Topics in Applied Physics
Volume 104

Available online at
SpringerLink.com

Topics in Applied Physics is part of the SpringerLink service. For all customers with standing orders for Topics in Applied Physics we offer the full text in electronic form via SpringerLink free of charge. Please contact your librarian who can receive a password for free access to the full articles by registration at:

springerlink.com → Orders

If you do not have a standing order you can nevertheless browse through the table of contents of the volumes and the abstracts of each article at:

springerlink.com → Browse Publications

Topics in Applied Physics

Topics in Applied Physics is a well-established series of review books, each of which presents a comprehensive survey of a selected topic within the broad area of applied physics. Edited and written by leading research scientists in the field concerned, each volume contains review contributions covering the various aspects of the topic. Together these provide an overview of the state of the art in the respective field, extending from an introduction to the subject right up to the frontiers of contemporary research.
Topics in Applied Physics is addressed to all scientists at universities and in industry who wish to obtain an overview and to keep abreast of advances in applied physics. The series also provides easy but comprehensive access to the fields for newcomers starting research.
Contributions are specially commissioned. The Managing Editors are open to any suggestions for topics coming from the community of applied physicists no matter what the field and encourage prospective editors to approach them with ideas.

David A. Drabold Stefan K. Estreicher (Eds.)

Theory of Defects in Semiconductors

With 60 Figures and 15 Tables

Springer

David A. Drabold
Department of Physics and Astronomy
Ohio University
Athens, OH 45701, USA
drabold@ohio.edu

Stefan K. Estreicher
Physics Department
Texas Tech University
Lubbock, TX 79409-1051, USA
stefan.estreicher@ttu.edu

Library of Congress Control Number: 2006934116

Physics and Astronomy Classification Scheme (PACS):
71.10.-w, 71.17.-m, 71.23.-k, 71.55.-i, 63.20Mt

ISSN print edition: 0303-4216
ISSN electronic edition: 1437-0859
ISBN-10 3-540-33400-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33400-2 Springer Berlin Heidelberg New York

Typesetting: DA-TEX · Gerd Blumenstein · www.da-tex.de
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper 57/3100/YL 5 4 3 2 1 0

Contents

This book is dedicated to Manuel Cardona, who has done so much for the field of defects in semiconductors over the past decades, and convinced so many theorists to calculate beyond what they thought possible.

Preface

Semiconductor materials emerged after World War II and their impact on our lives has grown ever since. Semiconductor technology is, to a large extent, the art of defect engineering. Today, defect control is often done at the atomic level. Theory has played a critical role in understanding, and therefore controlling, the properties of defects.

Conversely, the careful experimental studies of defects in Ge, Si, then many other semiconductor materials have generated a huge database of measured quantities that allowed theorists to test their methods and approximations.

Dramatic improvement in methodology, especially density-functional theory, along with inexpensive and fast computers, has impedance matched the experimentalist and theorist in ways unanticipated before the late 1980s. As a result, the theory of defects in semiconductors has become quantitative in many respects. Today, more powerful theoretical approaches are still being developed. More importantly perhaps, the tools developed to study defects in semiconductors are now being adapted to approach many new challenges associated with nanoscience, a very long list that includes quantum dots, buckyballs and buckytubes, spintronics, interfaces, and many others.

Despite the importance of the field, there have been no modern attempts to treat the computational science of the field in a coherent manner within a single treatise. This is the aim of the present volume.

This book brings together several leaders in theoretical research on defects in semiconductors. Although the treatment is tutorial, the level at which the various applications are discussed is today's state-of-the-art in the field.

The book begins with a "big picture" view from Manuel Cardona, and continues with a brief summary of the historical development of the subject in Chap. 1. This includes an overview of today's most commonly used method to describe defects.

We have attempted to create a balanced and tutorial treatment of the basic theory and methodology in Chaps. 3–6. They include detailed discussions of the approximations involved, the calculation of electrically active levels, and extensions of the theory to finite temperatures. Two emerging electronic structure methodologies of special importance to the field are discussed in Chaps. 7 (quantum Monte-Carlo) and 8 (the GW method). Then come two chapters on molecular dynamics (MD). In Chap. 9, a combination of high-

level and approximate MD is developed, with applications to the dynamics of extended defect. Chapter 10 deals with semiempirical treatments of microstructures, including issues such as wafer bonding. The book concludes with studies of defects and their role in the photoresponse of topologically disordered (amorphous) systems.

The intended audience for the book is graduate students as well as advanced researchers in physics, chemistry, materials science, and engineering. We have sought to provide self-contained descriptions of the work, with detailed references available when needed. The book may be used as a text in a practical graduate course designed to prepare students for research work on defects in semiconductors or first-principles theory in materials science in general. The book also serves as a reference for the active theoretical researcher, or as a convenient guide for the experimentalist to keep tabs on their theorist colleagues.

It was a genuine pleasure to edit this volume. We are delighted with the contributions provided in a timely fashion by so many busy and accomplished people. We warmly thank all the contributors and hope to have the opportunity to share some nice wine(s) with all of them soon. After all,

When Ptolemy, now long ago,
Believed the Earth stood still,
He never would have blundered so
Had he but drunk his fill.
He'd then have felt it circulate
And would have learnt to say:
The true way to investigate
Is to drink a bottle a day.

(author unknown)
published in Augustus de Morgan's *A Budget of Paradoxes*, (1866).

Athens, Ohio, *David A. Drabold*
Lubbock, Texas *Stefan K. Estreicher*
February 2006

Contents

This book is dedicated to Manuel Cardona, who has done so much for the field of defects in semiconductors over the past decades, and convinced so many theorists to calculate beyond what they thought possible.

Preface

Semiconductor materials emerged after World War II and their impact on our lives has grown ever since. Semiconductor technology is, to a large extent, the art of defect engineering. Today, defect control is often done at the atomic level. Theory has played a critical role in understanding, and therefore controlling, the properties of defects.

Conversely, the careful experimental studies of defects in Ge, Si, then many other semiconductor materials have generated a huge database of measured quantities that allowed theorists to test their methods and approximations.

Dramatic improvement in methodology, especially density-functional theory, along with inexpensive and fast computers, has impedance matched the experimentalist and theorist in ways unanticipated before the late 1980s. As a result, the theory of defects in semiconductors has become quantitative in many respects. Today, more powerful theoretical approaches are still being developed. More importantly perhaps, the tools developed to study defects in semiconductors are now being adapted to approach many new challenges associated with nanoscience, a very long list that includes quantum dots, buckyballs and buckytubes, spintronics, interfaces, and many others.

Despite the importance of the field, there have been no modern attempts to treat the computational science of the field in a coherent manner within a single treatise. This is the aim of the present volume.

This book brings together several leaders in theoretical research on defects in semiconductors. Although the treatment is tutorial, the level at which the various applications are discussed is today's state-of-the-art in the field.

The book begins with a "big picture" view from Manuel Cardona, and continues with a brief summary of the historical development of the subject in Chap. 1. This includes an overview of today's most commonly used method to describe defects.

We have attempted to create a balanced and tutorial treatment of the basic theory and methodology in Chaps. 3–6. They include detailed discussions of the approximations involved, the calculation of electrically active levels, and extensions of the theory to finite temperatures. Two emerging electronic structure methodologies of special importance to the field are discussed in Chaps. 7 (quantum Monte-Carlo) and 8 (the GW method). Then come two chapters on molecular dynamics (MD). In Chap. 9, a combination of high-

level and approximate MD is developed, with applications to the dynamics of extended defect. Chapter 10 deals with semiempirical treatments of microstructures, including issues such as wafer bonding. The book concludes with studies of defects and their role in the photoresponse of topologically disordered (amorphous) systems.

The intended audience for the book is graduate students as well as advanced researchers in physics, chemistry, materials science, and engineering. We have sought to provide self-contained descriptions of the work, with detailed references available when needed. The book may be used as a text in a practical graduate course designed to prepare students for research work on defects in semiconductors or first-principles theory in materials science in general. The book also serves as a reference for the active theoretical researcher, or as a convenient guide for the experimentalist to keep tabs on their theorist colleagues.

It was a genuine pleasure to edit this volume. We are delighted with the contributions provided in a timely fashion by so many busy and accomplished people. We warmly thank all the contributors and hope to have the opportunity to share some nice wine(s) with all of them soon. After all,

When Ptolemy, now long ago,
Believed the Earth stood still,
He never would have blundered so
Had he but drunk his fill.
He'd then have felt it circulate
And would have learnt to say:
The true way to investigate
Is to drink a bottle a day.

(author unknown)
published in Augustus de Morgan's *A Budget of Paradoxes*, (1866).

Athens, Ohio, *David A. Drabold*
Lubbock, Texas *Stefan K. Estreicher*
February 2006

Contents

This book is dedicated to Manuel Cardona, who has done so much for the field of defects in semiconductors over the past decades, and convinced so many theorists to calculate beyond what they thought possible.

Preface

Semiconductor materials emerged after World War II and their impact on our lives has grown ever since. Semiconductor technology is, to a large extent, the art of defect engineering. Today, defect control is often done at the atomic level. Theory has played a critical role in understanding, and therefore controlling, the properties of defects.

Conversely, the careful experimental studies of defects in Ge, Si, then many other semiconductor materials have generated a huge database of measured quantities that allowed theorists to test their methods and approximations.

Dramatic improvement in methodology, especially density-functional theory, along with inexpensive and fast computers, has impedance matched the experimentalist and theorist in ways unanticipated before the late 1980s. As a result, the theory of defects in semiconductors has become quantitative in many respects. Today, more powerful theoretical approaches are still being developed. More importantly perhaps, the tools developed to study defects in semiconductors are now being adapted to approach many new challenges associated with nanoscience, a very long list that includes quantum dots, buckyballs and buckytubes, spintronics, interfaces, and many others.

Despite the importance of the field, there have been no modern attempts to treat the computational science of the field in a coherent manner within a single treatise. This is the aim of the present volume.

This book brings together several leaders in theoretical research on defects in semiconductors. Although the treatment is tutorial, the level at which the various applications are discussed is today's state-of-the-art in the field.

The book begins with a "big picture" view from Manuel Cardona, and continues with a brief summary of the historical development of the subject in Chap. 1. This includes an overview of today's most commonly used method to describe defects.

We have attempted to create a balanced and tutorial treatment of the basic theory and methodology in Chaps. 3–6. They include detailed discussions of the approximations involved, the calculation of electrically active levels, and extensions of the theory to finite temperatures. Two emerging electronic structure methodologies of special importance to the field are discussed in Chaps. 7 (quantum Monte-Carlo) and 8 (the GW method). Then come two chapters on molecular dynamics (MD). In Chap. 9, a combination of high-

level and approximate MD is developed, with applications to the dynamics of extended defect. Chapter 10 deals with semiempirical treatments of microstructures, including issues such as wafer bonding. The book concludes with studies of defects and their role in the photoresponse of topologically disordered (amorphous) systems.

The intended audience for the book is graduate students as well as advanced researchers in physics, chemistry, materials science, and engineering. We have sought to provide self-contained descriptions of the work, with detailed references available when needed. The book may be used as a text in a practical graduate course designed to prepare students for research work on defects in semiconductors or first-principles theory in materials science in general. The book also serves as a reference for the active theoretical researcher, or as a convenient guide for the experimentalist to keep tabs on their theorist colleagues.

It was a genuine pleasure to edit this volume. We are delighted with the contributions provided in a timely fashion by so many busy and accomplished people. We warmly thank all the contributors and hope to have the opportunity to share some nice wine(s) with all of them soon. After all,

When Ptolemy, now long ago,
Believed the Earth stood still,
He never would have blundered so
Had he but drunk his fill.
He'd then have felt it circulate
And would have learnt to say:
The true way to investigate
Is to drink a bottle a day.

(author unknown)
published in Augustus de Morgan's *A Budget of Paradoxes*, (1866).

Athens, Ohio, *David A. Drabold*
Lubbock, Texas *Stefan K. Estreicher*
February 2006

Contents

This book is dedicated to Manuel Cardona, who has done so much for the field of defects in semiconductors over the past decades, and convinced so many theorists to calculate beyond what they thought possible.

Preface

Semiconductor materials emerged after World War II and their impact on our lives has grown ever since. Semiconductor technology is, to a large extent, the art of defect engineering. Today, defect control is often done at the atomic level. Theory has played a critical role in understanding, and therefore controlling, the properties of defects.

Conversely, the careful experimental studies of defects in Ge, Si, then many other semiconductor materials have generated a huge database of measured quantities that allowed theorists to test their methods and approximations.

Dramatic improvement in methodology, especially density-functional theory, along with inexpensive and fast computers, has impedance matched the experimentalist and theorist in ways unanticipated before the late 1980s. As a result, the theory of defects in semiconductors has become quantitative in many respects. Today, more powerful theoretical approaches are still being developed. More importantly perhaps, the tools developed to study defects in semiconductors are now being adapted to approach many new challenges associated with nanoscience, a very long list that includes quantum dots, buckyballs and buckytubes, spintronics, interfaces, and many others.

Despite the importance of the field, there have been no modern attempts to treat the computational science of the field in a coherent manner within a single treatise. This is the aim of the present volume.

This book brings together several leaders in theoretical research on defects in semiconductors. Although the treatment is tutorial, the level at which the various applications are discussed is today's state-of-the-art in the field.

The book begins with a "big picture" view from Manuel Cardona, and continues with a brief summary of the historical development of the subject in Chap. 1. This includes an overview of today's most commonly used method to describe defects.

We have attempted to create a balanced and tutorial treatment of the basic theory and methodology in Chaps. 3–6. They include detailed discussions of the approximations involved, the calculation of electrically active levels, and extensions of the theory to finite temperatures. Two emerging electronic structure methodologies of special importance to the field are discussed in Chaps. 7 (quantum Monte-Carlo) and 8 (the GW method). Then come two chapters on molecular dynamics (MD). In Chap. 9, a combination of high-

level and approximate MD is developed, with applications to the dynamics of extended defect. Chapter 10 deals with semiempirical treatments of microstructures, including issues such as wafer bonding. The book concludes with studies of defects and their role in the photoresponse of topologically disordered (amorphous) systems.

The intended audience for the book is graduate students as well as advanced researchers in physics, chemistry, materials science, and engineering. We have sought to provide self-contained descriptions of the work, with detailed references available when needed. The book may be used as a text in a practical graduate course designed to prepare students for research work on defects in semiconductors or first-principles theory in materials science in general. The book also serves as a reference for the active theoretical researcher, or as a convenient guide for the experimentalist to keep tabs on their theorist colleagues.

It was a genuine pleasure to edit this volume. We are delighted with the contributions provided in a timely fashion by so many busy and accomplished people. We warmly thank all the contributors and hope to have the opportunity to share some nice wine(s) with all of them soon. After all,

When Ptolemy, now long ago,
Believed the Earth stood still,
He never would have blundered so
Had he but drunk his fill.
He'd then have felt it circulate
And would have learnt to say:
The true way to investigate
Is to drink a bottle a day.

(author unknown)
published in Augustus de Morgan's *A Budget of Paradoxes*, (1866).

Athens, Ohio, *David A. Drabold*
Lubbock, Texas *Stefan K. Estreicher*
February 2006

Foreword

Manuel Cardona

Max-Planck-Institut für Festkörperforschung,
70569 Stuttgart, Germany
M.Cardona@fkf.mpg.de

Man sollte sich mit Halbleitern nicht beschäftigen,
das sind Dreckeffekte –
wer weiss, ob sie wirklich existieren.
Wolfgang Pauli, 1931

1 Early History and Contents of the Present Volume

This volume contains a comprehensive description of developments in the field of defects in semiconductors that have taken place during the past two decades. Although the field of defects in semiconductors is at least 60 years old, it had to wait, in order to reach maturity, for the colossal increase in computer power that has more recently taken place, following the predictions of *Moore's* law [1]. The ingenuity of computational theorists in developing algorithms to reduce the intractable many-body problem of defect and host to one that can be handled with existing and affordable computer power has also played a significant role: much of it is described in the present volume. As computational power grew, the simplifying assumptions of these algorithms, some of them hard to justify, were reduced. The predictive accuracy of the new calculations then took a great leap forward.

In the early days, the real-space structure of the defect had to be postulated in order to get on with the theory and self-consistency of the electronic calculations was beyond reach. During the past two decades emphasis has been placed in calculating the real-space structure of defect plus host and achieving self-consistency in the electronic calculations. The results of these new calculations have been a great help to experimentalists groping to interpret complicated data related to defects. I have added up the number of references in the various chapters of the book corresponding to years before 1990 and found that they amount to only 25 % of the total number of references. Many of the remaining 75 % of references are actually even more recent, having been published after the year 2000. Thus one can say that the contents of the volume represent the state-of-the-art in the field. Whereas most of the chapters are concerned with defects in crystalline semiconductors, Chaps. 10 and 11 deal with defects in amorphous materials, in particular amorphous silicon, a field about which much less information is available.

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 1–10 (2007)

The three aspects of the defect problem, real-space structure, electronic structure and vibrational properties are discussed in the various Chapters of the book, mainly from the theoretical point of view. Defects break the translational symmetry of a crystal, a property that already made possible rather realistic calculations of the host materials half a century ago. Small crystals and clusters with a relatively small number of atoms (including impurities and other defects), have become useful to circumvent, in theoretical calculations, the lack of translational symmetry in the presence of defects or in amorphous materials. The main source of uncertainty in the state-of-the-art calculations remains the small number of cluster atoms imposed by the computational strictures. This number is often smaller than that corresponding to real-world samples, even including nanostructures. Clusters with a number of atoms that can be accommodated by extant computers are then repeated periodically so as to obtain a crystal lattice, with a supercell and a mini-Brillouin zone. Although these lattices do not exactly correspond to physical reality, they enable the use of k-space techniques and are instrumental in keeping computer power to available and affordable levels. Another widespread approach is to treat the cluster in real space after passivating the fictitious surface with hydrogen atoms or the like. When using these methods it is good practice to check convergence with respect to the cluster size by performing similar calculations for at least two sets of clusters with numbers of atoms differing, say, by a factor of two.

