, 381
  thermodynamic principles, 374–375
Combustion energy computation, combustion calorimetry, enthalpies of formation computation *vs.,* 380–381
Compact-tension (CT) specimens, fracture toughness testing
  crack extension measurement, 309–311
  load-displacement curve measurement, 308
  sample preparation, 311–312
  stress intensity and J-integral calculations, 314–315
Compensator, ellipsometry, 738
Compliance calculations, fracture toughness testing, 315
Compton scattering
  diffuse scattering techniques, inelastic scattering background removal, 891–893
  x-ray absorption fine-structure (XAFS) spectroscopy, 875–877
  x-ray microfluorescence, 942
Computational analysis. *See also* specific computational techniques, e.g. Schrödinger equation
  basic principles, 71–74
  electronic structure methods
    basic principles, 74–77
    dielectric screening, 84–87
    Green's function Monte Carlo (GFMC), electronic structure methods, 88–89
    GW approximation, 84–85
      local-density approximation (LDA) + *U,* 87
    Hartree-Fock theory, 77
    local-density approximation (LDA), 77–84
    local-density approximation (LDA) + *U,* 86–87
    quantum Monte Carlo (QMC), 87–89
    SX approximation, 85–86
    variational Monte Carlo (VMC), 88
  low-energy electron diffraction (LEED), 1134–1135
  phase diagram prediction
    aluminum-lithium analysis, 108–110
    aluminum-nickel analysis, 107
    basic principles, 91
    cluster approach, 91–92, 96–99
    cluster expansion free energy, 99–101
    electronic structure calculations, 101–102
    ground-state analysis, 102–104
    mean-field approach, 92–96
    nickel-platinum analysis, 107–108
    nonconfigurational thermal effects, 106–107
    research background, 90–91
    static displacive interactions, 104–106
  research background, 71
  thermogravimetric (TG) analysis, 354–355
Computer interface, carrier lifetime measurement, free carrier absorption (FCA), 441
Computer x-ray analyzer (CXA), energy-dispersive spectrometry (EDS), 1137–1140
  automation, 1141
Concentration overpotentials, semiconductor materials, J-E behavior corrections, 611–612
Concentration variables, substitutional and interstitial metallic systems, 152
Concentration waves, diffuse intensities, metal alloys, 252–254
  as competitive strategy, 255
  density-functional theory (DFT), 260–262
  first-principles, electronic-structure calculations, 263–266
  multicomponent alloys, 257–260
Conductance, vacuum system design, 19–20
Conduction cooling, superconductors, electrical transport measurements, 478
Conductive substrates, Hall effect, semiconductor materials, 417
Conductivity measurements
  Hall effect, 411–412
  research background, 401–403
  theoretical background, 401
Configuration-interaction theory, electronic structure analysis, 75
Conservation laws, ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 724–725
Considere construction, nonuniform plastic deformation, 282
Constant current regime
  photoconductivity (PC), carrier lifetime measurement, 445
  scanning electrochemical microscopy (SECM), 643
Constant field regime, photoconductivity (PC), carrier lifetime measurement, 445
Constant fraction discriminators, heavy-ion backscattering spectrometry (HIBS), 1279
Constant Q scan, phonon analysis, triple-axis spectrometry, 1322–1323
Constant wavelength diffractometer, neutron powder diffraction, optical properties, 1303–1304
Constraint factors
  neutron powder diffraction, 1301
  static indentation hardness testing, 317
Construction materials, vacuum systems, 17
Contact materials
  superconductors, electrical transport measurement, 474
    sample contacts, 483
  superconductors, electrical transport measurements, properties, 477–478
  tribological and wear testing, 324–325
Contact size, Hall effect, semiconductor materials, 417
Continuity equation, chemical vapor deposition (CVD) model, hydrodynamics, 171
Continuous-current ramp measurements, superconductors, electrical transport measurements, 479
Continuous interfaces, liquid surface x-ray diffraction, reflectivity measurements, 1031–1033
Continuous-loop wire saw cutting, metallographic analysis, sample preparation, 65
Continuous magnetic fields, laboratory settings, 505
Continuous-wave (CW) experiments, electron paramagnetic resonance (EPR)
  calibration, 797
  error detection, 800–802
  microwave power, 796
  modulation amplitude, 796–797
  non-X-band frequencies, 795–796
  sensitivity, 797
  X-band with rectangular resonator, 794–795
Continuous wave nuclear magnetic resonance (NMR), magnetic field measurements, 510
Continuum field method (CFM), microstructural evolution
  applications, 122
  atomistic simulation, continuum process modeling and property calculation, 128–129
  basic principles, 117–118
  bulk chemical free energy, 119–120
  coarse-grained approximation, 118
  coarse-grained free energy formulation, 119
  coherent ordered precipitates, 122–126
  diffuse-interface nature, 119
  diffusion-controlled grain growth, 126–128
  elastic energy, 120–121
  external field energies, 121
  field kinetic equations, 121–122
  future applications, 128
  interfacial energy, 120
  limits of, 130
  numerical algorithm efficiency, 129–130
  research background, 114–115
  slow variable selection, 119
  theoretical basis, 118
Continuum process modeling, microstructural evolution modeling and, 128–129
Contrast images, scanning electron microscopy (SEM), 1054–1056
Control issues, tribological testing, 332–333
Controlled-rate thermal analysis (CRTA), mass loss, 345
Conventional transmission electron microscopy (CTEM). *See* Transmission electron microscopy (TEM)
Convergent-beam electron diffraction (CBED), transmission electron microscopy (TEM), selected-area diffraction (SAD), 1072–1073
Cooling options and procedures, superconductors, electrical transport measurements, 478–479
Coordinate transformation, resonant scattering techniques, 916–917
Copper-nickel-zinc alloys
  Fermi-surface nesting and van Hove singularities, 268–270
  ordering wave polarization, 270–271
Copper-platinum alloys, topological transitions, van Hove singularities, 271–273
Core-level absorption spectroscopy, ultraviolet photoelectron spectroscopy (UPS) and, 725–726
Core-level emission spectroscopy, ultraviolet photoelectron spectroscopy (UPS) and, 725–726
Core-level photoelectron spectroscopy, ultraviolet photoelectron spectroscopy (UPS) and, 725–726
Corrosion studies
  electrochemical quantification
    electrochemical impedance spectroscopy (EIS), 599–603
    linear polarization, 596–599
    research background, 592–593
    Tafel technique, 593–596
  scanning electrochemical microscopy (SECM), 643–644
  thermogravimetric (TG) analysis, gas-solid reactions, 355–356
Corrosive reaction gases, thermogravimetric (TG) analysis, instrumentation and apparatus, 348

Coster-Kronig transition
  Auger electron spectroscopy (AES), 1159–1160
  particle-induced x-ray emission (PIXE), 1211–1212
Coulomb interaction
  electronic structure analysis
    dielectric screening, 84
    local density approximation (LDA) + U theory, 87
  GW approximation and, 75
  ion-beam analysis (IBA), ERD/RBS equations, 1187–1189
  Mössbauer spectroscopy, isomer shift, 821–822
  particle scattering
    central-field theory, cross-sections, 60–61
    shadow cones, 58
Coulomb potential, metal alloy bonding, accuracy calculations, 139–141
Coulomb repulsion, metal alloy magnetism, 180
Counterflow mode, leak detection, vacuum systems, 21–22
Count rate range, energy-dispersive spectrometry (EDS)
  measurement protocols, 1156
  optimization, 1140
Coupling potential energy
  coherent ordered precipitates, microstructural evolution, 125–126
  magnetic x-ray scattering errors, 935–936
Cowley-Warren (CW) order parameter. *See* Warren-Cowley order parameter
Crack driving force *(G),* fracture toughness testing
  basic principles, 303–305
  defined, 302
Crack extension measurement
  fracture toughness testing, 308
    J-integral approach, 311
    SENB and CT specimens, 315
  hardness testing, 320
Crack instability and velocity, load-displacement curves, fracture toughness testing, 304
Crack propagation, hardness testing, 320
Crack tip opening displacement (CTOD) (δ), fracture toughness testing, 307
  errors, 313
  sample preparation, 312
Critical current density (J_c)
  superconductors, electrical transport measurement, 472
    data analysis and interpretation, 481
  superconductors, electrical transport measurements, current-carrying area, 476
Critical current (I_c)
  superconducting magnets, stability and losses, 501–502
  superconductors, electrical transport measurement, data analysis and interpretation, 481
  superconductors, electrical transport measurements, 472
    criteria, 474–475
    current-carrying area, 476
Critical temperature (T_c)
  ferromagnetism, 495
  superconductors, electrical transport measurement, 472
    data analysis and interpretation, 481
    magnetic measurements *vs.,* 473
Cromer-Liberman tabulation, diffuse scattering techniques, resonant scattering terms, 893–894
Cross-modulation, cyclotron resonance (CR), 812
Cross-sections
  diffuse scattering techniques, resonant scattering terms, 893–894
  ion-beam analysis (IBA), 1181–1184
    ERD/RBS equations, 1187–1189
    non-Rutherford cross-sections, 1189–1190
  magnetic neutron scattering, error detection, 1337–1338
  medium-energy backscattering, 1261
  metal alloy magnetism, atomic short range order (ASRO), 191
  nonresonant antiferromagnetic scattering, 929–930
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), Q values, 1205
  particle scattering, central-field theory, 60–61
  phonon analysis, 1318–1323
  resonant magnetic x-ray scattering, 924
  superconductors, electrical transport measurement, area definition, 482
  x-ray microfluorescence, 943
Crucible apparatus, thermogravimetric (TG) analysis, 351–352
Cryopumps
  applications, 9
  operating principles and procedures, 9–10
Crystallography
  crystal systems, 44
  lattices, 44–46
    Miller indices, 45–46
  metal alloy bonding, phase formation, 135
  neutron powder diffraction, refinement techniques, 1300–1302
  point groups, 42–43
  resonant x-ray diffraction, 905
  space groups, 46–50
    symbols, 50
  surface x-ray diffraction measurements, 1010–1011
    refinement limitations, 1018
  symmetry operators, 39–42
    improper rotation axes, 39–40
    proper rotation axes, 39
    screw axes and glide planes, 40–42
  symmetry principles, 39
  x-ray diffraction, 208–209
Crystal structure
  conductivity measurements, research background, 402–403
  diffuse scattering techniques, 885–889
  dynamical diffraction, applications, 225
  Mössbauer spectroscopy, defect analysis, 830–831
  photoluminescence (PL) spectroscopy, 685–686
  single-crystal neutron diffraction, quality issues, 1314–1315
  single-crystal x-ray structure determination
    refinements, 856–858
    symmetry properties, 854–856
  surface x-ray diffraction, alignment protocols, 1014–1015
  x-ray powder diffraction, lattice analysis, 840
Crystal truncation rods (CTR)
  liquid surface x-ray diffraction, non-specular scattering, 1036, 1038–1039
  surface x-ray diffraction
    basic properties, 1009
    error detection, 1018–1019
    interference effects, 219
    profiles, 1015
    reflectivity, 1015–1016
    silicon surface example, 1017–1018
  surface x-ray scattering, surface roughness, 1016
Crystral structure, materials characterization, 1
CSIRO Heavy Ion Analytical Facility (HIAF), trace element accelerator mass spectrometry (TEAMS) research at, 1245
Cubic lattice alloys
  diffuse scattering techniques, static displacement recovery, 897–898
  x-ray diffraction, local atomic correlation, 216–217
Curie law
  collective magnetism, 515
  paramagnetism, 493–494
  quantum paramagnetiic response, 521–522
Curie temperature
  band theory of magnetism, 515–516
  face-centered cubic (fcc) iron, moment alignment *vs.* moment formation, 194–195
  ferromagnetism, 523–524
  metal alloy magnetism, 181
  substitutional and interstitial metallic systems, 153–154
  thermogravimetric (TG) analysis, temperature measurement errors and, 358–359
Curie-Weiss law
  ferromagnetism, 524
  paramagnetism, 493–494
  thermomagnetic analysis, 543–544
Current carriers, illuminated semiconductor-liquid interface, J-E equations, 608
Current-carrying area, superconductors, electrical transport measurement, 476
Current density, semiconductor-liquid interface, 607–608
Current image tunneling sepctroscopy (CITS), image acquisition, 1114
Current-potential curve, corrosion quantification, linear polarization, 597–599
Current ramp rate, superconductors, electrical transport measurements, 479
Current sharing, superconductors, electrical transport measurement, other materials and, 474
Current supply, superconductors, electrical transport measurements, 476–477
Current transfer, superconductors, electrical transport measurement, 475–476
Current-voltage *(I-V)* measurement technique
  low-energy electron diffraction (LEED)
    quantitative analysis, 1130–1131
    quantitative measurement, 1124–1125
  pn junction characterization
    basic principles, 467–469
    competitive and complementary techniques, 466–467
    limitations of, 471
    measurement equipment sources and selection criteria, 471
    protocols and procedures, 469–470
    research background, 466–467
    sample preparation, 470–471
  superconductors, electrical transport measurement, baseline determination, 481–482
Cyclic voltammetry
  electrochemical profiling
    data analysis and interpretation, 586–587
    electrocatalytic mechanism, 587–588
    electrochemical cells, 584
    electrolytes, 586
    limitations, 590–591
    platinum electrode example, 588–590
    potentiostats and three-electrode chemical cells, 584–585
    reaction reversibility, 582–584
      non-reversible charge-transfer reactions, 583–584
      quasireversible reaction, 583
      total irreversible reaction, 583
      total reversible reaction, 582–583
    research background, 580–582

sample preparation, 590
specimen modification, 590
working electrodes, 585–586
feedback mode, scanning electrochemical microscopy (SECM), 638–640
Cyclotron resonance (CR)
basic principles, 806–808
cross-modulation, 812
data analysis and interpretation, 813–814
far-infrared (FIR) sources, 809
Fourier transform FIR magneto-spectroscopy, 809–810
laser far infrared (FIR) magneto-spectroscopy, 810–812
limitations, 814–815
optically-detected resonance (ODR) spectroscopy, 812–813
protocols and procedures, 808–813
quantum mechanics, 808
research background, 805–806
sample preparation, 814
semiclassical Drude model, 806–807
Cylinders, small-angle scattering (SAS), 220–221
Cylindrically-focused excitation pulse, *vs.* impulsive stimulated thermal scattering (ISTS), 745–746
Cylindrical mirror analyzer (CMA)
Auger electron spectroscopy (AES), 1160–1161
ultraviolet photoelectron spectroscopy (UPS), 729
x-ray photoelectron spectroscopy (XPS), 980–982

Dangling bonds, carrier lifetime measurement, surface passivation, free carrier absorption (FCA), 442–443
Dark current-potential characteristics, semiconductor-liquid interface, 607
Dark-field imaging
reflected-light optical microscopy, 676–680
transmission electron microscopy (TEM)
axial dark-field imaging, 1071
complementary bright-field and selected-area diffraction (SAD), 1082–1084
deviation vector and parameter, 1068
protocols and procedures, 1069–1071
thickness modification, deviation parameter, 1081–1082
Darwin width
neutron powder diffraction, peak shapes, 1292
two-beam diffraction, 232
dispersion equation solution, 233
Data reduction, thermal diffusivity, laser flash technique, 389
DC Faraday magnetometer
automation, 537
principles and applications, 534
Deadtime correction function
energy-dispersive spectrometry (EDS), 1137–1140
optimization, 1141
particle-induced x-ray emission (PIXE), 1213–1216
Debye-Grüneisen framework, phase diagram prediction, nonconfigurational thermal effects, 106–107
Debye length, *pn* junction characterization, 467–468
Debye-Scherrer geometry
liquid surface x-ray diffraction, instrumetation criteria, 1037–1038
neutron powder diffraction
axial divergence peak asymmetry, 1292–1293
quantitative phase analysis, 1297
x-ray powder diffraction, 838–839
error detection, 843
Debye temperature, semiconductor materials, Hall effect, 415–416
Debye-Waller factor
diffuse intensity computation, 901–904
liquid surface x-ray diffraction, grazing incidence and rod scans, 1036
magnetic neutron scattering, 1329–1330
diffraction techniques, 1333–1335
Mössbauer spectroscopy, recoil-free fraction, 821
phonon analysis, 1320
scanning transmission electron microscopy (STEM)
incoherent scattering, 1100–1101
strain contrast imaging, 1107–1108
single-crystal neutron diffraction and, 1310–1311
surface x-ray diffraction
crystallographic refinement, 1018
crystal truncation rod (CTR) profiles, 1015
thermal diffuse scattering (TDS), 211–214
x-ray absorption fine structure (XAFS) spectroscopy, disordered states, 873
x-ray powder diffraction, 837–838
Decomposition temperature, thermogravimetric (TG) analysis, 351–352
stability/reactivity measurements, 355
Deep level luminescence, carrier lifetime measurement, photoluminescence (PL), 450
Deep level transient spectroscopy (DLTS)
basic principles, 419–423
emission rate, 420–421
junction capacitance transient, 421–423
capacitance-voltage (C-V) analysis, comparisons, 401, 456–457
data analysis and interpretation, 425
limitations, 426
pn junction characterization, 467
procedures and automation, 423–425
research background, 418–419
sample preparation, 425–426
semiconductor defects, 418–419
Defect analysis
dynamical diffraction, applications, 225
photoluminescence (PL) spectroscopy, transitions, 683–684
transmission electron microscopy (TEM)
Kikuchi lines and deviation parameter, 1085
values for, 1084–1085
Deflection functions
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), particle filtering, 1204
particle scattering, central-field theory, 58–60
approximation, 59–60
central potentials, 59
hard spheres, 58–59
Density
definitions, 24–26
surface phenomena, molecular dynamics (MD) simulation, layer density and temperature variation, 164
Density-functional theory (DFT)
diffuse intensities, metal alloys
basic principles, 254
concentration waves, 260–262
first-principles calcuations, electronic structure, 263–266
limitations, 255
electronic structure analysis, local density approximation (LDA), 77–78
metal alloy bonding, accuracy calculations, 139–141
transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Density measurements
basic principles, 24
impulsive stimulated thermal scattering (ISTS), 749
indirect techniques, 24
mass, weight, and density definitions, 24–26
mass measurement process assurance, 28–29
materials characterization, 1
weighing devices, balances, 27–28
weight standards, 26–27
Dependent concentration variable, binary/multicomponent diffusion, 151
substitutional and interstitial metallic systems, 152
Depletion effects, Hall effect, semiconductor materials, 417
Depletion widths, capacitance-voltage (C-V) characterization, 458–460
Depth of focus, optical microscopy, 670
Depth-profile analysis
Auger electron spectroscopy (AES), 1165–1167
carrier lifetime measurement, free carrier absorption (FCA), 443
heavy-ion backscattering spectrometry (HIBS), 1276
ion-beam analysis (IBA), ERD/RBS techniques, 1191–1197
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1202–1203
nonresonant depth profiling, 1206
secondary ion mass spectrometry (SIMS) and, 1237
semiconductor materials, Hall effect, 415–416
surface x-ray diffraction measurements, grazing-incidence measurement, 1016–1017
trace element accelerator mass spectrometry (TEAMS)
data analysis protocols, 1251–1252
impurity measurements, 1250–1251
Depth resolution
ion beam analysis (IBA)
ERD/RBS techniques, 1183–1184
periodic table, 1177–1178
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), equations, 1209–1210
Derivative thermal analysis, defined, 337–338
Derivative thermogravity (DTG)
curve schematic, 345
defined, 338
kinetic theory, 352–354
procedures and protocols, 346–347
Design criteria, vacuum systems, 19–20
Detailed balance principle
carrier lifetime measurement, 429
deep level transient spectroscopy (DLTS), semiconductor materials, 420–421
Detection methods
Auger electron spectroscopy (AES), 1163
carrier lifetime measurement, free carrier absorption (FCA), 440–441
energy-dispersive spectrometry (EDS), limitations of, 1151–1152
ion-beam analysis (IBA), ERD/RBS geometries, 1185–1186
trace element accelerator mass spectrometry (TEAMS), high-energy beam transport, analysis, and detection, 1239
x-ray absorption fine structure (XAFS) spectroscopy, 875–877
x-ray magnetic circular dichroism (XMCD), 959
Detector criteria
diffuse scattering techniques, 894–897
energy-dispersive spectrometry (EDS), 1137–1140

Detector criteria (*Continued*)
solid angle protocols, 1156
standardless analysis, efficiency parameters, 1149
fluorescence analysis, 944
ion-beam analysis (IBA), ERD/RBS techniques, 1185–1186
magnetic x-ray scattering, 927–928
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1204
particle-induced x-ray emission (PIXE), 1212–1216
silicon detector efficiency, 1222–1223
photoluminescence (PL) spectroscopy, 684
scanning electron microscopy (SEM), 1054–1056
single-crystal neutron diffraction, 1313
x-ray photoelectron spectroscopy (XPS), 982–983
Deuteron acceleration, nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), equipment criteria, 1203–1204
Deviation vector and parameter, transmission electron microscopy (TEM), 1067–1068
Kikuchi lines, defect contrast, 1085
specimen orientation, 1077–1078
specimen thickness, 1081–1082
Device-related techniques, carrier lifetime measurement, 435
Diagonal disorders, diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 266
Diamagnetism
basic principles, 494
closed/core shell diamagnetism, dipole moments, atomic origin, 512
Diaphragm pumps
applications, 4
operating principles, 5
Dichroic signal, x-ray magnetic circular dichroism (XMCD), 953–955
basic theory, 955–957
data analysis, 963–964
Dielectric constants, ellipsometry, 737
Dielectric screening, electronic structure analysis, 84–87
Dielectric tensor, surface magneto-optic Kerr effect (SMOKE), 570–571
Difference equations, neutron powder diffraction, stacking faults, 1296
Differential capacitance measurements, semiconductor-liquid interfaces, 616–619
Differential cross-section, particle scattering, central-field theory, 60
Differential interference contrast (DIC), reflected-light optical microscopy, 679–680
Differential ion pump, operating principles, 11
Differential scanning calorimetry (DSC). *See also* Simultaneous thermogravimetry (TG)-differential thermal analysis/differential scanning calorimetry (TG-DTA/DSC)
automated procedures, 370
basic principles, 363–366
data analysis and interpretation, 370–371
defined, 337–339
limitations, 372
protocols and procedures, 366–370
applications, 367–370
calibration, 366–367
instrumentation, 367
zero-line optimization, 366
research background, 362–363
sample preparation, 371
specimen modification, 371–372
thermal analysis, 339
thermogravimetric (TG) analysis, 346
Differential scattering cross-sections, ion-beam analysis (IBA), 1181–1184
Differential thermal analysis (DTA). *See also* Simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA)
automated procedures, 370
basic principles, 363–366
data analysis and interpretation, 370–371
defined, 337–339
limitations, 372
protocols and procedures, 366–370
applications, 367–370
calibration, 366–367
instrumentation, 367
zero-line optimization, 366
research background, 362–363
sample preparation, 371
simultaneous techniques for gas analysis, research background, 392–393
specimen modification, 371–372
thermogravimetric (TG) analysis, 346
Diffraction techniques. *See also* specific techniques, e.g. X-ray diffraction
defined, 207
density definition and, 26
diffuse scattering and, 884
magnetic neutron scattering, 1332–1335
Mössbauer spectroscopy, 824–825
transmission electron microscopy (TEM)
bright-field/dark-field imaging, 1070–1071
lattice defect diffraction contrast, 1068–1069
pattern indexing, 1073–1074
Diffractometer components
liquid surface x-ray diffraction, instrumetation criteria, 1037–1038
neutron powder diffraction
angle-dispersive constant-wavelength diffractometer, 1289–1290
time-of-flight diffractomer, 1290–1291
surface x-ray diffraction, 1009–1010
alignment protocols, 1020–1021
angle calculations, 1021
five-circle diffractometer, 1013–1014
Diffuse intensities
metal alloys
atomic short-range ordering (ASRO)
principles, 256
concentration waves
density-functional approach, 260–262
first-principles, electronic-structure calculations, 263–266
multicomponent alloys, 257–260
mean-field results, 262
mean-field theory, improvement on, 262–267
pair-correlation functions, 256–257
sum rules and mean-field errors, 257
competitive and related techniques, 254–255
computational principles, 252–254
data analysis and interpretation, 266–273
Cu_2NiZn ordering wave polarization, 270–271
CuPt van Hove singularities, electronic topological transitions, 271–273
magnetic coupling and chemical order, 268
multicomponent alloys, 268–270
Ni-Pt hybridization and charge correlation, 266–268
temperature-dependent shift, ASRO peaks, 273
high-temperature experiments, effective interactions, 255–256
two-beam diffraction
boundary condition, 233
Bragg case, 234–235
integrated intensities, 235
Laue case, 233–234
Diffuse-interface approach, microstructural evolution, 119
Diffuse scattering techniques
intensity computations, 886–889
absolute calibration, measured intensities, 894
derivation protocols, 901–904
liquid surface x-ray diffraction
non-specular scattering, 1038–1039
simple liquids, 1041
surface x-ray scattering, 1016
x-ray and neutron diffuse scattering
applications, 889–894
automation, 897–898
bond distances, 885
chemical order, 884–885
comparisons, 884
competitive and related techniques, 883–884
crystalling solid solutions, 885–889
data analysis and interpretation, 894–896
diffuse x-ray scattering techniques, 889–890
inelastic scattering background removal, 890–893
limitations, 898–899
measured intensity calibration, 894
protocols and procedures, 884–889
recovered static displacements, 896–897
research background, 882–884
resonant scattering terms, 893–894
sample preparation, 898
Diffusion coefficients
binary/multicomponent diffusion, 146
multicomponent diffusion, frames of reference, 150–151
substitutional and interstitial metallic systems, 152–153
Diffusion-controlled grain growth, microstructural evolution, 126–128
Diffusion imaging sequence, magnetic resonance imaging (MRI), 769
Diffusion length
carrier lifetime measurement
free carrier absorption (FCA), limitations of, 443–444
methods based on, 434–435
surface recombination and diffusion, 432–433
semiconductor materials, semiconductor-liquid interfaces, 614–616
Diffusion potential, capacitance-voltage (C-V) characterization, *p-n* junctions, 457–458
Diffusion processes, binary/multicomponent diffusion
applications, 155
dependent concentration variable, 151
frames of reference and diffusion coefficients, 147–150
linear laws, 146–147
multicomponent alloys
Fick-Onsager law, 150
frames of reference, 150–151
research background, 145–146
substitutional and interstitial metallic systems, 152–155
Diffusion pumps
applications, 6
operating principles and procedures, 6–7
Dimagnetism, principles of, 511
Dimensionless quantitates, coherent ordered precipitates, microstructural evolution, 124–126
Dimer-adatom-stacking fault (DAS) model, surface x-ray diffraction, silicon surface example, 1017–1018
Diode structures, characterization
basic principles, 467–469

competitive and complementary techniques, 466–467
limitations of, 471
measurement equipment sources and selection criteria, 471
protocols and procedures, 469–470
research background, 466–467
sample preparation, 470–471
Dipole magnetic moments
atomic origin, 512–515
closed/core shell diamagnetism, 512
coupling theories, 519–527
antiferromagnetism, 524
ferrimagnetism, 524–525
ferromagnetism, 522–524
Heisenberg model and exchange interactions, 525–526
helimagnetism, 526–527
paragmagnetism, classical and quantum theories, 519–522
local atomic moments
collective magnetism, 515
ionic magnetism, 513–515
neutron powder diffraction, 1288–1289
nuclear quadrupole resonance (NQR), nuclear moments, 775–776
Dirac equation
computational analysis, applications, 71
Mössbauer effect, 819–820
Mössbauer spectroscopy, hyperfine magnetic field (HMF), 823–824
Dirac's delta function, phonon analysis, 1318
Direct configurational averaging (DCA), phase diagram prediction
cluster variation, 99
electronic structure, 101–102
ground-state analysis, 102–104
Direct detection procedures, nuclear quadrupole resonance (NQR), 780–781
Direct methods
single-crystal x-ray structure determination, 860–861
computations, 865–868
surface x-ray diffraction measurements, crystallographic refinement, 1018
x-ray absorption spectroscopy, 870
Discrete Fourier transform (DFT), thermal diffusivity, laser flash technique, 388–389
Disk-shaped electrodes, scanning electrochemical microscopy (SECM), 648–659
Disordered local moment (DLM)
face-centered cubic (fcc) iron, moment alignment *vs.* moment formation, 193–195
metal alloy magnetism
first-principles calculations, 188–189
gold-rich AuFe alloys, atomic short range order (ASRO), 198–199
iron-vanadium alloys, 196–198
local exchange splitting, 189–190
local moment fluctuation, 187–188
Disordered states, x-ray absorption fine structure (XAFS) spectroscopy, 872–873
Dispersed radiation measurment, Raman spectroscopy of solids, 707–708
Dispersion corrections
transmission electron microscopy (TEM), sample preparation, 1087
x-ray diffraction, scattering power and length, 210
Dispersion equation, two-beam diffraction, solution, 233
Dispersion surface
dynamical diffraction, basic principles, 228–229
two-beam diffraction, 230–231
boundary conditions and Snell's law, 230–231
hyperboloid sheets, 230
Poynting vector and energy flow, 231
wave field amplitude ratios, 230
Dispersive transmission measurement, x-ray absorption fine structure (XAFS) spectroscopy, 875–877
Displacement principle
density definition and, 26
diffuse scattering techniques
diffuse intensity computations, 902–904
static displacement recovery, 897–898
Distorted-wave Born approximation (DWBA)
grazing-incidence diffraction (GID), 244–245
liquid surface x-ray diffraction, non-specular scattering, 1034–1036
surface x-ray diffraction measurements, grazing incidence, 1011
Documentation protocols
in thermal analysis, 340–341
thermogravimetry (TG), 361–362
tribological testing, 333
Doniach-Sunjic peak shape, x-ray photoelectron spectroscopy (XPS), 1006
Donor-acceptor transitions, photoluminescence (PL) spectroscopy, 683
Doping dependence, carrier lifetime measurement, 431
Double-beam spectrometer, ultraviolet/visible absorption (UV-VIS) spectroscopy, 692–693
Double-counting corrections, diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 264–266
Double-crystal monochromator
liquid surface x-ray diffraction, error detection, 1044–1045
surface x-ray diffraction, beamline alignment, 1019–1020
Double-resonance experiments
electron-nuclear double resonance (ENDOR), 796
nuclear quadrupole resonance (NQR), frequency-swept Fourier-transform spectrometers, 780–781
Double sum techniques, diffuse intensity computations, 902–904
Dougle-critical-angle phenomenon, grazing-incidence diffraction (GID), 243
DPC microscopy, magnetic domain structure measurements, Lorentz transmission electron microscopy, 551–552
Drive stiffness, tribological and wear testing, 328–329
Drude model, cyclotron resonance (CR), 806–807
Drude-Sommerfeld "free electron" model, magnetotransport in metal alloys, 562–563
Dry gases, vacuum system principles, outgassing, 2–3
Dufour effect, chemical vapor deposition (CVD) model, hydrodynamics, 171
"Dummy" orientation matrix (DOM), surface x-ray diffraction
crystallographic alignment, 1014–1015
protocols, 1026
Dynamical diffraction
applications, 225–226
Bragg reflections, 225–226
crystals and multilayers, 225
defect topology, 225
grazing-incidence diffraction, 226
internal field-dependent phenomena, 225
basic principles, 227–229
boundary conditions, 229
dispersion surface, 228–229
fundamental equations, 227–228
internal fields, 229
grazing incidence diffraction (GID), 241–246
distorted-wave Born approximation, 244–245
evanescent wave, 242
expanded distorted-wave approximation, 245–246
inclinded geometry, 243
multlayers and superlattices, 242–243
specular reflectivity, 241–242
literature sources, 226–227
multiple-beam diffraction, 236–241
basic principles, 236–237
NBEAM theory, 237–238
boundary conditions, 238
D-field component eigenequation, 237
eigenequation matrix, 237–238
intensity computations, 238
numerical solution strategy, 238
phase information, 240
polarization density matrix, 241
polarization mixing, 240
second-order Born approximation, 238–240
standing waves, 240
three-beam interactions, 240
scanning transmission electron microscopy (STEM), Z-contrast imaging, 1101–1103
theoretical background, 224–225
two-beam diffraction, 229–236
anomalous transmission, 231–232
Darwin width, 232
diffuse intensities
boundary condition, 233
Bragg case, 234–235
integrated intensities, 235
Laue case, 233–234
dispersion equation solution, 233
dispersion surface properties, 230–231
boundary conditions, Snell's law, 230–231
hyperboloid sheets, 230
Poynting vector and energy flow, 231
wave field amplitude ratios, 230
Pendellösung, 231
standing waves, 235–236
x-ray birefringence, 232–233
x-ray standing waves (XSWs), 232
Dynamical processes, nuclear quadrupole resonance (NQR), 790
Dynamic random-access memory (DRAM) device
heavy-ion backscattering spectrometry (HIBS)
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1230
Dynamic stability conditions, superconducting magnets, 502
Dynamic thermomechanometry, defined, 339

Echo-planar imaging sequence, magnetic resonance imaging (MRI), 768–769
ECPSSR model, particle-induced x-ray emission (PIXE), 1212
Eddy-current measurements, non-contact techniques, 407–408
Effective cluster interactions (ECIs)
diffuse intensities, metal alloys, 255
high-temperature experiments, 255–256
mean-field results, 263
phase diagram prediction, ground-state analysis, 102–104
phase diagram predictions, cluster variational method (CVM), 98–99
Effective drift diffusion model, chemical vapor deposition (CVD), plasma physics, 170
Effectively probed volume, carrier lifetime measurement, free carrier absorption (FCA), probe selection criteria, 440
Effective mass approximation (EMA), cyclotron resonance (CR), 806
quantum mechanics, 808

Eigenequations
  dynamical diffraction, boundary conditions, 229
  multiple-beam diffraction
    D-field components, 237
    matrix form, 237–238
Elastic collision
  particle scattering, 51–52
    diagrams, 52–54
  phase diagram prediction, static displacive interactions, 105–106
Elastic constants
  impulsive stimulated thermal scattering (ISTS), 745–746
  local density approximation (LDA), 81–82
Elastic energy, microstructural evolution, 120–121
Elastic ion scattering
  composition analysis
    applications, 1191–1197
    basic concepts, 1181–1184
    detector criteria and detection geometries, 1185–1186
    equations, 1186–1189
    experimental protocols, 1184–1186
    limitations, 1189–1191
    nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1201–1202
    research background, 1179–1181
  Rutherford backscattering spectroscopy (RBS), 1178
Elastic properties
  impulsive stimulated thermal scattering (ISTS) analysis, 747–749
  tension testing, 279
Elastic recoil detection analysis (ERDA)
  composition analysis
    applications, 1191–1197
    basic concepts, 1181–1184
    detector criteria and detection geometries, 1185–1186
    equations, 1186–1189, 1199–1200
    experimental protocols, 1184–1186
    limitations, 1189–1191
    research background, 1179–1181
  elastic scattering, 1178
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1201–1202
Elastomers, O-ring seals, 17
Electrical and electronic measurements
  bulk measurements, 403–405
  conductivity measurement, 401–403
  microwave techniques, 408–410
  non-contact methods, 407–408
  research background, 401
  surface measurements, 405–406
Electrical feedthroughs, vacuum systems, 18
Electrical interference, thermogravimetric (TG) analysis, mass measurement errors and, 357–358
Electrical-resistance thermometer, operating principles, 36
Electrical resistivity, magnetotransport in metal alloys
  magnetic field behavior, 563–565
  research applications, 559
  transport equations, 560–561
  zero field behavior, 561–563
Electrical transport measurements, superconductors
  automation of, 480–481
  bath temperature fluctuations, 486–487
  competitive and related techniques, 472–473
  contact materials, 474
  cooling options and procedures, 478–479
  critical current criteria, 474–475
  current-carring area for critical current to current density, 476
  current contact length, 477
  current ramp rate and shape, 479
  current supply, 476–477
  current transfer and transfer length, 475–477
  data analysis and interpretation, 481–482
  electromagnetic phenomena, 479
  essential theory
    current sharing, 474
    four-point measurement, 473–474
    Ohm's law, 474
    power law transitions, 474
  generic protocol, 480
  instrumentation and data acquisition, 479–480
  lead shortening, 486
  magnetic field strength extrapolation and irreversibility field, 475
  maximum measurement current determination, 479
  probe thermal contraction, 486
  research background, 472–473
  sample handling/damage, 486
  sample heating and continuous-current measurements, 486
  sample holding and soldering/making contacts, 478
  sample preparation, 482–483
  sample quality, 487
  sample shape, 477
  self-field effects, 487
  signal-to-noise ratio parameters, 483–486
    current supply noise, 486
    grounding, 485
    integration period, 486
    pick-up loops, 483–484
    random noise and signal spikes, 485–486
    thermal electromotive forces, 484–485
  specimen modification, 483
  thermal cycling, 487
  troubleshooting, 480
  voltage tap placement, 477
  voltmeter properties, 477
  zero voltage definition, 474
Electric discharge machine (EDM), fracture toughness testing, sample preparation, 312
Electric-discharge machining, metallographic analysis, sample preparation, 65
Electric field gradient (EFG)
  Mössbauer spectroscopy
    data analysis and interpretation, 831–832
    electric quadrupole splitting, 822–823
    hyperfine interactions, 820–821
  nuclear couplings, nuclear quadrupole resonance (NQR), 776–777
  semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 606
  two-dimensional exchange NQR spectroscopy, 785
Electric phenomena, superconductors, electrical transport measurements, 479
Electric potential distributions, semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 606
Electric potential drop technique, fracture toughness testing, crack extension measurement, 308
Electric quadrupole splitting, Mössbauer spectroscopy, 822–823
Electric resistance *(R),* magnetotransport in metal alloys, basic principles, 559
Electrocatalytic mechanism, cyclic voltammetry, 587–588
Electrochemical analysis. *See also* Photoelectrochemistry
  capacitance-voltage (C-V) characterization, 462
  corrosion quantification
    electrochemical impedance spectroscopy (EIS), 599–603
    linear polarization, 596–599
    research background, 592–593
    Tafel technique, 593–596
  cyclic voltammetry
    data analysis and interpretation, 586–587
    electrocatalytic mechanism, 587–588
    electrochemical cells, 584
    electrolytes, 586
    limitations, 590–591
    platinum electrode example, 588–590
    potentiostats and three-electrode chemical cells, 584–585
    reaction reversibility, 582–584
      non-reversible charge-transfer reactions, 583–584
      quasireversible reaction, 583
      total irreversible reaction, 583
      total reversible reaction, 582–583
    research background, 580–582
    sample preparation, 590
    specimen modification, 590
    working electrodes, 585–586
  quartz crystal microbalance (QCM)
    automation, 659–660
    basic principles, 653–658
    data analysis and interpretation, 660
    equivalent circuit, 654–655
    film and solution effects, 657–658
    impedance analysis, 655–657
    instrumentation criteria, 648–659
    limitations, 661
    quartz crystal properties, 659
    research background, 653
    sample preparation, 660–661
    series and parallel frequency, 660
    specimen modification, 661
  research background, 579–580
  scanning electrochemical microscopy (SECM)
    biological and polymeric materials, 644
    competitive and related techniques, 637–638
    constant-current imaging, 643
    corrosion science applications, 643–644
    feedback mode, 638–640
    generation/collection mode, 641–642
    instrumentation criteria, 642–643
    limitations, 646–648
    reaction rate mode, 640–641
    research background, 636–638
    specimen modification, 644–646
    tip preparation protocols, 648–649
Electrochemical cell design, semiconductor materials, photocurrent/photovoltage measurements, 609–611
Electrochemical cells, cyclic voltammetry, 584
Electrochemical dissolution, natural oxide films, ellipsometric measurement, 742
Electrochemical impedance spectroscopy (EIS), corrosion quantification, 599–603
Electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 623–626
  bulk capacitance measurements, 624
  electrode bias voltage, 625–626
  electrolyte selection, 625
  experimental protocols, 626
  kinetic rates, 625
  limitations, 626
  measurement frequency, 625
  surface capacitance measurements, 624–625
  surface recombination capture coeffieicnt, 625