The epigraph above, attributed to Wolfgang Pauli, translates as *One should not keep busy with semiconductors, they are dirt effects – Who knows whether they really exist.* The 24 authors of this book, like many tens of thousands of other physicists and engineers, have fortunately not heeded Pauli's advice (given in 1931, 14 years before he received the Nobel Prize). Had they done so, not only would the world have missed a revolutionary and nowadays ubiquitous technology, but basic physical science would have lost some of the most fruitful, beautiful and successful applications of quantum mechanics. "Dreckeffekte" is often imprecisely translated as effects of dirt, i.e., as effects of impurities. However, effects of structural defects would also fall into the category of Dreckeffekte. In Pauli's days applications of semiconductors, including variation of resistivity through doping leading to photocells and rectifiers, had been arrived at purely empirically, through some sort of trial and error alchemy. I remember as a child using galena (PbS) detectors in crystal radio sets. I had lots of galena from various sources: some of it worked, some not but nobody seemed to know why. Sixty years later, only a few months ago, I was measuring PbS samples in order to characterize the number of carriers (of nonstoichiometric origin involving vacancies) and their type (n or p) so as to wrap up original research on this canonical material [2]. Today, GOOGLE lists 860 000 entries under the heading "defects in

semiconductors". The Web of Science (WoS) lists 3736 mentions in the title and abstract of source articles[1].

The modern science of defects in semiconductors is closely tied to the invention of the transistor at Bell Laboratories in 1948 (by *Bardeen, Shockley* and *Brattain* [3], Physics Nobel laureates for 1956). Early developments took place mainly in the United States, in particular at Bell Laboratories, the Lincoln Lab (MIT) and Purdue University. Karl Lark-Horovitz, an Austrian immigrant, started at Purdue a program to investigate the growth and doping (n- and p-type) of germanium and all sorts of electrical and optical properties of this element in crystalline form [4]. The initial motivation was the development of germanium detectors for radar applications. During the years 1928 till his untimely death in 1958 he built up the Physics Department at Purdue into the foremost center of academic semiconductor research. Work similar to that at Purdue for germanium was carried out at Bell Labs, also as a spinoff of the development of silicon rectifiers during World War II. At *Bell, Scaff* et al. [5] discovered that crystalline silicon could be made n- or p-type by doping with atoms of the fifth (P, As, Sb) or the third (B, Al, Ga, In) column of the periodic table, respectively. n-type dopants were called donors, p-type ones acceptors. *Pearson* and *Bardeen* performed a rather extensive investigation of the electrical properties of intrinsic and doped silicon [6]. These authors proposed the simplest possible expression for estimating the binding energy of the so-called hydrogenic energy levels of those impurities: The ionization energy of the hydrogen atom (13.6 eV) had to be divided by the square of the static dielectric constant ϵ ($\epsilon = 12$ for silicon) and multiplied by an effective mass (typical values $m^* \approx 0.1$) that simulated the presence of a crystalline potential. According to this Ansatz, all donors (acceptors) would have the same binding energy, a fact that we now know is only approximately true (see Fig. 5 of Chap. 3 for diamond).

The simple hydrogenic Ansatz applies to semiconductors with isotropic extrema, so that a unique effective mass can be defined (e.g., n-type GaAs). It does not apply to electrons in either Ge or Si because the conduction-band extrema are strongly anisotropic. The hydrogen-like Schrödinger equation can, however, be modified so as to include anisotropic masses, as apply to germanium and silicon [7]. The maximum of the valence bands of most diamond and zincblende-like semiconductors occurs at or very close to $k = 0$. It is fourfold degenerate in the presence of spin-orbit interaction and sixfold if such interaction is neglected [8]. The simple Schrödinger equation of the hydrogen atom must be replaced by a set of four coupled equations (with spin-orbit coupling) with effective-mass parameters to be empirically determined [9]. Extensive applications of Kohn's prescriptions were performed by several Italian theorists [10].

[1] A source article is one published in a Source Journal as defined by the ISI-Thomson Scientific. There are about 6000 such journals, including all branches of science.

In the shallow (hydrogenic) level calculations based on effective-mass Hamiltonians the calculated impurity eigenvalues are automatically referred to the corresponding band edges, thus obviating the need for using a marker, of the type discussed in Chap. 3 of this book. This marker was introduced in order to avoid errors inherent to the "first-principles" calculations, such as those related to the so-called "gap problem" found when using local-density functionals to represent many-body exchange and correlation. For a way to palliate this problem using the so-called GW approximation see Chap. 7, where defects at surfaces are treated.

We have discussed so far the electronic levels of shallow substitutional impurities. In this volume a number of other defects, such as vacancies, interstitial impurities, clusters, etc., will be encountered. Energy levels related to structural defects were first discussed by *Lark-Horovitz* and coworkers [11]. These levels were produced by irradiation with either deuterons, alpha particles or neutrons. After irradiation, the material became more p-type. It was thus postulated that the defect levels introduced by the bombardment were acceptors (vacancies?).

It was also discovered by *Lark-Horovitz* that neutron bombardment, followed by annealing in order to reduce structural damage, could be used to create electrically active impurities by nuclear transmutation [12]. The small amount of the ^{30}Si isotope ($\approx 4\,\%$) present in natural Si converts, by neutron capture, into radioactive ^{31}Si, which decays through β-emission into stable ^{31}P, a donor. This technique is still commercially used nowadays for producing very uniform doping concentrations.

Since Kohn–Luttinger perturbation theory predicts reasonably well the electronic levels of shallow impurities (except for the so-called central-cell corrections [13]) this book covers mainly deep impurity levels that not only are difficult to calculate for a given real-space structure but also require relaxation of the unperturbed host crystal around the defect. Among these deep levels, native defects such as vacancies and self-interstitials are profusely discussed. Most of these levels are related to transition-metal atoms, such as Mn, Cu, and Au (I call Au and Cu transition metals for obvious reasons). The solubility of these transition-metal impurities is usually rather low (less than $10^{15}\,\mathrm{cm}^{-3}$. Exceptions: $Cd_{1-x}Mn_xTe$ and related alloys). They can go into the host lattice either as substitutional or as interstitial atoms,[2] a point that can be clarified with EPR and also with ab initio total-energy calculations. These dopants were used in early applications in order to reduced the residual conductivity due to shallow levels (because of the fact that transition-metal impurities have levels close to the middle of the gap). One can even nowadays find in the market semi-insulating GaAs obtained by doping with chromium.

[2] The reader who tries to do a literature search for interstitial gold may be surprised by the existence of homonyms: Interstitial gold is important in the treatment of prostate cancer. It has, of course, nothing to do with our interstitial gold. See [14].

I remember having obtained in 1957 semi-insulating germanium and silicon (doped with either Mn or Au) with carrier concentrations lower than intrinsic (this makes a good exam question!). They were used for measurements of the low-frequency dielectric constants of these materials, in particular vs. temperature and pressure [15] while I was working at Harvard on my PhD under W. Paul.

Rough estimates of the positions of deep levels of many impurity elements in the gap of group IV and III–V semiconductors were obtained by *Hjalmarson* et al. [8, 16] using Green's function methods. In the case of GaAs and related materials, two kinds of defect complexes, involving structural changes and metastability have received a lot of attention because of technological implications: the so-called EL2 and DX centers. Searching the Web of Science for EL2 one finds 1055 mentions in abstracts and titles of source articles. Likewise 695 mentions are found for the DX centers. *Chadi* and coworkers have obtained theoretical predictions for the structure of these centers and their metastability [17, 18]. Although these theoretical models explain a number of observations related to these centers, there is not yet a general consensus concerning their structures.

An aspect of the defect problem that has not been dealt with explicitly in this volume is the errors introduced by using nonrelativistic Schrödinger equations, in particular the neglect of mass–velocity corrections and spin–orbit interaction (the latter, however, is explicitly included in the Kohn–Luttinger Hamiltonian, either in its 4×4 or its 6×6 version). Discrepancies between calculated and measured gaps are attributed to the "gap problem" inherent in the local-density approximation (LDA). However, already for relatively heavy atoms (Ge, GaAs) the mass–velocity correction decreases the s–like conduction levels and, together with the LDA gap problem. converts the semiconductor into a metal in the case of germanium. For GaAs it is stated several times in this volume that the LDA calculated gap is about half the experimental one. This is for a nonrelativistic Hamiltonian. Even a scalar relativistic one reduces the gap even further, to about 0.2 eV (experimental gap: 1.52 eV at 4 K) [19]. This indicates that the gap problem is more serious than previously thought on the basis of nonrelativistic LDA calculations.

Another relativistic effect is the spin-orbit coupling. For moderately heavy atoms such as Ge, Ga and As the spin-orbit splitting at the top of the valence bands (≈ 0.3 eV) is much larger than the binding energy of hydrogenic acceptors. Hence we can calculate the binding energies of the latter by solving the decoupled 4×4 ($J = 3/2$) and 2×2 ($J = 1/2$) effective-mass equations. This leads to two series of acceptor levels separated by a "spin-orbit" splitting basically equal to that of the band-edge states. In the case of silicon, however, the spin-orbit splitting at $k = 0$ ($\Delta = 0.044$ eV) is of the order of shallow-impurity binding energies. The impurity potential thus couples the $J = 3/2$ and $J = 1/2$ bands and the apparent spin-orbit splitting of the corresponding impurity series becomes smaller than that at the band edges [20, 21]. The difference between band-edge spin-orbit splitting ($\Delta = 0.014$ eV) and that of

the acceptor levels becomes even larger in diamond. Using a simple Green's functions technique and a Slater–Koster δ-function potential, the impurity-level splittings have been calculated and found to be indeed much smaller than $\Delta = 0.014\,\text{eV}$. This splitting depends strongly on the binding energy of the impurity.

Another aspect that has hardly been treated in the present volume (see, however, Chap. 10 for amorphous silicon) is the temperature dependence of the electronic energy levels that is induced by the electron–phonon interaction. Whenever this question appears in this volume, it is assumed that we are in the classical high-temperature limit, in which the corresponding renormalization of electronic gaps and states is proportional to temperature. At low temperatures, the electron–phonon interaction induces a zero-point renormalization of the electronic states that can be estimated from the measured temperature dependence. It is also possible to determine the zero-point renormalization of gaps by measuring samples with different isotopic compositions. The interested reader should consult the review by *Cardona* and *Thewalt* [22].

When an atom of the host lattice of a semiconductor has several stable isotopes (e.g., diamond, Si, Ge, Ga, Zn samples grown with natural material lose, strictly speaking, their translational symmetry. In the past 15 years a large number of semiconductors have been grown using isotopically pure elements (which have become available in macroscopic and affordable quantities after the fall of the Iron Curtain). A different isotope added to an isotopically pure sample can thus be considered as an impurity, probably the simplest kind of defect possible: Only the atomic mass of such an impurity differs from that of the host, the electronic properties remain nearly the same.[3] The main effect of isotope mass substitution is found in the vibrational frequencies of host as well as local vibrational modes: such frequencies are inversely proportional to the square root of the vibrating mass (see Chap. 4). Although this effect sounds rather trivial it often induces changes in phonon widths and in the zero-point anharmonic renormalizations (see [22]) that in some cases can be rather drastic and unexpected [23]. The structural relaxation around isotopic impurities is rather small. The main such effect corresponds to an increase of the lattice constant with increasing isotopic mass, about 0.015 % between ^{12}C and ^{13}C diamond. Its origin lies in the change in the zero-point renormalization of the lattice constant: ab initio calculations are available [24].

The third class of effects of the isotopic impurities refers to electronic states and energy gaps and their renormalization on account of the electron–phonon interaction. The zero-point renormalizations also vary as the inverse square root of the relevant isotopic mass. By measuring a gap energy at low temperatures for samples with two different isotopic masses, one can

[3] Except for the electron–phonon renormalization of the electronic states and gaps that is usually rather small. See [22].

extrapolate to infinite mass and thus determine the unrenormalized value of the gap. Renormalizations of around 60 meV have been found for Ge and Si. For diamond, however, this renormalization seems to be much larger [25], around 400 meV.[4] This large renormalization is a signature of strong electron–phonon interaction that seems to be responsible for the superconductivity recently observed in heavily boron doped (p-type) diamond (T_c higher than 10 K) [26,27]. Ab initio calculations of the electronic and vibronic structure of heavily boron doped diamond have been performed and used for estimating the critical temperature T_c [28].

2 Bibliometric Studies

In the previous section I have already discussed the number of times certain topics appear in titles, keywords and abstracts in *source journals* (about 6000 publications chosen by the ISI among $\approx 100\,000$, as those that contribute significantly to the progress of science). While titles go back to the present starting date of the source journal selection (the year 1900), abstracts and keywords have only been collected since 1990. In the Web of Science (WoS) one can completely eliminate the latter in order to avoid distortions but, for simplicity, I kept them in the qualitative survey presented here.

In this section, a more detailed bibliometric analysis will be performed using the WoS that draws on the citation index as the primary database. In order to get a feeling for the standing of the various contributors to this volume, we could simply perform a citations count (it can be done relatively easily within the WoS using the *cited reference mode*). However, a more tale-telling index has been recently suggested by *Hirsch* [29], the so-called h-index. This index is easily obtained for anyone with access to the WoS going back to the first publication of the authors under scrutiny (1974 for Nieminen and Shaw). How far back your access to the WoS goes depends on how much your institution is willing to pay to ISI-Thomson Scientific. The h-index is obtained by using the *general search* mode of the WoS and ordering the results of the search for a given individual according to the number of citations (there is a function key to order the author's contributions from most cited to less cited). You then go down the list till the order number of a paper equals its number of citations (you may have to take one more or less citation if equality does not exist). The number so obtained is the h-index. It rewards more continued, sustained well-cited publications rather than only a couple with a colossal number (such as those that deserve the Nobel Prize). Watch out for possible homonyms although, on the average, they appear seldom. They can be purged by hand if the number of terms is not too high. I had problems with homonyms only for five out of the 24 (excluding myself) contributors

[4] Theorists: beware (and be aware) of this large renormalization when comparing your fancy GW calculations of gaps with experimental data.

to this volume (Antonelli, Colombo, Hernández, Sanati and Shaw). I simply excluded them from the count.

The average h-factor of the remaining 19 authors is $h = 20$. Hirsch mentions in [29] that recently elected fellows of the American Physical Society have typically $h \approx 15$–20. Advancement of a physicist to full professor at a reputable US university corresponds to $h \geq 18$. The high average h already reveals the high standing of the authors of this book. In several cases, the authors involve a senior partner ($h \geq 20$, Chaps. 3, 4, 5, 7, 8, 10 and 11) and a junior colleague. I welcome this decision. It is a good procedure for introducing junior researchers to the intricacies and ordeals involved in writing a review article of such extent. In this connection, I should mention that the h-index is roughly proportional to the scientific age (counted from the first publication or the date of the PhD thesis). The values of h given above for faculty and NAS membership are appropriate to physicists and chemists. Biomedical scientists often have, everything else being equal, twice as large h-indices, whereas engineers and mathematicians (especially the latter) have much lower ones.

After having discussed the average h-index of our contributors, I would like to mention the range they cover without mentioning specific names.[5] The h-indices of our contributors cover the range $6 \leq h \leq 64$. Four very junior authors who have not yet had a chance of being cited have been omitted (one could have set $h = 0$ in their case). Hirsch mentions in his seminal article [29] that election to the National Academy of Sciences of the US is usually associated with $h = 45$. We therefore must have some *potential academicians* among our contributors.

Because of the ease in the use of the h-algorithm just described and its usefulness to evaluate the "impact" of a scientist's career, bibliometrists have been looking for other applications of the technique. Instead of people one can apply it to journals (provided they are not too large in terms of published articles), institutions, countries, etc. One has to keep in mind that the resulting h-number always reverts to an analysis of the citations of *individuals* that are attached to the investigated items (e.g., countries, institutions, etc.). One can also use the algorithm to survey the importance of keywords or title subjects. The present volume has 11 chapters and this gives it a certain (albeit small) statistical value to be of use in such a survey. We thus attach to each chapter title a couple of keywords and evaluate the corresponding h-index entering these under "topic" in the general search mode of the WoS. In the table below we list these words, the number of items we find for each set of them and the corresponding h-index. There is considerable arbitrariness in the procedure to choose the keywords but we must keep in mind that these applications are just exploratory and at their very beginning.

[5] Mentioning the h-indices of the authors, one by one, may be invidious. The interested reader with access to the WoS can do it by following the prescription given above.

We display in Table 1 the keywords we have assigned to the eleven chapters, the number of terms citing them and the corresponding h-index that weights them according to the number of times each citing term is cited. One can draw a number of conclusions from this table. Particularly interesting are the low values of n and h for empirical molecular dynamics, which probably signals the turn towards ab initio techniques. Amorphous semiconductors, including defect and the metastabilities induced by illumination plus possibly their applications to photovoltaics are responsible for the large values of n and h.