Electrochemical polishing, ellipsometry, surface preparations, 742
Electrode bias voltage, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 625–626
Electrode materials, cyclic voltammetry, 585–586
Electrolytes
  cyclic voltammetry, 586
  semiconductor materials, electrochemical photocapacitance spectroscopy (EPS), 625
Electrolytic polishing, metallographic analysis
  deformed high-purity aluminum, 69
  sample preparation, 66–67
Electromagnetic radiation, classical physics, Raman spectroscopy of solids, classical physics, 699–701
Electromagnetic spectrum, optical imaging, 666
Electromagnets
  applications, 496
  structure and properties, 499–500
  x-ray magnetic circular dichroism (XMCD), error sources, 966
Electron area detectors, single-crystal x-ray structure determination, 859–860
Electron beam induced current (EBIC) technique
  carrier lifetime measurement, diffusion-length-based methods, 434–435
  ion-beam-induced charge (IBIC) microscopy and, 1223–1224
    topographic analysis, 1226–1227
  pn junction characterization, 466–467
Electron beams, Auger electron spectroscopy (AES), 1160–1161
Electron deformation densities, nuclear quadrupole resonance (NQR), 777
Electron diffraction
  single-crystal neutron diffraction and, 1309
  single-crystal x-ray structure determination, 851
  x-ray diffraction *vs.,* 836
Electron diffuse scattering, transmission electron microscopy (TEM), Kikuchi lines, 1075
Electronegativities, metal alloy bonding, 137–138
Electron energy loss spectroscopy (EELS)
  scanning transmission electron microscopy (STEM)
    atomic resolution spectroscopy, 1103–1104
    comparisons, 1093
    instrumentation criteria, 1104–1105
  transmission electron microscopy (TEM) and, 1063–1064
  x-ray photoelectron spectroscopy (XPS), survey spectrum, 974
Electron excitation, energy-dispersive spectrometry (EDS), 1136
Electron flux equations, x-ray photoelectron spectroscopy (XPS), elemental composition analysis, 983–985
Electron guns
  low-energy electron diffraction (LEED), 1126–1127
  scanning electron microscopy (SEM)
    instrumentation criteria, 1053–1057
    selection criteria, 1061
  x-ray photoelectron spectroscopy (XPS), sample charging, 1000–1001
Electron holography, magnetic domain structure measurements, 554–555
Electronic balances, applications and operation, 28
Electronic density of states (DOS)
  diffuse intensities, metal alloys
    concentration waves, first-principles calcuations, electronic structure, 264–266
    hybridization in NiPt alloys, charge correlation effects, 267–268
  nickel-iron alloys, atomic long and short range order (ASRO), 193–196
  transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
Electronic phase transitions, ultraviolet photoelectron spectroscopy (UPS), 727
Electronic structure computational analysis
  basic principles, 74–77
  dielectric screening, 84–87
  diffuse intensities, metal alloys
    concentration waves, first-principles calculations, 263–266
    topological transitions, van Hove singularities in CuPt, 271–273
  Green's function Monte Carlo (GFMC), electronic structure methods, 88–89
  GW approximation, 84–85
    local-density approximation (LDA) + *U,* 87
  Hartree-Fock theory, 77
  local-density approximation (LDA), 77–84
  local-density approximation (LDA) + *U,* 86–87
  metal alloy magnetism, 184–185
    nickel-iron alloys, atomic long and short range order (ASRO), 193–196
  phase diagram prediction, 101–102
  quantum Monte Carlo (QMC), 87–89
  resonant x-ray diffraction, 905
  SX approximation, 85–86
  variational Monte Carlo (VMC), 88
Electron microprobe analysis (EMPA), micro-particle-induced x-ray emission (Micro-PIXE) and, 1210–1211
Electron multipliers, x-ray photoelectron spectroscopy (XPS), 982–983
Electron-nuclear double resonance (ENDOR)
  basic principles, 796
  error detection, 801
Electron paramagnetic resonance (EPR)
  automation, 798
  basic principles, 762–763, 793–794
  calibration, 797
  continuous-wave (CW) experiments
    microwave power, 796
    modulation amplitude, 796–797
    non-rectangular resonators, 796
    non-X-band frequencies, 795–796
    sensitivity parameters, 797
    X-band with rectangular resonator, 794–795
  data analysis and interpretation, 798
  electron-nuclear double resonance (ENDOR), 796
  instrumentation criteria, 804
  limitations, 800–802
  pulsed/Fourier transform EPR, 796
  research background, 792–793
  sample preparation, 798–799
  specimen modification, 799–800
Electron penetration, energy-dispersive spectrometry (EDS), 1155–1156
Electrons, magnetism, general principles, 492
Electron scattering
  energy-dispersive spectrometry (EDS), matrix corrections, 1145
  resonant scattering analysis, 907
Electron scattering quantum chemistry (ESQC), scanning tunneling microscopy (STM), 1116–1117
Electron sources, x-ray photoelectron spectroscopy (XPS), 978–980
Electron spectrometers, ultraviolet photoelectron spectroscopy (UPS), 729
  alignment protocols, 729–730
  automation, 731–732
Electron spectroscopy for chemical analysis (ESCA)
  Auger electron spectroscopy (AES) *vs.,* 1158–1159
  ultraviolet photoelectron spectroscopy (UPS) and, 725–726
  x-ray photoelectron spectroscopy (XPS), comparisons, 971
Electron spin echo envelope modulation (ESEEM), electron paramagnetic resonance (EPR) and, 801
Electron spin resonance (ESR), basic principles, 762–763
Electron stimulated desorption (ESD), hot cathode ionization gauges, 15–16
Electron stopping, energy-dispersive spectrometry (EDS), matrix corrections, 1145
Electron stripping, trace element accelerator mass spectrometry (TEAMS), secondary ion acceleration and, 1238–1239
Electron techniques
  Auger electron spectroscopy (AES)
    automation, 1167
    basic principles, 1159–1160
    chemical effects, 1162–1163
    competitive and related techniques, 1158–1159
    data analysis and interpretation, 1167–1168
    depth profiling, 1165–1167
    instrumentation criteria, 1160–1161
    ion-excitation peaks, 1171–1173
    ionization loss peaks, 1171
    limitations, 1171–1173
    line scan properties, 1163–1164
    mapping protocols, 1164–1165
    plasmon loss peaks, 1171
    qualitative analysis, 1161–1162
    quantitative analysis, 1169
    research background, 1157–1159
    sample preparation, 1170
    sensitivity and detectability, 1163
    specimen modification, 1170–1171
      alignment protocols, 1161
    spectra categories, 1161
    standards sources, 1169–1170, 1174
  energy-dispersive spectrometry (EDS)
    automation, 1141
    background removal, 1143–1144
    basic principles, 1136–1140
    collection optimization, 1140–1141
    deadtime correction, 1141
    energy calibration, 1140
    escape peaks, 1139
    limitations, 1153–1156
      electron preparation, 1155–1156
      ground loops, 1153–1154
      ice accumulation, 1155
      light leakage, 1154–1155
      low-energy peak distortion, 1155
      stray radiation, 1155
    matrix corrections, 1145–1147
      electron scattering, 1145
      electron stopping, 1145
      secondary x-ray fluorescence, 1145
      x-ray absorption, 1145
    measurement protocols, 1156–1157
    nonuniform detection efficiency, 1139–1140
    peak overlap, deconvolution, 1144–1145
    qualitative analysis, 1141–1143
    quantitative analysis, 1143
    research background, 1135–1136
    resolution/count rate range, 1140
    sample preparation, 1152
    specimen modification, 1152–1553
    spectrum distortion, 1140
    standardless analysis, 1147–1148
      accuracy testing, 1150
      applications, 1151

Electron techniques (*Continued*)
first-principles standardless analysis, 1148–1149
fitted-standards standardless analysis, 1149–1150
sum peaks, 1139
x-ray detection limits, 1151–1152
low-energy electron diffraction (LEED)
automation, 1127
basic principles, 1121–1125
complementary and related techniques, 1120–1121
data analysis and interpretation, 1127–1131
instrumentation criteria, 1125–1127
limitations, 1132
qualitative analysis
basic principles, 1122–1124
data analysis, 1127–1128
quantitative measurements
basic principles, 1124–1125
data analysis, 1128–1131
research background, 1120–1121
sample preparation, 1131–1132
specimen modification, 1132
research background, 1049–1050
scanning electron microscopy (SEM)
automation, 1057–1058
constrast and detectors, 1054–1056
data analysis and interpretation, 1058
electron gun, 1061
image formation, 1052–1053
imaging system components, 1061, 1063
instrumentation criteria, 1053–1054
limitations and errors, 1059–1060
research background, 1050
resolution, 153
sample preparation, 1058–1059
selection criteria, 1061–1063
signal generation, 1050–1052
specimen modification, 1059
techniques and innovations, 1056–1057
vacuum system and specimen handling, 1061
scanning transmission electron microscopy (STEM)
atomic resolution spectroscopy, 1103–1104
coherent phase-contrast imaging and, 1093–1097
competitive and related techniques, 1092–1093
data analysis and interpretation, 1105–1108
object function retrieval, 1105–1106
strain contrast, 1106–1108
dynamical diffraction, 1101–1103
incoherent scattering, 1098–1101
weakly scattered objects, 1111
limitations, 1108
manufacturing sources, 1111
probe formation, 1097–1098
protocols and procedures, 1104–1105
research background, 1090–1093
sample preparation, 1108
specimen modification, 1108
scanning tunneling microscopy (STM)
automation, 1115–1116
basic principles, 1113–1114
complementary and competitive techniques, 1112–1113
data analysis and interpretation, 1116–1117
image acquisition, 1114
limitations, 1117–1118
material selection and limitations, 1114–1115
research background, 1111–1113
sample preparation, 1117
transmission electron microscopy (TEM)
automation, 1080
basic principles, 1064–1069
deviation vector and parameter, 1066–1067
Ewald sphere construction, 1066
extinction distance, 1067–1068
lattice defect diffraction contrast, 1068–1069
structure and shape factors, 1065–1066
bright field/dark-field imaging, 1069–1071
data analysis and interpretation, 1080–1086
bright field/dark-field, and selected-area diffraction, 1082–1084
defect analysis values, 1084–1085
Kikuchi lines and deviation parameter, defect contrast, 1085–1086
shape factor effect, 1080–1081
specimen thickness and deviation parameter, 1081–1082
diffraction pattern indexing, 1073–1074
Kikuchi lines and specimen orientation, 1075–1078
deviation parameters, 1077–1078
electron diffuse scattering, 1075
indexing protocols, 1076–1077
line origins, 1075–1076
lens defects and resolution, 1078–1080
aperture diffraction, 1078
astigmatism, 1079
chromatic aberration, 1078
resolution protocols, 1079–1080
spherical aberration, 1078
limitations, 1088–1089
research background, 1063–1064
sample preparation, 1086–1087
dispersion, 1087
electropolishing and chemical thinning, 1086
ion milling and focused gallium ion beam thinning, 1086–1087
replication, 1087
ultramicrotomy, 1087
selected-area diffraction (SAD), 1071–1073
specimen modification, 1087–1088
tilted illumination and diffraction, 1071
tilting specimens and electron beams, 1071
Electron transfer rate, scanning electrochemical microscopy (SECM), reaction rate mode, 640–641
Electron yield measurements, x-ray magnetic circular dichroism (XMCD), 961–962
Electropolishing, transmission electron microscopy (TEM), sample preparation, 1086
Elemental composition analysis, x-ray photoelectron spectroscopy (XPS), 983–986
Ellipsoids, small-angle scattering (SAS), 220
Ellipsometry
automation, 739
intensity-measuring ellipsometer, 739
null elipsometers, 739
data analysis and interpretation, 739–741
optical properties from phase and amplitude changes, 740–741
polarizer/analyzer phase and amplitude changes, 739–740
dielectric constants, 737
limitations, 742
liquid surfaces and monomolecular layers, 1028
optical conductivity, 737
optical constants, 737
PQSA/PCSA component arrangement, 738
relative phase/amplitude readings, 740
protocols and procedures, 737–739
alignment, 738–739
compensator, 738
light source, 738
polarizer/analyzer, 738
reflecting surfaces, 735–737
reflectivity, 737
research background, 735
sample preparation, 741–742
Emanation thermal analysis, defined, 338
Embedded-atom method (EAM)
electronic structure analysis, phase diagram prediction, 101–102
molecular dynamics (MD) simulation, surface phenomena, 159
interlayer surface relaxation, 161
metal surface phonons, 161–162
Emissivity measurements
deep level transient spectroscopy (DLTS), semiconductor materials, 420–421
radiation thermometer, 37
Emittance measurements, radiation thermometer, 37
"Empty lattice" bands, ultraviolet photoelectron spectroscopy (UPS), 730
Endothermic reaction
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 370–371
thermal analysis and principles of, 342–343
Energy band dispersion
collective magnetism, 515
ultraviolet photoelectron spectroscopy (UPS), 723–725
automation, 731–732
solid materials, 727
Energy calibration
energy-dispersive spectrometry (EDS), optimization, 1140–1141
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE)
equations, 1209
resonance depth profiling, 1204–1205
semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 605–606
x-ray absorption fine structure (XAFS) spectroscopy, 877
Energy content, magnetic field effects and applications, 496
Energy-dispersive spectrometry (EDS)
automation, 1141
background removal, 1143–1144
basic principles, 1136–1140
collection optimization, 1140–1141
deadtime correction, 1141
energy calibration, 1140
escape peaks, 1139
ion-beam analysis (IBA) *vs.*, 1181
limitations, 1153–1156
electron preparation, 1155–1156
ground loops, 1153–1154
ice accumulation, 1155
light leakage, 1154–1155
low-energy peak distortion, 1155
stray radiation, 1155
matrix corrections, 1145–1147
electron scattering, 1145
electron stopping, 1145
secondary x-ray fluorescence, 1145
x-ray absorption, 1145
measurement protocols, 1156–1157
nonuniform detection efficiency, 1139–1140
peak overlap, deconvolution, 1144–1145
qualitative analysis, 1141–1143
quantitative analysis, 1143
research background, 1135–1136
resolution/count rate range, 1140
sample preparation, 1152
specimen modification, 1152–1553
spectrum distortion, 1140
standardless analysis, 1147–1148

accuracy testing, 1150
applications, 1151
first-principles standardless analysis, 1148–1149
fitted-standards standardless analysis, 1149–1150
sum peaks, 1139
x-ray detection limits, 1151–1152
Energy-dispersive spectroscopy (EDS)
transmission electron microscopy (TEM) and, 1063–1064
x-ray magnetic circular dichroism (XMCD), 962
x-ray photoelectron spectroscopy (XPS), comparisons, 971
Energy-filtered transmission electron microscopy (EFTEM)
basic principles, 1063–1064
bright-field/dark field imaging, 1069
Energy flow, two-beam diffraction, dispersion surface, 231
Energy precision, resonance spectroscopies, 761–762
Energy product $(BH)_{max}$, permanent magnets, 497–499
Energy resolution
ion-beam analysis (IBA), ERD/RBS techniques, 1191–1197
x-ray absorption fine structure (XAFS) spectroscopy, 877
Engineering strain, stress-strain analysis, 280–281
Enhanced backscattering spectrometry (EBS)
composition analysis
applications, 1191–1197
basic concepts, 1181–1184
detector criteria and detection geometries, 1185–1186
equations, 1186–1189
experimental protocols, 1184–1186
limitations, 1189–1191
research background, 1179–1181
ion-beam analysis (IBA)
ERD/RBS examples, 1194–1197
non-Rutherford cross-sections, 1189–1191
Enthalpy
chemical vapor deposition (CVD) model, hydrodynamics, 171
combustion calorimetry, 371–372
enthalpies of formation computation, 380–381
differential thermal analysis (DTA)/differential scanning calorimetry (DSC)
applications, 368–370
calibration protocols, 366–367
thermal analysis and principles of, 342–343
Entropy. *See also* Maximum entropy method (MEM)
phase diagram predictions, cluster variational method (CVM), 96–99
thermodynamic temperature scale, 31–32
Environmental testing, stress-strain analysis, 286
Equal arm balance
classification, 27
mass, weight, and density definitions, 25–26
Equilibrium charge density, semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 606
Equilibrium constant, thermal analysis and principles of, 342–343
Equivalent circuit model
electrochemical quartz crystal microbalance (EQCM), 654–655
semiconductor-liquid interfaces, differential capacitance measurements, 616–617
Equivalent positions
space groups, 48–50
symmetry operators, 41–42
Erosion mechanics, tribological and wear testing, 324–325
Error detection
Auger electron spectroscopy (AES), 1171–1173
carrier lifetime measurement, photoconductivity (PC) techniques, 446–447
combustion calorimetry, 381–382
corrosion quantification
electrochemical impedance spectroscopy (EIS), 602–603
linear polarization, 599
Tafel technique, 596
cyclic voltammetry, 590–591
cyclotron resonance (CR), 814–815
deep level transient spectroscopy (DLTS), semiconductor materials, 426
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 372
diffuse scattering techniques, 899–900
electrochemical quartz crystal microbalance (EQCM), 661
electron paramagnetic resonance (EPR), 800–802
ellipsometric measurement, 742
energy-dispersive spectrometry (EDS), 1153–1156
electron preparation, 1155–1156
ground loops, 1153–1154
ice accumulation, 1155
light leakage, 1154–1155
low-energy peak distortion, 1155
stray radiation, 1155
fracture toughness testing, 312–313
Hall effect, semiconductor materials, 417
hardness testing, 322
heavy-ion backscattering spectrometry (HIBS), 1281–1282
impulsive stimulated thermal scattering (ISTS) analysis, 757–758
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1233
liquid surface x-ray diffraction, 1043–1045
low-energy electron diffraction (LEED), 1132
magnetic neutron scattering, 1337–1338
magnetic x-ray scattering, 935
magnetometry, 538–539
magnetotransport in metal alloys, 567–568
mass measurement process assurance, 28–29
medium-energy backscattering and forward-recoil spectrometry, 1271–1272
neutron powder diffraction, 1300–1302
nuclear magnetic resonance, 772
nuclear quadrupole resonance (NQR), 789–790
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1207–1208
particle-induced x-ray emission (PIXE) analysis, 1220
phonon analysis, 1326–1327
photoluminescence (PL) spectroscopy, 687
pn junction characterization, 471
Raman spectroscopy of solids, 713–714
reflected-light optical microscopy, 680–681
resonant scattering, 914–915
scanning electrochemical microscopy (SECM), 646–648
scanning electron microscopy (SEM), 1059–1060
scanning transmission electron microscopy (STEM), 1108
scanning tunneling microscopy (STM), 1117–1118
simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 399
single-crystal neutron diffraction, 1314–1315
single-crystal x-ray structure determination, 863–864
surface magneto-optic Kerr effect (SMOKE) evolution, 575
surface x-ray diffraction measurements, 1018–1019
thermal diffusivity, laser flash technique, 390
thermogravimetric (TG) analysis
mass measurement errors, 357–358
temperature measurement errors, 358–359
trace element accelerator mass spectrometry (TEAMS), 1253
transmission electron microscopy (TEM), 1088–1089
tribological and wear testing, 335
ultraviolet photoelectron spectroscopy (UPS), 733
ultraviolet/visible absorption (UV-VIS) spectroscopy, 696
wear testing, 330–331
x-ray absorption fine structure (XAFS) spectroscopy, 880
x-ray magnetic circular dichroism (XMCD), 965–966
x-ray microfluorescence/microdiffraction, 950–951
x-ray photoelectron spectroscopy (XPS), 999–1001
x-ray powder diffraction, 842–843
Escape peaks, energy-dispersive spectrometry (EDS), 1139
qualitative analysis, 1142–1143
Estimated standard deviations, neutron powder diffraction, 1305–1306
Etching procedures, metallographic analysis
7075-T6 anodized aluminum alloy, 69
sample preparation, 67
4340 steel, 67
cadmium plating composition and thickness, 68
Ettingshausen effect, magnetotransport in metal alloys, transport equations, 561
Euler differencing technique, continuum field method (CFM), 129–130
Eulerian cradle mount, single-crystal neutron diffraction, 1312–1313
Euler-Lagrange equations, diffuse intensities, metal alloys, concentration waves, density-functional theory (DFT), 260–262
Euler's equation, crystallography, point groups, 42–43
Evanescent wave, grazing-incidence diffraction (GID), 242
Evaporation
ellipsometric measurements, vacuum conditions, 741
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 393
Everhart process measurement assurance program, mass measurement process assurance, 28–29
Everhart/Thornley (ET) detector, scanning electron microscopy (SEM), 1055–1056
Evolved gas analysis (EGA)
chromatography and, 396–397
defined, 338
simultaneous techniques
mass spectroscopy for thermal degradation studies, 395–396
research background, 393
Evolved gas detection (EGD)
chemical and physical principles, 398–399
defined, 338

Ewald sphere construction
  dynamical diffraction, basic principles, 228–229
  single-crystal x-ray structure determination, 853
  transmission electron microscopy (TEM), 1066–1067
    Kikuchi line orientation, 1078
    tilting specimens and electron beams, 1071
Ewald-von Laue equation, dynamical diffraction, 227–228
Excess carriers, carrier lifetime measurement, 428–429
  free carrier absorption (FCA), 438–439
Exchange current density, semiconductor-liquid interface, dark current-potential characteristics, 607
Exchange effects, resonant magnetic x-ray scattering, 922–924
Exchange-splitting
  band theory of magnetism, 515–516
  metal alloy paramagnetism, finite temperatures, 187
  ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 727
Excitation functions
  energy-dispersive spectrometry (EDS), standardless analysis, 1148–1149
  nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), resonant depth profiling, 1210
Exciton-polaritons, photoluminescence (PL) spectroscopy, 682
Excitons, photoluminescence (PL) spectroscopy, 682
Exothermic reaction
  differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 370–371
  thermal analysis and principles of, 342–343
Expanded distorted-wave approximation (EDWA), grazing-incidence diffraction (GID), 245–246
Extended programmable read-only memory (EPROM), ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1229–1230
Extended x-ray absorption fine structure (EXAFS) spectroscopy
  comparisons with other techniques, 870
  detection methods, 875–877
  diffuse scattering techniques, comparisons, 883–884
  disordered states, 873
  L edges, 873
  micro-XAFS, 943
  research background, 869–870
Extensometers, mechanical testing, 285
External field energy, continuum field method (CFM), 121
Extinction conditions
  neutron powder diffraction, microstrain broadening, 1295
  single-crystal x-ray structure determination, crystal symmetry, 856
  transmission electron microscopy (TEM), extinction distance, 1068–1069
Extreme temperatures, stress-strain analysis, 285–286
Extrinsic range of operation, capacitance-voltage (C-V) characterization, 460

Fabry factor, resistive magnets, power-field relations, 503
Face-centered cubic (fcc) cells
  diffuse intensities, metal alloys, multicomponent alloys, Fermi-surface nesting and van Hove singularities, 269–270
  iron magnetism, moment alignment *vs.* moment formation, 193–194
  phonon analysis, 1319–1323
  x-ray diffraction, structure-factor calculations, 209
Faraday coils, automated null ellipsometers, 739
Faraday constant, corrosion quantification, Tafel technique, 593–596
Faraday cup
  energy-dispersive spectrometry (EDS), stray radiation, 1155
  ion beam analysis (IBA), ERD/RBS techniques, 1184–1186
  low-energy electron diffraction (LEED), 1126–1127
Faraday effect. *See also* DC Faraday magnetometer
  magnetic domain structure measurements, magneto-optic imaging, 547–549
  surface magneto-optic Kerr effect (SMOKE), 570–571
  x-ray magnetic circular dichroism (XMCD), 953–955
Far field regime, scanning tunneling microscopy (STM), 1112–1113
Far-infrared (FIR) radiation, cyclotron resonance (CR)
  Drude model, 807
  error detection, 814–815
  Fourier transform FIR magneto-spectroscopy, 809–810
  laser magneto-spectroscopy (LMS), 810–812
  optically-detected resonance (ODR) spectroscopy, 812–813
  research background, 805–806
  sources, 809
Fast Fourier transform algorithm
  continuum field method (CFM), 129–130
  corrosion quantification, electrochemical impedance spectroscopy (EIS), 600–603
  nuclear quadrupole resonance spectroscopy, temperature, stress, and pressure measurements, 788
  scanning tunneling microscopy (STM), error detection, 1117–1118
Feedback mode, scanning electrochemical microscopy (SECM), 636
  basic principles, 638–640
FEFF program, x-ray absorption fine structure (XAFS) spectroscopy, data analysis, 879
Fermi contact interaction, Mössbauer spectroscopy, hyperfine magnetic field (HMF), 823–824
Fermi-Dirac probability
  deep level transient spectroscopy (DLTS), semiconductor materials, 420–421
  metal alloy paramagnetism, finite temperatures, 186–187
  semiconductor-liquid interface, flat-band potential measurements, 628–629
Fermi energy level
  band theory of magnetism, 516
  capacitance-voltage (C-V) characterization, trapping measurements, 465
  pn junction characterization, 467–468
  scanning tunneling microscopy (STM), 1113–1114
  semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 606
  transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
  ultraviolet photoelectron spectroscopy (UPS)
    automation, 731–732
    electronic phase transitions, 727
    energy band dispersion, 727
    full width at half maximum values, 730
    photoemission *vs.* inverse photoemission, 722–723
Fermi filling factor, diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 265–266
Fermi's golden rule
  resonant scattering analysis, 908–909
  x-ray absorption fine structure (XAFS) spectroscopy, single-scattering picture, 870–872
Fermi's pseudopotential, neutron powder diffraction, probe configuration, 1287–1289
Fermi-surface nesting
  electronic topological transitions, van Hove singularities in CuPt, 272–273
  multicomponent alloys, 268–270
Ferrimagnetism, principles and equations, 524–525
Ferroelectric materials, microbeam analysis, strain distribution, 948–949
Ferromagnetism
  basic principles, 494–495
  collective magnetism as, 515
  magnetic x-ray scattering, 924–925
    nonresonant scattering, 930
    resonant scattering, 932
  mean-field theory, 522–524
  permanent magnets, 497–499
  ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 725, 727
Feynman-Peierls' inequality, metal alloy magnetism
  atomic short range order (ASRO), 190–191
  first-principles calculations, 188–189
Fiberoptic thermometer, operating principles, 37
Fick-Onsager law, binary/multicomponent diffusion, 150
Fick's law
  binary/multicomponent diffusion
    basic equations, 146–147
    Fick-Onsager extension, 150
    general formulation, 147
    mobility, 147
    number-fixed frame of reference, 147–148
  chemical vapor deposition (CVD) model, hydrodynamics, 171
  semiconductor-liquid interface, laser spot scanning (LSS), 626–628
Field cooling, superconductor magnetization, 518–519
Field cycling methods, nuclear quadrupole resonance (NQR)
  indirect detection, 781
  spatially resolved imaging, 788–789
Field-dependent diffraction, dynamical diffraction, applications, 225
Field-effect transistor (FET), energy-dispersive spectrometry (EDS), 1137–1140
Field emission electron guns (FESEMs)
  applications, 1057
  scanning electron microscopy (SEM)
    instrumentation criteria, 1054
    resolution parameters, 1053
    selection criteria, 1061
Field factor, electromagnet structure and properties, 499–500
Field ion microscopy (FIM)
  diffuse scattering techniques, comparisons, 884
  transmission electron microscopy (TEM) and, 1064
Field kinetic equations, microstructural evolution, 121–122
Field simulation. *See also* Continuum field method (CFM)

microstructural evolution, 113–117
atomistic Monte Carlo method, 117
cellular automata (CA) method, 114–115
continuum field method (CFM), 114–115
conventional front-tracking, 113–114
inhomogeneous path probability method (PPM), 116–117
mesoscopic Monte Carlo method, 114–115
microscopic field model, 115–116
microscopic master equations, 116–117
molecular dynamics (MD), 117
Figure of merit
ultraviolet photoelectron spectroscopy (UPS) light sources, 734–735
x-ray magnetic circular dichroism (XMCD), circularly polarized x-ray sources, 969–970
Film properties
chemical vapor deposition (CVD) models, limitations, 175–176
electrochemical quartz crystal microbalance (EQCM), 657–658
impulsive stimulated thermal scattering (ISTS), 744–746, 745–746
applications, 749–752
automation, 753
competitive and related techniques, 744–746
data analysis and interpretation, 753–757
limitations, 757–758
procedures and protocols, 746–749
research background, 744–746
sample preparation and specimen modification, 757
ion-beam analysis (IBA), ERD/RBS techniques, 1192–1197
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1206
particle-induced x-ray emission (PIXE), 1213–1216
single-crystal x-ray structure determination, 859–860
surface phenomena, molecular dynamics (MD) simulation, 156–157
Filtering procedures, x-ray absorption fine structure (XAFS) spectroscopy, 878–879
Final-state effects, x-ray photoelectron spectroscopy (XPS), 974–976
Finite difference equation, continuum field method (CFM), 129–130
Finitei-pulse-time effect, thermal diffusivity, laser flash technique, 388–389
First Law of thermodynamics
combustion calorimetry and principles of, 374–375
radiation thermometer, 36–37
thermal analysis and principles of, 341–343
thermodynamic temperature scale, 31–32
First-principles calculations
diffuse intensities, metal alloys
basic principles, 254
concentration waves
density-functional theory (DFT), 260–262
electronic-structure calculations, 263–266
energy-dispersive spectrometry (EDS), standardless analysis, 1148–1149
metal alloy bonding, precision measurements, 141
metal alloy magnetism, 188–189
limitations, 201
paramagnetic Fe, Ni, and Co, 189
molecular dynamics (MD) simulation, surface phenomena, 158
Fitted-standards standardless analysis, energy-dispersive spectrometry (EDS), 1149–1150
Five-circle diffractometer, surface x-ray diffraction, 1013–1014
angle calculations, 1021
Fixed spin moment (FSM) calculations, transition metal magnetic ground state, itinerant magnetism at zero temperature, 183
Fixture size and properties, fracture toughness testing, 312–313
Flash diffusivity measurements, thermal diffusivity, laser flash technique, 390
Flash getter pump, applications, 12
Flat-band potential measurements, semiconductor-liquid interface, 628–630
Flow imaging sequence, magnetic resonance imaging (MRI), 769–770
Fluctuating Local Band (FLB) theory, metal alloy magnetism, local moment fluctuation, 187–188
FLUENT software, chemical vapor deposition (CVD) model, hydrodynamics, 174
Fluorescence microscopy
energy-dispersive spectrometry (EDS), matrix corrections, 1145
liquid surfaces and monomolecular layers, 1028
x-ray absorption fine-structure (XAFS) spectroscopy
basic principles, 875–877
errors, 879
x-ray magnetic circular dichroism (XMCD), 961–962
x-ray microfluorescence analysis, 942–945
automation, 949
background signals, 944
characteristic radiation, 942–943
data analysis, 950
detector criteria, 944
error detection, 950–951
fluorescence yields, 942
micro XAFS, 943
penetration depth, 943–944
photoabsorption cross-sections, 943
sample preparation, 949–950
x-ray microprobe
protocols and procedures, 941–942
research background, 939–941
Fluorescence yield, energy-dispersive spectrometry (EDS), standardless analysis, 1148
Flux density, permanent magnets, 497–499
Flux-gate magnetometer, magnetic field measurements, 506
Flux magnetometers
flux-integrating magnetometer, 533
properties of, 533–534
Flux sources, permanent magnets as, 499
Flying spot technique, carrier lifetime measurement, diffusion-length-based methods, 435
Focused ion beam (FIB) thinning, transmission electron microscopy (TEM), sample preparation, 1086–1087
Force field changes, surface phenomena, molecular dynamics (MD) simulation, 156–157
Force magnetometers, principles and applications, 534–535
Foreline traps, oil-sealed pumps, 4
Forward-recoil spectrometry
applications, 1268–1269
automation, 1269
backscattering data, 1269–1270
basic principles, 1261–1265
complementary and alternative techniques, 1261
data analysis and interpretation, 1269–1271
protocols for, 1270–1271
forward-recoil data, 1270–1271
instrumentation criteria, 1267–1268
limitations, 1271–1272
research background, 1259–1261
resolution, 1267
safety issues, 1269
sensitivity parameters, 1266–1267
spectrometer efficiency, 1265–1266
time-of-flight spectrometry (TOS), 1265
Foucault microscopy, magnetic domain structure measurements, Lorentz transmission electron microscopy, 551–552
FOURC software, resonant scattering, experimental design, 914
Fourier transform infrared (FTIR) spectrometry
gas analysis, 398
simultaneous techniques for gas analysis, research background, 393
Fourier transform magneto-spectroscopy (FTMS), cyclotron resonance (CR)
basic principles, 808–812
data analysis and interpretation, 813–814
error detection, 814–815
Fourier transform Raman spectroscopy (FTRS), principles and protocols, 708–709
Fourier transforms
chemical vapor deposition (CVD) model, radiation, 172–173
dynamical diffraction, 228
liquid surface x-ray diffraction, non-specular scattering, 1035–1036
low-energy electron diffraction (LEED), quantitative measurement, 1125
magnetic neutron scattering, 1329–1330
magnetic resonance imaging (MRI), 765–767
metal alloy magnetism, atomic short range order (ASRO), 191
Mössbauer effect, momentum space representation, 819–820
nuclear quadrupole resonance (NQR), data analysis and interpretation, 789
scanning transmission electron microscopy (STEM), probe configuration, 1098
single-crystal x-ray structure determination, direct method computation, 866–868
small-angle scattering (SAS), two-phase model, 220
x-ray absorption fine structure (XAFS) spectroscopy
data analysis and interpretation, 878–879
single scattering picture, 871–872
x-ray diffraction, crystal structure, 208–209
X-ray microdiffraction analysis, 944–945
Four-point bulk measurement
basic principles, 403
conductivity measurements, 402
protocols and procedures, 403–404
superconductors, electrical transport measurement, 473–474
Four-point probing, conductivity measurements, 402
Fracture toughness testing
basic principles, 302–307
crack driving force *(G),* 304–305
crack extension measurement, 308
crack tip opening displacement (CTOD) (δ), 307
data analysis and interpretation
sample preparation, 311–312
type A fracture, 308–309
type B fracture, 309–311
type C fracture, 311
errors and limitations, 312–313
J-integral, 306–307
load-displacement behaviors
measurement and recording apparatus, 307–308
notched specimens, 303–304
research background, 302
stress intensity factor *(K),* 305–306

Fracturing techniques, metallographic analysis, sample preparation, 65
Frames of reference and diffusion coefficients
  binary/multicomponent diffusion, 147–150
    Kirkendall effect and vacancy wind, 149
    lattice-fixed frame of reference, 148
    multicomponent alloys, 150–151
    number-fixed frame of reference, 147–148
    substitutional and interstitial metallic systems, 152
    tracer diffusion and self-diffusion, 149–150
    transformation between, 148–149
    volume-fixed frame of reference, 148
  multicomponent diffusion, 150–151
Frank-Kasper phase
  metal alloy bonding, size effects, 137
  transition metal crystals, 135
Free-air displacement, oil-sealed pumps, 3
Free carrier absorption (FCA), carrier lifetime measurement, 438–444
  automated methods, 441
  basic principles, 438–440
  carrier decay transient, 441
  computer interfacing, 441
  data analysis and interpretation, 441–442
  depth profiling, sample cross-sectioning, 443
  detection electronics, 440–441
  geometrical considerations, 441
  lifetime analysis, 441–442
  lifetime depth profiling, 441
  lifetime mapping, 441–442
  limitations, 443–444
  optical techniques, 434
  probe laser selection, 440
  processed wafers, metal and highly doped layer removal, 443
  pump laser selection, 440
  research background, 428
  sample preparation, 442–443
  virgin wafers, surface passivation, 442–443
Free-electron concentration, deep level transient spectroscopy (DLTS), semiconductor materials, 421
Free-electron laser, cyclotron resonance (CR), 808–809
Free electron model, magnetotransport in metal alloys, 562–563
Free energy calculations
  continuum field method (CFM)
    bulk chemical free energy, 119–120
    coarse-grained approximation, 118
    coarse-grained free energy formulation, 119
    diffuse-interface nature, 119
    elastic energy, 120–121
    external field energy, 121
    field kinetic equations, 121–122
    interfacial energy, 120
    slow variable selection, 119
  microstructural evolution, diffusion-controlled grain growth, 126–128
  phase diagram predictions, cluster variational method (CVM), 96–99, 99–100
Free excess carriers, carrier lifetime measurement, 438–439
Free induction decay (FID). *See also* Pseudo free induction decay
  magnetic resonance imaging (MRI), 765–767
  nuclear magnetic resonance, 765
  nuclear quadrupole resonance (NQR), data analysis and interpretation, 789
  nutation nuclear resonance spectroscopy, 784
Free molecular transport, chemical vapor deposition (CVD) model
  basic components, 169–170
  software tools, 174
Free-space radiation, microwave measurement techniques, 409
Free-to-bound transitions, photoluminescence (PL) spectroscopy, 683
Frequency-swept continuous wave detection, nuclear quadrupole resonance (NQR), 780–781
Frequency-swept Fourier-transform spectrometers, nuclear quadrupole resonance (NQR), 780–781
Fresnel microscope, magnetic domain structure measurements, Lorentz transmission electron microscopy, 551–552
Fresnel reflection coefficients
  ellipsometry
    reflecting surfaces, 736–737
    relative phase/amplitude calculations, 739–740
  grazing-incidence diffraction (GID), 241–242
    distorted-wave Born approximation (DWBA), 244–245
  liquid surface x-ray diffraction
    non-specular scattering, 1034–1036
    reflectivity measurements, 1029–1031
      multiple stepwise and continuous interfaces, 1031–1033
  surface magnetic scattering, 933–934
Friction, high-strain-rate testing, 299
Friction coefficient, tribological and wear testing, 325–326
  drive stiffness and, 328–329
  equipment and measurement techniques, 326–327
Friction tests, tribological and wear testing
  categories and classification, 328–329
  data analysis and interpretation, 334–335
Friedel's law
  d-band energetics, bonding-antibonding, 135
  single-crystal x-ray structure determination, 854
    crystal symmetry, 854–856
Fringe field sensors, magnetic field measurements, 506
Front-tracking simulation, microstructural evolution, 113–114
Frozen phonon calculations, phonon analysis, 1324–1325
Full potentials, metal alloy bonding, precision calculations, 143
FULLPROF software, neutron powder diffraction, Rietveld refinements, 1306
Full width at half-maximum (FWHM)
  diffuse scattering techniques, inelastic scattering background removal, 891–893
  energy-dispersive spectrometry (EDS), 1137–1140
  heavy-ion backscattering spectrometry (HIBS), 1277
  medium-energy backscattering, 1267
  neutron powder diffraction
    constant wavelength diffractometer, 1303–1304
    instrumentation criteria, 1292
    microstrain broadening, 1294–1295
    peak shapes, 1292
    Rietveld refinements, 1296
  particle-induced x-ray emission (PIXE) analysis, 1217–1218
  Raman spectroscopy of solids, 710–712
  ultraviolet photoelectron spectroscopy (UPS), 730
  x-ray microprobes, tapered capillary devices, 945
  x-ray photoelectron spectroscopy (XPS), 989
    background subtraction, 990–991
    peak position, 992, 1005
  x-ray powder diffraction, 837–838
Furnace windings, thermogravimetric (TG) analysis, 348–349

GaAs structures, capacitance-voltage (C-V) characterization, 462–463
Galvani potential, corrosion quantification, Tafel technique, 593–596
Gamma ray emission, Mössbauer effect, 818–820
Gärtner equation, semiconductor-liquid interfaces, diffusion length, 614–616
Gas analysis
  combustion calorimetry, 380
  simultaneous techniques for
    automated testing, 399
    benefits and limitations, 393–394
    chemical and physical methods, 398–399
    commercial TG-DTA equipment, 394–395
    evolved gas analysis and chromatography, 396–397
    gas and volatile product collection, 398
    infrared spectrometery, 397–398
    limitations, 399
    mass spectroscopy for thermal degradation and, 395–396
    research background, 392–393
    TG-DTA principles, 393
  thermal analysis with, 339
  time-resolved x-ray powder diffraction, 845–847
Gas chromatography (GC), simultaneous techniques for gas analysis, research background, 393
Gas density detector, evolved gas analysis (EGA) and chromatography, 397
Gas-filled thermometer, operating principles, 35–36
Gas flow, superconductors, electrical transport measurements, 478
Gas-liquid chromatography (GLC), evolved gas analysis (EGA) and, 397
Gas permeation, O-ring seals, 18
Gas-phase chemistry, chemical vapor deposition (CVD) model
  basic components, 168–169
  software tools, 173
Gas-sensing membrane electrodes, evolved gas detection (EGD), 399
Gas-solid reactions, thermogravimetric (TG) analysis, 355–356
Gaussian distribution
  liquid surface x-ray diffraction, simple liquids, 1040–1041
  x-ray absorption fine structure (XAFS) spectroscopy, 873
Gaussian peak shapes, x-ray photoelectron spectroscopy (XPS), 1005–1007
Gauss-Mehlaer quadrature, particle scattering, central-field theory, deflection function, 60
Gebhart absorption-factor method, chemical vapor deposition (CVD) model, radiation, 173
Gelfand-Levitan-Marchenko (GLM) method, liquid surface x-ray diffraction, scattering length density (SLD) analysis, 1039–1043
Generalized gradient approximations (GGAs)
  local density approximation (LDA), 78–79
    heats for formation and cohesive energies, 80–81
    structural properties, 78–79
  metal alloy bonding, accuracy calculations, 139–141
  metal alloy magnetism, 185–186
General-purpose interface bus (GPIB), thermomagnetic analysis, 541–544
Generation/collection (GC) mode, scanning electrochemical microscopy (SECM)

basic principles, 641–642
error detection, 647–648
properties and principles, 636–637
Generation lifetime, carrier lifetime measurement, 431
Geometrical issues, carrier lifetime measurement, free carrier absorption (FCA), 441
GEOPIXE software, particle-induced x-ray emission (PIXE) analysis, 1217–1218
Georgopoulos/Cohen (GC) technique, diffuse scattering techniques, 887–889
data interpretation, 896–897
Getter pumps
classification, 12
nonevaporable getter pumps (NEGs), 12
G-factor, resistive magnets, power-field relations, 503
Gibbs additive principle, continuum field method (CFM), bulk chemical free energy, 119–120
Gibbs energy activation barrier, substitutional and interstitial metallic systems, temperature and concentration dependence, 153
Gibbs free energy
combustion calorimetry, 371–372
phase diagram predictions, mean-field approximation, 92–96
thermal analysis and principles of, 342–343
Gibbs grand potential, metal alloy paramagnetism, finite temperatures, 186–187
Gibbs phase rule, phase diagram predictions, 91–92
Gibbs triangle, diffuse intensities, metal alloys
concentration waves, multicomponent alloys, 259–260
multicomponent alloys, Fermi-surface nesting and van Hove singularities, 269–270
Glass transition temperatures, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 368–369
Glide planes, symmetry operators, 40–42
Glow discharge mass spectrosscopy (GD0MS), heavy-ion backscattering spectrometry (HIBS) and, 1275
Gold impurities, deep level transient spectroscopy (DLTS), semiconductor materials, 419
Gold-rich AuFe alloys
atomic short range order (ASRO), 198–199
diffuse intensities, metal alloys, magnetic coupling and chemical order, 268
Gorsky-Bragg-Williams (GBW) free energies
diffuse intensities, metal alloys, mean-field results, 262
phase diagram predictions
cluster variational method (CVM), 99–100
mean-field approximation, 95–96
Gradient corrections, local density approximation (LDA), 78–79
Gradient energy term, interfacial energy, 120
Gradient-recalled echo sequence, magnetic resonance imaging (MRI), 768
Grain growth, microstructural evolution, diffusion-controlled grain growth, 126–128
Graphite furnace atomic absorption mass spectroscopy (GFAA-MS), heavy-ion backscattering spectrometry (HIBS) and, 1275
Grating criteria, spectrometers/monochromators, Raman spectroscopy of solids, 706–707
Grazing-incidence diffraction (GID), 241–246
applications, 226
distorted-wave Born approximation, 244–245
evanescent wave, 242
expanded distorted-wave approximation, 245–246
inclinded geometry, 243
liquid surface x-ray diffraction
alkane surface crystallization, 1043
components, 1038–1039
error sources, 1044–1045
Langmuir monolayers, 1042–1043
liquid metals, 1043
non-specular scattering, 1036
literature sources, 226
multlayers and superlattices, 242–243
specular reflectivity, 241–242
surface x-ray diffraction measurements
basic principles, 1011
measurement protocols, 1016–1017
Grease lubrication bearings, turbomolecular pumps, 7
Green's function
GW approximation and, 75
solid-solution alloy magnetism, 184
Green's function Monte Carlo (GFMC), electronic structure analysis, 88–89
Griffith criterion, fracture toughness testing, crack driving force *(G),* 305
Grounding, superconductors, electrical transport measurements, signal-to-noise ratio, 485
Ground loops, energy-dispersive spectrometry (EDS), 1153–1154
Ground-state analysis
atomic/ionic magnetism, ground-state multiplets, 514–515
phase diagram prediction, 102–104
transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Group theoretical analysis
Raman spectroscopy of solids, vibrational Raman spectroscopy, 702–704
character tables, 718
point groups and matrix representation, symmetry operations, 717–718
vibrational modes of solids, 720–722
vibrational selection rules, 716–720
resonant scattering, 909–910
Grüneisen theory, metal alloy magnetism, negative thermal expansion, 195–196
GSAS software, neutron powder diffraction
axial divergence peak asymmetry, 1293
Rietveld refinements, 1306
Guiner-Preston (GP) zones, transmission electron microscopy (TEM), 1080–1081
Guinier approximation, small-angle scattering (SAS), 221
GUPIX software, particle-induced x-ray emission (PIXE) analysis, 1217–1218
GW approximation, electronic structure analysis, 75
dielectric screening, 84–85
local density approximation (LDA) + U theory, 87
Gyromagnetic ratio
local moment origins, 514–515
nuclear magnetic resonance, 763–765

Half-mirror reflectors, reflected-light optical microscopy, 675–676
Half-rise times, thermal diffusivity, laser flash technique, 390
Half-width at half maximum (HWHM)
cyclotron resonance (CR), 814
x-ray photoelectron spectroscopy (XPS), peak position, 992
Hall effect
capacitance-voltage (C-V) characterization, 465
semiconductor materials
automated testing, 414
basic principles, 411–412
data analysis and interpretation, 414–416
equations, 412–413
limitations, 417
protocols and procedures, 412–414
research background, 411
sample preparation, 416–417
sensitivity, 414
sensors, magnetic field measurements, 506–508
surface magnetometry, 536
Hall resistance, magnetotransport in metal alloys
basic principles, 560
magnetic field behavior, 563–565
Hardenability properties, hardness testing, 320
Hardness testing
automated methods, 319
basic principles, 317–318
data analysis and interpretation, 319–320
limitations and errors, 322
procedures and protocols, 318–319
research background, 316–317
sample preparation, 320
specimen modification, 320–321
Hardness values
calculation of, 323
static indentation hardness testing, 317–318
Hard spheres, particle scattering, deflection functions, 58–59
Hardware components. *See also* Software tools
vacuum systems, 17–19
all-metal flange seals, 18
electrical feedthroughs, 18
O-ring flange seals, 17–18
rotational/translational feedthroughs, 18–19
valves, 18
Harmonic contamination
liquid surface x-ray diffraction, 1044–1045
magnetic x-ray scattering, 935
x-ray magnetic circular dichroism (XMCD), 965–966
Harmonic content, x-ray absorption fine structure (XAFS) spectroscopy, 877
Hartree-Fock (HF) theory
electronic structure analysis
basic components, 77
GW approximation, 84–85
summary, 75
metal alloy bonding, accuracy calculations, 140–141
Heat balance equation, thermal diffusivity, laser flash technique, 391–392
Heat-flux differential scanning calorimetry (DSC)
basic principles, 364–365
thermal analysis, 339
Heating curve determination, defined, 338
Heating rate
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 394
thermogravimetric (TG) analysis, 351–352
Heating-rate curves, defined, 338
Heat loss, thermal diffusivity, laser flash technique, 388–389
Heats of formation, local density approximation (LDA), 79–81
Heat transfer relations
chemical vapor deposition (CVD) model, radiation, 172–173
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 368–369
thermal diffusivity, laser flash technique, 385–386
thermogravimetric (TG) analysis, 350–352
Heavy-atom model, single-crystal x-ray structure determination, 860–861
computational techniques, 868–869

Heavy-ion backscattering spectrometry (HIBS)
applications, 1279–1280
automation, 1280
basic principles, 1275–1277
competitive, complementary, and alternative strategies, 1275
data analysis and interpretation, 1280–1281
areal density, 1281
background and sensitivity, 1281
mass measurements, 1281
elastic scattering, 1178
instrumentation criteria, 1277–1280
sensitivity and mass resolution, 1278–1279
limitations, 1281–1282
research background, 1273–1275
safety protocols, 1280
sample preparation, 1281
Heavy product, particle scattering, nuclear reactions, 57
Heisenberg-Dirac Hamiltonian, metal alloy magnetism, 180
Heisenberg exchange theory
collective magnetism, 515
ferromagnetism, 525–526
surface magneto-optic Kerr effect (SMOKE), 571
Heitler-London series, metal alloy bonding, accuracy calculations, 139–141
Helicopter effect, turbomolecular pumps, venting protocols, 8
Helimagnetism, principles and equations, 526–527
Helium gases, cryopump operation, 9–10
Helium mass spectrometer leak detector (HMSLD), leak detection, vacuum systems, 21–22
Helium vapor pressure, ITS-90 standard, 33
Helmholtz free energy
Landau magnetic phase transition theory, 529–530
semiconductor-liquid interfaces, differential capacitance measurements, 617–619
Hemi-spherical analysis, ultraviolet photoelectron spectroscopy (UPS), 729
Hemispherical sector analyzer (HSA), x-ray photoelectron spectroscopy (XPS), 980–982
Hermann-Mauguin symmetry operator
improper rotation axis, 39–40
single-crystal x-ray structure determination, crystal symmetry, 855–856
Heterodyne detection, impulsive stimulated thermal scattering (ISTS), 749
Heterogeneous broadening, nuclear quadrupole resonance (NQR), 790
Heterojunction bipolar transistors (HBTs), characterization
basic principles, 467–469
competitive and complementary techniques, 466–467
limitations of, 471
measurement equipment sources and selection criteria, 471
protocols and procedures, 469–470
research background, 466–467
sample preparation, 470–471
High-alumina ceramics, vacuum system construction, 17
High-energy beam transport, analysis, and detection, trace element accelerator mass spectrometry (TEAMS), 1239
High energy electron diffraction, transmission electron microscopy (TEM), deviation vector and parameter, 1068
High-energy ion beam analysis (IBA)
elastic recoil detection analysis (ERDA), 1178
heavy-ion backscattering, 1178
nuclear reaction analysis (NRA), 1178–1179
particle-induced x-ray emission, 1179
periodic table, 1177–1178
research background, 1176–1177
Rutherford backscattering spectrometry, 1178
Higher-order contamination, phonon analysis, 1326–1327
High-field sensors, magnetic field measurements, 506–510
Hall effect sensors, 506–508
magnetoresistive sensors, 508–509
nuclear magnetic resonance (NMR), 509–510
search coils, 509
High-frequency ranges
carrier lifetime measurement, automated photoconductivity (PC), 445–446
electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, 795–796
High level injection, carrier lifetime measurement, 428–429
Highly doped layers, carrier lifetime measurement, metal/highly doped layer removal, free carrier absorption (FCA), 443
High-pressure magnetometry, principles and applications, 535
High-resolution spectrum, x-ray photoelectron spectroscopy (XPS), 974–978
appearance criteria, 993–994
final-state effects, 974–976
initial-state effects, 976–978
silicon example, 997–998
High-resolution transmission electron microscopy (HRTEM)
astigmatism, 1079
basic principles, 1064
bright-field/dark-field imaging, 1071
data analysis, 1081
sample preparation, 1108
scanning transmission electron microscopy (STEM) *vs.,* 1090–1092
phase-contrast imaging, 1096–1097
High-strain-rate testing
basic principles, 290–292
data analysis and interpretation, 296–298
limitations, 299–300
method automation, 296
protocols for, 292–296
stress-state equilibrium, 294–295
temperature effects, 295–296
sample preparation, 298
specimen modification, 298–299
theoretical background, 288–290
High-temperature electron emitter, hot cathode ionization gauges, 15
High-temperature superconductors (HTSs)
electrical transport measurement
applications, 472
contact materials, 474
current-carrying area, 476
current sharing with other materials, 474
current supply, 476–477
powder-in-tube HTSs, sample quality issues, 486
magnetic measurements *vs.,* 473
superconducting magnets, 500–502
High-temperature testing, stress-strain analysis, 286
High-vacuum pumps, classification, 6–12
cryopumps, 9–10
diffusion pumps, 6–7
getter pumps, 12
nonevaporable getter pumps (NEGs), 12
sputter-ion pumps, 10–12
sublimation pumps, 12
turbomolecular pumps, 7–9
Holographic imaging, magnetic domain structure measurements, 554–555
Holographic notch filter (HNF), Raman spectroscopy of solids, optics properties, 706
Homogeneous materials
carrier lifetime measurement, 428–429
thermal diffusivity, laser flash technique, 391–392
Honl corrections, x-ray diffraction, 210
Hooke's law, elastic deformation, 281
Hopkinson bar technique
design specifications for, 292–294
high-strain-rate testing
historical background, 289
protocols for, 292–296
stress-state equilibrium, 294–295
temperature effects, 295–296
tensile/torsional variants, 289–290
Hot cathode ionization gauges
calibration stability, 16
electron stimulated desorption (ESD), 15–16
operating principles, 14–15
Hot corrosion, electrochemical impedance spectroscopy (EIS), 601–603
Hot zone schematics, thermogravimetric (TG) analysis, 348–349
Hund's rules
magnetic moments, 512
atomic and ionic magnetism, local moment origins, 513–515
metal alloy magnetism, 180–181
electronic structure, 184–185
Hybridization, diffuse intensities, metal alloys, charge correlation effects, NiPt alloys, 266–268
Hybrid magnets, structure and properties, 503–504
Hydraulic-driven mechanical testing system, fracture toughness testing, load-displacement curve measurement, 307–308
Hydrocarbon layer deposition, nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, error detection, 1207–1208
Hydrodynamics, chemical vapor deposition (CVD) model
basic components, 170–171
software tools, 174
Hyperboloid sheets, two-beam diffraction, dispersion surface, 230
Hyperfine interactions, Mössbauer spectroscopy
body-centered cubic (bcc) iron solutes, 828–830
data analysis and interpretation, 831–832
magnetic field splitting, 823–824
overview, 820–821
Hysteresis loop
magnetism principles, 494–495
permanent magnets, 497–499
surface magneto-optic Kerr effect (SMOKE), 569–570
experimental protocols, 571–572

Ice accumulation, energy-dispersive spectrometry (EDS), 1155
Ideal diode equation, *pn* junction characterization, 467–468
Ideality factor, *pn* junction characterization, 470
Ignition systems, combustion calorimetry, 376–377
Illuminated semiconductor-liquid interface
J-E equations, 608–609
monochromatic illumination, 610
polychromatic illumination, 610–611
Illumination modes, reflected-light optical microscopy, 676–680

Image-analysis measurement, hardness test equipment, automated reading, 319
Image-enhancement techniques, reflected-light optical microscopy, 676–680
Image formation protocols
  scanning electron microscopy (SEM), 1052–1053
    quality control, 1059–1060
    selection criteria, 1061, 1063
  scanning tunneling microscopy (STM), 1114
Imaging plate systems, single-crystal x-ray structure determination, 860
Impact parameters
  particle scattering, central-field theory, 57–58
  wear testing protocols, 331
Impedance analysis, electrochemical quartz crystal microbalance (EQCM), 655–657
Impingement of projectiles, materials characterization, 1
Impulse response, scanning electrochemical microscopy (SECM), feedback mode, 639–640
Impulsive stimulated thermal scattering (ISTS)
  applications, 749–752
  automation, 753
  competitive and related techniques, 744–746
  data analysis and interpretation, 753–757
  limitations, 757–758
  procedures and protocols, 746–749
  research background, 744–746
  sample preparation and specimen modification, 757
  vendor information, 759
Impurity measurements, trace element accelerator mass spectrometry (TEAMS)
  bulk analysis, 1247–1249
  depth-profiling techniques, 1250–1251
Inclined geometry, grazing-incidence diffraction (GID), 243
Incoherent imaging, scanning transmission electron microscopy (STEM)
  phase-contrast illumination *vs.*, 1094–1097
  probe configuration, 1098
  research background, 1091–1092
  scattering devices, 1098–1101
  weakly scattering objects, 1111
Indexing procedures
  low-energy electron diffraction (LEED), qualitative analysis, 1122–1124
  neutron powder diffraction, 1297–1298
  transmission electron microscopy (TEM)
    diffraction pattern indexing, 1073–1074
    Kikuchi line indexing, 1076–1077
Indirect detection techniques, nuclear quadrupole resonance (NQR), field cycling methods, 781
Indirect mass measurement techniques, basic principles, 24
Indirect structural technique, x-ray absorption spectroscopy, 870
Induced currents, magnetic field effects and applications, 496
Inductively coupled plasma (ICP) mass spectrometry
  heavy-ion backscattering spectrometry (HIBS) and, 1275
  particle-induced x-ray emission (PIXE), 1211
  trace element accelerator mass spectrometry (TEAMS) and, 1237
Inelastic scattering
  diffuse scattering subtraction, 890–893
  magnetic neutron scattering
    basic principles, 1331
    data analysis, 1335–1336
  particle scattering, 52
    diagrams, 52–54
  x-ray photoelectron spectroscopy (XPS), survey spectrum, 974
Inertia, high-strain-rate testing, 299
Inert/reactive atmosphere, thermogravimetric (TG) analysis, 354–355
Infrared absorption spectroscopy (IRAS), solids analysis, *vs.* Raman spectroscopy, 699
Infrared (IR) spectrometry
  gas analysis and, 397–398
  magnetic neutron scattering and, 1321
Inhomogeneity
  Hall effect, semiconductor materials, 417
  x-ray photoelectron spectroscopy (XPS), initial-state effects, 978
Inhomogeneous path probability method (PPM), microstructural evolution, 116–117
Initial-state effects, x-ray photoelectron spectroscopy (XPS), 976–978
In situ analysis
  time-dependent neutron powder diffraction, 1300
  transmission electron microscopy (TEM), specimen modification, 1088
Instrumentation criteria
  capacitance-voltage (C-V) characterization, 463–464
  combustion calorimetry, 375–377
  differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 367
    selection criteria, 373
  electrochemical quartz crystal microbalance (EQCM), 658–659
    automated procedures, 659–660
  gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), sample holder criteria, 394
  heavy-ion backscattering spectrometry (HIBS), 1279
  impulsive stimulated thermal scattering (ISTS), 749–752
  ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1228–1230
  liquid surface x-ray diffraction, 1036–1038
  low-energy electron diffraction (LEED), 1125–1126
  magnetic x-ray scattering, 934–935
  medium-energy backscattering, 1267–1268
  neutron powder diffraction, peak shape, 1292
  nuclear magnetic resonance, 770–771
  nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1203–1204
  optical microscopy, 668–671
  phonon analysis, triple-axis spectrometry, 1320–1323
  resonant scattering, 912–913
  scanning electrochemical microscopy (SECM), 642–643
  scanning electron microscopy (SEM), basic components, 1053–1057
  scanning transmission electron microscopy (STEM), sources, 1111
  single-crystal neutron diffraction, 1312–1313
  single-crystal x-ray structure determination, 859–860
  single event upset (SEU) microscopy, 1227–1228
  superconductors, electrical transport measurements, 476–477, 479–480
  surface x-ray diffraction, 1011–1015
    crystallographic alignment, 1014–1015
    five-circle diffractometer, 1013–1014
    laser alignment, 1014
    sample manipulator, 1012–1013
    vacuum system, 1011–1012
  thermal diffusivity, laser flash technique, 385–386
  trace element accelerator mass spectrometry (TEAMS), 1258
  ultraviolet photoelectron spectroscopy (UPS), 733
  x-ray magnetic circular dichroism (XMCD), 957–962
    circularly polarized sources, 957–959
    detector devicees, 959
    measurement optics, 959–962
  x-ray photoelectron spectroscopy (XPS), 978–983
    analyzers, 980–982
    detectors, 982
    electron detection, 982–983
    maintenance, 983
    sources, 978–980
Instrumentation errors, hardness testing, 322
Instrumented indentation testing, basic principles, 317
Integrated intensity
  magnetic x-ray scattering, nonresonant antiferromagnetic scattering, 929–930
  small-angle scattering (SAS), 222
  two-beam diffraction, 235
  x-ray powder diffraction, Bragg peaks, 840
Integrated vertical Hall sensor, magnetic field measurements, 507–508
Intensity computations
  Auger electron spectroscopy (AES), 1167–1168
  diffuse scattering techniques, 886–889
    absolute calibration, measured intensities, 894
    derivation protocols, 901–904
  low-energy electron diffraction (LEED), quantitative measurement, 1124–1125
  multiple-beam diffraction, NBEAM theory, 238
  particle-induced x-ray emission (PIXE), 1216–1217
    x-ray intensities and concentrations, 1221–1222
  surface x-ray diffraction, crystal truncation rods (CTR), 1009
Intensity-measuring ellipsometer, automation, 739
Interaction potentials
  molecular dynamics (MD) simulation, surface phenomena, 159
  particle scattering, central-field theory, 57
Interdiffusion, substitutional and interstitial metallic systems, 155
Interfacial energy, microstructural evolution, 120
Interference effects
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1205
  surface/interface x-ray diffraction, crystal truncation rods (CTR), 219
  x-ray absorption fine structure (XAFS) spectroscopy, single scattering picture, 871–872
  x-ray diffraction
    crystal structure, 208–209
    single-crystal x-ray structure determination, 851–858
Interlayer surface relaxation, molecular dynamics (MD) simulation, room temperature structure and dynamics, 161
Internal fields, dynamical diffraction, basic principles, 229
Internal magnetic field, ferromagnetism, 522–524
International Organization for Legal Metrology (OILM), weight standards, 26–27
International symmetry operators, improper rotation axis, 39–40
International temperature scales (ITS)
  IPTS-27 scale, 32–33
  IPTS-68 scale, 33
  ITS-90, 33

Interparticle interference, small-angle scattering (SAS), 222
Interplanar spacing, transmission electron microscopy (TEM), diffraction pattern indexing, 1073–1074
Interpolating gas thermometer, ITS-90 standard, 33
Inverse heating-rate curves, defined, 338
Inverse modeling, impulsive stimulated thermal scattering (ISTS), 753–754
Inverse photoemission, ultraviolet photoelectron spectroscopy (UPS)
  vs. photoemission, 722–723
  valence electron characterization, 723–724
Inverse photoemission (IPES), metal alloy magnetism, local exchange splitting, 189–190
Ion beam analysis (IBA)
  accelerator mass spectrometry, 1235
  elastic ion scattering for composition analysis
    applications, 1191–1197
    basic concepts, 1181–1184
    detector criteria and detection geometries, 1185–1186
    equations, 1186–1189
    experimental protocols, 1184–1186
    limitations, 1189–1191
    research background, 1179–1181
  forward-recoil spectrometry
    applications, 1268–1269
    automation, 1269
    backscattering data, 1269–1270
    basic principles, 1261–1265
    complementary and alternative techniques, 1261
    data analysis and interpretation, 1269–1271
    forward-recoil data, 1270–1271
    instrumentation criteria, 1267–1268
    limitations, 1271–1272
    research background, 1259–1261
    resolution, 1267
    safety issues, 1269
    sensitivity parameters, 1266–1267
    spectrometer efficiency, 1265–1266
    time-of-flight spectrometry (TOS), 1265
  heavy-ion backscattering spectrometry (HIBS)
    applications, 1279–1280
    automation, 1280
    basic principles, 1275–1277
    competitive, complementary, and alternative strategies, 1275
    data analysis and interpretation, 1280–1281
      areal density, 1281
      background and sensitivity, 1281
      mass measurements, 1281
    instrumentation criteria, 1277–1280
      sensitivity and mass resolution, 1278–1279
    limitations, 1281–1282
    research background, 1273–1275
    safety protocols, 1280
    sample preparation, 1281
  high-energy ion beam analysis (IBA)
    elastic recoil detection analysis (ERDA), 1178
    heavy-ion backscattering, 1178
    nuclear reaction analysis (NRA), 1178–1179
    particle-induced x-ray emission, 1179
    periodic table, 1177–1178
    research background, 1176–1177
    Rutherford backscattering spectrometry, 1178
  medium-energy backscattering
    applications, 1268–1269
    automation, 1269
    backscattering data, 1269–1270
    basic principles, 1261–1265
    complementary and alternative techniques, 1261
    data analysis and interpretation, 1269–1271
    forward-recoil data, 1270–1271
    instrumentation criteria, 1267–1268
    limitations, 1271–1272
    research background, 1258–1259, 1259–1261
    resolution, 1267
    safety issues, 1269
    sensitivity parameters, 1266–1267
    spectrometer efficiency, 1265–1266
    time-of-flight spectrometry (TOS), 1265
  nuclear reaction analysis (NRA)
    automation, 1205–1206
    background, interference, and sensitivity, 1205
    cross-sections and Q values, 1205
    data analysis, 1206–1207
    energy relations, 1209
    energy scanning, resonance depth profiling, 1204–1205
    energy spread, 1209–1210
    instrumentation criteria, 1203–1204
    limitations, 1207–1208
    nonresonant methods, 1202–1203
    research background, 1200–1202
    resonant depth profiling, 1203
    specimen modification, 1207
    standards, 1205
    unwanted particle filtering, 1204
  particle-induced x-ray emission (PIXE)
    automation, 1216
    basic principles, 1211–1212
    competing methods, 1210–1211
    data analysis, 1216–1218
    limitations, 1220
    protocols and techniques, 1213–1216
    research background, 1210–1211
    sample preparation, 1218–1219
    specimen modification, 1219–1220
  proton-induced gamma ray emission (PIGE)
    automation, 1205–1206
    background, interference, and sensitivity, 1205
    cross-sections and Q values, 1205
    data analysis, 1206–1207
    energy relations, 1209
    energy scanning, resonance depth profiling, 1204–1205
    energy spread, 1209–1210
    instrumentation criteria, 1203–1204
    limitations, 1207–1208
    nonresonant methods, 1202–1203
    research background, 1200–1202
    resonant depth profiling, 1203
    specimen modification, 1207
    standards, 1205
    unwanted particle filtering, 1204
  radiation effects microscopy
    basic principles, 1224–1228
    instrumentation criteria, 1128–1230
    ion-induced damage, 1232–1233
    limitations, 1233
    quantitative analysis, pulse height interpretations, 1225–1226
    research background, 1223–1224
    semiconductor materials, 1225
    SEU microscopy, 1227–1228
      static random-access memory (SRAM), 1230–1231
    specimen modification, 1231–1232
    topographical contrast, 1226–1227
  research background, 1175–1176
  trace element accelerator mass spectrometry (TEAMS)
    automation, 1247
    bulk analysis
      impurity measurements, 1247–1249
      measurement data, 1249–1250
    complementary, competitive and alternative methods, 1236–1238
      inductively coupled plasma mass spectrometry, 1237
      neutron activation-accelerator mass spectrometry (NAAMS), 1237
      neutron activation analysis (NAA), 1237
      secondary-ion mass spectrometry (SIMS), 1236–1237
      selection criteria, 1237–1238
      sputter-initiated resonance ionization spectrometry (SIRIS), 1237
    data analysis and interpretation, 1247–1252
      calibration of data, 1252
    depth-profiling
      data analysis, 1251–1252
      impurity measurements, 1250–1251
    facilities profiles, 1242–1246
      CSIRO Heavy Ion Analytical Facility (HIAF), 1245
      Naval Research Laboratory, 1245–1246
      Paul Scherrer Institute (PSI)/ETH Zurich Accelerator SIMS Laboratory, 1242–1244
      Technical University Munich Secondary Ion AMS Facility, 1245
      University of North Texas Ion Beam Modification and Analysis Laboratory, 1246
      University of Toronto IsoTrace Laboratory, 1244–1245
    facility requirements, 1238
    future applications, 1239
    high-energy beam transport, analysis, and detection, 1239
    historical evolution, 1246–1247
    impurity measurements
      bulk analysis, 1247–1249
      depth-profiling, 1250–1251
    instrumentation criteria, 1239–1247
      magnetic and electrostatic analyzer calibration, 1241–1242
      ultraclean ion source design, 1240–1241
    instrumentation specifications and suppliers, 1258
    limitations, 1253
    research background, 1235–1238
    sample preparation, 1252–1253
    secondary-ion acceleration and electron-stripping system, 1238–1239
    specimen modification, 1253
    ultraclean ion sources, negatively charged secondary-ion generation, 1238
Ion-beam-induced charge (IBIC) microscopy
  basic principles, 1224–1228
  instrumentation criteria, 1128–1230
  ion-induced damage, 1232–1233
  limitations, 1233
  quantitative analysis, pulse height interpretations, 1225–1226
  research background, 1223–1224
  semiconductor materials, 1225
  SEU microscopy, 1227–1228
    static random-access memory (SRAM), 1230–1231
  specimen modification, 1231–1232
  topographical contrast, 1226–1227
Ion burial, sputter-ion pump, 11
Ion excitation
  Auger electron spectroscopy (AES), peak error detection, 1171–1173
  energy-dispersive spectrometry (EDS), 1136
Ionic magnetism, local moment origins, 513–515
Ionic migration, energy-dispersive spectrometry (EDS), specimen modification, 1152–1153

Ion-induced damage, ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1232–1233
Ionization cross-section, energy-dispersive spectrometry (EDS), standardless analysis, 1148
Ionization gauges
  cold cathode type, 16
  hot cathode type
    calibration stability, 16
    electron stimulated desorption (ESD), 15–16
    operating principles, 14–15
Ionization loss peaks, Auger electron spectroscopy (AES), error detection, 1171
Ion milling, transmission electron microscopy (TEM), sample preparation, 1086–1087
Ion pumps, surface x-ray diffraction, 1011–1012
Ion scattering spectroscopy (ISS), Auger electron spectroscopy (AES) *vs.,* 1158–1159
Iron (Fe) alloys. *See also* Nickel-iron alloys
  magnetism
    atomic short range order (ASRO), 191
      gold-rich AuFe alloys, 198–199
    magnetocrystalline anisotropy energy (MAE), 193
    moment formation and bonding, FeV alloys, 196
  Mössbauer spectroscopy, body-centered cubic (bcc) solutes, 828–830
  paramagnetism, first-principles calculations, 189
Irreversibility field (H_{irr}), superconductors, electrical transport measurement
  applications, 472
  extrapolation, 475
  magnetic measurement *vs.,* 473
Ising energy, phase diagram prediction, static displacive interactions, 104–106
Isobaric mass change determination, defined, 338
Isolation valves, sputter-ion pump, 12
Isomers, combustion calorimetry, 377–378
Isomer shift, Mössbauer spectroscopy
  basic principles, 821–822
  hyperfine interactions, 820–821
Isothermal method, thermogravimetric (TG) analysis, kinetic theory, 352–354
Isotropic materials, reflected-light optical microscopy, 678–680
Itinerant magnetism, transition metal magnetic ground state, zero temperature, 181–183

J-E equations, semiconductor-liquid interface
  concentration overpotentials, 611–612
  illuminated cases, 608–609
  kinetic properties, 631–633
  series resistance overpotentials, 612
J-integral approach, fracture toughness testing
  basic principles, 306–307
  crack extension measurement, 311
  research background, 302
  sample preparation, 312
  SENB and CT speciments, 314–315
  unstable fractures, 309
Johnson noise, scanning tunneling microscopy (STM), 1113–1114
Joint leaks, vacuum systems, 20
Junction capacitance transient, deep level transient spectroscopy (DLTS), semiconductor materials, 421–423
Junction diodes, deep level transient spectroscopy (DLTS), semiconductor materials, 420
Junction field-effect transistors (JFETs), characterization
  basic principles, 467–469
  competitive and complementary techniques, 466–467
  limitations of, 471
  measurement equipment sources and selection criteria, 471
  protocols and procedures, 469–470
  research background, 466–467
  sample preparation, 470–471

Katharometer, evolved gas analysis (EGA) and chromatography, 396–397
Keating model, surface x-ray diffraction measurements, crystallographic refinement, 1018
Keithley-590 capacitance meter, capacitance-voltage (C-V) characterization, 460–461
Kelvin-Onsager relations, magnetotransport in metal alloys, 560–561
Kelvin relations, magnetotransport in metal alloys, basic principles, 560
Kerr effect magnetometry
  magnetic domain structure measurements, magneto-optic imaging, 547–549
  principles and applications, 535–536
  surface magneto-optic Kerr effect (SMOKE)
    automation, 573
    classical theory, 570–571
    data analysis and interpretation, 573–574
    limitations, 575
    multilayer formalism, 571
    phenomenologica origins, 570
    protocols and procedures, 571–573
    quantum theory, ferromagnetism, 571
    research background, 569–570
    sample preparation, 575
Kikuchi-Barker coefficients, phase diagram predictions, cluster variational method (CVM), 99–100
Kikuchi lines
  scanning transmission electron microscopy (STEM), instrumentation criteria, 1104–1105
  transmission electron microscopy (TEM)
    deviation vector and parameter, 1067–1068
      defect contrast, 1085
      specimen orientation, 1075–1078
        deviation parameters, 1077–1078
        electron diffuse scattering, 1075
        indexing protocols, 1076–1077
        line origins, 1075–1076
Kinematic theory
  diffuse scattering techniques, crystal structure, 885–889
  dynamical diffraction, theoretical background, 224–225
  ion beam analysis (IBA), 1176
  ion-beam analysis (IBA), elastic two-body collision, 1181–1184
  liquid surface x-ray diffraction, reflectivity measurements, 1032–1033
  particle scattering, 51–57
    binary collisions, 51
    center-of-mass and relative coordinates, 54–55
    elastic scattering and recoiling, 51–52
    inelastic scattering and recoiling, 52
    nuclear reactions, 56–57
    relativistic collisions, 55–56
    scattering/recoiling diagrams, 52–54
  single-crystal neutron diffraction and, 1310–1311
  transmission electron microscopy (TEM)
    deviation vector and parameter, 1068
    structure and shape factor analysis, 1065–1066
  x-ray diffraction
    crystalline material, 208–209
    lattice defects, 210
    local atomic arrangement - short-range ordering, 214–217
    research background, 206
    scattering principles, 206–208
    small-angle scattering (SAS), 219–222
      cylinders, 220–221
      ellipsoids, 220
      Guinier approximation, 221
      integrated intensity, 222
      interparticle interference, 222
      K=0 extrapolation, 221
      porod approximation, 221–222
      size distribution, 222
      spheres, 220
      two-phase model, 220
    structure factor, 209–210
    surface/interface diffraction, 217–219
      crystal truncation rods, 219
      two-dimensional diffraction rods, 218–219
    thermal diffuse scattering (TDS), 210–214
Kinetic energy
  Auger electron spectroscopy (AES), 1159–1160
  chemical vapor deposition (CVD) model
    basic components, 171
    software tools, 175
  electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 625
  low-energy electron diffraction (LEED), quantitative analysis, 1129–1131
  semiconductor-liquid interface
    steady-state J-E data, 631–633
    time-resolved photoluminescence spectroscopy (TRPL), 630–631
  thermal analysis and, 343
  thermogravimetric (TG) analysis, applications, 352–354
  ultraviolet photoelectron spectroscopy (UPS), photoemission process, 726–727
  x-ray photoelectron spectroscopy (XPS), 971–972
    analyzer criteria, 981–982
    background subtraction, 990–991
Kirchoff's law of optics, radiation thermometer, 36–37
Kirkendall effect, binary/multicomponent diffusion, vacancy wind and, 149
Kirkendall shift velocity, binary/multicomponent diffusion, frames of reference, transition between, 148–149
Kirkpatrick-Baez (KB) mirrors, x-ray microprobes, 946
KLM line markers, energy-dispersive spectrometry (EDS), qualitative analysis, 1142–1143
Knock-on damage, transmission electron microscopy (TEM), metal specimen modification, 1088
Knoop hardness testing
  automated methods, 319
  basic principles, 317–318
  data analysis and interpretation, 319–320
  hardness values, 317–318, 323
  limitations and errors, 322
  procedures and protocols, 318–319
  research background, 316–317
  sample preparation, 320
  specimen modification, 320–321
Knudsen diffusion coefficient, chemical vapor deposition (CVD) model, free molecular transport, 169–170
Knudsen effect
  combustion calorimetry, 371–372
  thermogravimetric (TG) analysis, instrumentation and apparatus, 349–350

Kohn-Sham equation
diffuse intensities, metal alloys, concentration waves, density-functional theory (DFT), 260–262
metal alloy bonding
accuracy calculations, 139–141
precision measurements, self-consistency, 141
metal alloy magnetism
competitive and related techniques, 185–186
magnetocrystalline anisotropy energy (MAE), 192–193
solid-solution alloy magnetism, 183–184
transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
Koopman's theorem, electronic structure analysis, Hartree-Fock (HF) theory, 77
Korringa-Kohn-Rostocker (KKR) method
band theory of magnetism, layer Korringa Kohn Rostoker (LKKR) technique, 516
diffuse intensities, metal alloys
concentration waves, first-principles calcuations, electronic structure, 264–266
hybridization in NiPt alloys, charge correlation effects, 267–268
metal alloy magnetism, magnetocrystalline anisotropy energy (MAE), 192–193
phase diagram prediction, electronic structure, 101–102
solid-solution alloy magnetism, 183–184
Kossel cones, transmission electron microscopy (TEM), Kikuchi line indexing, 1076–1077
Kovar, vacuum system construction, 17
k-points, metal alloy bonding, precision calculations, Brillouin zone sampling, 143
Kramers-Kronig relation, resonant scattering analysis, 907
Krivoglaz-Clapp-Moss formula, diffuse intensities, metal alloys
concentration waves, density-functional theory (DFT), 260–262
mean-field results, 262–263
Kronecker delta function, 209
Kruschov equation, wear testing, 334