Table 1. Keywords assigned (somewhat arbitrarily) to each of the 11 chapters in the book together with the corresponding number n of source articles citing them in abstract, keywords or title. Also, Hirsch number h that can be assigned to each of the chapters according to the keywords

Chapter	Keyword (topic in WoS)	n	h
1	defects and semiconductors	3735	76
2	supercell calculations	165	27
3	Gaussian orbitals	190	27
4	dynamical matrix	231	26
5	free energy and defect	494	36
6	quantum Monte Carlo	2551	71
7	point defect and surface	426	38
8	defect and molecular dynamics	2023	67
9	empirical molecular dynamics	23	7
10	defect and amorphous	4492	77
11	light and amorphous	5747	87

References

[1] G. E. Moore: Electronics **38**, 114 (1965) 1
[2] R. Sherwin, R. J. H. Clark, R. Lauck, M. Cardona: Solid State Commun. **134**, 265 (2005) 2
[3] J. Bardeen, W. Brattain: Phys. Rev. **74**, 230 (1948) 3
[4] K. Lark-Horovitz, V. A. Johnson: Phys. Rev. **69**, 258 (1946) 3
[5] J. A. Scaff, H. C. Theuerer, E. E. Schumacher: J. Metals: Trans. Am. Inst. Mining Metall. Eng. **185**, 383 (1949) 3
[6] G. L. Pearson, J. Bardeen: Phys. Rev. **75**, 865 (1949) 3
[7] W. Kohn, J. M. Luttinger: Phys. Rev. **97**, 1721 (1975) 3
[8] P. Y. Yu, M. Cardona: *Fundamentals of Semiconductors*, vol. 3 (Springer, Berlin, Heidelberg 2005) 3, 5
[9] W. Kohn, D. Schechter: Phys. Rev. **99**, 1903 (1955) 3
[10] A. Baldereschi, N. Lipari: Phys. Rev. B **9**, 1525 (1974) 3
[11] W. E. Johnson, K. Lark-Horovitz: Phys. Rev. **76**, 442 (1949) 4

[12] K. Lark-Horovitz: Nucleon-bombarded semiconductors, in *Semiconducting Materials* (Butterworths, London 1950) p. 47 4
[13] W. Kohn: *Solid State Physics*, vol. 5 (Academic, New York 1957) 4
[14] Lannon, et al.: Br. J. Urol. **72**, 782 (1993) 4
[15] M. Cardona, W. Paul, H. Brooks: J. Phys. Chem. Solids **8**, 204 (1959) 5
[16] H. P. Hjalmarson, P. Vogl, D. J. Wolford, J. D. Dow: Phys. Rev. Lett. **44**, 810 (1980) 5
[17] D. J. Chadi, K. J. Chang: Phys. Rev. Lett. **60**, 2187 (1988) 5
[18] S. B. Chang, D. J. Chadi: Phys. Rev. B **42**, 7174 (1990) 5
[19] M. Cardona, N. E. Christensen, G. Fasol: Phys. Rev. B **38**, 1806 (1988) 5
[20] N. O. Lipari: Solid State Commun. **25**, 266 (1978) 5
[21] J. Serrano, A. Wysmolek, T. Ruf, M. Cardona: Physica B **274**, 640 (1999) 5
[22] M. Cardona, M. L. V. Thewalt: Rev. Mod. Phys. **77**, 1173 (2005) 6, 10
[23] J. Serrano, F. J. Manjón, A. H. Romero, F. Widulle, R. Lauck, M. Cardona: Phys. Rev. Lett. **90**, 055510 (2003) 6
[24] P. Pavone, S. Baroni: Solid State Commun. **90**, 295 (1994) 6
[25] M. Cardona: Science and technology of advanced materials, in press. See also [22] 7
[26] E. A. Ekimov, V. A. Sidorov, E. D. Bauer, N. N. Mel'nik, N. J. Curro, J. D. Thompson, S. M. Stishov: Nature **428**, 542 (2004) 7
[27] Y. Takano, M. Nagao, I. Sakaguchi, M. Tachiki, T. Hatano, K. Kobayashi, H. Umezawa, H. Kawarada: Appl. Phys. Lett. **85**, 2851 (2004) 7
[28] L. Boeri, J. Kortus, O. K. Andersen: Phys. Rev. Lett. **93**, 237002 (2004) 7
[29] J. E. Hirsch: Proc. Nat. Acad. Sci. (USA) **102**, 16569 (2005) 7, 8

Index

Defect Theory: An Armchair History

David A. Drabold and Stefan K. Estreicher

[1] Dept. of Physics and Astronomy, Ohio University, Athens, OH 45701
drabold@ohio.edu
[2] Physics Department, Texas Tech University, Lubbock, TX 79409-1051
stefan.estreicher@ttu.edu

Abstract. This introductory chapter begins with a summary of the developments of the theory of defects in semiconductors in the past 50 years. This is followed by an overview of single-particle methods and today's first-principles approach, rooted in density-functional theory. Much more detail about this theory and the approximations it involves is found in subsequent chapters. The last section discusses the various contributions to this book.

1 Introduction

The voluntary or accidental manipulation of the properties of materials by including defects has been performed for thousands of years. The most ancient example we can think of is well over 5000 years old. It happened when someone realized that adding trace amounts of tin to copper lowers the melting temperature, increases the viscosity of the melt, and results in a metal considerably harder than pure copper: bronze. This allowed the manufacture of a variety of tools, shields and weapons. Not long afterwards, the early metallurgists realized that sand mixed with a metal is relatively easy to melt and produces glass. The Ancient Egyptians discovered that glass beads of various brilliant colors can be obtained by adding trace amounts of specific transition metals, such as gold for red or cobalt for blue [1].

Defect engineering is not something new. However, materials whose mechanical, electrical, optical, and magnetic properties are almost entirely controlled by defects are relatively new: semiconductors [2, 3]. Although the first publication describing the rectifying behavior of a contact dates back to 1874 [4], the systematic study of semiconductors began only during World War II. The first task was to grow high-quality Ge (then Si) crystals, that is removing as many defects as possible. The second task was to manipulate the conductivity of the material by adding selected impurities that control the type and concentration of charge carriers. This involved theory to understand as quantitatively as possible the physics involved. Thus, theory has played a key role since the very beginning of this field. These early developments have been the subject of several excellent reviews [5–8].

For a long time, theory has been trailing the experimental work. Approximations at all levels were too drastic to allow quantitative predictions.

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 11–28 (2007)

Indeed, modeling a perfect solid is relatively easy since the system is periodic. High-level calculations can be done in the primitive unit cell. This periodicity is lost when a defect is present. The perturbation to the defect-free material is often large, in particular when some of the energy eigenvalues of the defect are in the forbidden gap, far from band edges. However, in the past decade or so, theory has become quantitative in many respects. Today, theorists often predict geometrical configurations, binding, formation, and various activation energies, charge and spin densities, vibrational spectra, electrical properties, and other observable quantities with sufficient accuracy to be useful to experimentalists and sometimes device scientists.

Furthermore, the theoretical tools developed to study defects in semiconductors can be easily extended to other areas of materials theory, including many fields of nanoscience. It is the need to understand the properties of defects in semiconductors, in particular silicon, that has allowed theory to develop as much as it has. One key reason for this was the availability of microscopic experimental data, ranging from electron paramagnetic resonance (EPR) to vibrational spectroscopy, photoluminescence (PL), or electrical data, all of which provided critical tests for theory at every step.

The word "defect" means a native defect (vacancy, self-interstitial, antisite, ...), an impurity (atom of a different kind from the host atoms), or any combination of those isolated defects: small clusters, aggregates, or even larger defect structures such as precipitates, interfaces, grain boundaries, surfaces, etc. However, nanometer-size defects play many important roles and are the building blocks of larger defect structures. Therefore, understanding the properties of defects begins at the atomic scale.

There are many examples of the beneficial or detrimental roles of defects. Oxygen and nitrogen pin dislocations in Si and allow wafers to undergo a range of processing steps without breaking [9]. Small oxygen precipitates provide internal gettering sites for transition metals, but some oxygen clusters are unwanted donors that must be annealed out [10]. Shallow dopants are often implanted. They contribute electrons to the conduction band or holes to the valence band. Native defects, such as vacancies or self-interstitials, promote or prevent the diffusion of selected impurities, in particular dopants. Self-interstitial precipitates may release self-interstitials that in turn promote the transient enhanced diffusion of dopants [11]. Transition-metal impurities are often associated with electron–hole recombination centers. Hydrogen [12], almost always present at various stage of device processing, passivates the electrical activity of dopants and of many deep-level defects, or forms extended defect structures known as platelets. Mg-doped GaN must be annealed at rather high temperatures to break up the $\{\mathrm{Mg}, \mathrm{H}\}$ complexes that prevent p-type doping [13]. Magnetic impurities such as Mn can render a semiconductor ferromagnetic. The list goes on.

Much of the microscopic information about defects comes from electrical, optical, and/or magnetic experimental probes. The electrical data are often obtained from capacitance techniques such as deep-level transient spec-

troscopy (DLTS). The sensitivity of DLTS is very high and the presence of defects in concentrations as low as $10^{11}\,\mathrm{cm}^{-3}$ can be detected. However, even in conjunction with uniaxial stress experiments, these data provide little or no elemental and structural information and, by themselves, are insufficient to identify the defect responsible for electrical activity. Local vibrational mode (LVM) spectroscopy, Raman, and Fourier transform infrared absorption (FTIR), often give sharp lines characteristic of the Raman- or IR-active LVMs of impurities lighter than the host atoms. When uniaxial stress, annealing, and isotope substitution studies are performed, the experimental data provide a wealth of critical information about a defect. This information can be correlated, e.g., with DLTS annealing data. However, Raman and FTIR are not as sensitive as DLTS. In the case of Raman, over $10^{17}\,\mathrm{cm}^{-1}$ defect centers must be present in the surface layer exposed to the laser. In the case of FTIR, some $10^{16}\,\mathrm{cm}^{-3}$ defect centers are needed, although much higher sensitivities have been obtained from multiple-internal reflection FTIR [14]. Photoluminescence is much more sensitive, sometimes down to $10^{11}\,\mathrm{cm}^{-1}$, but the spectra can be more complicated to interpret [15]. Finally, magnetic probes such as EPR are wonderfully detailed and a lot of defect-specific data can be extracted: identification of the element(s) involved in the defect and its immediate surrounding, symmetry, spin density maps, etc. However, the sensitivity of EPR is rather low, of the order of $10^{16}\,\mathrm{cm}^{-3}$. Further, localized gap levels in semiconductors often prefer to be empty or doubly occupied as most defect centers in semiconductors are unstable in a spin-$\frac{1}{2}$ state. The sample must be illuminated in order to create an EPR-active version of the defect under study [16–18].

This introductory Chapter contains brief reviews of the evolution of theory [19, 20] since its early days and of the key ingredients of today's state-of-the-art theory. It concludes with an overview of the content of this book.

2 The Evolution of Theory

The first device-related problem that required understanding was the creation of electrons or holes by dopants. These (mostly substitutional) impurities are a small perturbation to the perfect crystal and are well described by effective mass theory (EMT) [21]. The Schrödinger equation for the nearly-free charge carrier, trapped very close to a parabolic band edge, is written in hydrogenic form with an effective mass determined by the curvature of the band. The calculated binding energy of the charge carrier is that of a hydrogen atom but reduced by the square of the dielectric constant. As a result, the associated wavefunction is substantially delocalized, with an effective Bohr radius some 100 times larger than that of the free hydrogen atom.

EMT has been refined in a variety of ways [22] and provided a basic understanding of doping. However, it cannot be extended to defects that have energy eigenvalues far from band edges. These so-called "deep-level" defects

are not weak perturbations to the crystal and often involve substantial relaxations and distortions. The first such defects to be studied were the byproducts of radiation damage, a hot issue in the early days of the cold war. EPR data became available for the vacancy [16, 23] and the divacancy [24], in silicon (the Si self-interstitial has never been detected). Transition-metal (TM) impurities, which are common impurities and active recombination centers, have also been studied by EPR [25].

In most charge states, the undistorted vacancy (T_d symmetry) or divacancy (D_{3d} symmetry) is an orbital triplet or doublet, respectively, and therefore should undergo Jahn–Teller distortions. The EPR studies showed that this is indeed the case. Although interstitial oxygen, the most common impurity in Czochralski-grown Si, was known to be at a puckered bond-centered site [26, 27], it was not realized how much energy is involved in relaxations and distortions. It was believed that the chemistry of defects in semiconductors is well described in first order by assuming high-symmetry, undistorted, lattice sites. Relaxations and distortions were believed to be a second-order correction. The important issue then was to correctly predict trends in the spin densities and electrical activities of specific defects centers in order to explain the EPR and electrical data (see, e.g., [28, 29]). The critical importance of carefully optimizing the geometry around defects and the magnitudes of the relaxation energies were not fully realized until the 1980s [30, 31]. The host-atom displacements can be of several tenths of an Å, and the chemical rebonding can lead to energy changes as large as several eV (undistorted vs. relaxed structures).

The first theoretical tool used to describe localized defects in semiconductors involved *Green's functions* [2, 19, 32, 33]. These calculations begin with the Hamiltonian H_0 of the perfect crystal. Its eigenvalues give the crystal's band structure and the eigenfunctions are Bloch or Wannier functions. In principle, the defect-free host crystal is perfectly described. The localized defect is represented by a Hamiltonian H' that includes the defect potential V. The Green's function is $G(E) = 1/(E - H')$. Therefore, the perturbed energies E coincide with its poles. The new eigenvalues include the gap levels of the defect and the corresponding eigenfunctions are the defect wavefunctions. In principle, Green's functions provide an ideal description of the defect in its crystalline environment. In practice, there are many difficulties associated with the Hamiltonian, the construction of perfect-crystal eigenfunctions that can be used as a basis set for the defect calculation [34, 35], and the construction of the defect potential itself. This is especially true for those defects that induce large lattice relaxations and/or distortions.

The first successful Green's functions calculations for semiconductors date back to the late 1970s [36–38]. They were used to study charged defects [39], calculate forces [40–42], total energies [43, 44], and LVMs [45, 46]. These calculations also provided important clues about the role of native defects in impurity diffusion [47]. However, while Green's functions do provide a near-ideal description of the defect in a crystal, their implementation is difficult

and not very intuitive. Clusters or supercells are much easier to use and provide a physically and chemically appealing description of the defect and its immediate surroundings. Green's functions have mostly been abandoned since the mid-1980s, but a rebirth within the GW formalism [48] is now taking place (see the Chapter by *Schindlmayr* and *Scheffler*).

In order to describe the distortions around a vacancy, *Friedel* et al. [49] completely ignored the host crystal and limited their description to rigid linear combinations of atomic orbitals (LCAO). *Messmer* and *Watkins* [50,51] expanded this approach to linear combinations of dangling-bond states. These simple quantum-chemical descriptions provided a much-needed insight and a correct, albeit qualitative, explanation of the EPR data. Here, the defect was assumed to be so localized that the entire crystal could be ignored in zeroth order.

The natural extension of this work was to include a few host atoms around the defect, thus defining a *cluster*. These types of calculations were performed in real space with basis sets consisting of localized functions such as Gaussians or LCAOs. The dangling bonds on the surface atoms must be tied up in some way, most often with H atoms. However, without the underlying crystal and its periodicity, the band structure is missing and the defect's energy eigenvalues cannot be placed within a gap. Further, the finite size of the cluster artificially confines the wavefunctions. This affects charged defects the most, as the charge tends to distribute itself on the surface of the cluster. However, the local covalent interactions are well described.

The Schrödinger equation for a cluster containing a defect can be solved using almost any electronic-structure method. The early work was empirical or semiempirical, with heavily approximated quantum-chemical methods. At first, extended Hückel theory [52, 53] was used then self-consistent semiempirical Hartree–Fock: CNDO [54], MNDO [55], MINDO [56]. Geometries could be optimized, albeit often with symmetry assumptions. The methods suffered from a variety of problems such as cluster size and surface effects, basis-set limitations, lack of electron correlation, and the use of adjustable parameters. Their values are normally fitted to atomic or molecular data, and transferability is a big issue.

In order to bypass the surface problem, cyclic clusters have been designed, mostly in conjunction with semiempirical Hartree–Fock. Cyclic clusters can be viewed as clusters to which Born–von-Karman periodic boundary conditions are applied [57, 58]. These boundary conditions can be difficult to handle, in particular when 3- and 4-center interactions are included [59].

DeLeo and coworkers [60, 61] extensively used the scattering-Xα method in clusters to study trends for interstitial TM impurities and hydrogen–alkali-metal complexes. The results provided qualitative insight into these issues. Ultimately, the method proved difficult to bring to self-consistency and the rather arbitrarily defined muffin-tin spheres rendered it poorly suited to the calculation of total energies vs. atomic positions.

The method of partial retention of diatomic differential overlap [62, 63] (PRDDO) was first used for defects in diamond and silicon in the mid-1980s. It is self-consistent, contains no semiempirical parameters, and allows geometry optimizations to be performed without symmetry assumptions. Convergence is very efficient and relatively large clusters (44 host atoms) could be used. However, PRDDO is a minimal basis-set technique and ignores electron correlation. Its earliest success was to demonstrate [30, 31] the stability of bond-centered hydrogen in diamond and silicon. It was not expected at all that an impurity as light as H could indeed force a Si–Si bond to stretch by over 1 Å. Substantial progress in the theory of defects in semiconductors occurred in the mid-1980s with the combination of periodic supercells to represent the host crystal, ab-initio-type pseudopotentials [64–66] for the core regions, DF theory for the valence regions, and ab-initio molecular dynamics (MD) simulations [67, 68] for nuclear motion. This combination is now referred to as "first-principles" as opposed to "semiempirical". There are parameters in the theory. They include the size of the supercell, k-point sampling, type and size of the basis set, chosen by the user, as well as the parameters associated with the basis sets and pseudopotentials. However, these parameters and user inputs are not fitted to an experimental database. Instead, some are determined self-consistently, other are calculated from first principles or obtained from high-level atomic calculations. Note that the first supercell calculations were done in the 1970s in conjunction with approximate electronic-structure methods [69–71]. PRDDO was used to study cluster size and surface effects [72] and many defects (see, e.g., [73, 74]). It provided good input geometries for single-point ab-initio Hartree–Fock calculations (see, e.g., [75]). However, it suffered from the problems associated with all Hartree–Fock techniques, such as unreasonably large gaps and inaccurate LVMs. A number of research groups have used Hartree–Fock and post-Hartree–Fock techniques [76–78] to study defects in clusters, but these efforts have now been mostly abandoned.