Laboratory balances, applications, 27
Lambert cosine law, radiation thermometer, 37
Lamb modes
impulsive stimulated thermal scattering (ISTS), 754–755
Mössbauer effect, 818–820
Landau levels, cyclotron resonance (CR), 805–806
quantum mechanics, 808
Landau-Lifshitz equation, resonant scattering analysis, 909–910
Landau theory, magnetic phase transitions, 529–530
thermomagnetic analysis, 544
Lande *g* factor, local moment origins, 514–515
Langevin function
diagmagnetism, dipole moments, atomic origin, 512
ferromagnetism, 524
paramagnetism, classical and quantum theories, 520–522
superparamagnetism, 522
Langmuir monolayers, liquid surface x-ray diffraction
data analysis, 1041–1043
grazing incidence and rod scans, 1036
Lanthanum hexaboride electron guns
Auger electron spectroscopy (AES), 1160–1161
scanning electron microscopy (SEM)
instrumentation criteria, 1054
selection criteria, 1061
Laplace equation
magnetotransport in metal alloys, 560–561
thermal diffusivity, laser flash technique, 388–389
Larmor frequency
closed/core shell diamagnetism, dipole moments, atomic origin, 512
nuclear magnetic resonance, 764–765
nuclear quadrupole resonance (NQR), Zeeman-perturbed NRS (ZNRS), 782–783
Laser components
impulsive stimulated thermal scattering (ISTS), 750–752
photoluminescence (PL) spectroscopy, 684
Raman spectroscopy of solids, radiation sources, 705
surface x-ray diffraction, alignment protocols, 1014
ultraviolet/visible absorption (UV-VIS) spectroscopy, 690
Laser flash technique, thermal diffusivity
automation of, 387
basic principles, 384–385
data analysis and interpretation, 387–389
limitations, 390
protocols and procedures, 385–387
research background, 383–384
sample preparation, 389
specimen modification, 389–390
Laser magneto-spectroscopy (LMS), cyclotron resonance (CR)
basic principles, 808–812
data analysis and interpretation, 813–814
error detection, 814–815
far-infrared (FIR) sources, 810–812
Laser spot scanning (LSS), semiconductor-liquid interface, photoelectrochemistry, 626–628
Lattice-fixed frame of reference
binary/multicomponent diffusion, 148
substitutional and interstitial metallic systems, 152
Lattice systems. *See also* Reciprocal lattices
conductivity measurements, research background, 401–403
crystallography, 44–46
Miller indices, 45–46
deep level transient spectroscopy (DLTS), 419
defects, x-ray diffraction, 210
density definition and, 26
local atomic arrangment - short-range ordering, 214–217
metal alloy magnetism, atomic short range order (ASRO), 191
molecular dynamics (MD) simulation, surface phenomena, 158–159
phonon analysis, basic principles, 1317–1318
Raman active vibrational modes, 709–710, 720–721
single-crystal x-ray structure determination, crystal symmetry, 854–856
thermal diffuse scattering (TDS), 212–214
transition metal magnetic ground state, itinerant magnetism at zero temperature, 183
transmission electron microscopy (TEM), defect diffraction contrast, 1068–1069
ultraviolet photoelectron spectroscopy (UPS), 730
x-ray powder diffraction, crystal lattice determination, 840
Laue conditions
diffuse scattering techniques, 887–889
low-energy electron diffraction (LEED), 1122
single-crystal neutron diffraction, instrumentation criteria, 1312–1313
single-crystal x-ray structure determination, 853
protocols and procedures, 858–860
two-beam diffraction
anomalous transmission, 231–232
diffracted intensities, 233–234
hyperboloid sheets, 230
Pendellösung, 231
x-ray standing wave (XWS) diffraction, 232
x-ray diffraction
crystal structure, 208–209
local atomic correlation, 215–217
Layer density, surface phenomena, molecular dynamics (MD) simulation, temperature variation, 164
Layer Korringa Kohn Rostoker (LKKR) technique, band theory of magnetism, 516
Lead zirconium titanate film, impulsive stimulated thermal scattering (ISTS), 751–752
Leak detection, vacuum systems, 20–22
Least-squares computation
energy-dispersive spectrometry (EDS), peak overlap deconvolution, multiple linear least squares (MLLS) method, 1144–1145
neutron powder diffraction
refinement algorithms, 1300–1301
Rietveld refinements, 1296
single-crystal x-ray structure determination, crystal structure refinements, 857–858
thermal diffusivity, laser flash technique, 389
L edges
resonant scattering
angular dependent tensors L=2, 916
L=2 measurements, 910–911
L=4 measurements, 911–912
x-ray absorption fine structure (XAFS) spectroscopy, 873
Lennard-Jones parameters, chemical vapor deposition (CVD) model, kinetic theory, 171
Lens defects and resolution
optical microscopy, 669
scanning electron microscopy (SEM), 1054
transmission electron microscopy (TEM), 1078–1080
aperture diffraction, 1078
astigmatism, 1079
chromatic aberration, 1078
resolution protocols, 1079–1080
spherical aberration, 1078
Lenz-Jensen potential, ion-beam analysis (IBA), non-Rutherford cross-sections, 1190
Lenz's law of electromagnetism
diamagnetism, 494
closed/core shell diamagnetism, dipole moments, atomic origin, 512
superconductor magnetization, 518–519
Level-crossing double resonance NQR nutation spectroscopy, 785
LHPM program, neutron powder diffraction
axial divergence peak asymmetry, 1293
Rietveld refinements, 1306
Lifetime analysis
carrier lifetime measurement, free carrier absorption (FCA), 441–442
x-ray photoelectron spectroscopy (XPS), 973–974
Lifetime characterization. *See* Carrier lifetime measurement
Lifetime depth profiling, carrier lifetime measurement, free carrier absorption (FCA), 441
Lifetime mapping, carrier lifetime measurement, free carrier absorption (FCA), 441–442
Light-emitting diodes (LEDs)
carrier lifetime measurement, photoluminescence (PL), 451–452
deep level transient spectroscopy (DLTS), semiconductor materials, 419

Light-ion backscattering, medium-energy backscattering, trace-element sensitivity, 1266–1267
Light leakage, energy-dispersive spectrometry (EDS), 1154–1155
Light product, particle scattering, nuclear reactions, 57
Light scattering, Raman spectroscopy of solids, semiclassical physics, 701–702
Light sources
  ellipsometry, 738
  ultraviolet photoelectron spectroscopy (UPS), 728–729
Linear combination of atomic orbitals (LCAOs), metal alloy bonding, precision calculations, 142–143
Linear dependence, carrier lifetime measurement, free carrier absorption (FCA), 438–439
Linear elastic fracture mechanics (LEFM)
  fracture toughness testing
    crack tip opening displacement (CTOD) (δ), 307
    stress intensity factor *(K),* 306
  fracture-toughness testing, 302
    basic principles, 303
Linear laws, binary/multicomponent diffusion, 146–147
Linearly variable differential transformer (LVDT), stress-strain analysis, low-temperature testing, 286
Linear muffin tin orbital method (LMTO), electronic topological transitions, van Hove singularities in CuPt, 272–273
Linear polarization, corrosion quantification, 596–599
Line scans, Auger electron spectroscopy (AES), 1163–1164
Lineshape analysis, surface x-ray diffraction, 1019
Liouville's theorem, ultraviolet photoelectron spectroscopy (UPS), figure of merit light sources, 734–735
Liquid alkanes, liquid surface x-ray diffraction, surface crystallization, 1043
Liquid-filled thermometer, operating principles, 35
Liquid metals, liquid surface x-ray diffraction, 1043
Liquid nitrogen (LN2) temperature trap, diffusion pump, 6–7
Liquid surfaces, x-ray diffraction
  basic principles, 1028–1036
  competitive and related techniques, 1028
  data analysis and interpretation, 1039–1043
    Langmuir monolayers, 1041–1043
    liquid alkane crystallization, 1043
    liquid metals, 1043
    simple liquids, 1040–1041
  non-specular scattering
    GID, diffuse scattering, and rod scans, 1038–1039
    reflectivity measurements, 1033
  p-polarized x-ray beam configuration, 1047
  reflectivity, 1029–1036
    Born approximation, 1033–1034
    distorted-wave Born approximation, 1034–1036
    Fresnel reflectivity, 1029–1031
    grazing incidence diffraction and rod scans, 1036
    instrumentation, 1036–1038
    multiple stepwise and continuous interfaces, 1031–1033
    non-specular scattering, 1033
  research background, 1027–1028
  specimen modification, 1043–1045
Literature sources
  dynamical diffraction, 226–227
  on thermal analysis, 343–344
Lithium models, computational analysis, 72–74
Load-displacement behaviors, fracture toughness testing
  J-integral approach, 306–307
  measurement and recording apparatus, 307–308
  notched specimens, 303–304
  stable crack mechanics, 309–311
  unstable fractures, 308–309
Loading modes, fracture toughness testing, 302–303
Load-lock procedures, surface x-ray diffraction, ultrahigh-vacuum (UHV) systems, 1023
Load-to-precision ratio (LPR), thermogravimetric (TG) analysis, 347–350
Local atomic correlation, x-ray diffraction, short-range ordering, 214–217
Local atomic dipole moments
  collective magnetism, 515
  ionic magnetism, 513–515
Local exchange splitting, metal alloy magnetism, 189–190
  atomic short range order (ASRO), 191
Localized material deposition and dissolution, scanning electrochemical microscopy (SECM), 645–646
Local measurements, carrier lifetime measurement, device-related techniques, 435
Local moment fluctuation, metal alloy magnetism, 187–188
Local spin density approximation (LDA)
  computational theory, 74
  diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 263–266
  electronic structure analysis
    basic components, 77–84
    elastic constants, 81
    extensions, 76–77
    gradient corrections, 78
    GW approximation, 84–85
    heats of formation and cohesive energies, 79–81
    implementation, 75–76
    magnetic properties, 81–83
    optical properties, 83–84
    structural properties, 78–79
    summary, 75
    SX approximation, 85–86
  metal alloy bonding, accuracy calculations, 139–141
  metal alloy magnetism, limitations, 201
Local spin density approximation (LDA) + U theory
  electronic structure analysis, 75
    dielectric screening, 86–87
  metal alloy magnetism, competitive and related techniques, 186
Local spin density approximation (LSDA)
  metal alloy magnetism, competitive and related techniques, 185–186
  transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Lock-in amplifiers/photon counters, photoluminescence (PL) spectroscopy, 685
Longitudinal Kerr effect, surface magneto-optic Kerr effect (SMOKE), 571–572
Longitudinal relaxation time, nuclear quadrupole resonance (NQR), 779–780
Long-range order (LRO), diffuse intensities, metal alloys, 253–254
Lorentz force
  electromagnet structure and properties, 499–500
  magnetic field effects and applications, 496
  Mössbauer spectroscopy, 817–818
  pulse magnets, 504–505
Lorentzian equation
  magnetic x-ray scattering, 934–935
  two-beam diffraction, diffracted intensities, 234
Lorentzian peak shape, x-ray photoelectron spectroscopy (XPS), 1005–1006
Lorentz transmission electron microscopy
  basic principles, 1064
  magnetic domain structure measurements, 551–552
Lossy samples, electron paramagnetic resonance (EPR), 799
Low-energy electron diffraction (LEED)
  automation, 1127
  basic principles, 1121–1125
  calculation protocols, 1134–1135
  coherence estimation, 1134
  complementary and related techniques, 1120–1121
  data analysis and interpretation, 1127–1131
  instrumentation criteria, 1125–1127
  limitations, 1132
  liquid surfaces and monomolecular layers, 1028
  qualitative analysis
    basic principles, 1122–1124
    data analysis, 1127–1128
  quantitative measurements
    basic principles, 1124–1125
    data analysis, 1128–1131
  research background, 1120–1121
  sample preparation, 1131–1132
  scanning tunneling microscopy (STM), sample preparation, 1117
  specimen modification, 1132
  surface x-ray diffraction and, 1007–1008
  ultraviolet photoelectron spectroscopy (UPS), sample preparation, 732–733
Low-energy electron microscopy (LEEM), magnetic domain structure measurements
  spin polarized low-energy electron microscopy (SPLEEM), 556–557
  x-ray magnetic circular dichroism (XMCD), 555
Low-energy ion scattering (LEIS)
  Auger electron spectroscopy (AES) *vs.,* 1158–1159
  ion beam analysis (IBA), 1175–1176
  medium-energy backscattering, 1259
Low-energy peak distortion, energy-dispersive spectrometry (EDS), 1155
Low-field sensors, magnetic field measurements, 506
Low-temperature photoluminescence (PL) spectroscopy, band identification, 686–687
Low-temperature superconductors (LTSs), electrical transport measurement
  applications, 472
  contact materials, 474
  current sharing with other materials, 474
Low-temperature testing, stress-strain analysis, 286
Lubricated materials, tribological and wear testing, 324–325
  properties of, 326
Luder's strain, stress-strain analysis, 282
Luggin capillaries, semiconductor electrochemical cell design, photocurrent/photovoltage measurements, 610
Lumped-circuit resonators, electron paramagnetic resonance (EPR), 796
Lumped deposition model, chemical vapor deposition (CVD), 167–168
  limitations of, 175–176

Macroscopic properties, computational analysis, 72–74
Madelung potential, x-ray photoelectron spectroscopy (XPS)
  chemical state information, 986
  initial-state effects, 977–978
Magnetic bearings, turbomolecular pumps, 7
Magnetic circular dichroism (MCD)
  magnetic domain structure measurements, 555–556
    automation of, 555–556
    basic principles, 555
    procedures and protocols, 555
  ultraviolet photoelectron spectroscopy (UPS) and, 726
Magnetic diffraction, magnetic neutron scattering, 1329–1330
Magnetic dipoles, magnetism, general principles, 492
Magnetic domain structure measurements
  bitter pattern imaging, 545–547
  holography, 554–555
  Lorenz transmission electron microscopy, 551–552
  magnetic force microscopy (MFM), 549–550
  magneto-optic imaging, 547–549
  scanning electron microscopy (Types I and II), 550–551
    polarization analysis, 552–553
  scanning Hall probe and scanning SQUID microscopes, 557
  spin polarized low-energy electron microscopy, 556–557
  theoretical background, 545
  x-ray magnetic circular dichroism, 555–556
Magnetic effects, diffuse intensities, metal alloys, chemical order and, 268
Magnetic field gradient
  Mössbauer spectroscopy, hyperfine magnetic field (HMF), 823–824
  spatially resolved nuclear quadruople resonance, 785–786
Magnetic fields
  continuous magnetic fields, 505
  effects and applications, 496
  electromagnets, 499–500
  generation, 496–505
  laboratories, 505
  magnetotransport in metal alloys
    free electron models, 563–565
    research applications, 559
    transport equations, 560–561
  measurements
    high-field sensors, 506–508
    low-field sensors, 506
    magnetoresistive sensors, 508–509
    nuclear magnetic resonance, 509–510
    research background, 505–506
    search coils, 509
  permanent magnets, 497–499
  pulse magnets, 504–505
    laboratories, 505
  research background, 495–496
  resistive magnets, 502–504
    hybrid magnets, 503–504
    power-field relations, 502–503
  superconducting magnets, 500–502
    changing fields stability and losses, 501–502
    protection, 501
    quench and training, 501
Magnetic force microscopy (MFM), magnetic domain structure measurements, 549–550
Magnetic form factor, resonant magnetic x-ray scattering, 922–924
Magnetic induction, principles of, 511
Magnetic measurements, superconductors, electric transport measurements *vs.*, 472–473
Magnetic moments
  band theory of solids, 515–516
  collective magnetism, 515
  dipole moment coupling, 519–527
    antiferromagnetism, 524
    ferrimagnetism, 524–525
    ferromagnetism, 522–524
    Heisenberg model and exchange interactions, 525–526
    helimagnetism, 526–527
    paragmagnetism, classical and quantum theories, 519–522
  dipole moments, atomic origin, 512–515
    closed/core shell diamagnetism, 512
    ionic magnetism, local atomic moment origins, 513–515
  magnetic field quantities, 511–512
  neutron powder diffraction, 1288–1289
  research background, 511
  spin glass and cluster magnetism, 516–517
  superconductors, 517–519
Magnetic neutron scattering
  automation, 1336
  data analysis, 1337
  diffraction applications, 1332–1335
  inelastic scattering
    basic principles, 1331
    protocols for, 1335–1336
  limitations, 1337–1338
  magnetic diffraction, 1329–1330
  neutron magnetic diffuse scattering, 904–905
  polarized beam technique, 1330–1331
  research background, 1328–1329
  sample preparation, 1336–1337
  subtraction techniques, 1330
Magnetic order, binary/multicomponent diffusion, substitutional and interstitial metallic systems, 153–154
Magnetic permeability
  defined, 492–493
  vacuum permeability, 511
Magnetic phase transition theory
  critical exponents, 530
  Landau theory, 529–530
  thermodynamics, 528–529
Magnetic phenomena, superconductors, electrical transport measurements, 479
Magnetic resonance imaging (MRI). *See also* Nuclear magnetic resonance (NMR)
  applications, 767–771
  basic principles and applications, 762–763
  permanent magnets, 496
  superconducting magnets, 500–502
  theoretical background, 765–767
Magnetic sector mass spectrometer, pressure measurments, 16–17
Magnetic short-range order (MSRO), face-centered cubic (fcc) iron, moment alignment *vs.* moment formation, 194–195
Magnetic susceptibility, defined, 492–493
Magnetic x-ray scattering
  data analysis and interpretation, 934–935
  hardware criteria, 925–927
  limitations, 935–936
  magnetic neutron scattering and, 1320–1321
  nonresonant scattering
    antiferromagnets, 928–930
    ferromagnets, 930
    research background, 918–919
    theoretical concepts, 920–921
  research background, 917–919
  resonant scattering
    antiferromagnets, 930–932
    ferromagnets, 932
    research background, 918–919
    theoretical concepts, 921–924
  sample preparation, 935
  spectrometer criteria, 927–928
  surface magnetic scattering, 932–934
  theoretical concepts, 919–925
    ferromagnetic scattering, 924–925
    nonresonant scattering, 920–921
    resonant scattering, 921–924
Magnetism. *See also* Paramagnetism
  antiferromagnetism, 494
  diamagnetism, 494
  ferromagnetism, 494–495
  local density approximation (LDA), 81, 83
  metal alloys
    anisotropy, 191–193
      MAE calculations, 192
      pure Fe, Ni, and Co, 193
    approximation techniques, 185–186
    atomic short range order, 190–191
    atoms-to-solids transition, 180–181
    bonding accuracy calculations, 140–141
    data analysis and interpretation, 193–200
      ASRO in FeV, 196–198
      ASRO in gold-rich AuFe alloys, 198–199
      atomic long- and short-range order, NiFe alloys, 193
      magnetic moments and bonding in FeV alloys, 196
      magnetocrystalline anisotropy, Co-Pt alloys, 199–200
      moment alignment *vs.* formation in fcc Fe, 193–195
      negative thermal expansion effects, 195–196
    electronic structure and Slater-Pauling curves, 184–185
    finite temperature paramagnetism, 186–187
    first-principles theories, 188–189
      paramagnetic Fe, Ni, and Co, 189
    limitations of analysis, 200–201
    local exchange splitting, 189–190
    local moment fluctuations, 187–188
    research background, 180
    solid-solution alloys, 183–184
    transition metal ground state, 181–183
  susceptibility and permeability, 492–493
  theory and principles, 491–492
  thermogravimetric (TG) analysis, mass measurement errors and, 357–358
  units, 492
Magnetization
  magnetic field effects and applications, 496
  magnetic moment
    band theory of solids, 515–516
    collective magnetism, 515
    dipole moment coupling, 519–527
      antiferromagnetism, 524
      ferrimagnetism, 524–525
      ferromagnetism, 522–524
      Heisenberg model and exchange interactions, 525–526
      helimagnetism, 526–527
      paragmagnetism, classical and quantum theories, 519–522
    dipole moments, atomic origin, 512–515
      closed/core shell diamagnetism, 512
      ionic magnetism, local atomic moment origins, 513–515
    magnetic field quantities, 511–512
    research background, 511
    spin glass and cluster magnetism, 516–517
    superconductors, 517–519
Magnetocrystalline anisotropy energy (MAE), metal alloy magnetism

calculation techniques, 192–193
Co-Pt alloys, 199–200
Magnetoelastic scattering, magnetic x-ray scattering errors, 935–936
Magnetometers/compound semiconductor sensors, Hall effect, 507
Magnetometry
automation, 537
limitations, 538
basic principles, 531–533
calibration procedures, 536
data analysis and interpretation, 537–538
flux magnetometers, 533–534
force magnetometers, 534–535
high-pressure magnetometry, 535
limitations, 538–539
research background, 531
rotating-sample magnetometer, 535
sample contamination, 538
sample preparation, 536–537
surface magnetometry, 535–536
vibrating-coil magnetometer, 535
Magneto-optic imaging
magnetic domain structure measurements, 547–549
surface magneto-optic Kerr effect (SMOKE)
automation, 573
classical theory, 570–571
data analysis and interpretation, 573–574
limitations, 575
multilayer formalism, 571
phenomenologica origins, 570
protocols and procedures, 571–573
quantum theory, ferromagnetism, 571
research background, 569–570
sample preparation, 575
Magneto-optic Kerr effect (MOKE)
surface magneto-optic Kerr effect (SMOKE) evolution, 569–570
superlattice structures, 574
x-ray magnetic circular dichroism (XMCD), 953–955
Magnetoresistive sensors, magnetic field measurements, 508–509
Magnetotransport properties, metal alloys
automated procedures, 565–566
data analysis and interpretation, 566
Hall resistance, 560
limitations, 567–568
magnetic field behaviors, 563–565
measured *vs.* intrinsic quantities, 559–560
research background, 559
sample preparation, 566–567
thermal conductance, 560
thermopower, 560
transport equations, 560–561
zero magnetic field behaviors, 561–563
Majority carrier currents, illuminated semiconductor-liquid interface, J-E equations, 608
Many-body interaction potentials
molecular dynamics (MD) simulation, surface phenomena, 158–159
transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Mapping techniques, Auger electron spectroscopy (AES), 1164–1165
Mass, definitions, 24–26
Mass loss, thermogravimetric (TG) analysis, 345
sample preparation, 357
Mass measurements
basic principles, 24
combustion calorimetry, 378–379
cyclotron resonance (CR), 806
electrochemical quartz crystal microbalance (EQCM), 653
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 394
heavy-ion backscattering spectrometry (HIBS), 1281
indirect techniques, 24
mass, weight, and density definitions, 24–26
mass measurement process assurance, 28–29
materials characterization, 1
thermogravimetric (TG) analysis, 345–346
error detection, 357–358
weighing devices, balances, 27–28
weight standards, 26–27
Mass resolution
heavy-ion backscattering spectrometry (HIBS), 1278–1279
ion beam analysis (IBA), 1178
ion-beam analysis (IBA), ERD/RBS techniques, 1191–1197
Mass spectrometers
evolved gas analysis (EGA), thermal degradation studies, 395–396
leak detection, vacuum systems, 21–22
pressure measurments, 16–17
Mass transfer, thermogravimetric (TG) analysis, 350–352
Material errors, hardness testing, 322
Materials characterization
common concepts, 1
particle scattering applications, 61
ultraviolet/visible absorption (UV-VIS) spectroscopy
competitive and related techniques, 690
interpretation, 694
measurable properties, 689
procedures and protocols, 691–692
Matrix corrections, energy-dispersive spectrometry (EDS), 1145–1147
Matrix factor, x-ray photoelectron spectroscopy (XPS), elemental composition analysis, 985–986
Matrix representations, vibrational Raman spectroscopy, symmetry operators, 717–720
Matrix techniques, liquid surface x-ray diffraction, reflectivity measurements, 1031–1033
Mattheissen's rule, magnetotransport in metal alloys, 562–563
Maximum entropy method (MEM)
neutron powder diffraction, structure factor relationships, 1298
nutation nuclear resonance spectroscopy, 784
rotating frame NQR imaging, 787–788
scanning transmission electron microscopy (STEM), data interpretation, 1105–1106
Maximum measurement current, superconductors, electrical transport measurements, 479
Maximum-sensitivity regime, photoconductivity (PC), carrier lifetime measurement, 445
Maxwell's equation
dynamical diffraction, 228
microwave measurement techniques, 408–409
Mean-field approximation (MFA)
diffuse intensities, metal alloys
atomic short-range ordering (ASRO) principles, errors, 257
basic principles, 254
concentration waves, 262
metal alloy magnetism, first-principles calculations, 188–189
phase diagram predictions, 92–96
aluminum-nickel alloys, 107
cluster variational method (CVM), 99–100
Mean-field theory
collective magnetism, 515
ferrimagnetism, 525
ferromagnetism, 522–524
Mean-square atomic displacement, surface phenomena, molecular dynamics (MD) simulation, high-temperature structure and dynamics, 162–163
Mean-square vibrational amplitudes, surface phenomena, molecular dynamics (MD) simulation, room temperature structure and dynamics, 160–161
Measured quantities, magnetotransport in metal alloys, intrinsic quantities *vs.,* 559–560
Measurement errors
energy-dispersive spectrometry (EDS), 1136
protocols, 1156–1157
hardness testing, 322
Measurement frequency, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 625
Measurement time, semiconductor materials, Hall effect, 416–417
Mechanical abrasion, metallographic analysis
7075-T6 anodized aluminum alloy, 68
deformed high-purity aluminum, 69
sample preparation, 66
Mechanical polishing
ellipsometric measurements, 741
metallographic analysis
7075-T6 anodized aluminum alloy, 68
deformed high-purity aluminum, 69
sample preparation, 66
4340 steel, 67
Mechanical testing. *See* specific testing protocols, e.g. Tension testing
data analysis and interpretation, 287–288
elastic deformation, 281
method automation, 286–287
nonuniform plastic deformation, 282
research background, 279
sample preparation, 287
stress/strain analysis, 280–282
curve form variables, 282–283
definitions, 281–282
material and microstructure, 283
temperature and strain rate, 283
yield-point phenomena, 283–284
tension testing
basic principles, 279–280
basic tensile test, 284–285
elastic properties, 279
environmental testing, 286
extreme temperatures and controlled environments, 285–286
high-temperature testing, 286
low-temperature testing, 286
plastic properties, 279–280
specimen geometry, 285
testing machine characteristics, 285
tribological testing
acceleration, 333
automation of, 333–334
basic principles, 324–326
control factors, 332–333
data analysis and interpretation, 334
equipment and measurement techniques, 326–327
friction coefficients, 326
friction testing, 328–329
general procedures, 326
limitations, 335
research background, 324
results analysis, 333
sample preparation, 335

Mechanical testing. *See* specific testing protocols, e.g. Tension testing (*Continued*)
test categories, 327–328
wear coefficients, 326–327
wear testing, 329–332
uniform plastic deformation, 282
Mechanical tolerances, tribological testing, 332–333
MEDIC algorithm, neutron powder diffraction, structure factor relationships, 1298
Medium-boundary/medium-propagation matrices, surface magneto-optic Kerr effect (SMOKE), 576–577
Medium-energy backscattering
applications, 1268–1269
automation, 1269
backscattering data, 1269–1270
basic principles, 1261–1265
complementary and alternative techniques, 1261
data analysis and interpretation, 1269–1271
forward-recoil data, 1270–1271
instrumentation criteria, 1267–1268
limitations, 1271–1272
research background, 1258–1259, 1259–1261
resolution, 1267
safety issues, 1269
sensitivity parameters, 1266–1267
spectrometer efficiency, 1265–1266
time-of-flight spectrometry (TOS), 1265
Medium energy ion scattering (MEIS), ion beam analysis (IBA), 1175–1176
Mega-electron-volt deuterons, high-energy ion beam analysis (IBA), 1176–1177
Meissner effect
diamagnetism, 494
superconductor magnetization, 517–519
Mercury-filled thermometer, operating principles, 35
Mercury probe contacts, capacitance-voltage (C-V) characterization, 461–462
Mesoscopic Monte Carlo method, microstructural evolution, 114–115
Mesoscopic properties, neutron powder diffraction, 1293–1296
microstrain broadening, 1294–1295
particle size effect, 1293–1294
stacking faults, 1295–1296
Metal alloys. *See also* Transition metals; specific alloys
bonding
limits and pitfalls, 138–144
accuracy limits, 139–141
all-electron *vs.* pseudopotential methods, 141–142
basis sets, 142–143
first principles *vs.* tight binding, 141
full potentials, 143
precision issues, 141–144
self-consistency, 141
structural relaxations, 143–144
phase formation, 135–138
bonding-antibonding effects, 136–137
charge transfer and electronegativities, 137–138
Friedel's *d*-band energetics, 135
size effects, 137
topologically close-packed phases, 135–136
transition metal crystal structures, 135
wave function character, 138
research background, 134–135
diffuse intensities
atomic short-range ordering (ASRO)
principles, 256
concentration waves
density-functional approach, 260–262
first-principles, electronic-structure calculations, 263–266
multicomponent alloys, 257–260
mean-field results, 262
mean-field theory, improvement on, 262–267
pair-correlation functions, 256–257
sum rules and mean-field errors, 257
competitive and related techniques, 254–255
computational principles, 252–254
data analysis and interpretation, 266–273
Cu_2NiZn ordering wave polarization, 270–271
CuPt van Hove singularities, electronic topological transitions, 271–273
magnetic coupling and chemical order, 268
multicomponent alloys, 268–270
Ni-Pt hybridization and charge correlation, 266–268
temperature-dependent shift, ASRO peaks, 273
high-temperature experiments, effective interactions, 255–256
liquid surface x-ray diffraction, liquid metals, 1043
magnetism
anisotropy, 191–193
MAE calculations, 192
pure Fe, Ni, and Co, 193
approximation techniques, 185–186
atomic short range order, 190–191
atoms-to-solids transition, 180–181
data analysis and interpretation, 193–200
ASRO in FeV, 196–198
ASRO in gold-rich AuFe alloys, 198–199
atomic long- and short-range order, NiFe alloys, 193
magnetic moments and bonding in FeV alloys, 196
magnetocrystalline anisotropy, Co-Pt alloys, 199–200
moment alignment *vs.* formation in fcc Fe, 193–195
negative thermal expansion effects, 195–196
electronic structure and Slater-Pauling curves, 184–185
finite temperature paramagnetism, 186–187
first-principles theories, 188–189
paramagnetic Fe, Ni, and Co, 189
limitations of analysis, 200–201
local exchange splitting, 189–190
local moment fluctuations, 187–188
research background, 180
solid-solution alloys, 183–184
transition metal ground state, 181–183
magnetotransport properties
automated procedures, 565–566
data analysis and interpretation, 566
Hall resistance, 560
limitations, 567–568
magnetic field behaviors, 563–565
measured *vs.* intrinsic quantities, 559–560
research background, 559
sample preparation, 566–567
thermal conductance, 560
thermopower, 560
transport equations, 560–561
zero magnetic field behaviors, 561–563
photoluminescence (PL) spectroscopy, broadening, 682
scanning tunneling microscopy (STM) analysis, 1115
surface phenomena, molecular dynamics (MD) simulation, metal surface phonons, 161–162
transmission electron microscopy (TEM), specimen modification, 1088
ultraviolet photoelectron spectroscopy (UPS), sample preparation, 732–733
Metallographic analysis, sample preparation
basic principles, 63–64
cadmium plating composition and thickness, 4340 steel, 68
etching procedures, 67
mechanical abrasion, 66
microstructural evaluation
7075-T6 anodized aluminum, 68–69
deformed high-purity aluminum, 69
4340 steel sample, 67–68
mounting protocols and procedures, 66
polishing procedures, 66–67
sectioning protocols and procedures, 65–66
strategic planning, 64–65
Metal organic chemical vapor deposition (MOCVD), *pn* junction characterization, 469–470
Metal-semiconductor (MS) junctions, capacitance-voltage (C-V) characterization, 457–458
Metamagnetic state, Landau magnetic phase transition, 530
Microbalance measurement, thermogravimetric (TG) analysis, error detection, 358–359
Microbeam applications
trace element accelerator mass spectrometry (TEAMS), ultraclean ion sources, 1240
x-ray microprobes
strain distribution
ferroelectric sample, 948–949
tensile loading, 948
trace element distribution, SiC nuclear fuel barrier, 947–948
Microchannel plates (MCP)
carrier lifetime measurement, photoluminescence (PL), 451–452
heavy-ion backscattering spectrometry (HIBS)
sensitivity and mass resolution, 1278–1279
time-of-flight spectrometry (TOS), 1278
semiconductor-liquid interfaces, transient decay dynamics, 621–622
Microfabrication of specimens, scanning electrochemical microscopy (SECM), 644–645
Microindentation hardness testing
applications, 319
case hardening, 318
Micro-particle-induced x-ray emission (Micro-PIXE)
applications, 1215–1216
basic principles, 1212
competitive methods, 1210–1211
research background, 1210
Microscopic field model (MFM), microstructural evolution, 115–117
Microscopic master equations (MEs), microstructural evolution, 116–117
Microstrain broadening, neutron powder diffraction, 1294–1296
Microstructural evaluation
fracture toughness testing, crack extension measurement, 311
metallographic analysis
aluminum alloys, 7075-T6, 68–69
deformed high-purity aluminum, 69
steel samples, 4340 steel, 67–68
stress-strain analysis, 283
Microstructural evolution
continuum field method (CFM)
applications, 122
atomistic simulation, continuum process modeling and property calculation, 128–129
basic principles, 117–118
bulk chemical free energy, 119–120

coarse-grained approximation, 118–119
coarse-grained free energy formulation, 119
coherent ordered precipitates, 122–126
diffuse-interface nature, 119
diffusion-controlled grain growth, 126–128
elastic energy, 120–121
external field energies, 121
field kinetic equations, 121–122
future applications, 128
interfacial energy, 120
limits of, 130
numerical algorithm efficiency, 129–130
research background, 114–115
slow variable selection, 119
theoretical basis, 118
field simulation, 113–117
atomistic Monte Carlo method, 117
cellular automata (CA) method, 114–115
continuum field method (CFM), 114–115
conventional front-tracking, 113–114
inhomogeneous path probability method (PPM), 116–117
mesoscopic Monte Carlo method, 114–115
microscopic field model, 115–116
microscopic master equations, 116–117
molecular dynamics (MD), 117
morphological patterns, 112–113
Microwave measurement techniques
basic principles, 408–409
carrier lifetime measurement, photoconductivity (PC) techniques, 447–449
electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, 796
protocols and procedures, 409–410
semiconductor-liquid interfaces, time-resolved microwave conductivity, 622–623
Microwave PC-decay technique, carrier lifetime measurement, optical techniques, 434
Microwave reflectance power, carrier lifetime measurement, photoconductivity (PC) techniques, 447–449
Micro-x-ray absorption fine structure (XAFS) spectroscopy, fluorescence analysis, 943
Miller-Bravais indices, lattice systems, 46
Miller indices
lattice systems, 45–46
magnetic neutron scattering, polarized beam technique, 1330–1331
phonon analysis, 1320
surface x-ray diffraction, 1008
x-ray powder diffraction, 838
single-crystal x-ray structure determination, 853
Mininum-detectable limits (MDLs), fluorescence analysis, 940–941
Minority carrier injection
carrier lifetime measurement, 428–429
semiconductor-liquid interface
illuminated J-E equations, 608
laser spot scanning (LSS), 626–628
Mirror planes
group theoretical analysis, vibrational Raman spectroscopy, 703–704
single-crystal x-ray structure determination, crystal symmetry, 855–856
Miscibility gap, phase diagram predictions, mean-field approximation, 94–96
Mixed-potential theory, corrosion quantification, Tafel technique, 593–596
Mobility measurements
binary/multicomponent diffusion
Fick's law, 147
substitutional and interstitial metallic systems, 152–153
substitutional and interstitial metallic systems, tracer diffusion, 154–155
Modulated differential scanning calorimetry (MDSC), basic principles, 365–366
Modulated free carrier absorption, carrier lifetime measurement, 440
Modulation amplitude, electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, 796–797
Modulation-type measurements, carrier lifetimes, 435–438
data interpretation issues, 437
limitations of, 437
Molecular beam epitaxy (MBE)
low-energy electron diffraction (LEED), 1121
surface magneto-optic Kerr effect (SMOKE), superlattice structures, 573–574
Molecular dissociation
microstructural evolution, 117
sputter-ion pump, 11
Molecular drag pump, application and operation, 5
Molecular dynamics (MD) simulation
electronic structure analysis, phase diagram prediction, 101–102
phase diagram prediction, static displacive interactions, 105–106
surface phenomena
data analysis and interpretation, 160–164
higher-temperature dynamics, 161–162
interlayer relaxation, 161
layer density and temperature variation, 164
mean-square atomic displacement, 162–163
metal surface phonons, 161–162
room temperature structure and dynamics, 160–161
thermal expansion, 163–164
limitations, 164
principles and practices, 158–159
related theoretical techniques, 158
research background, 156
surface behavior, 156–157
temperature effects on surface behavior, 157
Molecular flow, vacuum system design, 19
Molecular orbitals, ultraviolet photoelectron spectroscopy (UPS), 727–728
Molecular point group, group theoretical analysis, vibrational Raman spectroscopy, 703–704
Moliére approximation, particle scattering, central-field theory, deflection function, 60
Moment formation
metal alloy magnetism, 180
FeV alloys, 196
local moment fluctuation, 187–188
nickel-iron alloys, atomic long and short range order (ASRO), 195–196
metal alloy paramagnetism, finite temperatures, 187
Momentum matching, low-energy electron diffraction (LEED), sample preparation, 1131–1132
Momentum space representation, Mössbauer effect, 819–820
Monochromatic illumination, semiconductor-liquid interface, 610
Monochromators. *See* Spectrometers/monochromators
Monomolecular layers, x-ray diffraction
basic principles, 1028–1036
competitive and related techniques, 1028
data analysis and interpretation, 1039–1043
Langmuir monolayers, 1041–1043
liquid alkane crystallization, 1043
liquid metals, 1043
simple liquids, 1040–1041
non-specular scattering
GID, diffuse scattering, and rod scans, 1038–1039
reflectivity measurements, 1033
p-polarized x-ray beam configuration, 1047
reflectivity, 1029–1036
Born approximation, 1033–1034
distorted-wave Born approximation, 1034–1036
Fresnel reflectivity, 1029–1031
grazing incidence diffraction and rod scans, 1036
instrumentation, 1036–1038
multiple stepwise and continuous interfaces, 1031–1033
non-specular scattering, 1033
research background, 1027–1028
specimen modification, 1043–1045
Monotonic heating, thermal diffusivity, laser flash technique, 386–387
Monte Carlo simulation
chemical vapor deposition (CVD) model
free molecular transport, 170
radiation, 173
diffuse intensities, metal alloys
concentration waves, multicomponent alloys, 259–260
effective cluster interactions (ECIs), 255–256
energy-dispersive spectrometry (EDS), standardless analysis, backscatter loss correction, 1149
scanning electron microscopy (SEM), signal generation, 1050–1052
Mössbauer effect, basic properties, 818–820
Mössbauer spectroscopy
basic properties, 761–762
bcc iron alloy solutes, 828–830
coherence and diffraction, 824–825
crystal defects and small particles, 830–831
data analysis and interpretation, 831–832
diffuse scattering techniques, comparisons, 883–884
electric quadrupole splitting, 822–823
hyperfine interactions, 820–821
magnetic field splitting, 823–824
isomer shift, 821–822
magnetic neutron scattering and, 1321
Mössbauer effect, 818–820
nuclear excitation, 817–818
phase analysis, 827–828
phonons, 824
radioisotope sources, 825–826
recoil-free fraction, 821
relaxation phenomena, 824
research background, 816–817
sample preparation, 832
single-crystal x-ray structure determniation, 851
synchrotron sources, 826–827
valence and spin determination, 827
Mott detector, ultraviolet photoelectron spectroscopy (UPS), 729
Mott-Schottky plots
electrochemical photocapacitance spectroscopy (EPS), surface capacitance measurements, 625
semiconductor-liquid interface
differential capacitance measurements, 617–619
flat-band potential measurements, 628–629
Mounting procedures, metallographic analysis
7075-T6 anodized aluminum alloy, 68
deformed high-purity aluminum, 69
sample preparation, 66
4340 steel, 67