Density-functional (DF) theory [79–82] with local basis sets (see, e.g., [83]) in large clusters allowed more quantitative predictions. The DF-based AIMPRO code [84, 85] uses Gaussian basis sets and has been applied to many defect problems (this code handles periodic supercells as well). In addition to geometries and energetics, rather accurate LVMs for light impurities can be predicted [86, 87]. Large clusters have been used [88, 89] to study the distortions around a vacancy or divacancy in Si. However, all clusters suffer from the surface problem and lack of periodicity.

3 A Sketch of First-Principles Theory

The theoretical approach known as "first principles" has proven to be a revolutionary tool to predict quantitatively some key properties of defects. An elementary exposition of the theory follows.

3.1 Single-Particle Methods: History

After the Born–Oppenheimer approximation is made, so that electronic and ionic degrees of freedom are separated, we face the time-independent many-electron problem [82]:

$$\left[-\hbar^2/2m \sum_j \nabla_j^2 - \sum_{j,l} Z_l e^2/|\boldsymbol{r}_j - \boldsymbol{R}_l| + 1/2 \sum_{j \neq j'} e^2/|\boldsymbol{r}_j - \boldsymbol{r}_{j'}| - E\right] \Psi = 0\,. \tag{1}$$

Here, $\boldsymbol{r}_j$ are electron coordinates, $\boldsymbol{R}_l$ and Z_l are positions and atomic numbers of the nuclei and E is the energy. It is worth reflecting on the remarkable simplicity of this equation: it is exact (nonrelativistically), and the meaning of each term is entirely transparent. In one of the celebrated legends of science, P. A. M. Dirac is said to have implied that chemistry was just an application of (1), though he also acknowledged that the equation was intractable. It is true that the quantum mechanics of the many-electron problem is beautifully and succinctly represented in (1).

Kohn gives an interesting argument [82] stating that even in principle, (1) is hopeless as a practical tool for calculation if the number of electrons exceeds 10^3 or so. His argument is a development of a paper of *Van Vleck* [90][1] and points out that it appears to be fundamentally impossible to obtain an approximate many-electron function Ψ with significant overlap with the "true" many-body wavefunction for large systems. He further points out that the sheer dimensionality of the problem rapidly makes it unrepresentable on any conceivable computer. Thus, a credible case can be made that for large systems *it does not even make sense to estimate* Ψ *directly.* Kohn has named this the "exponential wall". To some extent this is disconcerting, because of the simplicity of the *form* of (1), but it points to the need for new concepts if we are to make sense of solids – to say nothing of defects!

Empirical experience with solids also suggests that the unfathomable complexity of the many-body wavefunction is unnecessary. If all of the information contained in Ψ was really required for estimating the properties of solids that we care about, e.g., experimental observables, molecular physics would reach exhaustion with tiny molecules, and solid-state physics would never get off the ground at all. The fact is that many characteristics of materials are independent of system size. For example, in a macroscopic sample, the electronic density of states has the same form for a system with N atoms and an identically prepared one with $2N$ atoms. Yet Ψ is immeasurably more complex for the second system than the first. Thus, the additional complexity

[1] In this prescient paper on Heisenberg's theory of ferromagnetism, Van Vleck introduces what later was called the "Van Vleck Catastrophe", and emphasizes the fundamentally nonlocal character of quantum mechanics as expressed in (1), the associated factorial growth in complexity, and its dire implications for attempts to compute many-particle states for large systems.

of the $2N$ system ψ must be completely irrelevant to our observable, in this case the density of states.

The saturation of complexity of the preceding paragraph is connected to *Kohn's* "principle of nearsightedness" [91], which states that in fact quantum mechanics in the solid state is intrinsically local (*how local* depends sensitively on the system [92, 93]). The natural gauge of this locality is the decay of the density operator in the position representation: $\rho(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}|\hat{\rho}|\boldsymbol{x}'\rangle$: a function of $|\boldsymbol{x} - \boldsymbol{x}'|$, decaying as a power law in metals and exponentially in systems with a gap. For systems with a gap, the exponential fall off enables accurate calculations of all local properties by undertaking a calculation in a finite volume determined by the rate of decay. The decay in semiconductors and insulators can be exploited to produce efficient order-N methods for computing total energies and forces, with computational cost scaling linearly with the number of atoms or electrons [94].

3.2 Direct Approaches to the Many-Electron Problem

While this book emphasizes single-particle methods, there is one important exception. *Needs* shows in his Chapter in this volume that remarkable progress can be made for the computation of *expectation values* of observables using quantum Monte Carlo methods, with no essential approximations to (1). This is a promising class of methods that offers the most accurate calculations available today for complex systems. Several groups are advancing these methods, and even the stochastic calculation of forces is becoming possible. One can be certain that quantum Monte Carlo will play an important role in systems needing the most accurate calculations available, and certainly this is the case for the theory of defects.

3.3 Hartree and Hartree–Fock Approximations

In 1928, *Hartree* [95] started with (1) with a view to extracting useful single-particle equations from it. He used the variational principle for the ground-state wavefunction adopting a simple product trial function: $\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \psi_1(\boldsymbol{x}_1)\psi_2(\boldsymbol{x}_2)\ldots\psi_n(\boldsymbol{x}_n)$. The product *ansatz* did not enforce Fermion antisymmetry requirements; this was built in with a Slater determinant ansatz as proposed by *Fock* in 1930 [96]: the Hartree–Fock method.

These methods map the many-body ground-state problem onto a set of challenging single-particle equations. The structure of the Hartree equations is a Schrödinger equation with a potential depending upon all the orbitals:

$$-\hbar^2/2m\nabla^2\psi_l + V_{\rm eff}\psi_l = \varepsilon_l\psi_l\,, \tag{2}$$

where the effective potential $V_{\rm eff}(\boldsymbol{r}) = \sum_{j\neq\ell}\int \mathrm{d}^3r'\psi_j^*(\boldsymbol{r}')\psi_\ell(\boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'|$, and the sum is over occupied states. This appealing equation prescribes an effective Coulomb field for electron l arising from all of the other electrons. Since

computing the potential requires knowledge of the other wavefunctions, the equation must be solved self-consistently with a scheme of iteration. While the equation is intuitive, it is *highly* approximate. Curiously, it will turn out that the equations of density-functional theory have the mathematical structure of (2), but with quite a different (and far more predictive) V_{eff} (see (7) below), derived from a very different point of view.

The Hartree–Fock approximation includes that part of the exchange energy implied by the exclusion principle and has well-known analytic problems at the Fermi surface in metals [94] and a tendency to exaggerate charge fluctuations at atomic sites in molecules or solids [97]. In molecular calculations, the correlation energy (roughly speaking, what is missing from Hartree–Fock) can be estimated in perturbation theory. This is computationally expensive: the most popular fourth-order perturbation theory (MP4) scales as n^3N^4, where n is the number of electrons and N the number of orbitals (basis set size). Such methods are important for molecular systems, but challenging to apply to condensed systems.

3.4 Density-Functional Theory

3.4.1 Thomas–Fermi Model

Not long after the dawn of quantum mechanics, *Thomas* and *Fermi* [98, 99] suggested a key role for the electron-density distribution as the determiner of the total energy of an inhomogeneous electron gas. This was the first serious attempt to express the energy as a functional of the electron density ρ:

$$E[\rho] = \int \mathrm{d}^3 r V(\boldsymbol{r})\rho(\boldsymbol{r}) + e^2/2 \int \mathrm{d}^3 r\, \mathrm{d}^3 r' \rho(\boldsymbol{r})\rho(\boldsymbol{r}')/|\boldsymbol{r} - \boldsymbol{r}'| + \alpha/m \times \int \mathrm{d}^3 r \rho^{5/3}(\boldsymbol{r})\,. \quad (3)$$

Here, α is a numerical constant, m and e are the electron mass and charge respectively, and V is an external potential. The first two terms are evidently obtained from classical electrostatics. Quantum mechanics appears only in the third term, derived from the kinetic energy of a homogeneous electron gas. Already this equation is making a "local-density approximation", forming an estimate for the inhomogeneous electron kinetic energy from the result from a homogeneous gas; an approximation that is expected to succeed if $|\nabla\rho|/\rho$ is sufficiently small. Variation of (3) with respect to ρ with fixed electron number leads to the coupled self-consistent Thomas–Fermi equations (see, e.g., the treatment by *Fulde* [97]). The coarse treatment of the kinetic energy greatly limits the predictive power of the method – for example, the shell structure of atoms is not predicted [100]. However, the Thomas–Fermi model is conceptually a density-functional theory.

3.4.2 Modern Density-Functional Theory

For atomistic force calculations on solids, the overwhelming method of choice is density-functional theory, due to *Kohn*, *Hohenberg* and *Sham* [101, 102]. This Chapter only sketches the concepts in broad outline, as a detailed treatment focused on defect calculations is available in this book (*Niemenen* Chapter). For additional discussion we strongly recommend the books of *Martin* (see [94]) and *Fulde* [97].

The following statements embody the foundation of zero-temperature density-functional theory:

1. The ground-state energy of a many-electron system is a functional of the electron density $\rho(\boldsymbol{x})$:

$$E[\rho] = \int \mathrm{d}^3 x V(\boldsymbol{x})\rho(\boldsymbol{x}) + F[\rho]\,, \tag{4}$$

   where V is an external potential (due, for example, to interaction with ions, external fields, e.g., *not* with electrons), and $F[\rho]$ is a *universal* functional of the density. The trouble is that $F[\rho]$ is not exactly known, though there is continuing work to determine it. The practical utility of this result derives from:
2. The functional $E[\rho]$ is minimized by the true ground-state electron density.

It remains to estimate the functional $F[\rho]$, which in conjunction with the variational principle 2, enables real calculations. To estimate $F[\rho]$, the usual procedure is to note that we already know some of the major contributions to $F[\rho]$, and decompose the functional in the form:

$$F[\rho] = e^2/2 \int \mathrm{d}^3 x\, \mathrm{d}^3 x' \rho(\boldsymbol{x})\rho(\boldsymbol{x}')/|\boldsymbol{x} - \boldsymbol{x}'| + T_{\mathrm{ni}}(\rho) + E_{\mathrm{xc}}(\rho)\,. \tag{5}$$

Here, the integral is just the electrostatic (Hartree) interaction of the electrons, T_{ni} is the kinetic energy of a *noninteracting* electron gas of density ρ, and $E_{\mathrm{xc}}(\rho)$ is yet another unknown functional, the exchange-correlation functional, which includes nonclassical effects of the interacting electrons. Equation (5) is difficult to evaluate directly in terms of ρ, because of the term T_{ni}. Thus, one introduces single-electron orbitals $|\chi_i\rangle$, for which $T_{\mathrm{ni}} = \sum_{i\mathrm{occ}} \langle\chi_i| - \hbar^2/2m\nabla^2|\chi_i\rangle$, and $\rho = \sum_{i\mathrm{occ}} |\chi_i(\boldsymbol{x})|^2$ is the charge density of the physically relevant *interacting* system. The value of this decomposition is that $E_{\mathrm{xc}}(\rho)$ is smooth and reasonably slowly varying, and therefore a functional that we can successfully approximate: we have included the most difficult and rapidly varying parts of F in T_{ni} and the Hartree integral, as can be seen from essentially exact many-body calculations on the homogeneous electron gas [103]. The Hartree and noninteracting kinetic energy terms are

easy to compute if one makes the local-density approximation, that is assuming that the electron density is *locally* uniform. With information about the homogeneous electron gas, the functional (5) is fully specified.

With noninteracting orbitals $|\chi_i\rangle$, (with $\rho(\boldsymbol{x}) = 2\sum_{i\mathrm{occ}} |\langle \boldsymbol{x}|\chi_i\rangle|^2$), then the minimum principle plus the constraint that $\langle\chi_i|\chi_j\rangle = \delta_{ij}$ can be translated into an eigenvalue problem for the $|\chi_i\rangle$:

$$\{-\hbar^2\nabla^2/2m + V_{\mathrm{eff}}[\rho(\boldsymbol{x})]\}|\chi_i\rangle = \varepsilon_i|\chi_i\rangle\,, \tag{6}$$

where the effective density-dependent (in practical calculations, orbital-dependent) potential V_{eff} is:

$$V_{\mathrm{eff}}[\rho(\boldsymbol{x})] = V(\boldsymbol{x}) + e^2\int \mathrm{d}^3x'\rho(\boldsymbol{x}')/|\boldsymbol{x}-\boldsymbol{x}'| + \delta\varepsilon_{\mathrm{xc}}/\delta\rho\,. \tag{7}$$

In this equation $\varepsilon_{\mathrm{xc}}$ is the parameterized exchange-correlation energy density from the homogeneous electron gas.

The quantities to be considered as physical in local-density functional calculations are: the total energy (electronic or system), the ground-state electronic charge density $\rho(\boldsymbol{x})$, and related ground-state properties like the forces. In particular, it is tempting to interpret the $|\chi_i\rangle$ and ε_i as genuine electronic eigenstates and energies, and indeed this can often be useful. Such identifications are not rigorous [94]. It is instructive to note that the starting point of density-functional theory was to depart from the use of orbitals and formulate the electronic-structure problem rigorously in terms of the electron density ρ; yet a practical implementation (which enables an accurate estimate of the electronic kinetic energy) led us immediately back to orbitals! This illustrates why it would be very worthwhile to know $F(\rho)$, or at least the kinetic energy functional since we would then have a theory with a structure close to the Thomas–Fermi form and would therefore be able to seek *one* function ρ rather than the cumbersome collection of orthonormal $|\chi_i\rangle$.

The initiation of modern first-principles theory and its development into a standard method with widespread application was due to *Car* and *Parrinello* [67]. They developed a powerful method for simultaneously solving the electronic problem and evolving the positions of the ions. The method is usually applied to plane-wave basis sets, though the key ideas are independent of the choice of representation. One of the early applications to defects in silicon was the diffusion of bond-centered hydrogen [104]. An alternative ab-initio approach to MD simulations, based on a tight-binding perspective, was proposed by *Sankey* and *Niklewski* [68]. Their basis sets consist of pseudoatomic orbitals with $s, p, d, \ldots$ symmetry. The wavefunctions are truncated beyond some radius and renormalized. The early version of this code was not self-consistent and was restricted to minimum basis sets. A more recent version [105] is self-consistent and can accommodate expanded and polarized basis sets. This is also the case for the highly flexible SIESTA code [106,107] (Spanish initiative for the electronic structure with thousands

of atoms). Although basis sets of local orbitals (typically, LCAOs) are highly intuitive and allow population analysis and other chemical information to be calculated, they are less complete than plane-wave basis sets. The latter can easily be checked for convergence. On the other hand, when an atom such as Si is given two sets of $3s$ and $3p$ plus a set of $3d$ orbitals, the basis set is sufficient to describe quite well virtually all the chemical interactions of this element, as the contribution of the $n = 4$ shell of Si is exceedingly small, except such under extreme conditions that a ground-state theory is not capable of handling anyway. Many details of the implementation of density-functional methods are given in the contribution of *Nieminen* in his Chapter in this volume.

The power of these methods is that they yield parameter-free estimates for the structure of defects and even topologically disordered systems, provide accurate estimates of total energies, formation energies, vibrational states, defect dynamics, and with suitable *caveats* information about electronic structure, defect levels, localization, etc. There are many subtle aspects to their applications to defect physics, partly because high accuracy is often required in such studies.

4 The Contributions

In his Chapter in this volume, *Nieminen* carefully discusses the use of periodic boundary conditions in supercell calculations, detailing both strengths and weaknesses, and other basic features of these calculations. In their Chapter, *Goss* and coworkers discuss the marker method to extract electronic energy levels, and include several important applications. In their Chapter, *Estreicher* and *Sanati* describe the calculation and remarkable utility of vibrational modes in systems containing defects, including novel analysis of finite-temperature properties of defects. Then, in their Chapter, *Hernandez* and coworkers work out a proper theory of free energies and phase diagrams for semiconductors. *Needs* describes in his Chapter the most rigorous attack of the book on the quantum many-body problem, as Needs explores quantum Monte Carlo methods and their promise for defects in semiconductors. *Schindlmayr* and *Scheffler* describe the theory and application of self-energy corrected density-functional theory, the "GW" approximation in their Chapter. Like the work of Needs, this technique has predictive power beyond DFT. *Csanyi* and coworkers present a multiscale modeling approach in their Chapter with applications. *Scheerschmidt* offers a comprehensive view of molecular dynamics in his Chapter, focusing on empirical methods, though much of his Chapter is applicable to first-principles simulation as well. In their Chapter, *Drabold* and *Abtew* discuss defects in amorphous semiconductors with a special emphasis on hydrogenated amorphous silicon. Last but not least, in their Chapter, *Simdyankin* and *Elliott* write on the theory of light-induced effects

in amorphous materials, an area of great basic and practical interest, which nevertheless depends very much upon the defects present in the material.

References

[1] J. L. Mass, M. T. Wypyski, R. E. Stone: Mater. Res. Soc. Bull. **26**, 38 (2001) 11
[2] A. M. Stoneham: *Theory of Defects in Solids* (Clarendon, Oxford 1985) 11, 14
[3] H. J. Queisser, E. E. Haller: Science **281**, 945 (1998) 11
[4] F. Braun: Ann. Phys. Chem. (Leipzig) **153**, 556 (1874) 11
[5] A. B. Fowler: Phys. Today **46**, 59 (1993) 11
[6] F. Seitz: Phys. Today **48**, 22 (1995) 11
[7] M. Riordan, I. Hoddeson: Phys. Today **50**, 42 (1997) 11
[8] I. M. Ross: Phys. Today **50**, 34 (1997) 11
[9] F. Shimura (Ed.): *Oxygen in Silicon*, Semicond. Semimet. **42** (Academic, Boston 1994) 12
[10] R. Jones (Ed.): *Early Stages of Oxygen Precipitation in Silicon* (Kluwer, Dordrecht 1996) 12
[11] R. Scholtz, U. Gösele, J. Y. Y. Huh, T. Y. Tan: Appl. Phys. Lett. **72**, 200 (1998) 12
[12] S. K. Estreicher: Mater. Sci. Eng. R **14**, 319 (1995) 12
[13] S. J. Pearton: *GaN and Related Materials* (Gordon and Breach, Amsterdam 1997) p. 333 12
[14] F. Jiang, M. Stavola, A. Rohatgi, D. Kim, J. Holt, H. Atwater, J. Kalejs: Appl. Phys. Lett. **83**, 931 (2003) 13
[15] G. Davies: Phys. Rep. **176**, 83 (1989) 13
[16] G. D. Watkins: *Deep Centers in Semiconductors* (Gordon and Breach, New York 1986) p. 147 13, 14
[17] Y. V. Gorelkinskii, N. N. Nevinnyi: Physica B **170**, 155 (1991) 13
[18] Y. V. Gorelkinskii, N. N. Nevinnyi: Mater. Sci. Eng. B. **36**, 133 (1996) 13
[19] S. T. Pantelides (Ed.): *Deep Centers in Semiconductors* (Gordon and Breach, New York 1986) 13, 14
[20] S. K. Estreicher: Mater. Today **6**, 26 (2003) 13
[21] W. Kohn: Solid State Phys. **5**, 257 (1957) 13
[22] S. T. Pantelides: Rev. Mod. Phys. **50**, 797 (1978) 13
[23] G. D. Watkins: *Radiation Damage in Semiconductors* (Dunod, Paris 1964) p. 97 14
[24] G. D. Watkins, J. W. Corbett: Phys. Rev. **138**, A543 (1965) 14
[25] G. W. Ludwig, H. H. Woodbury: Solid State Phys. **13**, 223 (1962) 14
[26] W. Kaiser, P. H. Keck, C. F. Lange: Phys. Rev. **101**, 1264 (1956) 14
[27] J. W. Corbett, R. S. McDonald, G. D. Watkins: J. Phys. Chem. Solids **25**, 873 (1964) 14
[28] G. G. DeLeo, G. D. Watkins, W. Beal Fowler: Phys. Rev. B **25**, 4972 (1982) 14
[29] A. Zunger, U. Lindefelt: Phys. Rev. B **26**, 5989 (1982) 14

[30] T. L. Estle, S. K. Estreicher, D. S. Marynick: Hyp. Inter. **32**, 637 (1986) 14, 16
[31] T. L. Estle, S. K. Estreicher, D. S. Marynick: Phys. Rev. Lett. **58**, 1547 (1987) 14, 16
[32] G. F. Koster, J. C. Slater: Phys. Rev. **95**, 1167 (1954) 14
[33] J. Callaway: J. Math. Phys. **5**, 783 (1964) 14
[34] J. Callaway, A. J. Hughes: Phys. Rev. **156**, 860 (1967) 14
[35] J. Callaway, A. J. Hughes: Phys. Rev. **164**, 1043 (1967) 14
[36] J. Bernholc, S. T. Pantelides: Phys. Rev. B **18**, 1780 (1978) 14
[37] J. Bernholc, N. O. Lipari, S. T. Pantelides: Phys. Rev. Lett. **41**, 895 (1978) 14
[38] G. A. Baraff, M. Schlüter: Phys. Rev. Lett. **41**, 892 (1978) 14
[39] G. A. Baraff, E. O. Kane, M. Schlüter: Phys. Rev. Lett. **43**, 956 (1979) 14
[40] M. Scheffler, J. P. Vigneron, G. B. Bachelet: Phys. Rev. Lett. **49**, 1756 (1982) 14
[41] U. Lindefelt: Phys. Rev. B **28**, 4510 (1983) 14
[42] U. Lindefelt, A. Zunger: Phys. Rev. B **30**, 1102 (1984) 14
[43] G. A. Baraff, M. Schlüter: Phys. Rev. B **28**, 2296 (1983) 14
[44] R. Car, P. J. Kelly, A. Oshiyama, S. T. Pantelides: Phys. Rev. Lett. **52**, 1814 (1984) 14
[45] D. N. Talwar, M. Vandevyver, M. Zigone: J. Phys. C **13**, 3775 (1980) 14
[46] R. M. Feenstra, R. J. Hauenstein, T. C. McGill: Phys. Rev. B **28**, 5793 (1983) 14
[47] R. Car, P. J. Kelly, A. Oshiyama, S. T. Pantelides: Phys. Rev. Lett. **54**, 360 (1985) 14
[48] F. Aryasetiawan, O. Gunnarsson: Rep. Prog. Phys. **61**, 237 (1998) 15
[49] J. Friedel, M. Lannoo, G. Leman: Phys. Rev. **164**, 1056 (1967) 15
[50] R. P. Messmer, G. D. Watkins: Phys. Rev. Lett. **25**, 656 (1970) 15
[51] R. P. Messmer, G. D. Watkins: Phys. Rev. B **7**, 2568 (1973) 15
[52] F. P. Larkins: J. Phys. C: Solid State Phys. **4**, 3065 (1971) 15
[53] R. P. Messmer, G. D. Watkins: *Proc. Reading Conf. on Radiation Damage and Defects in Semiconductors* (Insitute of Physics, Bristol 1973) p. 255 15
[54] A. Mainwood: J. Phys. C: Solid State Phys. **11**, 2703 (1978) 15
[55] J. W. Corbett, S. N. Sahu, T. S. Shi, L. C. Snyder: Phys. Lett. **93A**, 303 (1983) 15
[56] P. Deák, L. C. Snyder: Radiat. Eff. Def. Solids **111–112**, 77 (1989) 15
[57] P. Deák, L. C. Snyder: Phys. Rev. B **36**, 9619 (1987) 15
[58] P. Deák, L. C. Snyder, J. W. Corbett: Phys. Rev. B **45**, 11612 (1992) 15
[59] J. Miró, P. Deák, C. P. Ewels, R. Jones: J. Phys.: Condens. Matter **9**, 9555 (1997) 15
[60] G. G. DeLeo, G. D. Watkins, W. B. Fowler: Phys. Rev. B **23**, 1819 (1981) 15
[61] G. G. DeLeo, W. B. Fowler, G. D. Watkins: Phys. Rev. B **29**, 1851 (1984) 15
[62] T. A. Halgren, W. N. Lipscomb: J. Chem. Phys. **58**, 1569 (1973) 16
[63] D. S. Marynick, W. N. Lipscomb: Proc. Nat. Acad. Sci. (USA) **79**, 1341 (1982) 16
[64] D. R. Hamann, M. Schlüter, C. Chiang: Phys. Rev. Lett. **43**, 1494 (1979) 16

[65] G. B. Bachelet, D. R. Hamann, M. Schlüter: Phys. Rev. B **26**, 4199 (1982) 16
[66] L. Kleinman, D. M. Bylander: Phys. Rev. Lett. **48**, 1425 (1982) 16
[67] R. Car, M. Parrinello: Phys. Rev. Lett. **55**, 2471 (1985) 16, 21
[68] O. Sankey, D. J. Niklewski: Phys. Rev. B **40**, 3979 (1989) 16, 21
[69] A. Zunger, A. Katzir: Phys. Rev. B **11**, 2378 (1975) 16
[70] S. G. Louie, M. Schlüter, J. R. Chelikowsky, M. L. Cohen: Phys. Rev. B **13**, 1654 (1976) 16
[71] W. Pickett, M. L. Cohen, C. Kittel: Phys. Rev. B **20**, 5050 (1979) 16
[72] S. K. Estreicher, A. K. Ray, J. L. Fry, D. S. Marynick: Phys. Rev. Lett. **55**, 1976 (1985) 16
[73] S. K. Estreicher: Phys. Rev. B **41**, 9886 (1990) 16
[74] S. K. Estreicher: Phys. Rev. B **41**, 5447 (1990) 16
[75] S. K. Estreicher: Phys. Rev. B **60**, 5375 (1999) 16
[76] A. A. Bonapasta, A. Lapiccirella, N. Tomassini, M. Capizzi: Phys. Rev. B **36**, 6228 (1987) 16
[77] E. Artacho, F. Ynduráin: Solid State Commun. **72**, 393 (1989) 16
[78] R. Luchsinger, P. F. Meier, N. Paschedag, H. U. Suter, Y. Zhou: Philos. Trans. Roy. Soc. Lond. A **350**, 203 (1995) 16
[79] P. Hohenberg, W. Kohn: Phys. Rev. **136B**, 468 (1964) 16
[80] W. Kohn, L. J. Sham: Phys. Rev. **140A**, 1133 (1965) 16
[81] L. J. Sham, W. Kohn: Phys. Rev. **145**, 561 (1966) 16
[82] W. Kohn: Rev. Mod. Phys. **71**, 1253 (1999) 16, 17
[83] M. Saito, A. Oshiyama: Phys. Rev. B **38**, 10711 (1988) 16
[84] R. Jones, A. Sayyash: J. Phys. C **19**, L653 (1986) 16
[85] R. Jones, P. R. Briddon: The ab-initio cluster method and the dynamics of defects in semiconductors, in M. Stavola (Ed.): *Identification of Defects in Semiconductors*, vol. 51A, Semicond. Semimet. (Academic, Boston 1998) Chap. 6, p. 287 16
[86] J. D. Holbech, B. Bech Nielsen, R. Jones, P. Sitch, S. Öberg: Phys. Rev. Lett. **71**, 875 (1993) 16
[87] R. Jones, B. J. Coomer, P. R. Briddon: J. Phys.: Condens. Matter **16**, S2643 (2004) 16
[88] S. Ogüt, J. R. Chelikowsky: Phys. Rev. Lett. **83**, 3852 (1999) 16
[89] S. Ogüt, J. R. Chelikowsky: Phys. Rev. B **64**, 245206 (2001) 16
[90] J. H. Van Vleck: Phys. Rev. **49**, 232 (1936) 17
[91] W. Kohn: Phys. Rev. Lett. **76**, 3168 (1996) 18
[92] X. P. Li, R. W. Nunes, D. Vanderbilt: Phys. Rev. B **47**, 10891 (1993) 18
[93] S. N. Taraskin, D. A. Drabold, S. R. Elliott: Phys. Rev. Lett. **88**, 196405 (2002) 18
[94] R. M. Martin: *Electronic Structure, Basic Theory and Practical Applications* (Cambridge University Press, Cambridge 2004) 18, 19, 20, 21
[95] D. R. Hartree: Proc. Cambridge Philos. Soc. **24**, 89 (1928) 18
[96] F. Fock: Z. Phys. **61**, 126 (1930) 18
[97] P. Fulde: *Electron Correlations in Molecules and Solids* (Springer, Berlin, Heidelberg 1993) 19, 20
[98] E. Fermi: Z. Phys. **48**, 73 (1928) 19
[99] L. H. Thomas: Proc. Cambridge Philos. Soc. **23**, 542 (1927) 19

[100] R. G. Parr, W. Yang: *Density-Functional Theory of Atoms and Molecules* (Clarendon, Oxford 1989) 19
[101] P. Hohenberg, W. Kohn: Phys. Rev. **136**, B864 (1964) 20
[102] W. Kohn, L. J. Sham: Phys. Rev. **140**, A1133 (1965) 20
[103] D. M. Ceperley, G. J. Alder: Phys. Rev. Lett. **45**, 566 (1980) 20
[104] F. Buda, G. L. Chiarotti, R. Car, M. Parrinello: Phys. Rev. Lett. **63**, 4294 (1989) 21
[105] A. A. Demkov, J. Ortega, O. F. Sankey, M. P. Grumbach: Phys. Rev. B **53**, 10441 (1995) 21
[106] D. Sánchez-Portal, P. Ordejón, E. Artacho, J. M. Soler: Int. J. Quant. Chem. **65**, 453 (1997) 21
[107] E. Artacho, D. Sánchez-Portal, P. Ordejón, A. García, J. M. Soler: Phys. Stat. Sol. (b) **215**, 809 (1999) 21

Index

Supercell Methods for Defect Calculations

Risto M. Nieminen

COMP/Laboratory of Physics, Helsinki University of Technology, POB 1100, 02015 HUT, Finland
rni@fyslab.hut.fi

Abstract. Periodic boundary conditions enable fast density-functional-based calculations for defects and their complexes in semiconductors. Such calculations are popular methods to estimate defect energetics, structural parameters, vibrational modes and other physical characteristics. However, the periodicity introduces spurious defect–defect interactions and dispersion of the defect-induced electronic states. For charged defects, compensating background charging has to be introduced to avoid electrostatic divergences. These factors, together with the intrinsic limitation of standard density-functional theory for accurate estimation of semiconducting gaps, pose challenges for quantitatively accurate and properly controlled calculations. This chapter discusses these issues, including point sampling, electrostatic (Madelung) corrections, valence-band (reference energy) alignment, and other finite-size effects in supercell calculations. These are important for reliable estimation of formation and migration energies as well as ionization-level prediction. Moreover, the chapter discusses the various ways of generating transferable pseudopotentials, the choice of the exchange-correlation functional, and other topics related to total-energy calculations. Methods to calculate excitation energies and other spectroscopic properties as well as atomic motions are also discussed. Examples of applications of the supercell methods to a few selected semiconductor defects are presented.

1 Introduction

An important role for theory and computation in studies of defects in semiconductors is to provide means for reliable, robust *interpretation of defect fingerprints* observed by many different experimental techniques, such as deep-level transient spectroscopy, various methods based on positron annihilation (PA), local-vibrational-mode (LVM) spectroscopy, and spin-resonance techniques [1]. High-resolution studies can provide a dizzying zoo of defect-related features, the interpretation and assignment of which requires accurate calculations of both the defect electronic and atomic properties as well as quantitative theory and computation also for the probe itself.

Materials processing can also draw significant advantages of the predictive computational studies. The kinetics of defect diffusion and reactions during thermal treatment depend crucially on the atomic-scale energetics (formation energies, migration barriers, etc.), derived from the bond-breaking,

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 29–68 (2007)

bond-making and other chemical interactions between the atoms in question. Accurately calculated estimates for defect and impurity energetics can considerably facilitate the design of strategies for doping and thermal processing to achieve the desired materials properties. Parameter values calculated atomistically, from the electronic degrees of freedom, can be fed into *multiscale modeling* methods for defect evolution, such as kinetic Monte Carlo, cellular automaton, or phase-field simulation tools [2].

A defect in otherwise perfect material breaks the crystalline symmetry and introduces the possibility of localized electronic states in the fundamental semiconducting or insulating gap. Depending on the position of the host-material Fermi level of the host semiconductor, these states are either unoccupied or occupied by one or more electrons, depending on their degeneracy and the spin assignment. The defects thus appear in different charge states, with varying degrees of charge localization around the defect center. To achieve a new ground state, the neighboring atoms relax around the defect to new equilibrium positions. For a point defect a new point symmetry group can be defined.

The energy levels in the gap, also known as "ionization levels", "occupation energy levels" or "transition levels", correspond to the Fermi-level positions where the ground state of the system changes from one charge state to another. Their values can be computationally estimated by the "ΔSCF" approach, i.e., from total-energy differences between different charge states. These gap levels can be probed experimentally by temperature-dependent Hall conductivity measurements, deep-level transient spectroscopy (DLTS), and photoluminescence (PL). The nature of the defect-related gap-state wavefunctions can be examined with electron paramagnetic resonance (EPR) and electron-nuclear double-resonance (ENDOR) measurements.

Experimentally, the defect energy levels can often be located with the precision of the order of 0.01 eV with respect to host material band edges. The proper interpretation of the levels and the physical features associated with them pose demanding challenges for theory and computation. At present, the *computational* accuracy of the level position is typically a few tenths of an eV. It follows that sometimes different calculations for the same physical system lead to very different conclusions and interpretations of the experiments. Defect calculations may also fail to predict correctly the symmetry-breaking distortions around defects undisputedly revealed by experiments. While theoretical modeling has had considerable success and a major impact in semiconductor physics, there are still severe limitations to its capabilities. The purpose of this Chapter is to point out and discuss these.

Density-functional theory (DFT) [3, 4] is the workhorse of atomic-scale computational materials science and is also widely used to study defects in semiconductors, predict their structures and energetics, vibrational and diffusional dynamics, and elucidate their electronic and optical properties, as observed by the various experimental probes. The central quantity in DFT is the electron density. The ground-state total energy is a functional of the

electron density and can be obtained via variational minimization of the functional. The wave-mechanical kinetic energy part of the total-energy functional is obtained not from the density directly but through a mapping to a non-interacting Kohn–Sham system. The mean-field electrostatic Hartree energy is obtained from the density, as are in principle the remaining exchange and correlation terms, coded into the exchange-correlation functional.

The purpose of this Chapter is to examine critically the methodological status of calculations of defects in semiconductors, especially those based on the so-called supercell methods. As will be discussed in more detail below, there are three main sources of error in such calculations. The first is the proper quantum-mechanical treatment of electron–electron interactions, which at the DFT level is dependent on the choice of the exchange-correlation functional. The popular local or semilocal density approximations lead to underestimation, sometimes serious, of the semiconducting gap. The underlying reasons for this are the neglect of self-interactions and the unphysical continuity of the exchange-correlation energy functional as a function of level-occupation number. This has naturally serious consequences to the mapping of the calculated defect electronic levels onto the experimental energy gap. The second source of errors is related to the geometrical description of the defect region, often known as the finite-size effect. The third source of error is the numerical implementation of DFT, i.e., the self-consistent solution of the effective-particle Kohn–Sham equations and the evaluation of the total electronic energy.

2 Density-Functional Theory

The quantum physics of electrons in materials is governed by the Schrödinger equation. Density-functional theory (DFT) provides a parameter-free framework for casting the formidable many-electron problem to a numerically tractable form involving only the three spatial coordinates of the N interacting electrons (as opposed to $3N$ in the full solution of the Schrödinger equation).