Muffin-tin approximation
linear muffin tin orbital method (LMTO), electronic topological transitions, van Hove singularities in CuPt, 272–273
metal alloy bonding, precision calculations, 143
Multichannel analyzer (MCA)
energy-dispersive spectrometry (EDS), 1137–1140
heavy-ion backscattering spectrometry (HIBS), 1279
ion-beam analysis (IBA), 1182–1184
ERD/RBS examples, 1191–1197
surface x-ray diffraction
basic principles, 1010
crystallographic alignment, 1014–1015
Multicomponent alloys
concentration waves, 257–260
diffuse intensities, Fermi-surface nesting, and van Hove singularities, 268–270
Multilayer structures
dynamical diffraction, applications, 225
grazing-incidence diffraction (GID), 242–243
surface magneto-optic Kerr effect (SMOKE), 571
Multiphonon mechanism, carrier lifetime measurement, 429
Multiple-beam diffraction, 236–241
basic principles, 236–237
literature sources, 226
NBEAM theory, 237–238
boundary conditions, 238
D-field component eigenequation, 237
eigenequation matrix, 237–238
intensity computations, 238
numerical solution strategy, 238
phase information, 240
polarization density matrix, 241
polarization mixing, 240
second-order Born approximation, 238–240
standing waves, 240
three-beam interactions, 240
Multiple linear least squares (MLLS) method, energy-dispersive spectrometry (EDS), peak overlap deconvolution, 1144–1146
Multiple scattering
magnetic x-ray scattering
error detection, 935
nonresonant scattering, 929–930
x-ray absorption fine structure (XAFS) spectroscopy, 873–874
Multiple-specimen measurement technique, fracture toughness testing, crack extension measurement, 308
Multiple stepwise interfaces, liquid surface x-ray diffraction, reflectivity measurements, 1031–1033
Multiplet structures
metal alloy bonding, accuracy calculations, 140–141
x-ray photoelectron spectroscopy (XPS), final-state effects, 976
Multiple-wavelength anomalous diffraction (MAD), resonant scattering, 905
Multipole contributions, resonant scattering analysis, 909–910
Muon spin resonance, basic principles, 762–763

Nanotechnology, surface phenomena, molecular dynamics (MD) simulation, 156–157
Narrow-band thermometer, operating principles, 37
National Institutes of Standards and Technology (NIST), weight standards, 26–27
Natural oxide films, electrochemical dissolution, ellipsometric measurement, 742
Naval Research Laboratory, trace element accelerator mass spectrometry (TEAMS) research at, 1245–1246
NBEAM theory, multiple-beam diffraction, 237–238
boundary conditions, 238
D-field component eigenequation, 237
eigenequation matrix, 237–238
intensity computations, 238
numerical solution strategy, 238
Near-band-gap emission, carrier lifetime measurement, photoluminescence (PL), 450
Near-edge x-ray absorption fine structure (NEXAFS) spectroscopy. *See also* X-ray absorption near-edge structure (XANES) spectroscopy
micro-XAFS, 943
ultraviolet photoelectron spectroscopy (UPS) and, 726
Near-field regime, scanning tunneling microscopy (STM), 1112–1113
Near-field scanning optical microscopy (NSOM)
magnetic domain structure measurements, 548–549
pn junction characterization, 467
Near-surface effects
magnetic x-ray scattering errors, 936
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1202
Necking phenomenon
nonuniform plastic deformation, 282
stress-strain analysis, 280–281
Neel temperature, antiferromagnetism, 524
Nernst equation
cyclic voltammetry
quasireversible reaction, 583
totally reversible reaction, 582–583
semiconductor-liquid interface, photoelectrochemistry, thermodynamics, 605–606
Nernst-Ettingshausen effect, magnetotransport in metal alloys, transport equations, 561
Net magnetization, nuclear magnetic resonance, 765
Neutron activation-accelerator mass spectrometry, trace element accelerator mass spectrometry (TEAMS) and, 1237
Neutron activation analysis (NAA)
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1201–1202
trace element accelerator mass spectrometry (TEAMS) and, 1237
Neutron diffuse scattering
applications, 889–894
automation, 897–898
bond distances, 885
chemical order, 884–885
comparisons, 884
competitive and related techniques, 883–884
crystalling solid solutions, 885–889
data analysis and interpretation, 894–896
diffuse x-ray scattering techniques, 889–890
inelastic scattering background removal, 890–893
limitations, 898–899
magnetic diffuse scattering, 904–905
magnetic x-ray scattering, comparisons, 918–919
measured intensity calibration, 894
protocols and procedures, 884–889
recovered static displacements, 896–897
research background, 882–884
resonant scattering terms, 893–894
sample preparation, 898
Neutron powder diffraction
angle-dispersive constant-wavelength diffractometer, 1289–1290
constant wavelength diffractometer wavelength, 1304–1305
constraints, 1301
crystallographic refinements, 1300–1301
data analysis, 1291–1300
ab initio structure determination, 1297–1298
axial divergence and peak asymmetry, 1292–1293
Bragg reflection positions, 1291–1292
indexing techniques, 1297–1298
particle structure determination, 1297
peak shape, 1292
quantitative phase analysis, 1296–1297
structure factor extraction, 1298
structure solution, 1298
estimated standard deviations, 1305–1306
limitations, 1300–1302
mesoscopic properties, 1293–1296
microstrain broadening, 1294–1295
particle size effect, 1293–1294
stacking faults, 1295–1296
probe characteristics, 1286–1289
reliability factors, 1305
research background, 1285–1286
restraints, 1301–1302
Rietveld analysis
protocols, 1296
software programs, 1306–1307
sample preparation, 1291
single-crystal neutron diffraction and, 1307–1309
single-crystal x-ray structure determination, 850–851
time-dependent neutron powder diffraction, 1298–1300
time-of-flight diffractometers, 1290–1291
x-ray diffraction *vs.*, 836
Neutron sources, single-crystal neutron diffraction, 1312–1313
Neutron techniques
magnetic neutron scattering
automation, 1336
data analysis, 1337
diffraction applications, 1332–1335
inelastic scattering
basic principles, 1331
protocols for, 1335–1336
limitations, 1337–1338
magnetic diffraction, 1329–1330
polarized beam technique, 1330–1331
research background, 1328–1329
sample preparation, 1336–1337
subtraction techniques, 1330
neutron powder diffraction
angle-dispersive constant-wavelength diffractometer, 1289–1290
constant wavelength diffractometer wavelength, 1304–1305
constraints, 1301
crystallographic refinements, 1300–1301
data analysis, 1291–1300
ab initio structure determination, 1297–1298
axial divergence and peak asymmetry, 1292–1293
Bragg reflection positions, 1291–1292
indexing techniques, 1297–1298
particle structure determination, 1297
peak shape, 1292
quantitative phase analysis, 1296–1297
structure factor extraction, 1298
structure solution, 1298
estimated standard deviations, 1305–1306

limitations, 1300–1302
mesoscopic properties, 1293–1296
microstrain broadening, 1294–1295
particle size effect, 1293–1294
stacking faults, 1295–1296
probe characteristics, 1286–1289
reliability factors, 1305
research background, 1285–1286
restraints, 1301–1302
Rietveld analysis
protocols, 1296
software programs, 1306–1307
sample preparation, 1291
time-dependent neutron powder diffraction, 1298–1300
time-of-flight diffractometers, 1290–1291
phonon studies
automation, 1323
basic principles, 1317–1318
data analysis, 1324–1325
instrumentation and applications, 1318–1323
triple-axis spectrometry, 1320–1323
limitations, 1326–1327
research background, 1316–1317
sample preparation, 1325–1326
specimen modification, 1326
research background, 1285
single-crystal neutron diffraction
applications, 1311–1312
data analysis and interpretation, 1313–1314
neutron sources and instrumentation, 1312–1313
research background, 1307–1309
sample preparation, 1314–1315
theoretical background, 1309–1311
Newton-Raphson algorithm, diffuse intensities, metal alloys, mean-field theory and, 263
Newton's equations, molecular dynamics (MD) simulation, surface phenomena, 159
Newton's law of cooling, combustion calorimetry, 379
Nickel-iron alloys, atomic short range order (ASRO), energetics and electronic origins, 193–196
Nickel-platinum alloys
diffuse intensities, hybriziation in NiPt alloys, charge correlation effects, 266–268
magnetism
first-principles calculations, 189
paramagnetism, 189
magnetocrystalline anisotropy energy (MAE), 193
phase diagram predictions, cluster variational method (CVM), 107–108
Nonconfigurational thermal effects, phase diagram prediction, 106–107
Nonconsumable abrasive wheel cutting, metallographic analysis, sample preparation, 65
Non-contact measurements
basic principles, 407
protocols and procedures, 407–408
Nonevaporable getter pumps (NEGs), applications, operating principles, 12
Nonlinearity, fracture toughness testing
load-displacement curves, 303–304
stable crack mechanics, 309–311
Nonlinear least-squares analysis, thermal diffusivity, laser flash technique, 389
Nonlinear optical effects, cyclotron resonance (CR), far-infrared (FIR) radiation sources, 809
Nonohmic contacts, Hall effect, semiconductor materials, 417
Nonrelativistic collisions, particle scattering conversions, 62–63
fundamental and recoiling relations, 62
Nonresonant depth-profiling, nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and
basic principles, 1202
data analysis, 1206
Nonresonant magnetic x-ray scattering
antiferromagnets, 928–930
ferromagnets, 930
research background, 918–919
theoretical concepts, 920–921
Nonreversible charge-transfer reaactions, cyclic voltammetry, 583–584
Non-Rutherford cross-sections, ion-beam analysis (IBA)
ERD/RBS techniques, 1191–1197
error detection, 1189–1190
Non-specular scattering, liquid surface x-ray diffraction
GID, diffuse scattering, and rod scans, 1038–1039
measurement protocols, 1033–1036
Nonuniform detection efficiency, energy-dispersive spectrometry (EDS), 1139–1140
Nonuniform laser beam, thermal diffusivity, laser flash technique, 390
Nonuniform plastic deformation, stress-strain analysis, 282
Nordsieck's algorithm, molecular dynamics (MD) simulation, surface phenomena, 159
Normalization procedures, x-ray absorption fine structure (XAFS) spectroscopy, 878–879
Notch filter
microbeam analysis, 948
Raman spectroscopy of solids, optics properties, 705–706
Notch-toughness testing
research background, 302
sample preparation, 312
n-type metal-oxide semiconductor (NMOS), ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1229–1230
Nuclear couplings, nuclear quadrupole resonance (NQR), 776–777
Nuclear excitation, Mössbauer spectroscopy, 817–818
Nuclear magnetic resonance (NMR)
applications, 767–771
back-projection imaging sequence, 768
basic principles, 762–763
chemical shift imaging sequence, 770
data analysis and interpretation, 771
diffusion imaging sequence, 769
echo-planar imaging sequence, 768–769
flow imaging sequence, 769–770
gradient recalled echo sequence, 768
instrumentation setup and tuning, 770–771
limitations, 772
magnetic field effects, 496
measurement principles, 509–510
magnetic neutron scattering and, 1321
sample preparation, 772
single-crystal neutron diffraction and, 1309
solid-state imaging, 770
spin-echo sequence, 768
superconducting magnets, 500–502
theoretical background, 763–765
three-dimensional imaging, 770
Nuclear moments, nuclear quadrupole resonance (NQR), 775–776
Nuclear quadrupole resonance (NQR)
data analysis and interpretation, 789
direct detection techniques, spectrometer criteria, 780–781
dynamics limitations, 790
first-order Zeeman perturbation and line shapes, 779
heterogeneous broadening, 790
indirect detection, field cycling methods, 781
nuclear couplings, 776–777
nuclear magnetic resonance (NMR), comparisons, 761–762
nuclear moments, 775–776
one-dimensional Fourier transform NQR, 781–782
research background, 775
sensitivity problems, 789
spatially resolved NQR, 785–789
field cycling methods, 788–789
magnetic field gradient method, 785–786
rotating frame NQR imaging, 786–788
temperature, stress, and pressure imaging, 788
spin relaxation, 779–780
spurious signals, 789–790
Sternheimer effect and electron deformation densities, 777
two-dimensional zero-field NQR, 782–785
exchange NQR spectroscopy, 785
level-crossing double resonance NQR nutation spectroscopy, 785
nutation NRS, 783–784
Zeeman-perturbed NRS (ZNRS), 782–783
zero field energy levels, 777–779
higher spin nuclei, 779
spin 1 levels, 777–778
spin 3/2 levels, 778
spin 5/2 levels, 778–779
Nuclear reaction analysis (NRA)
automation, 1205–1206
background, interference, and sensitivity, 1205
cross-sections and Q values, 1205
data analysis, 1206–1207
energy relations, 1209
energy scanning, resonance depth profiling, 1204–1205
energy spread, 1209–1210
high-energy ion beam analysis (IBA) and, 1176–1177
instrumentation criteria, 1203–1204
limitations, 1207–1208
nonresonant methods, 1202–1203
particle scattering, kinematics, 56–57
research background, 1178–1179, 1200–1202
resonant depth profiling, 1203
specimen modification, 1207
standards, 1205
unwanted particle filtering, 1204
Null ellipsometry
automation, 739
defined, 735
schematic, 737–738
Null Laue method, diffuse scattering techniques, 887–889
data interpretation, 895–897
Number-fixed frame of reference, binary/multicomponent diffusion, 147–148
Numerical algorithms
microstructural evolution, 129–130
multiple-beam diffraction, NBEAM theory, 238
x-ray photoelectron spectroscopy (XPS), composition analysis, 994–996
Numerical aperture, optical microscopy, 669
Nutation nuclear resonance spectroscopy
level-crossing double resonance NQR nutation spectroscopy, 785
zero-field nuclear quadrupole resonance, 783–784
n value
semiconductor-liquid interface, current density-potential properties, 607–608

n value (*Continued*)
superconductors-to-normal (S/N) transition, electrical transport measurement, 472
determination, 482
Nyquist representation
corrosion quantification, electrochemical impedance spectroscopy (EIS), 599–603
semiconductor-liquid interfaces, differential capacitance measurements, 617–619

Object function retrieval, scanning transmission electron microscopy (STEM)
data interpretation, 1105–1106
incoherent imaging, weakly scattering objects, 1111
Objective lenses, optical microscopy, 668
Oblique coordinate systems, diffuse intensities, metal alloys, concentration waves, multicomponent alloys, 259–260
Oblique-light illumination, reflected-light optical microscopy, 676–680
"Off-axis pinhole" device, surface x-ray diffraction, diffractometer alignment, 1020–1021
Off-diagonal disorder, diffuse intensities, metal alloys, concentration waves, first-principles calcuations, electronic structure, 266
Offset, Hall effect sensors, 508
Ohmeter design, bulk measurements, 403–404
Ohm's law
electrochemical quartz crystal microbalance (EQCM), impedance analysis, 655–657
superconductors, electrical transport measurement, 473–474
Oil composition, diffusion pumps, selection criteria, 6
Oil-free (dry) pumps
classification, 4–5
diaphragm pumps, 4–5
molecular drag pump, 5
screw compressor, 5
scroll pumps, 5
sorption pumps, 5
Oil lubrication bearings, turbomolecular pumps, 7
Oil-sealed pumps
applications, 3
foreline traps, 4
oil contamination, avoidance, 3–4
operating principles, 3
technological principles, 3–4
One-body contact, tribological and wear testing, 324–325
One-dimensional Fourier transform NQR, implementation, 781–782
One-electron model, x-ray magnetic circular dichroism (XMCD), 956
dichroism principles and notation, 968–969
One-step photoemission model, ultraviolet photoelectron spectroscopy (UPS), 726–727
One-wave analysis, high-strain-rate testing
data analysis and interpretation, 296–298
Hopkinson bar technique, 291–292
Onsager cavity-field corrections, diffuse intensities, metal alloys, mean-field theory and, 262–263
Open circuit voltage decay (OCVD) technique, carrier lifetime measurement, 435
Optical beam induced current (OBIC) technique, carrier lifetime measurement, diffusion-length-based methods, 434–435
Optical conductivity, ellipsometry, 737
Optical constants, ellipsometry, 737
Optical imaging and spectroscopy
electromagnetic spectrum, 666
ellipsometry
automation, 739
intensity-measuring ellipsometer, 739
null elipsometers, 739
data analysis and interpretation, 739–741
optical properties from phase and amplitude changes, 740–741
polarizer/analyzer phase and amplitude changes, 739–740
dielectric constants, 737
limitations, 742
optical conductivity, 737
optical constants, 737
protocols and procedures, 737–739
alignment, 738–739
compensator, 738
light source, 738
polarizer/analyzer, 738
reflecting surfaces, 735–737
reflectivity, 737
research background, 735
sample preparation, 741–742
impulsive stimulated thermal scattering (ISTS)
applications, 749–752
automation, 753
competitive and related techniques, 744–746
data analysis and interpretation, 753–757
limitations, 757–758
procedures and protocols, 746–749
research background, 744–746
sample preparation and specimen modification, 757
optical microscopy
adjustment protocols, 671–674
basic components, 668–671
research background, 667–668
photoluminescence (PL) spectroscopy
alloy broadening, 682
automation, 686
band-to-band recombination, 684
bound excitons, 682
defect-level transitions, 683–684
donor-acceptor and free-to-bound transitions, 683
excitons and exciton-polaritons, 682
experimental protocols, 684–865
limitations, 686
low-temperature PL spectra
band interpretation, 686–687
crystal characterization, 685–686
phonon replicas, 682–683
research background, 681–682
room-temperature PL plating, 686
curve-fitting, 686
sample preparation, 686
specimen modification, 686
Raman spectroscopy of solids
competitive and related techniques, 699
data analysis and interpretation, 709–713
active vibrational modes, aluminum compounds, 709–710
carbon structure, 712–713
crystal structures, 710–712
dispersed radiation measurement, 707–708
electromagnetic radiation, classical physics, 699–701
Fourier transform Raman spectroscopy, 708–709
group theoretical analysis, vibrational Raman spectroscopy, 702–704
character tables, 718
point groups and matrix representation, symmetry operations, 717–718
vibrational modes of solids, 720–722
vibrational selection rules, 716–720
light scattering, semiclassical physics, 701–702
limitations, 713–714
optical alignment, 706
optics, 705–706
polarization, 708
quantitative analysis, 713
radiation sources, 704–705
research background, 698–699
spectrometer components, 706–707
theoretical principles, 699–700
research background, 665, 667
ultraviolet photoelectron spectroscopy (UPS)
alignment procedures, 729–730
atoms and molecules, 727–728
automation, 731
competitive and related techniques, 725–726
data analysis and interpretation, 731–732
electronic phase transitions, 727
electron spectrometers, 729
energy band dispersion, 723–725
solid materials, 727
light sources, 728–729
limitations, 733
photoemission process, 726–727
photoemission *vs.* inverse photoemission, 722–723
physical relations, 730
sample preparation, 732–733
sensitivity limits, 730–731
surface states, 727
valence electron characterization, 723–724
ultraviolet/visible absorption (UV-VIS) spectroscopy
applications, 692
array detector spectrometer, 693
automation, 693
common components, 692
competitive and related techniques, 690
dual-beam spectrometer, 692–693
limitations, 696
materials characterization, 691–692, 694–695
materials properties, 689
qualitative/quantitative analysis, 689–690
quantitative analysis, 690–691, 693–694
research background, 688–689
sample preparation, 693, 695
single-beam spectrometer, 692
specimen modification, 695–696
Optically-detected resonance (ODR) spectroscopy, cyclotron resonance (CR), 812–813
Optical microscopy
adjustment protocols, 671–674
basic components, 668–671
reflected-light optical microscopy
illumination modes and image-enhancement techniques, 676–680
limitations, 680–681
procedures and protocols, 675–680
research background, 674–675
sample preparation, 680
research background, 667–668
Optical parametric oscillation (OPO), cyclotron resonance (CR), far-infrared (FIR) radiation sources, 809
Optical properties
ellipsometric measurement, 740–741
local density approximation (LDA), 83–84
scanning electron microscopy (SEM), 1054
quality control, 1059–1060
x-ray magnetic circular dichroism (XMCD), 959–962
x-ray microprobes, 945–947
Optical pyrometry
ITS-90 standard, 33
operating principles, 37
Optical reciprocity theorem, grazing-incidence diffraction (GID), distorted-wave Born approximation (DWBA), 244–245

Optical techniques, carrier lifetime measurement, 434
Optics properties, Raman spectroscopy of solids, 705–706
Optimization protocols, energy-dispersive spectrometry (EDS), 1140–1141
Orbital angular momentum number, metal alloy bonding, wave function variation, 138
Orbital effects
  metal alloy bonding, accuracy calculations, 140–141
  x-ray magnetic circular dichroism (XMCD), 953–955
Orientation matrices, surface x-ray diffraction
  crystallographic alignment, 1014–1015
  protocols, 1026
O-ring seals
  configuration, 17–18
  elastomer materials for, 17
  rotational/translation motion feedthroughs, 18–19
Ornstein-Zernicke equation
  diffuse intensities, metal alloys, concentration waves, density-functional theory (DFT), 261–262
  metal alloy magnetism, atomic short range order (ASRO), 190–191
ORTEP program, single-crystal x-ray structure determination, 861–864
Outgassing
  Auger electron spectroscopy (AES), sample preparation, 1170
  hot cathode ionization gauges, 15
  vacuum system principles, 2–3
    pumpdown procedures, 19
  x-ray photoelectron spectroscopy (XPS), sample preparation, 998–999
Overcorrelation, diffuse intensities, metal alloys, mean-field results, 262
Overpotentials, semiconductor materials, J-E behavior, 611–612
  series resistance, 612
Oxidative mechanisms, tribological and wear testing, 324–325
Oxides, scanning tunneling microscopy (STM) analysis, 1115
Oxygen purity, combustion calorimetry, 377–378
  error detection and, 381–382

Pair correlations
  diffuse intensities, metal alloys
    atomic short-range ordering (ASRO) principles, 256–257
    basic definitions, 252–254
  diffuse scattering techniques, 888–889
Palladium tips, scanning tunneling microscopy (STM), 1114
Parabolic rate law, corrosion quantification, electrochemical impedance spectroscopy (EIS), 600–603
Parallel frequencies, electrochemical quartz crystal microbalance (EQCM), 660
Parallel geometry, carrier lifetime measurement, free carrier absorption (FCA), 439440
Paramagnetism
  classical and quantum theories, 519–522
  electron paramagnetic resonance (EPR), theoretical background, 793
  metal alloys, finite temperatures, 186–187
  nuclear magnetic resonance data analysis, 771–772
  principles of, 511–512
  spin glass materials, cluster magnetism of, 516–517
  theory and principles, 493–494
Parratt formalism, liquid surface x-ray diffraction, reflectivity measurements, 1031
Partial differential equations (PDEs), front-tracking simulation, microstructural evolution, 113–114
Partial pressure analyzer (PPA), applications, 16
Particle accelerator
  ion beam analysis (IBA), 1176
  nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1203–1204
Particle detection, medium-energy backscattering, 1260–1261
Particle erosion, wear testing and, 329–330
Particle filtering, nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1204
Particle-induced x-ray emission (PIXE)
  automation, 1216
  basic principles, 1211–1212
  competing methods, 1210–1211
  data analysis, 1216–1218
  limitations, 1220
  protocols and techniques, 1213–1216
  research background, 1179, 1210–1211
  sample preparation, 1218–1219
  specimen modification, 1219–1220
Particle scattering
  central-field theory, 57–61
    cross-sections, 60–61
    deflection functions, 58–61
      approximation, 59–60
      central potentials, 59
      hard spheres, 58–59
    impact parameter, 57–58
    interaction potentials, 57
    materials analysis, 61
    shadow cones, 58
  kinematics, 51–57
    binary collisions, 51
    center-of-mass and relative coordinates, 54–55
    elastic scattering and recoiling, 51–52
    inelastic scattering and recoiling, 52
    nuclear reactions, 56–57
    relativistic collisions, 55–56
    scattering/recoiling diagrams, 52–54
  nonrelativistic collisions
    conversions, 62–63
    fundamental and recoiling relations, 62
  research background, 51
Particle size distribution
  Mössbauer spectroscopy, 830–831
  neutron powder diffraction, 1293–1294
  small-angle scattering (SAS), 222
  thermogravimetric (TG) analysis, 357
Particle structure determination, neutron powder diffraction, 1297
Passive film deposition, corrosion quantification, Tafel technique, 596
Patterson mapping
  neutron powder diffraction
    stacking faults, 1296
    structure factor relationships, 1298
  single-crystal x-ray structure determination, 858–859
    heavy-atom computation, 868–869
  x-ray powder diffraction, candidate atom position search, 840–841
Pauli exclusion principle, metal alloy magnetism, 180
Pauli paramagnetism
  band theory of magnetism, 516
  defined, 493–494
  local moment origins, 513–515
Paul Scherrer Institute (PSI)/ETH Zurich Accelerator SIMS Laboratory, trace element accelerator mass spectrometry (TEAMS) research at, 1242–1244
PBE functionals, local density approximation (LDA), gradient corrections, 78
Peak current measurement, cyclic voltammetry, 587
Peak height measurments
  Auger electron spectroscopy (AES), 1169
  ion-beam analysis (IBA), 1183–1184
Peak identification
  Auger electron spectroscopy (AES), 1162
    ion excitation, 1171–1173
  energy-dispersive spectrometry (EDS), deconvolution protocols, 1144–1145
  x-ray photoelectron spectroscopy (XPS), 991–992
    reference spectra, 996–998
PEAK integration software, surface x-ray diffraction, data analysis, 1024–1025
Peak position, x-ray photoelectron spectroscopy (XPS), 992
Peak shape, neutron powder diffraction, mesoscopic properties, 1293–1296
  microstrain broadening, 1294–1295
  particle size effect, 1293–1294
  stacking faults, 1295–1296
Peak shapes
  neutron powder diffraction, 1292
    axial divergence asymmetry, 1292–1293
    instrumental contributions, 1292
  x-ray photoelectron spectroscopy (XPS), classification, 1004–1007
Peltier heat, magnetotransport in metal alloys
  basic principles, 560
  transport equations, 561
Pendellösung period, two-beam diffraction, 231
  diffracted intensities, 234
Pendry's *R* factor, low-energy electron diffraction (LEED), quantitative analysis, 1130–1131
Penetration depth, x-ray microfluorescence, 943–944
Penning ionization gauge (PIG), operating principles, 16
Periodic boundary conditions, molecular dynamics (MD) simulation, surface phenomena, 158–159
  limitations from, 164
Periodic heat flow, thermal diffusivity, laser flash technique, 383–384
Periodic table, ion beam analysis (IBA), 1177–1178
Permanent magnets
  applications, 496
  structures and properties, 497–499
Permittivity parameters
  microwave measurement techniques, 409
    cavity resonators, 409–410
    coaxial probes, 409
  surface x-ray diffraction, defined, 1028–1029
Perpendicular geometry, carrier lifetime measurement, free carrier absorption (FCA), 439–440
Phase-contrast imaging
  reflected-light optical microscopy, 679–680
  scanning transmission electron microscopy (STEM)
    coherence analysis, 1093–1097
    dynamical diffraction, 1101–1103
    research background, 1092
Phase diagrams, predictions
  aluminum-lithium analysis, 108–110
  aluminum-nickel analysis, 107
  basic principles, 91
  cluster approach, 91–92, 96–99
  cluster expansion free energy, 99–101
  electronic structure calculations, 101–102

Phase diagrams, predictions (*Continued*)
ground-state analysis, 102–104
mean-field approach, 92–96
nickel-platinum analysis, 107–108
nonconfigurational thermal effects, 106–107
research background, 90–91
static displacive interactions, 104–106
Phase equilibrium, basic principles, 91–92
Phase field method. *See also* Continuum field method (CFM)
Phase formation
metal alloy bonding, 135–138
bonding-antibonding effects, 136–137
charge transfer and electronegativities, 137–138
Friedel's *d*-band energetics, 135
size effects, 137
topologically close-packed phases, 135–136
transition metal crystal structures, 135
wave function character, 138
multiple-beam diffraction, 240
Phase retarders, x-ray magnetic circular dichroism (XMCD), 958–959
Phase shifts
low-energy electron diffraction (LEED), 1125
Mössbauer spectroscopy, 827–828
Raman spectroscopy of solids, 711–712
scanning transmission electron microscopy (STEM), research background, 1092
x-ray absorption fine structure (XAFS) spectroscopy, single scattering picture, 871–872
Phenomenological models, phonon analysis, 1324–1325
PHOENICS-CVD, chemical vapor deposition (CVD) model, hydrodynamics, 174
Phonons
magnetotransport in metal alloys, 562–563
Mössbauer spectroscopy, 824
neutron analysis
automation, 1323
basic principles, 1317–1318
data analysis, 1324–1325
instrumentation and applications, 1318–1323
triple-axis spectrometry, 1320–1323
limitations, 1326–1327
research background, 1316–1317
sample preparation, 1325–1326
specimen modification, 1326
photoluminescence (PL) spectroscopy, replicas, 682–683
spectral densities, metal surfaces, molecular dynamics (MD) simulation, 161–162
thermal diffuse scattering (TDS), 212–214
Photoabsorption, x-ray microfluorescence
basic principles, 942
cross-sections, 943
Photoconductivity (PC)
carrier lifetime measurement, 444–450
basic principles, 444
data analysis, 446–447
high-frequency range, automation, 445–446
limitations, 449–450
microwave PC decay, 447–449
optical techniques, 434
radio frequency PC decay, 449
research background, 428
sample preparation, 447
standard PC decay method, 444–445
steady-state methods, 435–438
cyclotron resonance (CR), 812
Photocurrent density-potential behavior, semiconductor-liquid junctions, 609
Photocurrent/photovoltage measurements, semiconductor materials, 605–613
charge transfer at equilibrium, 606–607
electrochemical cell design, 609–610
J-E equations, 608–609
concentration overpotentials, 611–612
series resistance overpotentials, 612
raw current-potential data, 611
sample preparation, 612–613
semiconductor-liquid interface, 610–611
current density-potential properties, 607–608
dark current-potential characteristics, 607
thermodynamics, 605–606
Photodiode array (PDA), Raman spectroscopy of solids, dispersed radiation measurment, 707–708
Photoelectrochemistry, semiconductor materials
electrochemical photocapacitance spectroscopy, 623–626
photocurrent/photovoltage measurements, 605–613
charge transfer at equilibrium, 606–607
electrochemical cell design, 609–610
J-E equations, 608–609
concentration overpotentials, 611–612
series resistance overpotentials, 612
raw current-potential data, 611
sample preparation, 612–613
semiconductor-liquid interface, 610–611
current density-potential properties, 607–608
dark current-potential characteristics, 607
junction thermodynamics, 605–606
research background, 605
semiconductor-liquid interface
band gap measurement, 613–614
differential capacitance measurements, 616–619
diffusion length, 614–616
flat-band potential measurements, 628–630
J-E data, kinetic analysis, 631–632
laser spot scanning, 626–628
photocurrent/photovoltage measurements, 610–611
current density-potential properties, 607–608
dark current-potential characteristics, 607
junction thermodynamics, 605–606
transient decay dynamics, 619–622
surface recombiation velocity, time-resolved microwave conductivity, 622–623
time-resolved photoluminescence spectroscopy, interfacial charge transfer kinetics, 630–631
Photoelectromagnetic (PEM) effect, carrier lifetime measurement, steady-state methods, 436
Photoelectron detectors, x-ray photoelectron spectroscopy (XPS), 982–983
Photoemission
angle-resolved x-ray photoelectron spectroscopy (ARXPS), 987–988
metal alloy magnetism, local exchange splitting, 189–190
ultraviolet photoelectron spectroscopy (UPS)
automation, 731–732
vs. inverse photoemission, 722–723
limitations, 733
process, 726–727
valence electron characterization, 723–724
x-ray photoelectron spectroscopy (XPS), 971–972
survey spectrum, 973–974
Photoluminescence (PL) spectroscopy
alloy broadening, 682
automation, 686
band-to-band recombination, 684
bound excitons, 682
carrier lifetime measurement, 450–453
automated procedures, 451–452
data analysis and interpretation, 452–453
deep level luminescence, 450–451
limitations, 453
near-band-gap emission, 450
photon recycling, 451
research background, 428
shallow impurity emission, 450
steady-state methods, 436
defect-level transitions, 683–684
donor-acceptor and free-to-bound transitions, 683
excitons and exciton-polaritons, 682
experimental protocols, 684–865
limitations, 686
low-temperature PL spectra
band interpretation, 686–687
crystal characterization, 685–686
phonon replicas, 682–683
research background, 681–682
room-temperature PL plating, 686
curve-fitting, 686
sample preparation, 686
specimen modification, 686
Photometric ellipsometry, defined, 735
Photomultiplier tubes (PMTs)
magnetic x-ray scattering, detector criteria, 927–928
photoluminescence (PL) spectroscopy, 684
Raman spectroscopy of solids, dispersed radiation measurment, 707–708
semiconductor-liquid interfaces, transient decay dynamics, 621–622
Photon energy
energy-dispersive spectrometry (EDS), 1136, 1137–1140
ultraviolet photoelectron spectroscopy (UPS), 728–729
light sources, 728–729
wavelength parameters, 730
Photon recycling, carrier lifetime measurement, photoluminescence (PL), 451
Physical quantities, carrier lifetime measurement, 433
Pick-up loops, superconductors, electrical transport measurements, signal-to-noise ratio, 483–484
Piezo-birefringent elements, polarization-modulation ellipsometer, 739
Piezoelectric behavior, electrochemical quartz crystal microbalance (EQCM), 653–654
Pirani gauge, applications and operation, 14
Placement effects, Hall effect, semiconductor materials, 417
Planar devices
magnetic x-ray scattering, beamline properties, 927
x-ray magnetic circular dichroism (XMCD), 958–959
Planck's constant
chemical vapor deposition (CVD) model, gas-phase chemistry, 169
deep level transient spectroscopy (DLTS), semiconductor materials, 421
neutron powder diffraction, probe configuration, 1286–1289
phonon analysis, 1318
radiation thermometer, 36–37
Plane wave approach
electronic structure analysis, 76
phase diagram prediction, 101–102
metal alloy bonding, precision calculations, 142–143
Plasma physics
chemical vapor deposition (CVD) model
basic components, 170

software tools, 174
ion beam analysis (IBA), ERD/RBS techniques, 1184–1186
Plasmon loss
Auger electron spectroscopy (AES), error detection, 1171
x-ray photoelectron spectroscopy (XPS), final-state effects, 974–978
Plasticity
fracture toughness testing
crack driving force *(G),* 305
load-displacement curves, 304
semibrittle materials, 311
nonuniform plastic deformation, 282
tension testing, 279–280
uniform plastic deformation, 282
Platinum alloys. *See also* Nickel-platinum alloys
cyclic voltammetry
experimental protocols, 588–590
working electrode, 585–586
electrodes, scanning electrochemical microscopy (SECM), 649
low-energy electron diffraction (LEED), qualitative analysis, 1127–1128
magnetocrystalline anisotropy energy (MAE), Co-Pt alloys, 199–200
scanning tunneling microscopy (STM)
tip configuration, 1113–1114
tip preparation, 1117
Platinum resistance thermometer, ITS-90 standard, 33
pn junctions
capacitance-voltage (C-V) characterization
basic principles, 457–458
data analysis, 462–463
electrochemical profiling, 462
instrument limitations, 463–464
mercury probe contacts, 461–462
profiling equipment, 460–461
protocols and procedures, 458–460
research background, 456–457
sample preparation, 463
trapping effects, 464–465
characterization
basic principles, 467–469
competitive and complementary techniques, 466–467
limitations of, 471
measurement equipment sources and selection criteria, 471
protocols and procedures, 469–470
research background, 466–467
sample preparation, 470–471
deep level transient spectroscopy (DLTS), semiconductor materials, 425–426
semiconductor-liquid interface, current density-potential properties, 607–608
Point groups
crystallography, 42–43
crystal systems, 44
group theoretical analysis, vibrational Raman spectroscopy, 703–704
matrix representations, 717–720
Point spread function (PSF), scanning electrochemical microscopy (SECM), feedback mode, 639–640
Poisson's equation
capacitance-voltage (C-V) characterization, 457–458
neutron powder diffraction, refinement algorithms, 1300–1301
Poisson's ratio
fracture-toughness testing, lnear elastic fracture mechanics (LEFM), 303
high-strain-rate testing
data interpretation and analysis, 296–298
inertia, 299
impulsive stimulated thermal scattering (ISTS), 753–755
Polarimeter devices, magnetic x-ray scattering, 928
Polarimetric spectroscopy. *See* Ellipsometry
Polarizability of materials, Raman spectroscopy of solids, electromagnetic radiation, 701
Polarization analysis
copper-nickel-zinc alloys, ordering wave polarization, 270–271
multiple-beam diffraction, mixing, 240
Raman spectroscopy of solids, 708
resonant magnetic x-ray scattering, 922–924
resonant scattering, 910–912
experimental deesign, 914
instrumentation, 913
ultraviolet photoelectron spectroscopy (UPS), automation, 731–732
x-ray absorption fine structure (XAFS) spectroscopy, 873
Polarization density matrix, multiple-beam diffraction, 241
Polarization-modulation ellipsometer, automation of, 739
Polarized beam technique, magnetic neutron scattering, 1330–1331
error detection, 1337–1338
Polarized-light microscopy, reflected-light optical microscopy, 677–680
Polarizers
ellipsometry, 738
relative phase/amplitude calculations, 739–740
magnetic x-ray scattering, beamline properties, 925–927
Polar Kerr effect, surface magneto-optic Kerr effect (SMOKE), 571–572
Polishing procedures, metallographic analysis
sample preparation, 66–67
4340 steel, cadmium plating composition and thickness, 68
Polychromatic illumination, semiconductor-liquid interface, 610
Polymeric materials
impulsive stimulated thermal scattering (ISTS) analysis, 746–749
Raman spectroscopy of solids, 712
scanning electrochemical microscopy (SECM), 644
transmission electron microscopy (TEM), specimen modification, 1088
x-ray photoelectron spectroscopy (XPS), 988–989
Polynomial expansion coefficients, coherent ordered precipitates, microstructural evolution, 123–124
"Pop-in" phenomenon, load-displacement curves, fracture toughness testing, 304
Porod approximation, small-angle scattering (SAS), 221–222
Porous materials, thermal diffusivity, laser flash technique, 390
Position-sensitive detector (PSD)
liquid surface x-ray diffraction, non-specular scattering, 1039
surface x-ray diffraction, 1010
crystallographic alignment, 1014–1015
diffractometer alignment, 1021
grazing-incidence measurement, 1016–1017
time-dependent neutron powder diffraction, 1299
Positive-intrinsic-negative (PIN) diode detector
photoluminescence (PL) spectroscopy, 684
x-ray powder diffraction, 838–839
Potential energy, fracture toughness testing, crack driving force *(G),* 305
Potentiometric measurements, electrochemical profiling, 579–580
Potentiostats
cyclic voltammetry, 584–585
electrochemical profiling, 579–580
scanning electrochemical microscopy (SECM), 642–643
Powder diffraction file (PDF), x-ray powder diffraction, 844
Power-compensation differential scanning calorimetry (DSC)
basic principles, 364–365
thermal analysis, 339
Power-field relations, resistive magnets, 502–503
Power law transitions, superconductors, electrical transport measurement, 474
Poynting vector
Raman spectroscopy of solids, electromagnetic radiation, 700–701
two-beam diffraction, dispersion surface, 231
P-polarized x-ray beam, liquid surface x-ray diffraction, 1047
Preamplifier design, ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1228–1230
Precision measurements, metal alloy bonding, 141–144
all-electron *vs.* pseudopotential methods, 141–142
basis sets, 142–143
Brillouin zone sampling, 143
first principles *vs.* tight binding, 141
full potentials, 143
self-consistency, 141
structural relaxation, 143–144
Pressure measurements
capacitance diaphragm manometers, 13
leak detection, vacuum systems, 21–22
mass spectrometers, 16–17
nuclear quadrupole resonance spectroscopy, 788
thermal conductivity gauges
applications, 13
ionization gauge
cold cathode type, 16
hot cathode type, 14–16
operating principles and procedures, 13–14
Pirani gauge, 14
thermocouple gauges, 14
total and partial gauges, 13
Primitive unit cells
lattices, 44–46
x-ray diffraction, structure-factor calculations, 209
Principal component analysis (PCA), x-ray photoelectron spectroscopy (XPS), 992–993
Probe configuration
contact problems, *pn* junction characterization, 471
neutron powder diffraction, 1286–1289
nuclear quadrupole resonance (NQR), frequency-swept Fourier-transform spectrometers, 780–781
scanning transmission electron microscopy (STEM), 1097–1098
transmission electron microscopy (TEM), lens resolution, 1079–1080
Probe lasers, carrier lifetime measurement, free carrier absorption (FCA), selection criteria, 440
Processed wafers, carrier lifetime measurement, metal/highly doped layer removal, free carrier absorption (FCA), 443
Processing operations, vacuum systems, 19
PROFIL program, neutron powder diffraction, 1306