The fundamental starting point is the DFT expression for the total energy E_{tot} of the electronic system:

$$
\begin{aligned}
E_{\mathrm{tot}}\left[\{\psi_{i\sigma}(\boldsymbol{r}),\{\boldsymbol{R}_\alpha\}\right] = {} & \frac{\hbar^2}{2m}\sum_{i,\sigma}\left\langle \psi_{i\sigma}(\boldsymbol{r})\left|-\frac{1}{2}\nabla^2\right|\psi_{i\sigma}(\boldsymbol{r})\right\rangle \\
& + \frac{e^2}{2}\int\int \mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}'\frac{n(\boldsymbol{r})n(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} \\
& + \int \mathrm{d}\boldsymbol{r} V_{\mathrm{ion}}(\boldsymbol{r})n(\boldsymbol{r}) + \frac{1}{2}\sum_{\alpha,\beta;\alpha\neq\beta}\frac{Z_\alpha Z_\beta}{|\boldsymbol{R}_\alpha-\boldsymbol{R}_\beta|} \\
& + E_{\mathrm{xc}}\left[n_\uparrow(\boldsymbol{r}),n_\downarrow(\boldsymbol{r})\right]\,, \qquad (1)
\end{aligned}
$$

where the total electronic density $n(\boldsymbol{r}) = \sum_{i,\sigma} f_{i\sigma} \left|\psi_{i\sigma}(\boldsymbol{r})\right|^2$ is expressed as a summation of the one-electron states $\psi_{i\sigma}(\boldsymbol{r})$, dependent on the spin index $\sigma = \downarrow, \uparrow$ and with the occupation number $f_{i\sigma}$. Above, $V_{\rm ion}$ is the nuclear (ionic) potential, and Z_α denotes the charge of ion α at $\boldsymbol{R}_\alpha$.

The coupled Kohn–Sham eigenvalue equations

$$-\frac{\hbar^2}{2m}\nabla^2\psi_{i\sigma}(\boldsymbol{r}) + V_{\rm eff}\left[n_\uparrow(\boldsymbol{r}), n_\downarrow(\boldsymbol{r})\right]\psi_{i\sigma}(\boldsymbol{r}) = \varepsilon_{i\sigma}\psi_{i\sigma}(\boldsymbol{r}) \tag{2}$$

resulting from the minimization of $E_{\rm tot}$ with respect to the density recast the complex many-body interactions into an effective single-electron potential $V_{\rm eff}$ via the exchange-correlation functional $E_{\rm xc}$. In practice, the functional $E_{\rm xc}$ has to be approximated. The popular choices are functionals where exchange and correlation are treated as functions of the local density and/or its gradients. This sounds drastic, because electronic Pauli exchange is a manifestly nonlocal object, as demonstrated by the Hartree–Fock theory. However, there is a partial cancelation of errors, and these methods are surprisingly robust and accurate. The generalization to spin degrees of freedom is also straightforward in scalar-relativistic DFT.

A useful primer for designing DFT calculations has been written by *Mattsson* et al. [5]. The design of any calculation, whether using periodic boundary conditions, finite clusters or embedding techniques, involves the choice of the exchange-correlation functional. The choice, whether a particular flavor of local-density (LDA) or local spin-density (LSDA) approximation, a generalized-gradient approximation (GGA), full Hartree–Fock-type exchange, screened exchange, or a "hybrid" between a nonlocal orbital-based functional and a density-dependent functional, defines the physical accuracy of the calculation. The numerical accuracy of the calculation depends on such technical things as basis-set completeness, accuracy of integration in both real and reciprocal space, convergence criteria, etc. The model accuracy depends naturally on how faithfully the chosen supercell, cluster or embedding methods describe the desired situation.

In the context of defects in semiconductors, the ground-state properties can be obtained via the minimization of the total energy $E_{\rm tot}$ with respect to the charge density $n(\boldsymbol{r})$ and the ionic positions $\boldsymbol{R}_\alpha$ through the Hellmann–Feynman interatomic forces. This enables the calculation of not only the formation energy of a defect in a given lattice position, but also its low-energy vibrational excitations, and any observables accessible via the charge and spin densities, such as positron-annihilation parameters [6] or hyperfine fields [7–9].

3 Supercell and Other Methods

A popular way to calculate defect energetics from first principles (i.e., from electronic degrees of freedom) is based on the *supercell* idea. In this method

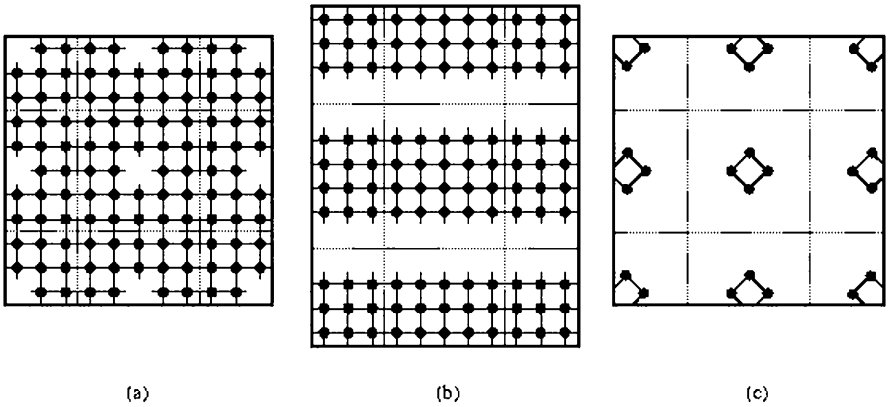

Fig. 1. Schematic presentation of the supercell construction. (**a**) lattice vacancy in a solid, (**b**) surfaces in a stack of slabs, (**c**) isolated molecules

one confines the atoms defining the defect area of interest into an otherwise arbitrary box, which is then repeated infinitely in one or more spatial directions (see Fig. 1). In other words, this box (the supercell) now becomes the new unit cell of the system, and periodic boundary conditions are applied at one or more of its boundaries. For a point-defect assembly in a three-dimensional solid, the system now becomes a three-dimensional periodic defect array. For a line defect (e.g., a dislocation) the result is a regular line-defect network. For a surface, the system becomes a sandwich of material slabs interlaced by vacuum regions. Typical sizes of the supercell are nowadays 64 atoms or larger in a cubic system, and the increasing computational power has enabled supercells with hundreds of atoms.

The supercell method is one of the three main approaches to defect calculations. The other two are the *finite-cluster* methods [10, 11] and the *Green's function embedding* techniques [12–14]. In the former the interesting defect is simply incorporated in a finite atomic cluster (with often atoms added to the cluster surface to saturate any dangling electron bonds). The finite-cluster method is well suited for studies of local defect properties (such as vibrational modes), provided that the cluster is large enough and thus the surface effects small. However, care should be taken to make sure that manipulating the cluster surface does not interfere with the defect-localized electrons. *Estreicher* and *Marinyck* [15] have shown how this can influence, for example, the defect-related hyperfine fields.

The embedding techniques, in turn, match the perturbed defect region to the known DFT Green's function of the unperturbed host material, and offer in principle the best way to study isolated defects. However, the numerical implementation of the Green's function method is challenging for accurate interatomic forces and long-range atomic relaxations, as it requires a well-

localized defect potential and usually short-range basis functions. Thus, supercell methods have surpassed the Green's function methods in popularity.

The great advantage of supercell calculations is that the periodic boundary conditions allow the utilization of the many efficient techniques derived for the quantum physics of periodic systems. The wavevectors of the Brillouin zone (BZ) in the reciprocal space of the supercell are good quantum numbers, and the standard "band-structure methods" of periodic solids can be applied in full force. Particularly fast computation methods can be performed by using Fourier analysis (plane-wave basis sets), as they adopt naturally to periodic boundary conditions and offer spatially uniform resolution.

The supercell approach enables the full relaxation of the structure to minimize total energy, and the calculation of the formation and migration energies of defects in different charge states as a function of the Fermi-level position of the host semiconductor, and as a function of the chemical potentials of the atoms building up the material. Thus, the method can be conveniently adapted to describe such effects as the host material stoichiometry on the energetics of isolated defects in compound semiconductors [16]. From supercell calculations one can extract several useful physical properties, such as: 1. the probabilities of certain types of defect to form under given chemical and thermodynamical conditions during the growth, 2. the basic nature (acceptor, donor, deep level) of the defect electronic states, 3. the migration and recombination barriers, 4. vibrational modes, hyperfine fields, positron-annihilation parameters and other ground-state properties, and 5. within certain limitations, excited-state (for example optical) properties as well. The ground-state properties are those most reliably accessible. For example, the defect-induced electronic energy levels in the fundamental gap are obtained as differences between formation energies for different charge states.

4 Issues with the Supercell Method

The obvious drawback of the supercell methods for defect calculations is that the periodicity is artificial and can lead to spurious interactions between the defects. They have a finite density, and do not necessarily mimic the true physical situation with an aperiodic, very low density defect distribution. The supercell method, widely used and with demonstrated success in defect studies, requires a critical examination of the finite-size and periodicity effects.

The first consequence of the finite-size supercell approximation is the *broadening* of the defect-induced electronic levels to "defect bands", with a bandwidth of the order of 0.1 eV for 64-atom supercells. This translates into a difficulty in placing accurately the ionization levels, also in the total-energy based ΔSCF approach. The one-electron Kohn–Sham states associated with the defect are often assigned a position by averaging over the supercell Brillouin zone, or just values at the Brillouin zone origin (the Γ point) are dis-

cussed as they exhibit the full symmetry of the defect. It can be argued that the defect-state dispersion has a smaller effect on the total energies that include summations and integrations of the Kohn–Sham states over the entire Brillouin zone. Thus, ionization levels determined from the total-energy differences would also be less sensitive to the defect-band dispersion. However, this is hard to prove systematically.

Another source of difficulty is the accurate determination of the host crystal *band edges* (valence-band maximum, conduction-band minimum), which are the natural reference energies for the defect-induced gap states. In the defect-containing supercell, the band-edge states themselves are affected by the defect. The band-edge positions are often determined by aligning a chosen reference level between the defect-containing and perfect solid. For example, the effective potential in a localized region far from the defect can be aligned with the potential in the same region in a perfect crystal. The calculated band-edge distances from the reference energy in a perfect crystal can then be used to align the band edges in the system containing the defect. Alternatively, one can consider a deeper, core-level energy of an atom as a local reference. Another possibility is to define as the "crystal zero" the electrostatic potential at a Wigner–Seitz cell boundary far away from the defect. This is particularly convenient when the atomic-sphere approximation (ASA) is used in the context of minimal basis sets, such as the LMTO [17].

Another, more difficult problem arises with *charged defects*. In order to avoid divergences in electrostatic energies, a popular solution is to introduce a homogeneous neutralizing background charge to the supercell array, which enables the evaluation of electrostatic (Coulomb) energies. This, however, introduces an electrostatic interaction between the periodic charge distribution in the supercells and the background, which vanishes only at the limit of infinitely large supercells. The influence of the fictitious charge has to be subtracted in the end, and this is a highly nontrivial task, as discussed below.

Point defects induce elastic stress in the host lattice, which is relieved by ion displacements, i.e., lattice relaxation. The lattice-relaxation pattern is restrained by the supercell geometry. The argument, often used in supercell calculations, is that the ion displacements vanish near the borders of the supercell. However, this does not necessarily guarantee that the long-range ionic relaxations are correctly described, as the supercell symmetry itself may fix the positions of the border ions. According to elastic continuum theory, the strain field at large distances from the point defect should fall off as $|\boldsymbol{r}|^{-3}$ and the ionic displacements should fall off as $|\boldsymbol{r}|^{-2}$. In tetrahedrally coordinated covalent materials the distances between the periodic images along the rigid [110] zigzag chains are typically more important to the convergence of lattice relaxations than the volume relaxations. A case in point is the vacancy in silicon [18], where the sense of the lattice relaxation changes from outward to inward only at large enough cell sizes. A similar behavior is also observed in finite-cluster calculations [19].

The finite size of the supercell thus restricts also the *ionic relaxations* around the defect. The relaxation pattern is truncated midway between a defect and its nearest periodic replica. In the case of long-range relaxations this cutoff may be reflected dramatically close to the defect, as was demonstrated by *Puska* et al. [18] by detailed calculations for vacancies in silicon.

Further comparison of the supercell to the finite-cluster and Green's function methods reveals the following. The supercell and Green's function methods have a well-defined electron chemical potential (they are coupled to a reservoir of electrons) whereas the cluster method does not. This has consequences for the treatment of ionization levels, which is somewhat problematic for the cluster method. The finite-size effect is there in the cluster method as well, as the clusters can contain spurious surface-related effects. The Green's function method is mathematically elegant and in principle the best for isolated defects, but its computational implementation is difficult in view of the accuracy required for total electronic energies and derived quantities (such as interatomic Hellmann–Feynman forces).

Finally, it is important to note that defect and impurity calculations should, as a rule, be carried out using the theoretical lattice constant, optimized for the *bulk* unit cell. This is crucial in order to avoid spurious elastic interactions with defects or impurities in the neighboring supercells. The purpose is to investigate properties of isolated defects or impurities in the dilute limit. If the volume of the defect-containing supercell is relaxed (in addition to relaxing the positions of the atoms near the defect), the calculation would in fact correspond to finding the lattice parameter of the system containing an ordered array of defects at a high concentration.

5 The Exchange-Correlation Functionals and the Semiconducting Gap

The fundamental semiconductor gap E_{g} is defined as the difference between the ionization energy I and the electron affinity E_{A} of the bulk material,

$$E_{\mathrm{g}} = I - E_{\mathrm{A}} \,. \tag{3}$$

As both quantities can be written in terms of total energies of systems with different numbers of electrons, the value of the gap E_{g} is, in principle, a ground-state property and can thus be calculated within the Kohn–Sham DFT. The fundamental gap can also be written as

$$E_{\mathrm{g}} = E_{\mathrm{g}}^{\mathrm{KS}} + \Delta_{\mathrm{xc}} \,. \tag{4}$$

$E_{\mathrm{g}}^{\mathrm{KS}}$ is the difference between the highest occupied and lowest unoccupied Kohn–Sham eigenvalues $\varepsilon_{i\sigma}$. Δ_{xc} is the discontinuity (as a function of the occupation number) of the exchange-correlation potential at the Fermi level.

The estimation of semiconductor bandgaps and positions is one of the classic failures of local(-spin-)density (L(S)DA) formulation of the Kohn–Sham DFT. There are two origins of this error. First is the *self-interaction* error inherent in the LDA potential, and the second is the *vanishing of the discontinuity* Δ_{xc} of the local-density exchange potential as a function of the level occupation at the Fermi level. They lead to both wrong absolute energy positions and too small or entirely absent bandgaps for many materials. This is not remedied by semilocal approximations for exchange and correlation, such as the generalized-gradient approximations (GGA) where the discontinuity also vanishes. In fact, the experience is that GGA approximations tend to make the gap even smaller than local-density approximations [20]. This can be traced to the property of GGA methods increasing the interatomic distances and lattice constants compared to LSDA. It is another well-known flaw of local-density methods that they *overbind* molecules and solids and *underestimate* bond and lattice distances. GGA methods correct and sometimes overcorrect this, but give even smaller bandgap values.

A qualitatively opposite problem is encountered with the Hartree–Fock method, which is known to considerably overestimate the bandgaps. Moreover, its application to metallic systems leads to well-known problems, such as the vanishing density of states at the Fermi level. Thus, intuitively, a hybrid method "interpolating" between the two seems an attractive approach.

Nonlocal descriptions of exchange are expected to improve the Kohn–Sham bandgaps as they remove the self-exchange error and exhibit discontinuity at the Fermi level. In the "exact exchange" (EXX) formulation one evaluates the exchange energy and potential using the Kohn–Sham (rather than Hartree–Fock) wavefunctions. *Städele* et al. [21] have shown that the discontinuity Δ_x in the exchange potential is much larger than the bandgap for several semiconductors, which points to the existence of a large cancellation between the exchange Δ_x and the corresponding correlation Δ_c.

The EEX method has been reported to give very good values for the bandgap and many other properties for *sp* semiconductors [22]. However, when applied to rare-gas solids, EXX fails to reproduce the experimental bandgaps [23]. A related problem is the position of the *d* band in many solids. Again, the self-interaction error in local-density methods displaces the *d* states and contributes to the bandgap error. The *d* band positions can be improved by the explicit removal of self-interactions [24], but the application of the nonlocal EXX does not seem to lead to a consistent improvement of *both d* band positions and bandgap values.

This points to the important interplay between treatment of nonlocal exchange and the description of the lower-lying and core states in semiconductors and insulators. It has recently been pointed out [25] that the lack of explicitly treated core-valence interactions can lead to an anomalously good agreement of the calculated Kohn–Sham gap with the experimental gap in *sp* semiconductors. The inclusion of these interactions worsens the agreement, while leading to a consistent treatment of semiconductors and insulators. It

thus appears that the EXX alone does not solve the bandgap problem, although it is a marked improvement over LDA or GGA. A major drawback is also the high computational cost of EEX, which has so far prohibited its widespread use in defect calculations.

In the screened-exchange LDA (sx-LDA) scheme [26] the Kohn–Sham single-particle orbitals are used to construct a nonlocal exchange operator, again improving the description compared to the local-density approximation. The many-body correlations and screening missing from the Hartree–Fock method are treated in a form of model screening and LDA correlation. This approach can solve the gap problem, but contains a phenomenological constant (e.g., the Thomas–Fermi screening constant). *Lento* and *Nieminen* [27] have applied this method for a prototypical supercell defect calculation (vacancy in Si). While the method can in principle give accurate total energies and thus ionization levels (and the correct gap), its full exploitation is still prohibitively expensive computing-wise. As with other hybrid functionals and with EXX, the computational cost is typically two orders of magnitude larger than with LDA or GGA. Thus the calculations with full lattice relaxation have so far been limited to supercells not large enough to allow the correct pattern to emerge (see Sect. 16 below).