Projectiles
heavy-ion backscattering spectrometry (HIBS), 1276–1277
impingement, materials characterization, 1
ion beam analysis (IBA), ERD/RBS techniques, 1184–1186
Property calculation, microstructural evolution modeling and, 128–129
Proton-induced gamma ray emission (PIGE)
automation, 1205–1206
background, interference, and sensitivity, 1205
cross-sections and *Q* values, 1205
data analysis, 1206–1207
energy relations, 1209
energy scanning, resonance depth profiling, 1204–1205
energy spread, 1209–1210
instrumentation criteria, 1203–1204
limitations, 1207–1208
nonresonant methods, 1202–1203
research background, 1200–1202
resonant depth profiling, 1203
specimen modification, 1207
standards, 1205
unwanted particle filtering, 1204
Proton scattering, ion-beam analysis (IBA), ERD/RBS examples, 1195–1197
Pseudo free induction decay, rotating frame NQR imaging, 786–788
Pseudopotential approach
electronic structure analysis, 75
metal alloy bonding
precision measurements, self-consistency, 141–142
semiconductor compounds, 138
phonon analysis, 1324–1325
Pulsed Fourier transform NMR
basic principles, 765
electron paramagnetic resonance (EPR), 796
Pulsed magnetic fields, laboratory settings, 505
Pulsed MOS capacitance technique
carrier lifetime measurement, 431, 435
Pulsed step measurements, superconductors, electrical transport measurements, current ramp rate, 479
Pulsed-type measurement techniques, carrier lifetime measurement, 436–438
Pulse height measurements, ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1225–1226
Pulse laser
impulsive stimulated thermal scattering (ISTS), 750–752
thermal diffusivity, laser flash technique, 385–386
Pulse magnets
cyclotron resonance (CR), laser magneto-spectroscopy (LMS), 811–812
structure and properties, 504–505
Pulse nclear magnetic resonance (NMR), magnetic field measurements, 510
Pump beam photons, carrier lifetime measurement, free carrier absorption (FCA), 439
Pumpdown procedures, vacuum systems, 19
Pumping speed
sputter-ion pump, 11
vacuum system principles, outgassing, 2–3
Pump lasers, carrier lifetime measurement, free carrier absorption (FCA), selection criteria, 440
Purity of materials, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 367–368
Quadrupolar moments
Mössbauer spectroscopy, electric quadrupole splitting, 822–823
nuclear quadrupole resonance (NQR), 776
resonant magnetic x-ray scattering, 922–924
Quadrupole mass filter, applications and operating principles, 17
Qualitative analysis
Auger electron spectroscopy (AES), 1161–1162
energy-dispersive spectrometry (EDS), 1141–1143
low-energy electron diffraction (LEED)
basic principles, 1122–1124
data analysis, 1127–1128
mass measurement process assurance, 28–29
ultraviolet/visible absorption (UV-VIS) spectroscopy, comptetitive and related techniques, 689–690
Quantitative analysis
Auger electron spectroscopy (AES), 1169
differential thermal analysis, defined, 363
energy-dispersive spectrometry (EDS), 1143
standardless analysis, 1147–1148
accuracy testing, 1150
applications, 1151
first-principles standardless analysis, 1148–1149
fitted-standards standardless analysis, 1149–1150
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, pulse height interpretation, 1225–1226
low-energy electron diffraction (LEED)
basic principles, 1124–1125
data analysis, 1128–1131
medium-energy backscattering, 1271–1272
neutron powder diffraction, 1296–1297
Raman spectroscopy of solids, 713
ultraviolet/visible absorption (UV-VIS) spectroscopy
comptetitive and related techniques, 689–690
interpretation, 693–694
procedures and protocols, 690–691
x-ray powder diffraction, phase analysis, 842, 844–845
Quantum mechanics
cyclotron resonance (CR), 808
low-energy electron diffraction (LEED), quantitative analysis, 1129–1131
resonant scattering analysis, 908–909
surface magneto-optic Kerr effect (SMOKE), 571
Quantum Monte Carlo (QMC), electronic structure analysis, 75
basic principles, 87–89
Quantum paramagnetiic response, principles and equations, 521–522
Quarter wave plates, magnetic x-ray scattering, beamline properties, 927
Quartz crystal microbalance (QCM)
basic principles, 24
electrochemical analysis
automation, 659–660
basic principles, 653–658
data analysis and interpretation, 660
equivalent circuit, 654–655
film and solution effects, 657–658
impedance analysis, 655–657
instrumentation criteria, 648–659
limitations, 661
quartz crystal properties, 659
research background, 653
sample preparation, 660–661
series and parallel frequency, 660
specimen modification, 661
Quasi-Fermi levels, carrier lifetime measurement, trapping, 431–432
Quasireversible reaction, cyclic voltammetry, 583
Quasi-steady-state-type measurement technique, carrier lifetime measurement, 436–437
Quench detection
electrical transport measurements, superconductors, 480–481
superconducting magnets, 501
Q values
heavy-ion backscattering spectrometry (HIBS), 1275–1276
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), cross-sections and, 1205
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), particle filtering, 1204

Radiation effects microscopy
basic principles, 1224–1228
instrumentation criteria, 1128–1230
ion-induced damage, 1232–1233
limitations, 1233
quantitative analysis, pulse height interpretations, 1225–1226
research background, 1223–1224
semiconductor materials, 1225
SEU microscopy, 1227–1228
static random-access memory (SRAM), 1230–1231
specimen modification, 1231–1232
topographical contrast, 1226–1227
Radiation sources
chemical vapor deposition (CVD) model
basic components, 172–173
software tools, 175
energy-dispersive spectrometry (EDS), stray radiation, 1155
ion-beam analysis (IBA), hazardous exposure, 1190
liquid surface x-ray diffraction, damage from, 1044–1045
Raman spectroscopy of solids, 704–705
x-ray microfluorescence, 942–943
Radiation thermometer, operating principles, 36–37
Radiative process, carrier lifetime measurement, 430
Radiofrequency (RF)
magnetic resonance imaging (MRI), 765–767
nuclear quadrupole resonance (NQR), research background, 775
PC-decay, carrier lifetime measurement
optical techniques, 434
principles and procedures, 449
Radiofrequency (RF) quadrupole analyzer, secondary ion mass spectrometry (SIMS) and, 1236–1237
Radioisotope sources, Mössbauer spectroscopy, 825–826
Radiolysis, transmission electron microscopy (TEM), polymer specimen modification, 1088
Radius of gyration, small-angle scattering (SAS), 221
Raman active vibrational modes
aluminum crystals, 709–710
solids, 720–721
Raman spectroscopy of solids
competitive and related techniques, 699
data analysis and interpretation, 709–713
active vibrational modes, aluminum compounds, 709–710
carbon structure, 712–713
crystal structures, 710–712
dispersed radiation measurement, 707–708
electromagnetic radiation, classical physics, 699–701

Fourier transform Raman spectroscopy, 708–709
group theoretical analysis, vibrational Raman spectroscopy, 702–704
character tables, 718
point groups and matrix representation, symmetry operations, 717–718
vibrational modes of solids, 720–722
vibrational selection rules, 716–720
light scattering, semiclassical physics, 701–702
limitations, 713–714
magnetic neutron scattering and, 1321
optical alignment, 706
optics, 705–706
polarization, 708
quantitative analysis, 713
radiation sources, 704–705
research background, 698–699
spectrometer components, 706–707
theoretical principles, 699–700
Random noise, superconductors, electrical transport measurements, signal-to-noise ratio, 485
Random-phase approximation (RPA), electronic structure analysis, 84–87
Rapid-scan FT spectrometry, cyclotron resonance (CR), 810
Rare earth ions, atomic/ionic magnetism, ground-state multiplets, 514–515
Ratio thermometer, operating principles, 37
Rayleigh scattering
impulsive stimulated thermal scattering (ISTS) analysis, 754–755
Raman spectroscopy of solids, optics properties, 705–706
scanning transmission electron microscopy (STEM)
incoherent scattering, 1099–1101
phase-contrast illumination, 1094–1097
transmission electron microscopy (TEM), aperture diffraction, 1078
R-curve behavior, fracture toughness testing, 302
crack driving force *(G),* 305
sample preparation, 311–312
Reaction onset temperatures, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 368–369
Reaction rate mode, scanning electrochemical microscopy (SECM), 640–641
Reaction reversibility. *See* Reversible reactions
Reactive gases
cryopump, 9–10
turbomolecular pumps, bearing failure and, 8
"Reactive sticking coefficient" concept, chemical vapor deposition (CVD) model, surface chemistry, 168
Readout interpretation
magnetic resonance imaging (MRI), readout gradient, 768
thermometry, 35
Rear-face temperature rise, thermal diffusivity, laser flash technique, 384–385
equations for, 391–392
error detection, 390
Rebound hardness testing, basic principles, 316
Reciprocal lattices
phonon analysis, basic principles, 1317–1318
surface x-ray scattering, mapping protocols, 1016
transmission electron microscopy (TEM)
deviation vector and parameter, 1067–1068
diffraction pattern indexing, 1073–1074
tilting specimens and electron beams, 1071
x-ray diffraction, crystal structure, 209
Recoil-free fraction, Mössbauer spectroscopy, 821
Recoiling, particle scattering
diagrams, 52–54
elastic recoiling, 51–52
inelastic recoiling, 52
nonrelativistic collisions, 62
Recombination mechanisms
carrier lifetime measurement, 429–431
limitations, 438
photoluminescence (PL), 452–453
surface recombination and diffusion, 432–433
photoluminescence (PL) spectroscopy, band-to-band recombination, 684
pn junction characterization, 468–469
Rectangular resonator, electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, 794–795
Redhead Extractor gauge, operating principles, 14–15
Redox potentials, semiconductor-liquid interface, current density-potential properties, 607–608
Reduced mass, particle scattering, kinematics, 54
Reference electrodes, semiconductor electrochemical cell design, photocurrent/photovoltage measurements, 610
Reference materials, tribological testing, 333
Refinement algorithms, neutron powder diffraction, 1300–1301
Reflected-light optical microscopy
basic principles, 667
illumination modes and image-enhancement techniques, 676–680
limitations, 680–681
procedures and protocols, 675–680
research background, 674–675
sample preparation, 680
Reflection high-energy electron diffraction (RHEED)
low-energy electron diffraction (LEED), comparisons, 1121
surface magneto-optic Kerr effect (SMOKE), 573–574
surface x-ray diffraction and, 1007–1008
ultraviolet photoelectron spectroscopy (UPS), sample preparation, 732–733
Reflection polarimetry. *See* Ellipsometry
Reflectivity
ellipsometry, 735–737
defined, 737
liquid surface x-ray diffraction, 1029–1036
Born approximation, 1033–1034
distorted-wave Born approximation, 1034–1036
error sources, 1044–1045
Fresnel reflectivity, 1029–1031
grazing incidence diffraction and rod scans, 1036
instrumentation, 1036–1038
multiple stepwise and continuous interfaces, 1031–1033
non-specular scattering, 1033–1036
simple liquids, 1040–1041
surface x-ray diffraction, 1015–1016
Refractive lenses, x-ray microprobes, 946
REITAN program, neutron powder diffraction, Rietveld refinements, 1306
Relative amplitude
ellipsometric measurement
angle of incidence errors, 742
basic principles, 735
optical properties, 740–741
polarizer/analyzer readings, 739–740
ellipsometry, reflecting surfaces, 736–737
Relative coordinates, particle scattering, kinematics, 54–55
Relative energy, particle scattering, kinematics, 54
Relative phase, ellipsometric measurement
angle of incidence errors, 742
basic principles, 735
optical properties, 740–741
polarizer/analyzer readings, 739–740
reflecting surfaces, 736–737
Relative sensitivity factors (RSFs), secondary ion mass spectrometry (SIMS) and, 1236–1237
Relativistic collisions, particle scattering, kinematics, 55–56
Relaxation measurements
Mössbauer spectroscopy, 824
non-contact techniques, 408
x-ray photoelectron spectroscopy (XPS), final-state effects, 975–976
Relaxation time approximation (RTA), semiconductor materials, Hall effect, 412
interpretation, 416
Reliability factors
low-energy electron diffraction (LEED), quantitative analysis, 1130–1131
neutron powder diffraction, 1305
Renninger peaks, multiple-beam diffraction, 236–237
Renormalized forward scattering (RFS), low-energy electron diffraction (LEED), 1135
Repeated-cycle deformation, tribological and wear testing, 324–325
Replication process, transmission electron microscopy (TEM), sample preparation, 1087
Residual gas analyzer (RGA), applications, 16
Resistive magnets, structure and properties, 502–504
hybrid magnets, 503–504
power-field relations, 502–503
Resistivity. *See* Conductivity measurements; Electrical resistivity
Resolution ellipsoid, phonon analysis, triple-axis spectrometry, 1322–1323
Resolution parameters
energy-dispersive spectrometry (EDS), 1137–1140
measurement protocols, 1156
optimization, 1140
medium-energy backscattering, 1267
optical microscopy, 669
scanning electron microscopy (SEM), 1053
surface x-ray diffraction, diffractometer components, 1010
transmission electron microscopy (TEM), lens defects and, 1079–1080
x-ray photoelectron spectroscopy (XPS), 993–994
Resonance methods. *See also* Nonresonant techniques
cyclotron resonance (CR)
basic principles, 806–808
cross-modulation, 812
data analysis and interpretation, 813–814
far-infrared (FIR) sources, 809
Fourier transform FIR magneto-spectroscopy, 809–810
laser far infrared (FIR) magneto-spectroscopy, 810–812
limitations, 814–815
optically-detected resonance (ODR) spectroscopy, 812–813
protocols and procedures, 808–813
quantum mechanics, 808
research background, 805–806
sample preparation, 814
semiclassical Drude model, 806–807
electron paramagnetic resonance (EPR)
automation, 798
basic principles, 762–763, 793–794
calibration, 797

Resonance methods. *See also* Nonresonant techniques (*Continued*)
continuous-wave experiments, X-band with rectangular resonator, 794–795
data analysis and interpretation, 798
electron-nuclear double resonance (ENDOR), 796
instrumentation criteria, 804
limitations, 800–802
microwave power, 796
modulation amplitude, 796–797
non-rectangular resonators, 796
non-X-band frequencies, 795–796
pulsed/Fourier transform EPR, 796
research background, 792–793
sample preparation, 798–799
sensitivity parameters, 797
specimen modification, 799–800
magnetic resonance imaging (MRI), theoretical background, 765–767
Mössbauer spectroscopy
basic properties, 761–762
bcc iron alloy solutes, 828–830
coherence and diffraction, 824–825
crystal defects and small particles, 830–831
data analysis and interpretation, 831–832
electric quadrupole splitting, 822–823
hyperfine interactions, 820–821
magnetic field splitting, 823–824
isomer shift, 821–822
Mössbauer effect, 818–820
nuclear excitation, 817–818
phase analysis, 827–828
phonons, 824
radioisotope sources, 825–826
recoil-free fraction, 821
relaxation phenomena, 824
research background, 816–817
sample preparation, 832
synchrotron sources, 826–827
valence and spin determination, 827
nuclear magnetic resonance imaging (NMRI)
applications, 767–771
back-projection imaging sequence, 768
basic principles, 762–763
chemical shift imaging sequence, 770
data analysis and interpretation, 771
diffusion imaging sequence, 769
echo-planar imaging sequence, 768–769
flow imaging sequence, 769–770
gradient recalled echo sequence, 768
instrumentation setup and tuning, 770–771
limitations, 772
sample preparation, 772
solid-state imaging, 770
spin-echo sequence, 768
theoretical background, 763–765
three-dimensional imaging, 770
nuclear quadrupole resonance (NQR)
data analysis and interpretation, 789
direct detection techniques, spectrometer criteria, 780–781
dynamics limitations, 790
first-order Zeeman perturbation and line shapes, 779
heterogeneous broadening, 790
indirect detection, field cycling methods, 781
nuclear couplings, 776–777
nuclear magnetic resonance (NMR), comparisons, 761–762
nuclear moments, 775–776
one-dimensional Fourier transform NQR, 781–782
research background, 775
sensitivity problems, 789
spatially resolved NQR, 785–789
field cycling methods, 788–789
magnetic field gradient method, 785–786
rotating frame NQR imaging, 786–788
temperature, stress, and pressure imaging, 788
spin relaxation, 779–780
spurious signals, 789–790
Sternheimer effect and electron deformation densities, 777
two-dimensional zero-field NQR, 782–785
exchange NQR spectroscopy, 785
level-crossing double resonance NQR nutation spectroscopy, 785
nutation NRS, 783–784
Zeeman-perturbed NRS (ZNRS), 782–783
zero field energy levels, 777–779
higher spin nuclei, 779
spin 1 levels, 777–778
spin 3/2 levels, 778
spin 5/2 levels, 778–779
research background, 761–762
Resonant depth profiling
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1203
data analysis, 1206–1207
energy calibration, 1204–1205
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), excitation functions, 1210
Resonant magnetic x-ray scattering
antiferromagnets, 930–932
ferromagnets, 932
anomalous dispersion effects, 925
research background, 918–919
theoretical concepts, 921–924
Resonant nuclear reaction analysis (RNRA), research background, 1178–1179
Resonant Raman scattering, diffuse scattering techniques, 889
inelastic scattering background removal, 890–893
Resonant scattering techniques
angular dependent tensors, 916
calculation protocols, 909–912
L=2 measurements, 910–911
L=4 measurements, 911–912
comparisons with other methods, 905–906
coordinate transformation, 916–917
data analysis and interpretation, 914–915
experiment design, 914
instrumentation criteria, 912–913
magnetic x-ray scattering
antiferromagnets, 930–932
ferromagnets, 932
research background, 918–919
theoretical concepts, 921–924
materials properties measurments, 905
polarization analysis, 913
research background, 905
sample preparation, 913–914
tensor structure factors, transformation, 916
x-ray diffraction principles
classical mechanics, 906–907
quantum mechanics, 908–909
theory, 906
Response function, particle-induced x-ray emission (PIXE), 1222–1223
Restraint factors, neutron powder diffraction, 1301–1302
Reverse-bias condition, deep level transient spectroscopy (DLTS), semiconductor materials, 421–422
Reverse currents, *pn* junction characterization, 470
Reverse recovery (RR) technique, carrier lifetime measurement, 435
Reversible reactions
chemical vapor deposition (CVD) model, gas-phase chemistry, 169
cyclic voltammetry, 582–584
non-reversible charge-transfer reactions, 583–584
quasireversible reaction, 583
total irreversible reaction, 583
total reversible reaction, 582–583
Rice-Rampsberger-Kassel-Marcus (RRKM) theory, chemical vapor deposition (CVD) model, gas-phase chemistry, 169
Rietveld refinements
neutron powder diffraction
axial divergence peak asymmetry, 1292–1293
data analysis, 1296
estimated standard deviations, 1305–1306
particle structure determination, 1297
quantitative phase analysis, 1296–1297
reliability factors, 1305
software tools, 1306–1307
single-crystal x-ray structure determination, 850–851
time-resolved x-ray powder diffraction, 845–847
x-ray powder diffraction
candidate atom position search, 840–841
estimated standard deviations, 841–842
final refinement, 841
quantitative phase analysis, 842
structural detail analysis, 847–848
Rising-temperature method, thermogravimetric (TG) analysis, kinetic theory, 353–354
RKKY exchange, ferromagnetism, 526–527
Rockwell hardness testing
automated methods, 319
basic principles, 317–318
data analysis and interpretation, 319–320
hardness values, 317–318, 323
limitations and errors, 322
procedures and protocols, 318–319
research background, 316–317
sample preparation, 320
specimen modification, 320–321
ROD program, surface x-ray diffraction
crystal truncation rod (CTR) profiles, 1015
data analysis, 1025
silicon surface example, 1017–1018
Rod scans. *See* Crystal truncation rods (CTR)
Rolling friction coefficient, tribological and wear testing, 326
Room-temperature photoluminescence (PL) mapping
applications, 686
curve fitting procedures, 687
Root-mean-square techniques, diffuse scattering techniques, inelastic scattering background removal, 892–893
1ro parameter, microstructural evolution
coherent ordered precipitates, 124–126
field kinetic equations, 121–122
Rotating-analyzer ellipsometer, automation of, 739
Rotating frame NQR imaging, spatially resolved nuclear quadruople resonance, 786–788
Rotating-sample magnetometer (RSM), principles and applications, 535
Rotating seal design, surface x-ray diffraction, 1013
Rotational/translation motion feedthroughs, vacuum systems, 18–19
Rotation axes
single-crystal x-ray structure determination, crystal symmetry, 854–856
surface x-ray diffraction, five-circle diffractometer, 1013–1014

symmetry operators
improper rotation axis, 39–40
proper rotation axis, 39
Roughing pumps
oil-free (dry) pumps, 4–5
oil-sealed pumps, 3–4
technological principles, 3
RUMP program, ion-beam analysis (IBA), ERD/RBS equations, 1189
Rutherford backscattering spectroscopy (RBS). *See also* Medium-energy backscattering
elastic scattering, 1178
heavy-ion backscattering spectrometry (HIBS) and, 1275–1276
high-energy ion beam analysis (IBA), 1176–1177
ion-beam composition analysis
applications, 1191–1197
basic concepts, 1181–1184
detector criteria and detection geometries, 1185–1186
equations, 1186–1189, 1199–1200
experimental protocols, 1184–1186
limitations, 1189–1191
research background, 1179–1181
low-energy electron diffraction (LEED), comparisons, 1121
medium-energy backscattering, 1261–1262
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1201–1202
Rutherford scattering
fluorescence and diffraction analysis, 940
particle scattering, central-field theory, cross-sections, 60–61
scanning transmission electron microscopy (STEM)
dynamical diffraction, 1101–1103
research background, 1091–1092

Safety protocols
heavy-ion backscattering spectrometry (HIBS), 1280
medium-energy backscattering, 1269
nuclear magnetic resonance, 771
Sample-and-hold technique, deep level transient spectroscopy (DLTS), semiconductor materials, 424–425
Sample manipulator components, surface x-ray diffraction, 1012–1013
Sample preparation
Auger electron spectroscopy (AES), 1170
bulk electronic measurements, 404
capacitance-voltage (C-V) characterization, 463
carrier lifetime measurement
free carrier absorption (FCA), 442–443
cross-sectioning, depth profiling, 443
photoconductivity (PC) techniques, 447
combustion calorimetry, 381
cyclic voltammetry, 590
cyclotron resonance (CR), 814
deep level transient spectroscopy (DLTS), semiconductor materials, 425–426
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 371
diffuse scattering techniques, 899
electrochemical quartz crystal microbalance (EQCM), 660–661
electron paramagnetic resonance (EPR), 798–799
ellipsometric measurements, 741–742
energy-dispersive spectrometry (EDS), 1152
fracture toughness testing, 311–312
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 394
hardness testing, 320
heavy-ion backscattering spectrometry (HIBS), 1281
high-strain-rate testing, 298
impulsive stimulated thermal scattering (ISTS) analysis, 757
low-energy electron diffraction (LEED)
cleaning protocols, 1125–1126
protocols, 1131–1132
magnetic domain structure measurements
bitter pattern imaging, 545–547
holography, 554–555
Lorentz transmission electron microscopy, 552
magnetic force microscopy (MFM), 550
magneto-optic imaging, 547–549
scanning electron microscopy with polarization analysis (SEMPA), 553
spin polarized low-energy electron microscopy (SPLEEM), 557
magnetic neutron scattering, 1336–1337
magnetic x-ray scattering, 935
magnetometry, 536–537
contamination problems, 538
magnetotransport in metal alloys, 566–567
mechanical testing, 287
metallographic analysis
basic principles, 63–64
cadmium plating composition and thickness, 4340 steel, 68
etching procedures, 67
mechanical abrasion, 66
microstructural evaluation
7075-T6 anodized aluminum, 68–69
deformed high-purity aluminum, 69
4340 steel sample, 67–68
mounting protocols and procedures, 66
polishing procedures, 66–67
sectioning protocols and procedures, 65–66
strategic planning, 64–65
micro-particle-induced x-ray emission (micro-PIXE) analysis, 1219
Mössbauer spectroscopy, 832
nuclear magnetic resonance, 772
particle-induced x-ray emission (PIXE) analysis, 1218–1219
phonon analysis, 1325–1326
photoluminescence (PL) spectroscopy, 687
pn junction characterization, 470–471
reflected-light optical microscopy, 680
resonant scattering, 913–914
scanning electron microscopy (SEM), 1058–1059
vacuum systems, 1061
scanning transmission electron microscopy (STEM), 1108
scanning tunneling microscopy (STM), 1117
semiconductor materials
Hall effect, 416–417
photoelectrochemistry, 612–613
single-crystal neutron diffraction, 1314
single-crystal x-ray structure determination, 862–863
superconductors, electrical transport measurements
contact issues, 483
geometric options, 482
handling and soldering/making contacts, 478
holder specifications, 478, 489
quality issues, 486
sample shape, 477
support issues, 482–483
surface magneto-optic Kerr effect (SMOKE) evolution, 575
surface x-ray diffraction, crystallographic alignment, 1014–1015
thermal diffusivity, laser flash technique, 389
thermogravimetric (TG) analysis, 350–352, 356–357
thermomagnetic analysis, 541–544
trace element accelerator mass spectrometry (TEAMS), 1252–1253
transmission electron microscopy (TEM), 1086–1087
dispersion, 1087
electropolishing and chemical thinning, 1086
ion milling and focused gallium ion beam thinning, 1086–1087
replication, 1087
ultramicrotomy, 1087
tribological and wear testing, 335
ultraviolet photoelectron spectroscopy (UPS), 732–733
ultraviolet/visible absorption (UV-VIS) spectroscopy, 693, 695
x-ray absorption fine structure (XAFS) spectroscopy, 879–880
x-ray magnetic circular dichroism (XMCD)
general techniques, 964–965
magnetization, 962–963
x-ray microfluorescence/microdiffraction, 949–950
x-ray photoelectron spectroscopy (XPS), 998–999
composition analysis, 994–996
sample charging, 1000–1001
x-ray powder diffraction, 842
Saturation magnetization, basic principles, 495
Sauerbrey relationship, electrochemical quartz crystal microbalance (EQCM), piezoelectric behavior, 654
Sawing techniques, metallographic analysis, sample preparation, 65
Sayre equation, single-crystal x-ray structure determination, direct method computation, 865–868
Scaling laws, magnetic phase transition theory, 530
Scanning acoustic microscopy (SAM), transmission electron microscopy (TEM) and, 1064
Scanning categories, surface x-ray diffraction, 1027
Scanning electrochemical microscopy (SECM)
biological and polymeric materials, 644
competitive and related techniques, 637–638
constant-current imaging, 643
corrosion science applications, 643–644
feedback mode, 638–640
generation/collection mode, 641–642
instrumentation criteria, 642–643
limitations, 646–648
localized material deposition and dissolution, 645–646
reaction rate mode, 640–641
research background, 636–638
specimen modification, 644–646
tip preparation protocols, 648–649
Scanning electron microscopy (SEM)
Auger electron spectroscopy (AES)
depth-profile analysis, 1166–1167
speciment alignment, 1161
automation, 1057–1058
constrast and detectors, 1054–1056
data analysis and interpretation, 1058
electron gun, 1061
energy-dispersive spectrometry (EDS), automation, 1141
fracture toughness testing, crack extension measurement, 308
image formation, 1052–1053
imaging system components, 1061, 1063
instrumentation criteria, 1053–1054

Scanning electron microscopy (SEM) (*Continued*)
limitations and errors, 1059–1060
magnetic domain structure measurements, type I and II protocols, 550–551
metallographic analysis, 7075-T6 anodized aluminum alloy, 69
research background, 1050
resolution, 153
sample preparation, 1058–1059
scanning transmission electron microscopy (STEM) *vs.,* 1092–1093
selection criteria, 1061–1063
signal generation, 1050–1052
specimen modification, 1059
techniques and innovations, 1056–1057
vacuum system and specimen handling, 1061
x-ray photoelectron spectroscopy (XPS), sample charging, 1000–1001
Scanning electron microscopy with polarization analysis (SEMPA), magnetic domain structure measurements, 552–553
Scanning monochromators, x-ray magnetic circular dichroism (XMCD), 960–962
Scanning probe microscopy (SPM), scanning tunneling microscopy (STM) *vs.,* 1112
Scanning profilometer techniques, wear testing, 332
Scanning transmission electron microscopy (STEM)
automation, 1080
transmission electron microscopy (TEM), comparison, 1063–1064, 1090–1092
x-ray absorption fine structure (XAFS) spectroscopy and, 875–877
Z-contrast imaging
atomic resolution spectroscopy, 1103–1104
coherent phase-contrast imaging and, 1093–1097
competitive and related techniques, 1092–1093
data analysis and interpretation, 1105–1108
object function retrieval, 1105–1106
strain contrast, 1106–1108
dynamical diffraction, 1101–1103
incoherent scattering, 1098–1101
weakly scattered objects, 1111
limitations, 1108
manufacturing sources, 1111
probe formation, 1097–1098
protocols and procedures, 1104–1105
research background, 1090–1093
sample preparation, 1108
specimen modification, 1108
Scanning transmission ion microscopy (STIM), particle-induced x-ray emission (PIXE) and, 1211
Scanning tunneling microscopy (STM)
automation, 1115–1116
basic principles, 1113–1114
complementary and competitive techniques, 1112–1113
data analysis and interpretation, 1116–1117
image acquisition, 1114
limitations, 1117–1118
liquid surfaces and monomolecular layers, 1028
low-energy electron diffraction (LEED), 1121
material selection and limitations, 1114–1115
research background, 1111–1113
sample preparation, 1117
scanning electrochemical microscopy (SECM) and, 637
scanning transmission electron microscopy (STEM) *vs.,* 1093
surface x-ray diffraction and, 1007–1008
Scattered radiation. *See also* Raman spectroscopy of solids
thermal diffusivity, laser flash technique, 390
Scattering analysis
cyclotron resonance (CR), 805–806
low-energy electron diffraction (LEED), 1135
Scattering length density (SLD)
liquid surface x-ray diffraction
data analysis and interpretation, 1039–1043
Langmuir monolayers, 1041–1043
liquid alkane crystallization, 1043
liquid metals, 1043
simple liquids, 1040–1041
error sources, 1044–1045
non-specular scattering, 1033–1036
reflectivity measurements, 1028–1033
surface x-ray diffraction, defined, 1028–1029
Scattering power and length
tribological and wear testing, 326–327
x-ray diffraction, 210
Scattering theory
defined, 207
kinematic principles, 207–208
Scherrer equation, neutron powder diffraction, microstrain broadening, 1294–1295
Scherrer optimum aperture, scanning transmission electron microscopy (STEM), probe configuration, 1097–1098
Schönflies notation, symmetry operators, improper rotation axis, 39–40
Schottky barrier diodes
capacitance-voltage (C-V) characterization
basic principles, 457–458
data analysis, 462–463
electrochemical profiling, 462
instrument limitations, 463–464
mercury probe contacts, 461–462
profiling equipment, 460–461
protocols and procedures, 458–460
research background, 456–457
sample preparation, 463
trapping effects, 464–465
deep level transient spectroscopy (DLTS), semiconductor materials, 425–426
Schrödinger equation
computational analysis, applications, 71
cyclotron resonance (CR), 808
electronic structure, 74–75
dielectric screening, 84
phase diagram prediction, 101–102
Mössbauer spectroscopy, isomer shift, 821–822
transition metal magnetic ground state, itinerant magnetism at zero temperature, 182–183
x-ray diffraction, scattering power and length, 210
Scintillation counters, single-crystal x-ray structure determination, 859–860
Scintillation detectors, single-crystal x-ray structure determination, 859–860
Scratch testing, basic principles, 316–317
Screening function, particle scattering, central-field theory, deflection function approximation, 59–60
Screw axes
single-crystal x-ray structure determination, crystal symmetry, 854–856
symmetry operators, 40–42
Screw compressor, application and operation, 5
Screw-driven mechanical testing system, fracture toughness testing, load-displacement curve measurement, 307–308
Scroll pumps, application and operation, 5
Search coils, magnetic field measurements, 508
Secondary electrons (SEs), scanning electron microscopy (SEM)
contamination issues, 1059–1060
contrast images, 1055–1056
data analysis, 1058
signal generation, 1051–1052
Secondary ion accelerator mass spectrometry (SIAMS), trace element accelerator mass spectrometry (TEAMS) and, 1235–1237
Secondary ion generation, trace element accelerator mass spectrometry (TEAMS)
acceleration and electron-stripping system, 1238–1239
ultraclean ion sourcees, 1238
Secondary ion mass spectroscopy (SIMS)
Auger electron spectroscopy (AES) *vs.,* 1158–1159
composition analysis, nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1201–1202
heavy-ion backscattering spectrometry (HIBS) and, 1275
ion beam analysis (IBA) and, 1175–1176, 1181
medium-energy backscattering, 1261
pn junction characterization, 467
scanning tunneling microscopy (STM) *vs.,* 1113
trace element accelerator mass spectrometry (TEAMS) and, 1235–1237
bulk impurity measurements, 1248–1249
ultraclean ion source design, 1241
Secondary x-ray fluorescence, energy-dispersive spectrometry (EDS), matrix corrections, 1145
Second-harmonic effects
magnetic domain structure measurements, magneto-optic imaging, 548–549
surface magneto-optic Kerr effect (SMOKE), 569–570
Second Law of thermodynamics
thermal analysis and principles of, 342–343
thermodynamic temperature scale, 31–32
Second-order Born approximation, multiple-beam diffraction, 238–240
Sectioning procedures, metallographic analysis
cadmium plating composition and thickness, 4340 steel, 68
deformed high-purity aluminum, 69
microstructural evaluation
4340 steel, 67
7075-T6 anodized aluminum alloy, 68
sample preparation, 65–66
Selected-area diffraction (SAD), transmission electron microscopy (TEM)
basic principles, 1071–1073
complementary bright-field and dark-field techniques, 1082–1084
data analysis, 1081
diffraction pattern indexing, 1073–1074
Selection rules, ultraviolet photoelectron spectroscopy (UPS), photoemission process, 726–727
Self-consistency, metal alloy bonding, precision measurements, 141
Self-diffusion, binary/multicomponent diffusion, 149–150
Self-field effects, superconductors, electrical transport measurements, signal-to-noise ratio, 486
Self-interaction correction, electronic structure analysis, 75
Semibrittle materials, fracture toughness testing, plasticity, 311
Semiclassical physics
cyclotron resonance (CR), Drude model, 806–807
Raman spectroscopy of solids, light scattering, 701–702
Semiconductor-based thermometer, operating principles, 36
Semiconductor materials

capacitance-voltage (C-V) characterization
basic principles, 457–458
data analysis, 462–463
electrochemical profiling, 462
instrument limitations, 463–464
mercury probe contacts, 461–462
profiling equipment, 460–461
protocols and procedures, 458–460
research background, 456–457
sample preparation, 463
trapping effects, 464–465
carrier lifetime measurement
characterization techniques, 433–435
device related techniques, 435
diffusuion-length-based methods, 434–435
optical techniques, 434
free carrier absorption, 438–444
automated methods, 441
basic principles, 438–440
carrier decay transient, 441
computer interfacing, 441
data analysis and interpretation, 441–442
depth profiling, sample cross-sectioning, 443
detection electronics, 440–441
geometrical considerations, 441
lifetime analysis, 441–442
lifetime depth profiling, 441
lifetime mapping, 441–442
limitations, 443–444
probe laser selection, 440
processed wafers, metal and highly doped layer removal, 443
pump laser selection, 440
sample preparation, 442–443
virgin wafers, surface passivation, 442–443
generation lifetime, 431
photoconductivity, 444–450
basic principles, 444
data analysis, 446–447
high-frequency range, automation, 445–446
limitations, 449–450
microwave PC decay, 447–449
radio frequency PC decay, 449
sample preparation, 447
standard PC decay method, 444–445
photoluminescence, 450–453
automated procedures, 451–452
data analysis and interpretation, 452–453
deep level luminescence, 450–451
limitations, 453
near-band-gap emission, 450
photon recycling, 451
shallow impurity emission, 450
physical quantities, 433
recombination mechanisms, 429–431
selection criteria, characterization methods, 453–454
steady-state, modulated, and transient methods, 435–438
data interpretation problems, 437
limitations, 437–438
modulation-type method, 436
pulsed-type methods, 437–438
quasi-steady-state-type method, 436–437
surface recombination and diffusion, 432–433
theoretical background, 401, 427–429
trapping techniques, 431–432
conductivity measurements, research background, 401–403
deep level transient spectroscopy (DLTS), 418–419
basic principles
emission rate, 420–421
junction capacitance transient, 421–423
data analysis and interpretation, 425
limitations, 426
procedures and automation, 423–425
research background, 418–419
sample preparation, 425–426
semiconductor defects, 418–419
electronic measurement, 401
energy-dispersive spectrometry (EDS), basic principles, 1136–1140
Hall effect
automated testing, 414
basic principles, 411–412
data analysis and interpretation, 414–416
equations, 412–413
limitations, 417
protocols and procedures, 412–414
research background, 411
sample preparation, 416–417
sensitivity, 414
heavy-ion backscattering spectrometry (HIBS), 1279–1280
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1225, 1229–1230
photoelectrochemistry
electrochemical photocapacitance spectroscopy, 623–626
photocurrent/photovoltage measurements, 605–613
charge transfer at equilibrium, 606–607
electrochemical cell design, 609–610
J-E equations, 608–609
concentration overpotentials, 611–612
series resistance overpotentials, 612
raw current-potential data, 611
sample preparation, 612–613
semiconductor-liquid interface, 610–611
current density-potential properties, 607–608
dark current-potential characteristics, 607
junction thermodynamics, 605–606
research background, 605
semiconductor-liquid interface
band gap measurement, 613–614
differential capacitance measurements, 616–619
diffusion length, 614–616
flat-band potential measurements, 628–630
J-E data, kinetic analysis, 631–632
laser spot scanning, 626–628
photocurrent/photovoltage measurements, 610–611
current density-potential properties, 607–608
dark current-potential characteristics, 607
junction thermodynamics, 605–606
transient decay dynamics, 619–622
surface recombiation velocity, time-resolved microwave conductivity, 622–623
time-resolved photoluminescence spectroscopy, interfacial charge transfer kinetics, 630–631
Raman spectroscopy of solids, 712
scanning tunneling microscopy (STM) analysis, 1114–1115
ultraviolet photoelectron spectroscopy (UPS), surface analysis, 727–728
x-ray photoelectron spectroscopy (XPS), 988–989
Sensing elements, thermometry, 34–35
Sensitive-tint plate, reflected-light optical microscopy, 678–680
Sensitivity measurements
Auger electron spectroscopy (AES), 1163
quantitative analysis, 1169
carrier lifetime measurement, microwave photoconductivity (PC) techniques, 448–449
electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, 797
heavy-ion backscattering spectrometry (HIBS), 1276, 1278–1279
background and, 1281
ion beam analysis (IBA), periodic table, 1177–1178
medium-energy backscattering, 1266–1267
nuclear quadrupole resonance (NQR), 789
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 1205
semiconductor materials, Hall effect, 414
ultraviolet photoelectron spectroscopy (UPS), 730–731
Series resistance
capacitance-voltage (C-V) characterization, 464
electrochemical quartz crystal microbalance (EQCM), 660
overpotentials, semiconductor materials, J-E behavior corrections, 612
SEXI protocol, rotating frame NQR imaging, 787–788
Shadow conees, particle scattering, central-field theory, 58
Shake-up/shake-off process, x-ray photoelectron spectroscopy (XPS), final-state effects, 976
Shallow impurity emission, carrier lifetime measurement, photoluminescence (PL), 450
Shape factor analysis, transmission electron microscopy (TEM), 1065–1066
data analysis, 1080–1081
deviation vector and parameter, 1067–1068
Sharp crack, fracture toughness testing, stress field parameters, 314
Shearing techniques, metallographic analysis, sample preparation, 65
SHELX direct method procedure, single-crystal x-ray structure determination, 862
computational techniques, 867–868
Shirley background, x-ray photoelectron spectroscopy (XPS), 990–991
Shockley-Reed-Hall (SRH) mechanism, carrier lifetime measurement, 429–430
free carrier absorption (FCA), 442
photoluminescence (PL), 452–453
Short-range ordering (SRO). *See also* Atomic short range ordering (ASRO)
diffuse scattering techniques, 886–889
absolute calibration, measured intensities, 894
x-ray scattering measurements, 889–890
metal alloy magnetism, local moment fluctuation, 187–188
neutron magnetic diffuse scattering, 904–905
x-ray diffraction
local atomic arrangement, 214–217
local atomic correlation, 214–217
Signal generation
ion beam analysis (IBA), 1175–1176
scanning electron microscopy (SEM), 1050–1052
Signal-to-noise ratio parameters
cyclic voltammetry, 591
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1233
nuclear magnetic resonance, 772
nuclear quadrupole resonance (NQR), data analysis and interpretation, 789
superconductors, electrical transport measurements, 483–486
current supply noise, 486
grounding, 485
integration period, 486