There are other suggested approaches to overcome the failure of DFT to produce accurate bandgaps and excited-state properties. However, currently no method is available that would go beyond standard DFT and yet be feasible for total-energy calculations for the large supercells required to investigate defects. The class of methods, such as the GW approach [28] which are mainly aimed at calculating the (quasiparticle) band structure and properties derived thereof, are prohibitively expensive for large supercells. Presently, such calculations can be carried out with supercells with less than twenty atoms. However, this size is perfectly adequate for treating well-localized excitations, such as self-trapped excitons [29].

The gap problem may sometimes make calculations for certain defect levels impossible. The too-narrow gap may induce the defect level to be a resonance in the conduction or valence band. In this case some charge states would not be accessible at all. It should be noted that this can also happen in just some part of the Brillouin zone because of the defect-level dispersion in the supercell approach. A possible remedy is then to simply occupy the defect level, ignoring the fact that the state lies above the conduction-band minimum at other $\boldsymbol{k}$ points.

Apart from underestimating the bandgap and thus hampering the positioning of the possible ionization levels, the local (or quasilocal)-density approximation for exchange and correlation can sometimes lead to other difficulties with the defect-localized electron states. This has recently been discussed by *Laegsgaard* and *Stokbro* [30]. They considered a substitutional Al impurity on the Si site in SiO_2. LDA and GGA calculations predict a structure that has the full tetrahedral symmetry. However, ENDOR experiments show unambiguously that the defect wavefunction is localized on just one of

the four neighboring oxygen atoms, as the hyperfine splittings are radically different from those expected for four equivalent neighbors. The source of this problem can again be traced to the self-interaction error in LDA and GGA. In LDA/GGA, there is a penalty for localizing a single electron as the cancelation between self-exchange/correlation and self-Coulomb interactions is not complete. Thus, one overestimates the Coulomb energy of a localized state, which then leads to charge delocalization. In the true situation, there is just a *single* electron state contributing to the Coulomb energy. Thus, the calculated charge delocalization is completely due to self-interaction. If exchange is treated correctly as in the Hartree–Fock theory, self-interaction is eliminated and one obtains the correct distorted structure with the wavefunction localized on a single oxygen atom. As noted above, the problem with Hartree–Fock is that it seriously overestimates the bandgap and spin splittings, and is as such usually not a good tool for calculations of semiconductor defects.

A self-interaction-corrected (SIC) scheme for the case of hole trapping in SiO_2 was recently suggested by *d'Aveszac* et al. [31]. In their approach the Coulomb energy arising from the charge density of a single defect orbital and the associated contribution to the exchange-correlation energy are explicitly subtracted. This is technically straightforward for methods using localized basis sets, but can also be implemented using projection techniques with plane-wave methods. The physical idea of this self-interaction correction is that delocalized states do not suffer as much from self-interaction error as localized states do. In the standard SIC [32, 33] approaches there is some ambiguity in choosing the states to which the correction should be applied. For defect calculations, the localized state in the gap is well defined and the application of the correction is quite straightforward.

The LDA/GGA gap error is a persistent and rather harmful problem in semiconductor defect physics. The Kohn–Sham eigenvalues of (2) are Lagrange parameters used to orthogonalize the states in total-energy minimization, but cannot be interpreted as true electron-excitation energies. These could be obtained from quasiparticle theories such as *Hedin's* GW approach [34] or time-dependent density-functional theory [35]. The GW method is computationally expensive and cannot be applied to defect supercells with many atoms. Its ability to give converged total energies is not clear. However, it is a powerful approach for doing perturbative calculations for optical and other excitation spectra, and can be generalized to include the final-state electron–hole interactions through the Bethe–Salpeter equation [36].

One simple, obvious and often used possibility is to use the "scissor operator", i.e., to correct the bandgap by simply shifting the unoccupied and occupied states further away from each other. This, of course, destroys the self-consistency between Kohn–Sham eigenvalues and eigenfunctions, which may be detrimental for observables calculated from them. For defect physics, the question arises whether to shift the *defect levels* or not in such a process.

Intuitively, one would expect acceptor-like levels to move with the valence-band edge and donor-like states with the conduction-band edge, while levels deep in the bandgap would not move at all. This reasoning is, of course, quite arbitrary and unsatisfactory. A recent example of these difficulties is the work on the oxygen vacancy in ZnO, where the LDA gap is again considerably in error. Two calculations [37, 38] have addressed this system using the LDA+U method [39], where an onsite Coulomb repulsion U is added "by hand" when calculating the Zn d-states. This shifts the occupied d bands and subsequently the valence-band minimum downwards. The defect level associated with the oxygen vacancy then moves deeper into the gap from the valence-band maximum. If one leaves the defect level in place and shifts just the conduction bands up, there is a single level in the lower half of the gap. If, on the other hand, one decides from LDA+U calculation how much valence/conduction band character the state has, one ends up moving the state further up in the gap, into the upper half. As the experiments are ambiguous at this time, the matter is undecided and again calls for a better solution to the gap problem.

Finally, one should mention the quantum Monte Carlo (QMC) methods that look promising for addressing total-energy calculations for defects and impurities. They approach the problem of interacting electrons by addressing the correlations explicitly and build in the Pauli exchange by starting from a Slater determinant. The fixed-node QMC method has been applied to study self-interstitials in Si [40]. The formation energy of the split-$\langle 110 \rangle$ interstitial defect was found to be significantly higher with QMC than LDA, which can be explained by the upward shift of the defect-induced level in the bandgap. Further work includes QMC studies of optical and diffusive properties of vacancies in diamond [41]. If QMC methods can be extended to large enough supercells, to force and relaxation calculations, and to secured convergence with respect to $\boldsymbol{k}$-point summation, they can provide a most useful set of benchmarks for accurate total-energy calculations for defects in semiconductors.

The above discussion may give an overly pessimistic impression of the capabilities of DFT calculations for defects in semiconductors. The other side of the coin is much brighter. There is a vast inventory of "standard" DFT calculations that have produced robust and reliable results for semiconductor defects and have thus decisively contributed to their identification. An illustrative example is the extensive work on III–V nitride semiconductors, recently reviewed by *Van de Walle* and *Neugebauer* [42].

6 Core and Semicore Electrons: Pseudopotentials and Beyond

The widespread use and success of DFT in predicting various materials properties stems largely from the successful application of the pseudopotential (PP) concept. Only relatively few "valence" electrons of an atom are

chemically active. The replacement of chemically inert "core" electrons with an effective core potential both reduces the number of electronic degrees of freedom and avoids the numerical challenge posed by rapidly varying wavefunctions near the nucleus. Efficient numerical techniques such as the fast Fourier transform (FFT) can be used. The replacement of the all-electron potential with a PP is, however, a nontrivial task where a balance has to be struck between optimal transferability (accurate reproduction of all-electron atom behavior) and computational efficiency (slow spatial variability).

Several methods are popular for generating PPs. They include the generalized norm-conserving pseudopotentials of *Hamann* [43], *Troullier* and *Martins* [44], *Hartwigsen* et al. [45] and the ultrasoft, non-norm-conserving PPs of *Vanderbilt* [46]. It is important to realize that the true figure of merit of a PP is not how well its results match experiment, but how well it reproduces the results of all-electron calculations when using otherwise similar methods.

The separable PPs commonly used in the context of plane-wave calculations, such as those of the Kleinman–Bylander [47] form, are built from norm-conserving PPs. The separation process can sometimes add complications, for example, in the form of so-called ghost states. The construction of reliable ultrasoft pseudopotentials [46] or projected-augmented-wave (PAW) potentials [48, 49] presents additional challenges, but the latter is quickly gaining in popularity as it has proven a particularly robust and transparent approximation.

To compare calculated values reported in the literature for defects in semiconductors, specific details of the PP generation are important for the reproduction of a given result. Several parameters usually enter the PP construction, such as the core radii for different angular-momentum channels and the core-matching radii. The effectiveness of a given PP needs to be checked and validated in every new chemical environment before being used in quantitative studies.

Constructing PPs thus involves compromises, and in reporting computational results it is important to state those compromises clearly. In particular, the choice of a given state as an "active" valence state vs. its treatment via perturbative, nonlinear core-valence corrections [50] are crucial. For example, it has been demonstrated [51] that a nonlinear core-valence correction is necessary to get spin-polarized states of first-row atoms O and N properly described. Unlike previously assumed, the 1s core in these atoms is not so deep as to be uninvolved in the chemistry important in defining the valence-electron potential.

An interesting approach that yields bulk band structures in good agreement with experiment is based on the idea of self-interaction and relaxation-corrected pseudopotentials (SIRC) [52]. The removal of self-interactions has important consequences for ionization levels in the semiconducting gap, as discussed above. However, it seems that the SIRC approach does not allow a simple self-consistent calculation of total energies.

Finally, it should be emphasized that consistency in the choice of the exchange-correlation functional for pseudopotential construction and their application is highly advisable. The problems of using a different functional for the two have been demonstrated, especially in the cases where polarizable semicore states are present [53].

7 Basis Sets

The use of DFT is not limited to static electronic-structure problems such as total energies and bond distances. The potential-energy (Born–Oppenheimer) surface defined by the positions of the nuclei defines the interatomic forces, and in recent years there has been rapid growth in the use of first-principles molecular dynamics (FPMD) within the finite-temperature DFT framework for studying many materials properties, including atomic vibration and migration, and the full equation of state (EOS). If the force fields or pressure or stress tensor of the system are of interest, additional care must be taken to ensure that the calculated quantities are of sufficient numerical accuracy. Whereas errors in the total energy are second order with respect to errors in the electron density or Kohn–Sham eigenfunctions, errors in the interatomic forces and pressure are first order. The errors are significantly affected by the completeness and accuracy of the basis set.

A popular choice is the plane-wave expansion for the Kohn–Sham states,

$$\psi_{i\sigma}(\boldsymbol{r}) = \psi_{n\sigma\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} \psi_{n\sigma\boldsymbol{k}}(\boldsymbol{G}) \exp[-i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}]\,. \tag{5}$$

The summation is over the reciprocal lattice vectors $\boldsymbol{G}$ of the superlattice.

In plane-wave calculations the accuracy of the basis set is determined by the chosen kinetic energy cutoff or maximum wavevector $E_{\mathrm{cut}} = \frac{\hbar^2(\boldsymbol{k}+\boldsymbol{G}_{\max})^2}{2m}$. It is important to realize that while a lower cutoff may be sufficient for the convergence of total energies, it may be totally unacceptable for calculating interatomic force constants or pressure/stress tensors.

The reason that forces are more sensitive to the plane-wave cutoff than total energies is that the real-space force formula contains the derivatives with respect to coordinates (see (6)). This derivative will introduce an extra factor of the reciprocal lattice vectors $\boldsymbol{G}$ into the reciprocal-space formulae and the maximum value $\boldsymbol{G}_{\max}$ is determined by the cutoff. It also follows that exchange-correlation functionals involving gradients of the density (such as GGA) are more sensitive to the cutoff than LDA.

The issue of basis-set sufficiency is even more critical when using local basis sets, for example linear combinations of numerical atomic orbitals [54, 55] or Gaussians [56]. A great advantage of local-orbital methods is the much smaller size of the basis (typically up to twenty basis functions per atom),

reflecting their economy when fitting rapidly varying wavefunctions. However, as a discrete set the convergence of a local basis is not simply controlled by a single parameter. Local basis sets always require "optimization" and experience, and tested basis sets are collected into libraries with limited possibilities for modification. The words of caution about plane-wave basis-set convergence apply even more strongly to local basis sets. While a well-designed basis set (for example, the "double-zeta plus polarization" (DZP)) does match the accuracy of a fully converged plane-wave calculation, the omission of the polarization term or calculation with single s and p functions can lead to large errors. This is particularly true for forces. For atoms far from equilibrium (e.g., under large strain) it may be necessary to increase variational freedom by expanding the basis to triple functions to achieve convergence. In some cases, especially for systems with large open volumes, it is advisable to augment the basis set by introducing "floating orbitals", i.e., basis functions not centered at atoms.

Finally, one should mention the real-space methods quickly gaining popularity [57,58]. They use finite-difference approximations for the kinetic-energy operator (finite elements are also a possibility), and the quality of the basis set can be controlled simply by the choice of the spacing of the real-space grid. The great advantage of real-space techniques is their easy adaptability to different boundary conditions (periodic, open, etc.), which is especially useful for studies of defects in low-symmetry nanostructures, where supercell methods can be quite wasteful.

8 Time-Dependent and Finite-Temperature Simulations

Within the Born–Oppenheimer approximation, the electronic and ionic motion and timescales are well separated. First-principles molecular dynamics (FPMD) can be implemented on the Born–Oppenheimer hypersurface by calculating an instantaneous value for the electronic total energy $E_{\mathrm{tot}}[\{\boldsymbol{R}_{\mathrm{i}}\}]$ as determined by all the ionic positions $\{\boldsymbol{R}_{\mathrm{i}}\}$. The forces acting on the atoms are then evaluated as

$$\boldsymbol{F}_{\mathrm{i}} = -\frac{\partial E_{\mathrm{tot}}}{\partial \boldsymbol{R}_{\mathrm{i}}} \tag{6}$$

and Newton's equations of motion are solved for the ionic degrees of freedom, using standard algorithms for the second-order temporal differential equations. Periodic boundary conditions are routinely implemented, and allow simulations of disordered amorphous or liquid phases. A fixed-volume supercell with conserved total energy corresponds to the microcanonical statistical ensemble. The constant-temperature (canonical) ensemble can be simulated using the Nosé–Hoover thermostat [59, 60]. Allowing the supercell volume to change as a dynamical variable leads to the constant-pressure ensemble,

and structural phase transformations are enabled by making the directional vectors spanning the supercell also dynamical [61].

The molecular-dynamics techniques also enable *free-energy* simulations. The thermodynamic integration formula for the (Helmholtz) free energy reads

$$\Delta F = \int_0^1 \mathrm{d}\lambda \, \langle H_1 - H_0 \rangle_\lambda \,, \tag{7}$$

where ΔF is the free-energy difference between the actual system with the Hamiltonian H_1 and a reference system with the Hamiltonian H_0, and $\langle \ldots \rangle_\lambda$ denotes a temporal average along an isobaric-isothermal trajectory generated from the Hamiltonian

$$H(\lambda) = \lambda H_1 + (1 - \lambda) H_0 \,. \tag{8}$$

Under ergodic conditions this is equivalent to an ensemble average. If the free energy of the reference system is known, (7) allows one to compute the free energy of the actual system. The FPMD method is particularly well adaptable to the supercell geometry.

The quantum dynamics associated with the electronic degrees of freedom can be described by the time-dependent Schrödinger equation. Within the density-functional formulation of many-body systems, this leads to the powerful formulation of time-dependent density-functional theory [62]. It can be cast in the form of time-dependent Kohn–Sham equations with a generalized, time-dependent exchange-correlation potential. These equations can be explicitly solved by time-propagation algorithms and starting from an appropriately prepared initial state. The time evolution of the wavefunctions and the density can then be analyzed to obtain the desired physical quantities, such as the full response of the system to an external, polarizing field or exciting pulse.

The Kohn–Sham equations with an external time-dependent perturbation can also be treated in the linear-response limit if the perturbation is weak. As the Kohn–Sham states are noninteracting, the linear-response function of the system can be constructed. The linearization leads to integrodifferential equations with an exchange-correlation kernel derived from the description of E_{xc}. Linear-response TDDFT is a popular method [63] to estimate excitation energies, polarizabilities and inelastic-loss properties (such as photoabsorption cross sections), especially for finite systems. The treatment of extended systems (periodic boundary conditions) within TDDFT requires the consistent handling of electronic currents in the system, and is best characterized as time-dependent current-density-functional theory (TCDFT) [64].

9 Jahn–Teller Distortions in Semiconductor Defects

The Jahn–Teller effect is the removal of electronic degeneracy by spontaneous lowering of spatial symmetry of atoms surrounding a defect, which leads to

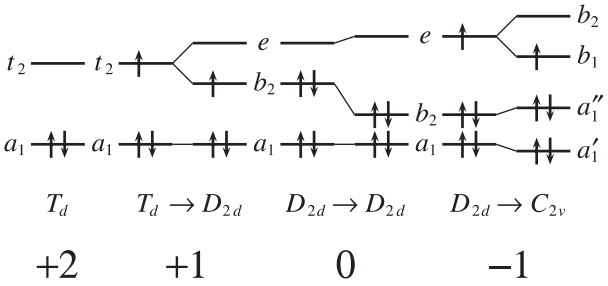

Fig. 2. The splitting of the defect-related gap states for a vacancy in Si. The point symmetry group is denoted for defects in different charge states. The *arrows* denote the population of the two spin states

the lowering of total energy. This is the result of a general theorem [65], with several well-known consequences in solid-state physics.

It is easiest to discuss the Jahn–Teller mechanism by using the monovacancy and a substitutional metal impurity in a tetrahedrally bonded semiconductor (Si) as examples. These same systems will later be presented as examples of state-of-the-art supercell DFT calculations in Sect. 16.

9.1 Vacancy in Silicon

In the pioneering linear combination of atomic-orbitals (LCAO) model by *Watkins* [66] (see Fig. 2), the electronic configuration of a neutral vacancy in Si (V_{Si}) with full cubic (T_d) symmetry is $a_1^2 t_1^2$, analogous to a Si atom ($[Ne]3s^2 2p^2$). The a_1 state lies in the valence band (is a resonance) and is doubly occupied. The other state t_1 is in the fundamental gap and can contain up to six electrons. It is pushed into the gap by the repulsive potential due to the removal of a positive ion in the lattice. In compound semiconductors, therefore, the states are expected to lie higher for anion vacancies with a larger number of removed valence electrons than for cation vacancies.