Signal-to-noise ratio parameters (*Continued*)
pick-up loops, 483–484
random noise and signal spikes, 485–486
thermal electromotive forces, 484–485
x-ray photoelectron spectroscopy (XPS), 993–994
Silicon Hall sensors, magnetic field measurements, 507
Silicon surfaces
low-energy electron diffraction (LEED), qualitative analysis, 1127–1128
particle-induced x-ray emission (PIXE), detector criteria, 1222–1223
surface x-ray diffraction, 1017–1018
SI magnetic units, general principles, 492–493
Simple liquids, liquid surface x-ray diffraction, 1040–1041
Simulation techniques, tribological and wear testing, 325
Simultaneous techniques
gas analysis
automated testing, 399
benefits and limitations, 393–394
chemical and physical methods, 398–399
commercial TG-DTA equipment, 394–395
evolved gas analysis and chromatography, 396–397
gas and volatile product collection, 398
infrared spectrometery, 397–398
limitations, 399
mass spectroscopy for thermal degradation and, 395–396
research background, 392–393
TG-DTA principles, 393
thermal analysis, 339
Simultaneous thermogravimetry (TG)-differential thermal analysis/differential scanning calorimetry (TG-DTA/DSC), gas analysis, 394
Simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA)
commercial sources, 394–395
gas analysis, 393
limitations, 399
Sine integral function
scanning transmission electron microscopy (STEM), incoherent scattering, 1100
x-ray absorption fine structure (XAFS) spectroscopy, single scattering picture, 871–872
Single-beam spectrometer, ultraviolet/visible absorption (UV-VIS) spectroscopy, 692
Single-crystal neutron diffraction
applications, 1311–1312
data analysis and interpretation, 1313–1314
neutron sources and instrumentation, 1312–1313
research background, 1307–1309
sample preparation, 1314–1315
theoretical background, 1309–1311
x-ray diffraction *vs.*, 836
Single-crystal x-ray diffraction
basic properties, 836
single-crystal neutron diffraction and, 1307–1309
Single-crystal x-ray structure determination
automation, 860
competitive and related techniques, 850–851
derived results interpretation, 861–862
initial model structure, 860–861
limitations, 863–864
nonhydrogen atom example, 862
protocols and procedures, 858–860
data collection, 859–860
research background, 850–851
sample preparation, 862–863
specimen modification, 863
x-ray crystallography principles, 851–858
crystal structure refinement, 856–858
crystal symmetry, 854–856
Single-cycle deformation, tribological and wear testing, 324–325
Single-domain structures, low-energy electron diffraction (LEED), qualitative analysis, 1123–1124
Single-edge notched bend (SENB) specimens, fracture toughness testing
crack extension measurement, 308
J-integral approach, 311
sample preparation, 311–312
stress intensity and J-integral calculations, 314–315
unstable fractures, 308–309
Single-event upset (SEU) imaging
basic principles, 1224–1228
instrumentation criteria, 1128–1230
ion-induced damage, 1232–1233
limitations, 1233
microscopic instrumentation, 1227–1228
static random-access memory (SRAM), 1230–1231
quantitative analysis, pulse height interpretations, 1225–1226
research background, 1223–1224
semiconductor materials, 1225
specimen modification, 1231–1232
static random-access memory (SRAM) devices, 1230–1231
topographical contrast, 1226–1227
Single-pan balance, classification, 27
Single-phase materials, Mössbauer spectroscopy, recoil-free fraction, 821
Single-scattering representation, x-ray absorption fine structure (XAFS) spectroscopy, 870–872
Size effects
Auger electron spectroscopy (AES), sample preparation, 1170
diffuse intensities, metal alloys, hybridization in NiPt alloys, charge correlation effects, 268
metal alloy bonding, 137
semiconductor compounds, 138
semiconductor materials, Hall effect, 416–417
Slab analysis, ion-beam analysis (IBA), ERD/RBS equations, 1186–1189
Slater determinants
electronic structure analysis
Hartree-Fock (HF) theory, 77
local density approximation (LDA), 77–78
metal alloy bonding, precision calculations, 142–143
Slater-Pauling curves
band theory of magnetism, 516
metal alloy magnetism, electronic structure, 184–185
Sliding friction coefficient, tribological and wear testing, 326
Sliding models, wear testing protocols, 331
Slip line field theory, static indentation hardness testing, 317
Slow-scan FT spectroscopy, cyclotron resonance (CR), 810
Slow variables, continuum field method (CFM), 119
Small-angle neutron scattering (SANS)
magnetic neutron scattering and, 1320
neutron powder diffraction and, 1288–1289
Small-angle scattering (SAS)
local atomic correlation, short-range ordering, 217
x-ray diffraction, 219–222
cylinders, 220–221
ellipsoids, 220
Guinier approximation, 221
integrated intensity, 222
interparticle interference, 222
K=0 extrapolation, 221
porod approximation, 221–222
size distribution, 222
spheres, 220
two-phase model, 220
Small-angle x-ray scattering (SAXS)
diffuse scattering techniques, absolute calibration, measured intensities, 894
low-energy electron diffraction (LEED), comparisons, 1120–1121
Small-particle magnetism, Mössbauer spectroscopy, 830–831
Small-spot imaging
scanning tunneling microscopy (STM), 1113
x-ray photoelectron spectroscopy (XPS), 982
Smoothing routines, x-ray photoelectron spectroscopy (XPS), 994
Snell's law
dynamical diffraction, boundary conditions, 229
ellipsometry, reflecting surfaces, 736–737
liquid surface x-ray diffraction, reflectivity measurements, 1029–1031
two-beam diffraction, dispersion surface, 230–231
Software tools
chemical vapor deposition (CVD) models
free molecular transport, 174
gas-phase chemistry, 173
hydrodynamics, 174
kinetic theory, 175
limitations of, 175–176
plasma physics, 174
radiation, 175
surface chemistry, 173
diffuse scattering techniques, 898–899
electrical transport measurements, superconductors, 480–481
electron paramagnetic resonance (EPR), 798
Mössbauer spectroscopy, 831–832
neutron powder diffraction
axial divergence peak asymmetry, 1293
indexing procedures, 1298
Rietveld refinements, 1306–1307
particle-induced x-ray emission (PIXE) analysis, 1217–1218
phonon analysis, triple-axis spectrometry, 1320–1323
scanning electrochemical microscopy (SECM), 642–643
single-crystal x-ray structure determination, 858–860
superconductors, electrical transport measurement, data analysis, 482
surface x-ray diffraction
crystal truncation rod (CTR) profiles, 1015
data analysis, 1024–1025
lineshape analysis, 1019
Solenoids, electromagnet structure and properties, 499–500
Solids analysis
Raman spectroscopy
competitive and related techniques, 699
data analysis and interpretation, 709–713
active vibrational modes, aluminum compounds, 709–710
carbon structure, 712–713
crystal structures, 710–712
dispersed radiation measurement, 707–708
electromagnetic radiation, classical physics, 699–701
Fourier transform Raman spectroscopy, 708–709

group theoretical analysis, vibrational Raman spectroscopy, 702–704
character tables, 718
point groups and matrix representation, symmetry operations, 717–718
vibrational modes of solids, 720–722
vibrational selection rules, 716–720
light scattering, semiclassical physics, 701–702
limitations, 713–714
optical alignment, 706
optics, 705–706
polarization, 708
quantitative analysis, 713
radiation sources, 704–705
research background, 698–699
spectrometer components, 706–707
theoretical principles, 699–700
ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 727
x-ray photoelectron spectroscopy (XPS), research background, 970–972
Solid-solid reactions, thermogravimetric (TG) analysis, 356
Solid-solution alloys, magnetism, 183–184
Solid-state detectors, fluorescence analysis, 944
Solid-state imaging
nuclear magnetic resonance, 770
single-crystal x-ray structure determination, 851
Solution resistance
corrosion quantification, Tafel technique, 596
electrochemical quartz crystal microbalance (EQCM), 657–658
Soret effect, chemical vapor deposition (CVD) model, hydrodynamics, 171
Sorption pump, application and operation, 5
Source-to-specimen distance, x-ray diffraction, 206
Space groups
crystallography, 46–50
single-crystal x-ray structure determination, crystal symmetry, 855–856
x-ray powder diffraction, crystal lattice determination, 840
Spatially resolved nuclear quadruople resonance
basic principles, 785–789
field cycling methods, 788–789
magnetic field gradient method, 785–786
rotating frame NQR imiaging, 786–788
temperature, stress, and pressure imaging, 788
Specimen geometry
fracture toughness testing
load-displacement curves, 304
stress intensity factor *(K),* 306
stress-strain analysis, 285
thermogravimetric (TG) analysis, 351–352
transmission electron microscopy (TEM), deviation parameter, 1077–1078
Specimen modification
Auger electron spectroscopy (AES)
alignment protocols, 1161
basic principles, 1170–1171
cyclic voltammetry, 590
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 371–372
electrochemical quartz crystal microbalance (EQCM), 661
electron paramagnetic resonance (EPR), 799–800
energy-dispersive spectrometry (EDS), 1152–1153
position protocols, 1156
fracture toughness testing, alignment, 312–313
high-strain-rate testing, 298–299
impulsive stimulated thermal scattering (ISTS) analysis, 757
ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1231–1232
liquid surface x-ray diffraction, 1043–1045
low-energy electron diffraction (LEED), 1132
magnetic domain structure measurements
Lorentz transmission electron microscopy, 552
magneto-optic imaging, 548–549
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1207
particle-induced x-ray emission (PIXE)
error detection, 1219–1220
protocols, 1212–1216
phonon analysis, 1326
photoluminescence (PL) spectroscopy, 687
scanning electrochemical microscopy (SECM), 644–645
scanning electron microscopy (SEM), 1059
scanning transmission electron microscopy (STEM), 1108
single-crystal x-ray structure determination, 863
superconductors, electrical transport measurement, 483
thermal diffusivity, laser flash technique, 389
trace element accelerator mass spectrometry (TEAMS), 1253
transmission electron microscopy (TEM)
basic principles, 1087–1088
thickness modification, deviation parameter, 1081–1082
tribological testing, 332–333
ultraviolet/visible absorption (UV-VIS) spectroscopy, 695–696
x-ray absorption fine structure (XAFS) spectroscopy, 880
x-ray photoelectron spectroscopy (XPS), 999
Specimen-to-detector distance, x-ray diffraction, 206
Spectral resolution
Auger electron spectroscopy (AES), 1161
nuclear quadrupole resonance (NQR), data analysis and interpretation, 789
Spectroelectrochemical quartz crystal microbalance (SEQCM), instrumentation criteria, 659
Spectrometers/monochromators
diffuse scattering techniques, 894–987
electron spectrometers, ultraviolet photoelectron spectroscopy (UPS), 729
alignment protocols, 729–730
automation, 731–732
liquid surface x-ray diffraction, instrumetation criteria, 1036–1038
magnetic x-ray scattering, 925, 927–928
medium-energy backscattering, efficiency protocols, 1265–1266
nuclear quadrupole resonance (NQR), direct detection techniques, 780–781
phonon analysis, triple-axis spectrometry, 1320–1323
photoluminescence (PL) spectroscopy, 684
Raman spectroscopy of solids, criteria for, 706–707
surface x-ray diffraction, beamline alignment, 1019–1020
ultraviolet/visible absorption (UV-VIS) spectroscopy, components, 692
x-ray absorption fine structure (XAFS) spectroscopy
detection methods, 875–877
glitches, 877–878
x-ray magnetic circular dichroism (XMCD), 959–962
glitches, 966
Spectrum channel width and number, energy-dispersive spectrometry (EDS), measurement protocols, 1156
Spectrum distortion, energy-dispersive spectrometry (EDS), 1140
Spectrum subtraction, Auger electron spectroscopy (AES), 1167–1168
SPECT software
diffuse scattering techniques, 898–899
x-ray microfluorescence/microdiffraction, 949
Specular reflectivity
grazing-incidence diffraction (GID), 241–242
distorted-wave Born approximation (DWBA), 244–245
liquid surface x-ray diffraction, measurement protocols, 1028–1033
surface x-ray diffraction, crystal truncation rod (CTR) profiles, 1015–1016
Spheres, small-angle scattering (SAS), 220
Spherical aberration, transmission electron microscopy (TEM), lens defects and resolution, 1078
Spherical sector analysis, Auger electron spectroscopy (AES), 1160–1161
"SPICE" models, chemical vapor deposition (CVD), plasma physics, 170
Spin density functional theory (SDFT)
metal alloy magnetism
atomic short range order (ASRO), 190–191
competitive and related techniques, 185–186
first-principles calculations, 188–189
limitations, 200–201
transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Spin determination, Mössbauer spectroscopy, 827
Spin-echo sequence, magnetic resonance imaging (MRI)
basic principles, 768
theoretical background, 767
Spin 1 energy level, nuclear quadrupole resonance (NQR), 777–778
Spin 3/2 energy level, nuclear quadrupole resonance (NQR), 778
Spin 5/2 energy level, nuclear quadrupole resonance (NQR), 778
Spin glass materials, cluster magnetism, 516–517
Spin-lattice relaxation, nuclear magnetic resonance, 764–765
Spin moments
nuclear quadrupole resonance (NQR), 776
x-ray magnetic circular dichroism (XMCD), 953–955
Spin-orbit interaction
metal alloy bonding, accuracy calculations, 140–141
surface magneto-optic Kerr effect (SMOKE), 571
x-ray photoelectron spectroscopy (XPS), 973–974
Spin packets, nuclear magnetic resonance, 764–765
Spin-polarization effects, metal alloy bonding, accuracy calculations, 140–141
Spin polarized low-energy electron microscopy (SPLEEM), magnetic domain structure measurements, 556–557
Spin-polarized photoemission (SPPE) measurements, x-ray magnetic circular dichroism (XMCD) comparison, 955
Spin properties, nuclear magnetic resonance theory, 763–765
Spin relaxation, nuclear quadrupole resonance (NQR), 779–780
Spin-spin coupling, nuclear magnetic resonance, 764–765

Split Hopkinson pressure bar technique, high-strain-rate testing, 290
automation procedures, 296
data analysis and interpretation, 297–298
limitations, 299–300
sample preparation, 298
specimen modification, 298–299
stress-state equilibrium, 294–295
temperature effects, 295–296
Split-Pearson VII function, neutron powder diffraction, axial divergence peak asymmetry, 1293
Spontaneous magnetization
collective magnetism, 515
ferromagnetism, 523–524
Spot profile analysis LEED (SPA-LEED)
instrumentation criteria, 1126–1127
surface x-ray diffraction and, 1008
Spot profile analysis RHEED (SPA-RHEED), surface x-ray diffraction and, 1008
Spurious signaling
nuclear quadrupole resonance (NQR), 789–790
phonon analysis, 1326–1327
Sputter-initiated resonance ionization spectrometry (SIRIS)
secondary ion mass spectrometry (SIMS) and, 1236–1237
trace element accelerator mass spectrometry (TEAMS) and, 1235, 1237
Sputter-ion pump
applications, 10
operating principles and procedures, 10–12
Sputter profiling
angle-resolved x-ray photoelectron spectroscopy (ARXPS), 986–988
Auger electron spectroscopy (AES), depth-profile analysis, 1165–1167
ion beam analysis (IBA), ERD/RBS techniques, 1184–1186
Stability/reactivity, thermogravimetric (TG) analysis, 355
Stable crack mechanics, fracture toughness testing, 309–311
Stacking faults, neutron powder diffraction, 1295–1296
Stainless steel
vacuum system construction, 17
weight standards, 26–27
Staircase step measurements, superconductors, electrical transport measurements, current ramp rate, 479
Standardless analysis, energy-dispersive spectrometry (EDS), 1147–1148
accuracy testing, 1150
applications, 1151
first-principles standardless analysis, 1148–1149
fitted-standards standardless analysis, 1149–1150
Standard reference material
Auger electron spectroscopy (AES), sources, 1169–1170, 1174
energy-dispersive spectrometry (EDS), matrix corrections, 1146–1147
particle-induced x-ray emission (PIXE) analysis, 1218
Standards
mass measurement process assurance, 28–29
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), 10205
in thermal analysis, 340
weight, 26–27
Standard state corrections, combustion calorimetry, 379–380
Standing waves
multiple-beam diffraction, 240
two-beam diffraction, 235–236
Startup procedures, turbomolecular pumps, 8
Static displacive interactions
diffuse scattering techniques, recovered displacements, 897–898
phase diagram prediction, 104–106
Static indentation hardness testing
applications and nomenclature, 318–319
automated methods, 319
basic principles, 317–318
data analysis and interpretation, 319–320
limitations and errors, 322
procedures and protocols, 318–319
research background, 316–317
sample preparation, 320
specimen modification, 320–321
Static magnetic field, nuclear magnetic resonance, instrumentation and tuning, 770–771
Static random-access memory (SRAM) devices, ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1230–1231
Statistics-insensitive nonlinear peak-clipping (SNIP) algorithm, particle-induced x-ray emission (PIXE) analysis, 1217–1218
"Steady-state" carrier profile, carrier lifetime measurement, surface recombination and diffusion, 432–433
Steady-state techniques, carrier lifetime measurement, 435–438
data interpretation issues, 437
limitations of, 437–438
quasi-steady-state-type method, 436–437
Steel samples, metallographic analysis, 4340 steel
cadmium plating composition and thickness, 68
microstructural evaluation, 67–68
Sternheimer effect, nuclear quadrupole resonance (NQR), 777
Stoichiometry, combustion calorimetry, 378
Stokes-Poincaré parameters, multiple-beam diffraction, polarization density matrix, 241
Stokes Raman scattering, Raman spectroscopy of solids, semiclassical physics, 701–702
Stoner enhancement factor, band theory of magnetism, 516
Stoner theory
band theory of magnetism, 515–516
Landau magnetic phase transition, 529–530
Stoner-Wohlfarth theory, transition metal magnetic ground state, itinerant magnetism at zero temperature, 181–183
Stopping power, energy-dispersive spectrometry (EDS), standardless analysis, 1148–1149
Strain contrast imaging, scanning transmission electron microscopy (STEM), data interpretation, 1106–1108
Strain distribution, microbeam analysis
ferroelectric sample, 948–949
tensile loading, 948
Strain-gauge measurements, high-strain-rate testing, 300
Strain rate
fracture toughness testing, crack driving force *(G),* 305
stress-strain analysis, 283
Stress field, fracture toughness testing, sharp crack, 314
Stress intensity factor *(K),* fracture toughness testing
basic principles, 305–306
defined, 302
SENB and CT speciments, 314–315
stable crack mechanics, 309–311
two-dimensional representation, 303
unstable fractures, 309
Stress measurements
impulsive stimulated thermal scattering (ISTS), 745–746
nuclear quadrupole resonance spectroscopy, 788
Stress-state equilibrium, high-strain-rate testing, 294–295
Stress-strain analysis
data analysis and interpretation, 287–288
elastic deformation, 281
fracture toughness testing, crack driving force *(G),* 304–305
high-strain-rate testing, Hopkinson bar technique, 290–292
mechanical testing, 280–282
curve form variables, 282–283
definitions, 281–282
material and microstructure, 283
temperature and strain rate, 283
yield-point phenomena, 283–284
Stretched exponentials, carrier lifetime measurement, 432
Structural phase transformation, thermomagnetic analysis, 542–544
Structural relaxation, metal alloy bonding, precision calculations, 143–144
Structure-factor calculations, x-ray diffraction
crystal structure, 208–209
principles and examples, 209–210
Structure factor relationships
neutron powder diffraction
individual extraction, 1298
solution techniques, 1298
resonant scattering, 909–910
tensor structure factors, 916
single-crystal neutron diffraction and, 1309–1311
single-crystal x-ray structure determination, 852–853
direct method computation, 866–868
surface x-ray diffraction, 1010–1011
error detection, 1018–1019
transmission electron microscopy (TEM) analysis, 1065–1066
x-ray powder diffraction, 845
detail analysis, 847–848
Structure inversion method (SIM), phase diagram predictions, cluster variational method (CVM), 99
Sublimation pumps
applications, 12
operating principles, 12
surface x-ray diffraction, 1011–1012
Substitutional and interstitial metallic systems, binary/multicomponent diffusion, 152–155
B2 intermetallics, chemical order, 154–155
frame of reference and concentration variables, 152
interdiffusion, 155
magnetic order, 153–154
mobilities and diffusivities, 152–153
temperature and concentration dependence of mobilities, 153
Substrate generation/tip collection (SG/TC), scanning electrochemical microscopy (SECM), 641–642
Subtraction technique, magnetic neutron scattering, 1330
error detection, 1337–1338
Sum peaks, energy-dispersive spectrometry (EDS), 1139
qualitative analysis, 1142–1143
Sum rules
diffuse intensities, metal alloys, atomic short-range ordering (ASRO) principles, 257

x-ray magnetic circular dichroism (XMCD), 956–957
data analysis, 964
error detection, 966
Super-Borrmann effect, multiple-beam standing waves, 240
Super command format, surface x-ray diffraction, 1025–1026
Superconducting magnets
changing fields stability and losses, 501–502
protection, 501
quench and training, 501
structure and properties, 500–502
Superconducting quantum interference device (SQUID)
magnetometry
automation of, 537
components, 532
principles and applications, 534
nuclear quadrupole resonance (NQR)
detector criteria, 781
nuclear moments, 776
thermomagnetic analysis, 540–544
Superconductors
diamagnetism, 494
electrical transport measurements
automation of, 480–481
bath temperature fluctuations, 486–487
competitive and related techniques, 472–473
contact materials, 474
cooling options and procedures, 478–479
critical current criteria, 474–475
current-carrying area for critical current to current density, 476
current contact length, 477
current ramp rate and shape, 479
current supply, 476–477
current transfer and transfer length, 475–477
data analysis and interpretation, 481–482
electromagnetic phenomena, 479
essential theory
current sharing, 474
four-point measurement, 473–474
Ohm's law, 474
power law transitions, 474
generic protocol, 480
instrumentation and data acquisition, 479–480
lead shortening, 486
magnetic field strength extrapolation and irreversibility field, 475
maximum measurement current determination, 479
probe thermal contraction, 486
research background, 472–473
sample handling/damage, 486
sample heating and continuous-current measurements, 486
sample holding and soldering/making contacts, 478
sample preparation, 482–483
sample quality, 487
sample shape, 477
self-field effects, 487
signal-to-noise ratio parameters, 483–486
current supply noise, 486
grounding, 485
integration period, 486
pick-up loops, 483–484
random noise and signal spikes, 485–486
thermal electromotive forces, 484–485
specimen modification, 483
thermal cycling, 487
troubleshooting, 480
voltage tap placement, 477
voltmeter properties, 477
zero voltage definition, 474
magnetization, 517–519
permanent magnets in, 496
Superlattice structures
grazing-incidence diffraction (GID), 242–243
magnetic neutron scattering, diffraction, 1334–1335
surface magneto-optic Kerr effect (SMOKE), 573–574
Superparamagnetism, principles of, 522
Surface analysis
Auger electron spectroscopy (AES), 1158–1159
automation, 406
basic principles, 405–406
limitations, 406
nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE), film thickness, 1206
protocols and procedures, 406
Surface barrier detectors (SBDs)
heavy-ion backscattering spectrometry (HIBS), 1276
ion-beam analysis (IBA) experiments, 1181–1184
ERD/RBS techniques, 1185–1186
medium-energy backscattering, backscattering techniques, 1262–1265
Surface capacitance measurements, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 624–625
Surface degradation, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 626
Surface extended x-ray absorption fine structure (EXAFS) analysis, low-energy electron diffraction (LEED), 1121
Surface magnetic scattering, protocols and procedures, 932–934
Surface magnetometry, principles and applications, 535–536
Surface magneto-optic Kerr effect (SMOKE)
automation, 573
classical theory, 570–571
data analysis and interpretation, 573–574
limitations, 575
medium-boundary and medium-propagation matrices, 576–577
multilayer formalism, 571
phenomenologica origins, 570
protocols and procedures, 571–573
quantum theory, ferromagnetism, 571
research background, 569–570
sample preparation, 575
ultrathin limit, 577
Surface photovoltage (SPV) technique, carrier lifetime measurement, diffusion-length-based methods, 435
Surface preparation
carrier lifetime measurement
passivation, free carrier absorption (FCA), 442–443
recombination and diffusion, 432–433
chemical vapor deposition (CVD) model
basic components, 167–168
software tools, 173
ellipsometry
polishing protocols, 741–742
reflecting surfaces, 735–737
molecular dynamics (MD) simulation
data analysis and interpretation, 160–164
higher-temperature dynamics, 161–162
interlayer relaxation, 161
layer density and temperature variation, 164
mean-square atomic displacement, 162–163
metal surface phonons, 161–162
room temperature structure and dynamics, 160–161
thermal expansion, 163–164
limitations, 164
principles and practices, 158–159
related theoretical techniques, 158
research background, 156
surface behavior, 156–157
temperature effects on surface behavior, 157
pn junction characterization, 468
surface/interface x-ray diffraction, 217–219
crystal truncation rods, 219
two-dimensional diffraction rods, 218–219
ultraviolet photoelectron spectroscopy (UPS), 727
sample preparation, 732–733
Surface recombination capture coefficient, electrochemical photocapacitance spectroscopy (EPS), semiconductor materials, 625
Surface recombination velocity
carrier lifetime measurements, time-resolved microwave conductivity, 622–623
semiconductor-liquid interfaces, transient decay dynamics, 620–622
Surface reconstruction, molecular dynamics (MD) simulation, 156–157
Surface relaxation, molecular dynamics (MD) simulation, 156–157
Surface roughness
angle-resolved x-ray photoelectron spectroscopy (ARXPS), 988
liquid surface x-ray diffraction, reflectivity measurements, 1033
surface x-ray scattering, 1016
temperature effects, molecular dynamics (MD) simulation, 157
Surface x-ray diffraction
angle calculations, 1021
basic principles, 1008–1011
crystallographic measurements, 1010–1011
grazing incidence, 1011
measurement instrumentation, 1009–1010
surface properties, 1008–1009
beamline alignment, 1019–1020
competitive and related strategies, 1007–1008
data analysis and interpretation, 1015–1018
crystal truncation rod (CTR) profiles, 1015
diffuse scattering, 1016
grazing incidence measurements, 1016–1017
reciprocal lattice mapping, 1016
reflectivity, 1015–1016
silicon surface analysis, 1017–1018
software programs, 1024–1025
diffractometer alignment, 1020–1021
instrumentation criteria, 1011–1015
crystallographic alignment, 1014–1015
five-circle diffractometer, 1013–1014
laser alignment, 1014
sample manipulator, 1012–1013
vacuum system, 1011–1012
limitations, 1018–1019
liquid surfaces
basic principles, 1028–1036
competitive and related techniques, 1028
data analysis and interpretation, 1039–1043
Langmuir monolayers, 1041–1043
liquid alkane crystallization, 1043
liquid metals, 1043
simple liquids, 1040–1041
non-specular scattering
GID, diffuse scattering, and rod scans, 1038–1039
reflectivity measurements, 1033

Surface x-ray diffraction (*Continued*)
p-polarized x-ray beam configuration, 1047
reflectivity, 1029–1036
Born approximation, 1033–1034
distorted-wave Born approximation, 1034–1036
Fresnel reflectivity, 1029–1031
grazing incidence diffraction and rod scans, 1036
instrumentation, 1036–1038
multiple stepwise and continuous interfaces, 1031–1033
non-specular scattering, 1033
research background, 1027–1028
specimen modification, 1043–1045
research background, 1007–1008
scan categories, 1027
super command format, 1025–1026
ultrahigh-vacuum (UHV) systems
bakeout procedure, 1023–1024
basic principles, 1011–1012
load-lock procedure, 1023
protocols and procedures, 1022–1024
venting procedures, 1023
Survey spectrum
Auger electron spectroscopy (AES), 1163–1164
x-ray photoelectron spectroscopy (XPS), 972–974
analyzer criteria, 981–982
post-processing of, 1000–1001
SX approximation, electronic structure analysis, 85–86
Symbols, in thermal analysis, 339–340
Symmetry analysis
crystallography, 39–42
improper rotation axes, 39–40
proper rotation axes, 39
screw axes and glide planes, 40–42
in crystallography, 39
resonant scattering, 906
resonant scattering analysis, 909–910
single-crystal x-ray structure determination, crystal symmetry, 854–856
surface x-ray diffraction, error detection, 1019
transmission electron microscopy (TEM), diffraction pattern indexing, 1073–1074
vibrational Raman spectroscopy
group theoretical analysis, 718–720
matrix representations, 717–720
Synchrotron radiation
diffuse scattering techniques, 882–884
magnetic x-ray scattering
beamline properties, 925–927
hardware criteria, 925–926
Mössbauer spectroscopy, radiation sources, 826–827
surface x-ray diffraction, diffractometer components, 1009–1010
ultraviolet photoelectron spectroscopy (UPS), 728–729
commercial sources, 734
x-ray absorption fine structure (XAFS) spectroscopy, 877
x-ray photoelectron spectroscopy (XPS), 980
x-ray powder diffraction, 839

Tafel technique
corrosion quantification, 593–596
limitations, 596
linear polarization, 596–599
protocols and procedures, 594–596
cyclic voltammetry, quasireversible reaction, 583
Tapered capillary devices, x-ray microprobes, 945
Taylor expansion series
coherent ordered precipitates, microstructural evolution, 123
small-angle scattering (SAS), 221
thermal diffuse scattering (TDS), 211–214
Taylor rod impact test, high-strain-rate testing, 290
Technical University Munich Secondary Ion AMS Facility, trace element accelerator mass spectrometry (TEAMS) research at, 1245
Temperature cross-sensitivity, Hall effect sensors, 508
Temperature-dependent Hall (TDH) measurements, semiconductor materials, 414–416
Temperature factors, neutron powder diffraction, 1301
Temperature gradients, thermogravimetric (TG) analysis, 350–352
Temperature measurement
binary/multicomponent diffusion, substitutional and interstitial metallic systems, 153
combustion calorimetry, 378–379
control of, 37–38
corrosion quantification, electrochemical impedance spectroscopy (EIS), 600–603
defined, 30–31
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 367–368
fixed points, ITS-90 standard, 34
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 394
high-strain-rate testing, 295–296
international temperature scale (ITS), development of, 32–34
measurement resources, 38
metal alloy paramagnetism, finite temperatures, 186–187
nuclear quadrupole resonance spectroscopy, 788
stress-strain analysis, 283
extreme temperatures and controlled environments, 285–286
surface phenomena, molecular dynamics (MD) simulation
high-temperature structury and dynamics, 161–164
room temperature structure and dynamics, 160–161
variations caused by, 157
thermal diffusivity, laser flash technique, 387–389
thermogravimetric (TG) analysis
error detection, 358–359
instrumentation and apparatus, 349–350
Tensile-strength measurement, hardness testing, 320
Tensile test
basic principles, 284–285
high-strain-rate testing, Hopkinson bar technique, 289–290
microbeam analysis, strain distribution, 948
Tension testing
basic principles, 279–280
basic tensile test, 284–285
elastic properties, 279
environmental testing, 286
extreme temperatures and controlled environments, 285–286
high-temperature testing, 286
low-temperature testing, 286
plastic properties, 279–280
specimen geometry, 285
stress/strain analysis, 280–282
curve form variables, 282–283
definitions, 281–282
material and microstructure, 283
temperature and strain rate, 283
yield-point phenomena, 283–284
testing machine characteristics, 285
Tensor elements, magnetotransport in metal alloys, transport equations, 560–561
Tensor structure factors, resonant scattering, transformation, 916
Test machines
automated design, 286–287
mechanical testing, 285
Thermagnetometry/differential thermomagnetometry/DTA, temperature measurement errors and, 358–359
Thermal analysis
data interpretation and reporting, 340–341
definitions and protocols, 337–339
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 395
kinetic theory, 343
literature sources, 343–344
nomenclature, 337–339
research background, 337
scanning tunneling microscopy (STM) analysis, thermal drift, 1115
standardization, 340
symbols, 339–340
thermodynamic theory, 341–343
Thermal conductivity
gauges, pressure measurements
applications, 13
ionization gauge
cold cathode type, 16
hot cathode type, 14–16
operating principles and procedures, 13–14
Pirani gauge, 14
thermocouple gauges, 14
magnetotransport in metal alloys
basic principles, 560
magnetic field behavior, 563–565
research applications, 559
zero field behavior, 561–563
Thermal cycling, superconductors, electrical transport measurements, signal-to-noise ratio, 486
Thermal degradation
evolved gas analysis (EGA), mass spectrometry for, 395–396
gas analysis, simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA), 393
Thermal diffuse scattering (TDS)
diffuse scattering techniques, data interpretation comparisons, 895–897
neutron diffuse scattering, comparisons, 885
surface x-ray diffraction, crystallographic alignment, 1014–1015
x-ray diffraction, 210–214
Thermal diffusivity, laser flash technique
automation of, 387
basic principles, 384–385
data analysis and interpretation, 387–389
limitations, 390
protocols and procedures, 385–387
research background, 383–384
sample preparation, 389
specimen modification, 389–390
Thermal effects, phase diagram prediction, nonconfigurational thermal effects, 106–107
Thermal electromotive forces, superconductors, electrical transport measurements, signal-to-noise ratio, 484–485
Thermal expansion
metal alloy magnetism, negative effect, 195–196

surface phenomena, molecular dynamics (MD) simulation, 163–164
Thermal neutron scattering, time-dependent neutron powder diffraction, 1299–1300
Thermal vibrations, scanning transmission electron microscopy (STEM), incoherent scattering, 1100
Thermal volatilization analysis (TVA), gas analysis and, 398
Thermistor, operating principles, 36
Thermoacoustimetry, defined, 339
Thermobalance apparatus, thermogravimetric (TG) analysis, 347–350
Thermocouple feedthroughs, vacuum systems, 18
Thermocouple gauge, applications and operation, 14
Thermocouple thermometer, operating principles, 36
Thermocoupling apparatus
  differential thermal analysis (DTA), 364–365
  thermogravimetric (TG) analysis, 349–350
Thermodilatatometry, defined, 339
Thermodynamics
  combustion calorimetry and principles of, 374–375
  magnetic phase transition theory, 528–529
  semiconductor photoelectrochemistry, semiconductor-liquid interface, 605–606
  thermal analysis and principles of, 341–343
Thermodynamic temperature scale, principles of, 31–32
Thermoelectrometry, defined, 339
Thermogravimetry (TG). *See also* Simultaneous thermogravimetry (TG)-differential thermal analysis (TG-DTA)
  apparatus, 347–350
  applications, 352–356
    computational analysis, 354–355
    gas-solid reactions, 355–356
    kinetic studies, 352–354
    solid-solid reactions, 356
    thermal stability/reactivity, 355
  automated procedures, 356
  basic principles, 346–347
  data analysis and interpretation, 356
  defined, 338, 344–345
  documentation protocols, 361–362
  experimental variables, 350–352
  limitations, 357–359
    mass measurement errors, 357–358
    temperature measurement errors, 358–359
  research background, 344–346
  sample preparation, 356–357
  simultaneous techniques for gas analysis, research background, 392–393
Thermomagnetic analysis
  automation, 541–542
  basic principles, 540–541
  data analysis and interpretation, 543–544
  research background, 540
  structural phase transformations, 542–543
Thermomagnetic errors, Hall effect, semiconductor materials, 417
Thermomagnetometry, defined, 339
Thermomechanical analysis (TMA), defined, 339
Thermometry
  combustion calorimetry, 376–377
  definitions, 30–31
  electrical-resistance thermometers, 36
  fixed temperature points, 34
  gas-filled thermometers, 35–36
  helium vapor pressure, 33
  international temperature scales, 32–34
    IPTS-27 scale, 32–33
    ITS-90 development, 33–34
    pre-1990s revisions, 33
  interpolatinc gas thermometer, 33
  liquid-filled thermometers, 35
  optical pyrometry, 34
  platinum resistance thermometer, 33–34
  radiation thermometers, 36–37
  readout interpretation, 35
  semiconductor-based thermometers, 36
  sensing elements, 34–35
  temperature controle, 37–38
  temperature-measurement resources, 38
  thermistors, 36
  thermocouple thermometers, 36
  thermodynamic temperature scale, 31–32
Thermoparticulate analysis, defined, 338
Thermopower *(S),* magnetotransport in metal alloys
  basic principles, 560
  magnetic field behavior, 563–565
  research applications, 559
  transport equations, 561
  zero field behavior, 561–563
Thermoptometry, defined, 339
Thermosonimetry, defined, 339
Thickness parameters
  medium-energy backscattering, 1261
  scanning transmission electron microscopy (STEM)
    dynamical diffraction, 1102–1103
    instrumentation criteria, 1105
  transmission electron microscopy (TEM), deviation parameters, 1081–1082
  x-ray absorption fine structure (XAFS) spectroscopy, 879
  x-ray magnetic circular dichroism (XMCD), 965–966
Thin-detector techniques, nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), particle filtering, 1204
Thin-film structures
  conductivity measurements, 402
  Hall effect, semiconductor materials, depletion effects, 417
  impulsive stimulated thermal scattering (ISTS), 744–746
    applications, 749–752
    automation, 753
    competitive and related techniques, 744–746
    data analysis and interpretation, 753–757
    limitations, 757–758
    procedures and protocols, 746–749
    research background, 744–746
    sample preparation and specimen modification, 757
  ion-beam analysis (IBA), ERD/RBS techniques, 1191–1197
  medium-energy backscattering
    applications, 1268–1269
    backscattering techniques, 1261–1265
  particle-induced x-ray emission (PIXE), 1213–1216
  ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 725
  x-ray photoelectron spectroscopy (XPS)
    reference spectra, 996–998
    research background, 970–972
Third Law of thermodynamics, thermodynamic temperature scale, 31–32
Thomson coefficient, magnetotransport in metal alloys, basic principles, 560
Thomson scattering
  magnetic x-ray scattering, classical theory, 920
  resonant scattering analysis, 906–907
    quantum mechanics, 908–909
  x-ray diffraction, 210
Three-beam interactions, dynamic diffraction, 240
Three-dimensional imaging
  magnetic resonance, 770
  scanning electron microscopy (SEM), 1056–1057
  ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 725
Three-dimensional lattices
  crystal systems, 45–46
  low-energy electron diffraction (LEED), 1122
Three-dimensional standing wave, multiple-beam diffraction, 240
Three-electrode electrochemical cells, cyclic voltammetry, potentiostats, 584–585
Three-step photoemission model, ultraviolet photoelectron spectroscopy (UPS), 726–727
Three-wave analysis, high-strain-rate testing
  data analysis and interpretation, 296–298
  Hopkinson bar technique, 291–292
Tight binding (TB) method
  electronic structure analysis, phase diagram prediction, 101–102
  metal alloy bonding, precision measurements *vs.* first-principles calculations, 141
Tilted illumination and diffraction, transmission electron microscopy (TEM), axial dark-field imaging, 1071
Tilting specimens, transmission electron microscopy (TEM), diffraction techniques, 1071
Time-dependent Ginzburg-Landau (TDGL) equation
  continuum field method (CFM), 114
    diffusion-controlled grain growth, 127–128
    numerical algorithms, 129–130
  microscopic field model (MFM), 116
  microstructural evolution, field kinetics, 122
Time-dependent junction capacitance, deep level transient spectroscopy (DLTS), semiconductor materials, 422–423
Time-dependent neutron powder diffraction, applications and protocols, 1298–1300
Time-dependent wear testing, procedures and protocols, 330
Time-of-flight diffractomer, neutron powder diffraction, 1290–1291
  Bragg reflection positions, 1291–1292
Time-of-flight spectrometry (TOS)
  heavy-ion backscattering spectrometry (HIBS), 1278
  medium-energy backscattering, 1261
    backscattering techniques, 1262–1265
    error detection, 1271–1272
    protocols and procedures, 1265
  nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1204
  secondary ion mass spectrometry (SIMS) and, 1236–1237
Time-resolved microwave conductivity, semiconductor-liquid interfaces, surface recombination velocity, 622–623
Time-resolved photoluminescence (TRPL)
  semiconductor interfacial charge transfer kinetics, 630–631
  semiconductor-liquid interfaces, transient decay dynamics, 620–622
Time-resolved x-ray powder diffraction, protocols, 845–847
Time-to-amplitude converter (TAC)
  carrier lifetime measurement, photoluminescence (PL), 451–452
  heavy-ion backscattering spectrometry (HIBS), 1279
  medium-energy backscattering, 1268
Time-to-digital converter (TDC), medium-energy backscattering, 1268

Tip position modulation (TPM)
  scanning electrochemical microscopy (SECM)
    constant current regime, 643
    protocols, 648–659
  scanning tunneling microscopy (STM), 1117
Top-loading balance, classification, 27–28
Topographic analysis
  dynamical diffraction, applications, 225
  ion-beam-induced charge (IBIC)/single event upset (SEU) microscopy, 1226–1227
  scanning tunneling microscopy (STM), 1116–1117
Topological distribution
  diffusion-controlled grain growth, 128
  van Hove singularities, copper-platinum alloys, 271–273
Topologically close-packed phases (tcps), metal alloy bonding, 135–136
  size effects, 137
Toroidal analyzers, ultraviolet photoelectron spectroscopy (UPS), 729
Torsional testing, high-strain-rate testing, Hopkinson bar technique, 289–290
Total current under illumination, illuminated semiconductor-liquid interface, J-E equations, 608
Totally irreversible reaction, cyclic voltammetry, 583
Totally reversible reaction, cyclic voltammetry, 582–583
Total-reflection x-ray fluorescence (TXRF) spectroscopy, heavy-ion backscattering spectrometry (HIBS) and, 1275
  semiconductor manufacturing, 1280
Tougaard background, x-ray photoelectron spectroscopy (XPS), 990–991
Toughness property, tension testing, 280
Trace element accelerator mass spectrometry (TEAMS)
  automation, 1247
  bulk analysis
    impurity measurements, 1247–1249
    measurement data, 1249–1250
  complementary, competitive and alternative methods, 1236–1238
    inductively coupled plasma mass spectrometry, 1237
    neutron activation-accelerator mass spectrometry (NAAMS), 1237
    neutron activation analysis (NAA), 1237
    secondary-ion mass spectrometry (SIMS), 1236–1237
    selection criteria, 1237–1238
    sputter-initiated resonance ionization spectrometry (SIRIS), 1237
  data analysis and interpretation, 1247–1252
    calibration of data, 1252
  depth-profiling
    data analysis, 1251–1252
    impurity measurements, 1250–1251
  facilities profiles, 1242–1246
    CSIRO Heavy Ion Analytical Facility (HIAF), 1245
    Naval Research Laboratory, 1245–1246
    Paul Scherrer Institute (PSI)/ETH Zurich Accelerator SIMS Laboratory, 1242–1244
    Technical University Munich Secondary Ion AMS Facility, 1245
    University of North Texas Ion Beam Modification and Analysis Laboratory, 1246
    University of Toronto IsoTrace Laboratory, 1244–1245
  facility requirements, 1238
  future applications, 1239
  high-energy beam transport, analysis, and detection, 1239
  historical evolution, 1246–1247
  impurity measurements
    bulk analysis, 1247–1249
    depth-profiling, 1250–1251
  instrumentation criteria, 1239–1247
    magnetic and electrostatic analyzer calibration, 1241–1242
    ultraclean ion source design, 1240–1241
  instrumentation specifications and suppliers, 1258
  limitations, 1253
  research background, 1235–1238
  sample preparation, 1252–1253
  secondary-ion acceleration and electron-stripping system, 1238–1239
  specimen modification, 1253
  ultraclean ion sources, negatively charged secondary-ion generation, 1238
Trace element distribution, microbeam analysis, 947–948
Trace-element sensitivity, medium-energy backscattering, 1266–1267
Tracer diffusion
  binary/multicomponent diffusion, 149–150
  substitutional and interstitial metallic systems, 154–155
Training effect, superconducting magnets, 501
Transfer length, superconductors, electrical transport measurement, 475–476
Transfer pumps, technological principles, 3
Transient decay dynamics, semiconductor-liquid interfaces, 619–622
Transient gating (TG) methods, impulsive stimulated thermal scattering (ISTS), 744
Transient ion-beam-induced charge (IBIC) microscopy, basic principles, 1228
Transient measurement techniques, carrier lifetime measurement, 435–438
  data interpretation, 437
  free carrier absorption (FCA), 439
  limitations, pulsed-type methods, 437–438
  selection criteria, 453–454
Transient response curves, thermal diffusivity, laser flash technique, 387–389
Transition metals
  atomic/ionic magnetism, ground-state multiplets, 514–515
  bonding-antibonding, 136–137
  crystal structure, tight-binding calculations, 135
  magnetic ground state, itinerant magnetism at zero temperature, 181–183
  surface phenomena, molecular dynamics (MD) simulation, 159
Transition operator, resonant scattering analysis, quantum mechanics, 908–909
Transition temperature measurement, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 367–368
Transmission electron microscopy (TEM)
  automation, 1080
  basic principles, 1064–1069
    deviation vector and parameter, 1066–1067
    Ewald sphere construction, 1066
    extinction distance, 1067–1068
    lattice defect diffraction contrast, 1068–1069
    structure and shape factors, 1065–1066
  bright field/dark-field imaging, 1069–1071
  data analysis and interpretation, 1080–1086
    bright field/dark-field, and selected-area diffraction, 1082–1084
    defect analysis values, 1084–1085
    Kikuchi lines and deviation parameter, defect contrast, 1085–1086
    shape factor effect, 1080–1081
    specimen thickness and deviation parameter, 1081–1082
  diffraction pattern indexing, 1073–1074
  ion-beam analysis (IBA) *vs.*, 1181
  Kikuchi lines and specimen orientation, 1075–1078
    deviation parameters, 1077–1078
    electron diffuse scattering, 1075
    indexing protocols, 1076–1077
    line origins, 1075–1076
  lens defects and resolution, 1078–1080
    aperture diffraction, 1078
    astigmatism, 1079
    chromatic aberration, 1078
    resolution protocols, 1079–1080
    spherical aberration, 1078
  limitations, 1088–1089
  magnetic domain structure measurements, Lorentz transmission electron microscopy, 551–552
  pn junction characterization, 467
  research background, 1063–1064
  sample preparation, 1086–1087
    dispersion, 1087
    electropolishing and chemical thinning, 1086
    ion milling and focused gallium ion beam thinning, 1086–1087
    replication, 1087
    ultramicrotomy, 1087
  scanning electron microscopy (SEM) *vs.*, 1050
    phase-contrast illumination, 1093–1097
  scanning transmission electron microscopy (STEM) *vs.*, 1090–1092
  scanning tunneling microscopy (STM) *vs.*, 1112–1113
  selected-area diffraction (SAD), data analysis, 1080–1081
  specimen modification, 1087–1088
  tilted illumination and diffraction, 1071
  tilting specimens and electron beams, 1071
Transmission lines, microwave measurement techniques, 409
Transmission measurements
  x-ray absorption fine structure (XAFS) spectroscopy, 875–877
  x-ray photoelectron spectroscopy (XPS), elemental composition analysis, 984–985
Transmission x-ray magnetic circular dichroism (TXMCD), magnetic domain structure measurements, 555–556
Transmitted-light microscopy, optical microscopy and, 667
Transport equations, magnetotransport in metal alloys, 560–561
Transport measurements. *See also* Magnetotransport properties
  theoretical background, 401
Transport phenomena model, chemical vapor deposition (CVD), basic components, 167
Transverse Kerr effect, surface magneto-optic Kerr effect (SMOKE), 571–572
Transverse magnetization, nuclear magnetic resonance, 764–765
Transverse magnetoresistance effects, magnetotransport in metal alloys, 564–566
Transverse relaxation time, nuclear quadrupole resonance (NQR), 779–780

Trap-assisted Auger recombination, carrier lifetime measurement, 430–431
Trapping mechanisms
  capacitance-voltage (C-V) characterization, 464–465
  carrier lifetime measurement, 431–432
Traveling wave model, x-ray diffraction, 206
Tribological testing
  acceleration, 333
  automation of, 333–334
  basic principles, 324–326
  control factors, 332–333
  data analysis and interpretation, 334
  equipment and measurement techniques, 326–327
  friction coefficients, 326
  friction testing, 328–329
  general procedures, 326
  limitations, 335
  research background, 324
  results analysis, 333
  sample preparation, 335
  test categories, 327–328
  wear coefficients, 326–327
  wear testing, 329–332
Triclinic systems, crystallography, space groups, 50
Triode ionization gauge, 14–15
Triode sputter-ion pump, operating principles, 11
Triple-axis spectrometry, phonon analysis, 1320–1323
  error detection, 1326–1327
TRISO-coated fuel particles, microbeam analysis, 947–948
Trouton's rule, thermal analysis and, 343
Tungsten electron guns, scanning electron microscopy (SEM)
  instrumentation criteria, 1054–1057
  selection criteria, 1061
Tungsten filaments
  Auger electron spectroscopy (AES), 1160–1161
  hot cathode ionization gauges, 15
Tungsten tips, scanning tunneling microscopy (STM), 1117
Tuning protocols, nuclear magnetic resonance, 770–771
Tunneling mechanisms. *See also* Scanning tunneling microscopy (STM)
  pn junction characterization, 469
Turbomolecular pumps
  applications, 7
  bearings for, 7
  operating principles and procedures, 7–9
  reactive gas problems, 8
  startup procedure, 8
  surface x-ray diffraction, 1011–1012
  venting protocols, 8
Two-beam diffraction, 229–236
  anomalous transmission, 231–232
  Darwin width, 232
  diffuse intensities
    boundary condition, 233
    Bragg case, 234–235
    integrated intensities, 235
    Laue case, 233–234
  dispersion equation solution, 233
  dispersion surface properties, 230–231
    boundary conditions, Snell's law, 230–231
    hyperboloid sheets, 230
    Poynting vector and energy flow, 231
    wave field amplitude ratios, 230
  Pendellösung, 231
  standing waves, 235–236
  x-ray birefringence, 232–233
  x-ray standing waves (XSWs), 232
Two-body contact, tribological and wear testing, 324–325
  categories for, 327–328
  equipment and measurement techniques, 326–327
  wear testing categories, 329–332
Two-dimensional exchange NQR spectroscopy, basic principles, 785
Two-dimensional Fourier transform
  nuclear magnetic resonance data analysis, 771–772
  nutation nuclear resonance spectroscopy, 783–784
  Zeeman-perturbed NRS (ZNRS), 782–783
Two-dimensional lattices
  crystal systems, 44–46
  low-energy electron diffraction (LEED), 1122
  surface/interface x-ray diffraction, 218–219
Two-dimensional states, ultraviolet photoelectron spectroscopy (UPS), energy band dispersion, 724–725
Two-dimensional zero-field NQR
  level-crossing double resonance NQR nutation spectroscopy, 785
  Zeeman-perturbed nuclear resonance spectroscopy, 782–783
  zero-field nutation NRS, 783–784
Two-phase model, small-angle scattering (SAS), 220
Two-point measurement techniques
  bulk measurements, 403
    protocols and procedures, 403–404
  conductivity measurements, 402
Two-wave analysis, high-strain-rate testing, Hopkinson bar technique, 291–292

Ultraclean ion sources, trace element accelerator mass spectrometry (TEAMS)
  design criteria, 1239–1240
  negatively charged secondary ion generation, 1238
Ultrahigh-vacuum (UHV) systems
  all-metal flange seals, 18
  assembly, processing and operation, 19
  basic principles, 1–2
  bellows-sealed feedthroughs, 19
  construction materials, 17
  hot cathode ionization gauges, 15
  low-energy electron diffraction (LEED), comparisons, 1121
  O-ring seals, limits of, 18
  scanning tunneling microscopy (STM), 1117
  sputter-ion pump, 10–11
  surface magneto-optic Kerr effect (SMOKE), 573
  surface x-ray diffraction
    bakeout procedure, 1023–1024
    basic principles, 1011–1012
    load-lock procedure, 1023
    protocols and procedures, 1022–1024
    venting procedures, 1023
  valve construction, 18
  x-ray absorption fine structure (XAFS), 876–877
  x-ray photoelectron spectroscopy (XPS)
    instrumentation, 978–983
    research background, 970–972
Ultralarge-scale integrated circuits, heavy-ion backscattering spectrometry (HIBS), 1280
Ultra microbalances, thermogravimetric (TG) analysis, 347–350
Ultramicroelectrode (UME), scanning electrochemical microscopy (SECM)
  feedback mode, 638–640
  properties and applications, 636
Ultramicrotomy, transmission electron microscopy (TEM), sample preparation, 1087
Ultrasonic microhardness testing, basic principles, 316
Ultrathin limit, surface magneto-optic Kerr effect (SMOKE), 577
Ultraviolet photoelectron spectroscopy (UPS)
  alignment procedures, 729–730
  atoms and molecules, 727–728
  automation, 731
  competitive and related techniques, 725–726
  data analysis and interpretation, 731–732
  electronic phase transitions, 727
  electron spectrometers, 729
  energy band dispersion, 723–725
    solid materials, 727
  light sources, 728–729
    figure of merit, 734–735
  limitations, 733
  photoemission process, 726–727
  photoemission *vs.* inverse photoemission, 722–723
  physical relations, 730
  sample preparation, 732–733
  sensitivity limits, 730–731
  surface states, 727
  synchrotron light sources, 734
  valence electron characterization, 723–724
  x-ray photoelectron spectroscopy (XPS), comparisons, 971
Ultraviolet/visible absorption (UV-VIS) spectroscopy
  applications, 692
  array detector spectrometer, 693
  automation, 693
  common components, 692
  competitive and related techniques, 690
  dual-beam spectrometer, 692–693
  limitations, 696
  materials characterization, 691–692, 694–695
  materials properties, 689
  qualitative/quantitative analysis, 689–690
  quantitative analysis, 690–691, 693–694
  research background, 688–689
  sample preparation, 693, 695
  single-beam spectrometer, 692
  specimen modification, 695–696
Uncertainty principle
  diffuse scattering techniques, 887–889
  resonance spectroscopies, 761–762
  x-ray photoelectron spectroscopy (XPS), composition analysis, 994–996
Uniaxial flow stress, static indentation hardness testing, 317
Uniform plastic deformation, stress-strain analysis, 282
Units (magnetism), general principles, 492–493
Universal gas constant, corrosion quantification, Tafel technique, 593–596
University of North Texas Ion Beam Modification and Analysis Laboratory, trace element accelerator mass spectrometry (TEAMS) research at, 1246
University of Toronto IsoTrace Laboratory, trace element accelerator mass spectrometry (TEAMS) at, 1244–1245
Unloading compliance technique, fracture toughness testing, crack extension measurement, 308
Unstable crack growth, fracture toughness testing, load-displacement behavior, 304–305
Unstable fractures, fracture toughness testing, 308–309

Upper critical field strength (H_{c2}), superconductors
electrical transport measurements, extrapolation, 475
magnetic measurements, 473
magnetization, 519
superconductors-to-normal (S/N) transition, electrical transport measurement, estimation protocols, 482

Vacancy wind, binary/multicomponent diffusion, Kirkendall effect, 149
Vacuum systems
basic principles, 1–3
outgassing, 2–3
leak detection, 20–22
load-lock antechamber, basic principles, 1–2
nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1203–1204
pumps
surface x-ray diffraction, 1011–1012
technological principles, 3
scanning electron microscopy (SEM), 1061
scanning transmission electron microscopy (STEM), error detection, 1108
standards and units, 1
surface x-ray diffraction, basic properties, 1011–1012
system construction and design
assembly, processing and operation, 19
design criteria, 19–20
hardware components, 17–19
materials, 17
technological principles
cryopumps, 9–10
diffusion pumps, 6–7
getter pumps, 12
high-vacuum pumps, 5–6
nonevaporable getter pumps (NEGs), 12
oil-free (dry) pumps, 4–5
oil-sealed pumps, 3–4
roughing pumps, 3
scroll pumps, 5
sorption pumps, 5
sputter-ion pumps, 10–12
sublimation pumps, 12
thermal conductivity gauges for pressure measurement, 13–17
total/partial pressure measurement, 13
turbomolecular pumps, 7–9
vacuum pumps, 3
ultraviolet photoelectron spectroscopy (UPS), automation, 731–732
Valence bands, x-ray photoelectron spectroscopy (XPS), 973–974
Valence electron characterization
Mössbauer spectroscopy, 827
ultraviolet photoelectron spectroscopy (UPS), 723–724
Valence potential, x-ray photoelectron spectroscopy (XPS), initial-state effects, 977–978
Valves, vacuum systems, 18
Van de Graaff particle accelerator, nuclear reaction analysis (NRA)/proton-induced gamma ray emission (PIGE), 1203–1204
Van der Pauw method
conductivity measurements, 402
Hall effect, semiconductor materials, relaxation time approximation (RTA), 413
semiconductor materials, Hall effect, automated apparatus, 414
surface measurements, 406
Van der Waals forces
liquid surface x-ray diffraction, simple liquids, 1040–1041
neutron powder diffraction, 1302
vacuum system principles, outgassing, 2–3
Van Hove singularities, diffuse intensities, metal alloys
copper-platinum alloys, 271–273
multicomponent alloys, 268–270
Van't Hoff equation
differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 368
thermal analysis and principles of, 342–343
Vapor phase deposition (VPD), heavy-ion backscattering spectrometry (HIBS), 1274
Vapor phase epitaxy (VPE) reactor, capacitance-voltage (C-V) characterization, 462–463
Variability in test results, tribological and wear testing, 335
Variable frequency Hewlett-Packard meters, capacitance-voltage (C-V) characterization, 460–461
Variable-pressure scanning electron microscopy (SEM)
basic principles, 1056–1057
specimen modification, 1059
Variational Monte Carlo (VMC), electronic structure analysis, 88–89
Venting protocols
surface x-ray diffraction, ultrahigh-vacuum (UHV) systems, 1023
turbomolecular pumps, 8
Vertical illuminators, reflected-light optical microscopy, 675–676
Vibrating-coil magnetometer (VCM), principles and applications, 535
Vibrating-sample magnetometer (VSM)
automation, 537
properts and applications, 533–534
thermomagnetic analysis, 540–544
Vibrational Raman spectroscopy, group theoretical analysis, 702–704
character tables, 718
point groups and matrix representation, symmetry operations, 717–718
vibrational modes of solids, 720–722
vibrational selection rules, 716–720
Vibrational selection rules, group theoretical analysis, 716–717
symmetry operators, 718–720
Vibrational spectra, group theoretical analysis, 717–720
Vickers hardness testing
automated methods, 319
basic principles, 317–318
data analysis and interpretation, 319–320
hardness values, 317–318, 323
limitations and errors, 322
procedures and protocols, 318–319
research background, 316–317
sample preparation, 320
specimen modification, 320–321
Virgin wafers, carrier lifetime measurement, free carrier absorption (FCA), 442–443
Virtual leaks, vacuum systems, 20–22
Viscous flow
oil-sealed pumps, oil contamination, avoidance, 4
vacuum system design, 20
Voigt function, x-ray photoelectron spectroscopy (XPS), peak shape, 1006
Volatile products
energy-dispersive spectrometry (EDS), loss mechanisms, 1153
gas analysis and, 398
Voltage spikes, superconductors, electrical transport measurements, signal-to-noise ratio, 485
Voltage tap, superconductors, electrical transport measurements, placement protocols, 477
Voltmeter properties, superconductors, electrical transport measurements, 477
Volume deformation (VD), phase diagram prediction, static displacive interactions, 104–106
Volume-fixed frame of reference, binary/multicomponent diffusion, 148
Vortex melting transition, superconductor magnetization, 519

Warren-Cowley order parameter
diffuse intensities, metal alloys
atomic short-range ordering (ASRO) principles, pair correlations, 257
basic definitions, 252–254
concentration waves, density-functional theory (DFT), 261–262
copper-nickel-zinc alloys, 269–270
effective cluster interactions (ECIs), 255–256
diffuse scattering techniques, 886–889
protocols and procedures, 889–894
x-ray diffraction, local atomic correlation, 215–217
Wave equations, high-strain-rate testing, Hopkinson bar technique, 290–292
Wave field amplitude ratios, two-beam diffraction, dispersion surface, 230
Wave function character
metal alloy bonding, 138
ultraviolet photoelectron spectroscopy (UPS), photoemission process, 726–727
Waveguide mode velocities, impulsive stimulated thermal scattering (ISTS) analysis, 756–757
Wave-length-dispersive x-ray spectrometry (WDS), energy-dispersive spectrometry (EDS), matrix corrections, 1145–1147
Wavelength properties
fluorescence analysis, dispersive spectrometers, 944
impulsive stimulated thermal scattering (ISTS), 746–749
Wave measurements, thermal diffusivity, laser flash technique *vs.,* 383–384
Weak-beam dark-field (WBDF) imaging, transmission electron microscopy (TEM), 1071
Weak scattering sources
magnetic x-ray scattering, 935
scanning transmission electron microscopy (STEM), incoherent imaging, weakly scattering objects, 1111
Wear coefficients, tribological and wear testing, 326–327
Wear curves, wear testing protocols, 330–331
Wear-rate-*vs.*-usage curve, wear testing protocols, 331
Wear testing
acceleration, 333
automation of, 333–334
basic principles, 317, 324–326
classification of, 324–325
control factors, 332–333
data analysis and interpretation, 334
equipment and measurement techniques, 326–327
friction coefficients, 326
friction testing, 328–329
general procedures, 326
limitations, 335

research background, 324
results analysis, 333
sample preparation, 335
techniques and procedures, 329–332
test categories, 327–328
wear coefficients, 326–327
Weight
definitions, 24–26
standards for, 26–27
Weilbull distribution, fracture toughness testing, brittle materials, 311
Weiss molecular field constant
ferromagnetism, 523–524
surface magneto-optic Kerr effect (SMOKE), 571
Wigner-Eckart theorem, resonant scattering analysis, 909–910
Williamson-Hall plot, neutron powder diffraction, microstrain broadening, 1294–1295
Wollaston prism, reflected-light optical microscopy, 679–680
Wood notation, low-energy electron diffraction (LEED), qualitative analysis, 1123–1124
Working distance, optical microscopy, 670
Working electrode, cyclic voltammetry, 585–586

X-band frequencies, electron paramagnetic resonance (EPR), continuous-wave (CW) experiments, rectangular resonator, 794–795
XCOM database, particle-induced x-ray emission (PIXE), 1212
X-ray absorption, energy-dispersive spectrometry (EDS), matrix corrections, 1145
X-ray absorption fine structure (XAFS) spectroscopy
automation, 878
data analysis and interpretation, 878–879
detection methods, 875–877
energy calibration, 877
energy resolution, 877
harmonic content, 877
limitations, 880
micro-XAFS, fluorescence analysis, 943
monochromator glitches, 877–878
related structures, 870
research background, 869–870
sample preparation, 879–880
simple picture components, 872–874
disorder, 872–873
L edges, 873
multiple scattering, 873–874
polarization, 873
single-crystal x-ray structure determination, 851
single-scattering picture, 870–871
specimen modification, 880
synchrotron facilities, 877
x-ray photoelectron spectroscopy (XPS), 971
X-ray absorption near-edge structure (XANES) spectroscopy. *See also* Near-edge x-ray absorption fine structure (NEXAFS) spectroscopy
comparisons x-ray absorption (XAS), 870
research background, 874–875
ultraviolet photoelectron spectroscopy (UPS) and, 726
X-ray birefringence, two-beam diffraction, 232–233
X-ray crystallography, single-crystal x-ray structure determination, 851–858
crystal structure refinement, 856–858
crystal symmetry, 854–856
X-ray diffraction
kinematic theory
crystalline material, 208–209
lattice defects, 210
local atomic arrangement - short-range ordering, 214–217
research background, 206
scattering principles, 206–208
small-angle scattering (SAS), 219–222
cylinders, 220–221
ellipsoids, 220
Guinier approximation, 221
integrated intensity, 222
interparticle interference, 222
K=0 extrapolation, 221
porod approximation, 221–222
size distribution, 222
spheres, 220
two-phase model, 220
structure factor, 209–210
surface/interface diffraction, 217–219
crystal truncation rods, 219
two-dimensional diffraction rods, 218–219
thermal diffuse scattering (TDS), 210–214
liquid surfaces and monomolecular layers
basic principles, 1028–1036
competitive and related techniques, 1028
data analysis and interpretation, 1039–1043
Langmuir monolayers, 1041–1043
liquid alkane crystallization, 1043
liquid metals, 1043
simple liquids, 1040–1041
non-specular scattering
GID, diffuse scattering, and rod scans, 1038–1039
reflectivity measurements, 1033
p-polarized x-ray beam configuration, 1047
reflectivity, 1029–1036
Born approximation, 1033–1034
distorted-wave Born approximation, 1034–1036
Fresnel reflectivity, 1029–1031
grazing incidence diffraction and rod scans, 1036
instrumentation, 1036–1038
multiple stepwise and continuous interfaces, 1031–1033
non-specular scattering, 1033
research background, 1027–1028
specimen modification, 1043–1045
low-energy electron diffraction (LEED), comparisons, 1120–1121
neutron powder diffraction *vs.,* 1286–1289
resonant scattering techniques
calculation protocols, 909–912
L=2 measurements, 910–911
L=4 measurements, 911–912
classical mechanics, 906–907
comparisons with other methods, 905–906
data analysis and interpretation, 914–915
experiment design, 914
instrumentation criteria, 912–913
materials properties measurments, 905
polarization analysis, 913
quantum mechanics, 908–909
research background, 905
sample preparation, 913–914
theory, 906
single-crystal x-ray structure determination
automation, 860
competitive and related techniques, 850–851
derived results interpretation, 861–862
initial model structure, 860–861
limitations, 863–864
nonhydrogen atom example, 862
protocols and procedures, 858–860
data collection, 859–860
research background, 850–851
sample preparation, 862–863
specimen modification, 863
x-ray crystallography principles, 851–858
crystal structure refinement, 856–858
crystal symmetry, 854–856
surface x-ray diffraction
angle calculations, 1021
basic principles, 1008–1011
crystallographic measurements, 1010–1011
grazing incidence, 1011
measurement instrumentation, 1009–1010
surface properties, 1008–1009
beamline alignment, 1019–1020
competitive and related strategies, 1007–1008
data analysis and interpretation, 1015–1018
crystal truncation rod (CTR) profiles, 1015
diffuse scattering, 1016
grazing incidence measurements, 1016–1017
reciprocal lattice mapping, 1016
reflectivity, 1015–1016
silicon surface analysis, 1017–1018
diffractometer alignment, 1020–1021
instrumentation criteria, 1011–1015
crystallographic alignment, 1014–1015
five-circle diffractometer, 1013–1014
laser alignment, 1014
sample manipulator, 1012–1013
vacuum system, 1011–1012
limitations, 1018–1019
research background, 1007–1008
transmission electron microscopy (TEM) and, 1064
X-ray diffuse scattering
applications, 889–894
automation, 897–898
bond distances, 885
chemical order, 884–885
comparisons, 884
competitive and related techniques, 883–884
crystalling solid solutions, 885–889
data analysis and interpretation, 894–896
diffuse x-ray scattering techniques, 889–890
inelastic scattering background removal, 890–893
limitations, 898–899
measured intensity calibration, 894
protocols and procedures, 884–889
recovered static displacements, 896–897
research background, 882–884
resonant scattering terms, 893–894
sample preparation, 898
X-ray fluorescence analysis. *See* X-ray microfluorescence analysis
X-ray magnetic circular dichroism (XMCD)
automation, 963
basic principles, 955–957
circularly polarized x-ray sources, 957–959
competitive/related techniques, 954–955
data analysis and interpretation, 963–964
detection protocols, 959
limitations, 965–966
magnetic domain structure measurements, 555–556
automation of, 555–556
basic principles, 555
procedures and protocols, 555
magnetic x-ray scattering, comparisons, 919
nonresonant ferromagnetic scattering, 930
measurement optics, 959–962
research background, 953–955
sample magnetization, 962–963
sample preparation, 964–965

X-ray magnetic linear dichroism (XMLD), magnetic domain structure measurements, 555
X-ray microdiffraction
  analytic applications, 944–945
  basic principles, 941–942
  data analysis, 950
  error detection, 950–951
  sample preparation, 949–950
X-ray microfluorescence analysis, 942–945
  background signals, 944
  characteristic radiation, 942–943
  detector criteria, 944
  fluorescence yields, 942
  micro XAFS, 943
  particle-induced x-ray emission (PIXE) and, 1210–1211
  penetration depth, 943–944
  photoabsorption cross-sections, 943
  scanning transmission electron microscopy (STEM), atomic resolution spectroscopy, 1103–1104
X-ray microprobe
  protocols and procedures, 941–942
  research background, 939–941
X-ray microprobes
  microbeam applications
    strain distribution
      ferroelectric sample, 948–949
      tensile loading, 948
    trace element distribution, SiC nuclear fuel barrier, 947–948
  protocols and procedures, 941–942
  research background, 939–941
  sources, 945
X-ray photoelectron spectroscopy (XPS)
  applications, 983–988
    chemical state, 986
    elemental composition analysis, 983–986
    materials-specific issues, 988–989
  Auger electron spectroscopy (AES) *vs.,* 1158–1159
  automation, 989
  basic principles, 971–978
    final-state effects, 974–976
    high-resolution spectrum, 974–978
    initial-state effects, 976–978
    survey spectrum, 972–974
  competitive and related techniques, 971
  data analysis, 989–994
    appearance protocols, 993–994
    background subtraction, 990–991
    peak integration and fitting, 991–992
    peak position and half-width, 992
    principal components, 992–993
  instrumentation criteria, 978–983
    analyzers, 980–982
    detectors, 982
    electron detection, 982–983
    maintenance, 983
    sources, 978–980
  interpretation protocols, 994–998
    chemical state information, 996–998
    composition analysis, 994–996
  ion-beam analysis (IBA) *vs.,* 1181
  limitations, 999–1001
  low-energy electron diffraction (LEED), sample preparation, 1125–1126
  nuclear reaction analysis (NRA) and proton-induced gamma ray emission (PIGE) and, 1201–1202
  research background, 970–971
  sample preparation, 998–999
  specimen modification, 999
  ultraviolet photoelectron spectroscopy (UPS) and, 725–726
X-ray powder diffraction
  ab initio structure determination, 845
  basic principles, 836–838
  Bragg peak integrated intensity extraction, 840
  candidate atom positions, 840–841
  competitive and related techniques, 836
  crystal lattice and space group determination, 840
  data analysis and interpretation, 839–842
  database comparisons of known materials, 843–844
  limitations, 842–843
  protocols and procedures, 838–839
  quantitative phase analysis, 842, 844–845
  research background, 835–836
  Rietveld refinements, 841–842
  sample preparation, 842
  single-crystal neutron diffraction and, 1307–1309
  structure determination in known analogs, 845
  structure determination protocols, 847–848
  time-resolved x-ray powder diffraction, 845–847
X-ray resonant exchange scattering (XRES), x-ray magnetic circular dichroism (XMCD) comparison, 955
X-ray scattering and spectroscopy
  diffuse scattering techniques, measurement protocols, 889–894
  magnetic x-ray scattering
    data analysis and interpretation, 934–935
    hardware criteria, 925–927
    limitations, 935–936
    nonresonant scattering
      antiferromagnets, 928–930
      ferromagnets, 930
      theoretical concepts, 920–921
    research background, 917–919
    resonant scattering
      antiferromagnets, 930–932
      ferromagnets, 932
      theoretical concepts, 921–924
    sample preparation, 935
    spectrometer criteria, 927–928
    surface magnetic scattering, 932–934
    theoretical concepts, 919–925
      ferromagnetic scattering, 924–925
      nonresonant scattering, 920–921
      resonant scattering, 921–924
  research background, 835
X-ray self-absorption, energy-dispersive spectrometry (EDS), standardless analysis, 1149
X-ray standing wave (XWS) diffraction
  literature sources, 226
  low-energy electron diffraction (LEED), comparisons, 1120–1121
  two-beam diffraction, 232, 235–236
XRS-82 program, neutron powder diffraction, 1306

Yafet-Kittel triangular configuration, ferromagnetism, 525–526
Yield-point phenomena, stress-strain analysis, 283–284
Yield strength
  hardness testing, 320
  heavy-ion backscattering spectrometry (HIBS), 1277
  ion-beam analysis (IBA), ERD/RBS equations, 1186–1189
Young's modulus
  fracture-toughness testing, lnear elastic fracture mechanics (LEFM), 303
  high-strain-rate testing, Hopkinson bar technique, 291–292
  impulsive stimulated thermal scattering (ISTS), 753–755
  stress-strain analysis, 280–281

Zachariasen formalism, single-crystal neutron diffraction, 1314–1315
ZBL approximation, particle scattering, central-field theory, deflection function, 60
Z-contrast imaging
  diffuse scattering techniques, comparisons, 884
  scanning transmission electron microscopy (STEM)
    atomic resolution spectroscopy, 1103–1104
    coherent phase-contrast imaging and, 1093–1097
    competitive and related techniques, 1092–1093
    data analysis and interpretation, 1105–1108
      object function retrieval, 1105–1106
      strain contrast, 1106–1108
    dynamical diffraction, 1101–1103
    incoherent scattering, 1098–1101
      weakly scattered objects, 1111
    limitations, 1108
    manufacturing sources, 1111
    probe formation, 1097–1098
    protocols and procedures, 1104–1105
    research background, 1090–1093
    sample preparation, 1108
    specimen modification, 1108
  transmission electron microscopy (TEM), 1063–1064
Zeeman effects
  nuclear quadrupole resonance (NQR), perturbation and line shapes, 779
  splitting
    Pauli paramagnetism, 493–494
    quantum paramagnetiic response, 521–522
Zeeman-perturbed NRS (ZNRS), nuclear quadrupole resonance (NQR), basic principles, 782–783
Zero field cooling (ZFC), superconductor magnetization, 518–519
Zero-line optimization, differential thermal analysis (DTA)/differential scanning calorimetry (DSC), 366
Zero magnetic field
  magnetotransport in metal alloys, 561–563
  nuclear quadrupole resonance (NQR)
    energy levels, 777–779
      higher spin nuclei, 779
      spin 1 levels, 777–778
      spin 3/2 levels, 778
      spin 5/2 levels, 778–779
    two-dimensional zero-field NQR, 782–785
      exchange NQR spectroscopy, 785
      level-crossing double resonance NQR nutation spectroscopy, 785
      nutation NRS, 783–784
      Zeeman-perturbed NRS (ZNRS), 782–783
Zeroth Law of Thermodynamics, thermodynamic temperature scale, 31–32
Zero voltage definition, superconductors, electrical transport measurement, 474
Zinc alloys. *See also* Copper-nickel-zinc alloys
  x-ray diffraction, structure-factor calculations, 209
Zone plates, x-ray microprobes, 945–946

This page intentionally left blank