The filling of the gap states with rising Fermi level is associated with the relaxations of the atoms neighboring the defect. If there are no electrons in these levels, the T_d symmetry of the ideal vacancy (ideal crystal) is preserved. The only possible relaxation is of the breathing-mode symmetry around the defect center. Adding one or more electrons to the levels splits the degeneracy through the Jahn–Teller effect, and lowers the symmetry through atomic relaxation. The main component of this distortion is tetragonal, giving D_{2d} symmetry. This can occur in two senses, giving either a short, broad ("type A") or long, thin ("type B") shape for the box bounding the nearest-neighbor atoms (see Fig. 3). This determines the order of the resulting b_2 and e states. For V_{Si}, the latter option turns out to be valid.

One electron added to the gap level thus results in a singly positively charged vacancy with D_{2d} symmetry. Adding more electrons usually implies

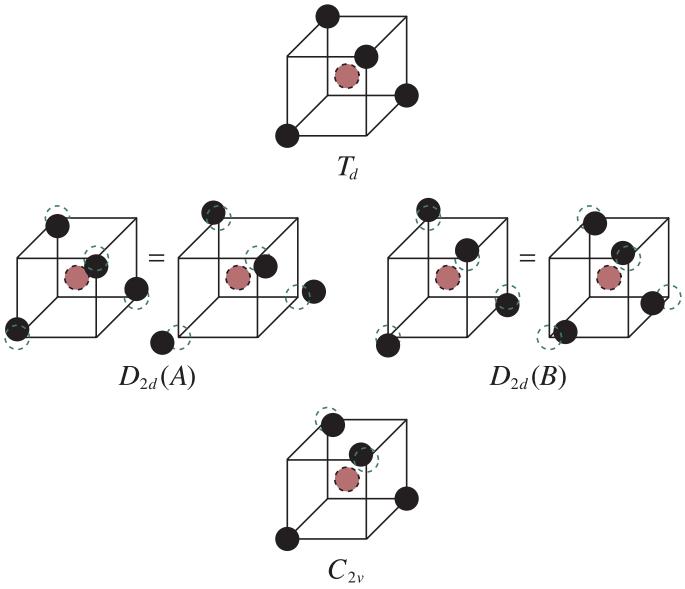

Fig. 3. The bounding boxes for the Si vacancy with distortions of different point-symmetry orientation

higher total energies because of the intraelectron Coulomb repulsion between the localized electrons in the gap states. However, adding a second electron to the vacancy in Si can enhance the pairing distortions (while retaining D_{2d} symmetry) and the Jahn–Teller splitting, possibly outweighing the increased Coulomb repulsion. This is the *negative-U* phenomenon [67] often encountered in semiconductor defects. Another way of stating the effect is to say that the energy to remove *two* electrons from V_{Si}^{0} is less than to remove *one* electron. Thus V_{Si}^{+} would be unstable.

A trigonal distortion from T_d symmetry occurs when a third electron is added and the vacancy is in the charge state $q = -1$. This lowers the symmetry to or C_{2v} or C_{3v} and splits the e state to b_2 and b_1. The quantitative results of state-of-the-art supercell calculations for the charge-state-dependent relaxation and how they match the Watkins model are given in Sect. 16.

9.2 Substitutional Copper in Silicon

Neutral $\mathrm{Cu_{Si}}$ [68] is isoelectronic with the negatively charged V_{Si} (see Fig. 4). The triplet state t_2^3 is partially occupied, in the bandgap, and susceptible to the Jahn–Teller effect. If onsite electronic repulsion prevails, the high-spin $S = 3/2$ states would be stabilized, as for example in the case of negatively charged substitutional Ni in diamond or the silicon vacancy in SiC [69]. Spin-orbit coupling also affects the orbital degeneracy, especially for the heavier impurity elements.

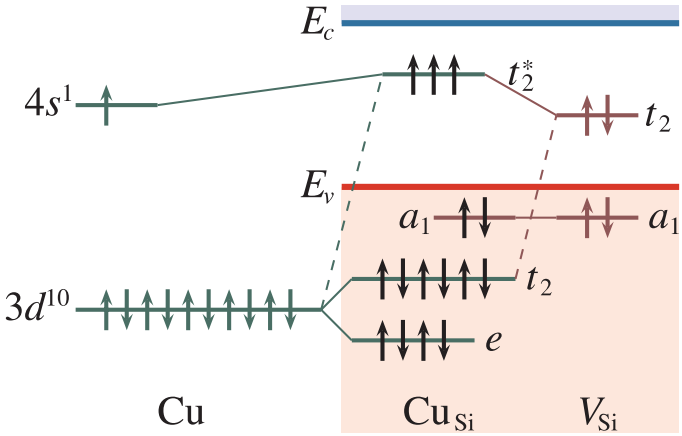

Fig. 4. The evolution of the defect-related states of substitutional Cu in Si from those for a Cu atom and Si vacancy

In general, it is in principle possible to reach higher positive and negative charge states by adding/removing electrons to/from the localized levels. The stability of a given charge state requires accurate calculation of its total energy, and is subject to the computational limitations discussed above and further expanded upon in Sects. 13–15.

10 Vibrational Modes

The phonon modes of a vibrating solid can be calculated using the powerful linear-response formalism [70], where the DFT total-energy hypersurface enters as the ground state and determines the elastic response. This formalism has been implemented using different basis sets for bulk calculations, including short-range local orbitals [71]. For defects in semiconductors, it is, however, the localized vibrational modes (LVMs) associated with defects that are often useful fingerprints for defect identification. In the case of defect-associated localized modes, a more economical route than the full phonon response is provided by considering only a single supercell. This is because the interaction between the defect replicas can be made small by using a large enough cell, whereby the vibrations do not mix and there is no phonon dispersion. The total energy of the supercell can be written as a Taylor series around the equilibrium positions in the harmonic approximation as

$$E_{\mathrm{tot}}[\{R_{\alpha i}+s_{\alpha i}\}] = E_{\mathrm{tot}}\Big[\{R_{\alpha i}\} + \frac{1}{2}\sum_{\alpha i,\beta j}\frac{\partial^2 E_{\mathrm{tot}}}{\partial R_{\alpha i}\partial R_{\beta j}}s_{\alpha i}s_{\beta j} + \cdots, \qquad (9)$$

where $s_{\alpha i}$ is the ith Cartesian component of the atomic displacement of the ion α. The coupling constants

$$\Phi_{\beta j}^{\alpha i} \equiv \frac{\partial}{\partial R_{\alpha i}} \Big(\frac{\partial E_{\mathrm{tot}}}{\partial R_{\beta j}} \Big) \tag{10}$$

are conveniently obtained as numerical derivatives of the (Hellmann–Feynman) forces on the atoms as they are shifted a small distance from their equilibrium positions around the defect ("frozen phonons"). The frequencies ω and amplitudes u of the localized vibrations can then be obtained as the normal-mode solutions to the equations

$$-\omega^2 u_{\alpha i} + \sum_{\beta j} D_{\alpha i}^{\beta j} \, , u_{\beta j} = 0 \, , \tag{11}$$

where $D_{\alpha i}^{\beta j}$ is the dynamical matrix. If M_α is the mass of ion α,

$$D_{\alpha i}^{\beta j} \equiv \frac{1}{\sqrt{M_\alpha M_\beta}} \Phi_{\beta j}^{\alpha i} \, . \tag{12}$$

In practical LVM calculations, selected atoms in the supercell are displaced to all three Cartesian directions, and after each displacement the ground-state electronic total energy is obtained and the Hellmann–Feynman forces calculated. This is done for all atoms that are a priori considered important for the description of the local vibrational modes around the defect. The dynamical matrix is then calculated by the finite-difference approximation using these forces and displacements. The normal modes and the corresponding vibrational frequencies can then be obtained by diagonalizing the dynamical matrix. The harmonic modes can easily be quantized and the zero-point vibrational motion estimated. Within the harmonic approximation, the phonon spectrum can be used to estimate the vibrational entropy of a given defect, which can then be entered as the prefactor when estimating absolute defect concentrations in thermal equilibrium.

Alternatively, the phonon spectrum, including the localized modes, can be obtained by MD simulation. The simulation time has to be long enough so that all the relevant vibrational modes have undergone several oscillation periods. The vibrational frequencies can then be obtained by Fourier transforming the velocity autocorrelation function of the moving ions. This usually requires that the LVM is well separated from the density of states for host-lattice vibrations. The MD approach includes also the anharmonic contributions to the vibrations, but not the zero-point motion.

11 Ionization Levels

The formation energy of a defect or impurity X in the charge state q is

$$E^{\mathrm{f}} \left[X^q \right] = E_{\mathrm{tot}} \left[X^q \right] - E_{\mathrm{tot}} \left[\mathrm{bulk} \right] - \sum_i n_i \mu_i + q \left[\mu_{\mathrm{e}} + E_{\mathrm{v}} \right] \, , \tag{13}$$

where $E_{\rm tot}\,[X^q]$ is the total energy from a supercell calculation with a defect or an impurity in the cell, and $E_{\rm tot}\,[\text{bulk}]$ is the total energy of the equivalent bulk supercell. n_i indicates the number of atoms of type i (host or impurity), either added ($n_i > 0$) or removed ($n_i < 0$) from the defected supercell, and μ_i denote the corresponding atomic chemical potentials. They represent the energy of the reservoirs with which atoms are exchanged when assembling the crystal in question. $\mu_{\rm e}$ is the electron chemical potential (Fermi level), referenced here to the valence-band maximum $E_{\rm v}$ in the bulk material. The valence-band maximum of defect supercell has to be aligned with that in the bulk (see below).

For a monoatomic solid, the formation energy for a defect in the charge state q reads

$$E_q^{\rm f} = E_q^{\rm def} - N\mu + q(E_{\rm v} + \mu_{\rm e})\,, \tag{14}$$

where $E_q^{\rm def}$ is the total energy of the defect-containing supercell and N the total number of atoms in it.

The *thermodynamic* ionization (transition) level of a given defect $E_{\rm d}(q/q')$ is defined as the Fermi-level position where the charge states q and q' have equal total energy. In experiments where the final charge state can relax to its equilibrium configuration after the transition (excitation), this is the energy level that would be observed. The ionization levels are therefore observed in DLTS experiments or (for shallow levels near band edges) as the thermal ionization energies derived from temperature-dependent conductivity or Hall effect data.

The *optical* energy $E_{\rm d}^{\rm opt}(q/q')$ associated with a transition between charge states q and q' is defined similarly to the thermodynamic transition energy, but now the final state with charge q' is calculated using the atomic geometry of the initial state q. The optical level is observed in "vertical" absorption experiments where the final charge state cannot relax to its equilibrium. In emission experiments, on the other hand, the initial excited state has evolved towards its minimum-energy configuration, which is different from the ground-state structure. Optical emission and absorption peaks are thus separated by the so-called Stokes shift. This poses challenges for both the experimental assignment of peaks and the theoretical analysis.

For a two-component (compound) semiconductor, (13) can be rewritten [16] in a form that reflects the stoichiometry of the host material. It is gauged by the chemical potential difference $\Delta\mu$ of an AB compound as

$$\Delta\mu = \mu_{\rm A} - \mu_{\rm B} - [\mu_{\rm A}(\text{bulk}) - \mu_{\rm B}(\text{bulk})]\,, \tag{15}$$

with an allowed range

$$-\Delta H \le \Delta\mu \le \Delta H\,, \tag{16}$$

where the lower limit corresponds to A-rich and the upper limit B-rich material. ΔH is the heat of formation of the compound from its bulk constituents,

$$\Delta H = \mu_{\mathrm{A}}(\mathrm{bulk}) + \mu_{\mathrm{B}}(\mathrm{bulk}) - \mu_{\mathrm{AB}}(\mathrm{bulk})\,. \tag{17}$$

By defining

$$\begin{aligned} E_1^{\mathrm{f}}\left[X^q\right] \equiv\; & E_{\mathrm{tot}}\left[X^q\right] - \tfrac{1}{2}(n_{\mathrm{A}} + n_{\mathrm{B}})\mu_{\mathrm{AB}}(\mathrm{bulk}) \\ & - \tfrac{1}{2}(n_{\mathrm{A}} - n_{\mathrm{B}})[\mu_{\mathrm{A}}(\mathrm{bulk}) - \mu_{\mathrm{B}}(\mathrm{bulk})] + qE_{\mathrm{v}} \end{aligned} \tag{18}$$

one can write (13) in the short form

$$E^{\mathrm{f}}\left[X^q\right] = E_1^{\mathrm{f}}\left[X^q\right] + q\mu_{\mathrm{e}} - \tfrac{1}{2}(n_{\mathrm{A}} - n_{\mathrm{B}})\Delta\mu\,. \tag{19}$$

Equations (14) or (19) provide a basis for evaluating the ionization level between charge states q and $q+1$ as the value of μ_{e} where the two formation energies are equal. This is a widely used technique for defects in semiconductors. It requires the accurate determination of the supercell total energies, and uses the valence-band maximum as the absolute reference energy (alternatively, the electron chemical potential can be measured from the conduction-band minimum). Although it makes no explicit reference to the eigenvalues of the Kohn–Sham gap states, the method can suffer from the DFT underestimation of the bandgap, especially if the state in question is far from the reference energy, close to the opposite band edge.

12 The Marker Method

The problems of bandgap underestimation and valence-band alignment can to some extent be avoided in using the so-called *marker* method for calculating ionization-level positions [72]. Let us define the *configuration* energy for a defect labeled d as the total energy difference per unit charge between two charge states

$$C_d(q/q') = \left(E_{\mathrm{tot}}\left[X^{q'}\right] - E_{\mathrm{tot}}\left[X^q\right]\right)\Big/(q - q')\,. \tag{20}$$

If there is a measured or otherwise accurately known reference ("marker") defect m with the ionization level $E_m(p/p')$ between charge states p and p' in the region of interest in the semiconductor bandgap, one can calculate its configuration energy $C_m(p/p')$ and the configuration energy difference of d from the marker as

$$D_d(q/q') = C_d(q/q') - C_m(p/p')\,. \tag{21}$$

Then the ionization level of interest can be estimated as

$$E_d(q/q') \approx D_d(q/q') + E_m(p/p')\,. \tag{22}$$

The idea is to achieve systematic cancelation of computational errors. This happens best when the defects d and m are similar and all computational parameters are the same.

The reference energy may also be taken as the valence-band maximum and the conduction-band bottom of the *bulk* supercell with the same number of atoms as the defect supercell, whereby

$$C_m^{\mathrm{bulk}}(0/+1) = E_{\mathrm{v}} \quad \text{and} \quad C_m^{\mathrm{bulk}}(-1/0) = E_{\mathrm{c}}\,, \tag{23}$$

and the bandgap is

$$E_{\mathrm{g}} = C_m^{\mathrm{bulk}}(-1/0) - C_m^{\mathrm{bulk}}(0/+1)\,. \tag{24}$$

The marker method can provide a useful alternative for ionization-level determination, but requires the knowledge of a suitable known transition level in the vicinity of the level under scrutiny.

13 Brillouin-Zone Sampling

For a perfect (no defects) lattice, the convergence of electronic properties can be achieved either by increasing the number of $\boldsymbol{k}$ points in the BZ or the product of the number of atoms (N) in the calculational unit cell and the number of $\boldsymbol{k}$ points. The computational cost increases linearly with the number of $\boldsymbol{k}$ points but is proportional to at least N^3. Thus increasing simply the number of $\boldsymbol{k}$ points would seem most economical. However, for defect calculations the spurious defect–defect interactions are a more subtle issue. For defects in metals already quite small supercells may give well-converged results if the number of $\boldsymbol{k}$ points is large enough. For defects in semiconductors the situation is more difficult. Their description involves localized gap states. The description of these is not straightforwardly improved as the number of $\boldsymbol{k}$ points increases, i.e., the detailed choice of the $\boldsymbol{k}$-point sampling may affect the convergence.

In supercell calculations, the evaluation of a ground-state property P of the system, such as total energy, requires integration over the Brillouin zone (BZ) of the reciprocal cell

$$P = \left(\frac{1}{V_{\mathrm{BZ}}}\right) \int_{\mathrm{BZ}} \mathrm{d}^3\boldsymbol{k} \sum_n p_n(\boldsymbol{k}) f[\varepsilon_n(\boldsymbol{k})]\,, \tag{25}$$

where V_{BZ} is the volume of the BZ, n enumerates Kohn–Sham states with the wavevector $\boldsymbol{k}$ and eigenvalue $\varepsilon_n(\boldsymbol{k})$, and p_n defines the physical property. $f[\varepsilon_n(\boldsymbol{k})]$ is the occupation number of the Kohn–Sham state, given by the Fermi–Dirac distribution around the chemical potential μ_{e} as

$$f[\varepsilon_n(\boldsymbol{k})] = \frac{1}{e^{\frac{\mu_{\mathrm{e}} - \varepsilon_n(\boldsymbol{k})}{k_{\mathrm{B}}T}} + 1}\,. \tag{26}$$

For semiconductors with a finite gap between occupied and unoccupied states, the integrand in (25) is continuous, and consequently the integral can, in principle, be replaced by a finite sampling of discrete $\boldsymbol{k}$ points. The computational cost increases linearly with the number of $\boldsymbol{k}$ points, and "special point" schemes are popular to reduce the computational cost to obtain the desired accuracy [73]. In particular, the uniform $\boldsymbol{k}$-point mesh approach suggested by *Monkhorst* and *Pack* [74] has been widely used in practical calculations. For very large actual supercell sizes, the BZ shrinks towards a point (the Γ-point, the $\boldsymbol{k} = 0$ of the BZ). Γ-point sampling offers additional savings in computing as the Kohn–Sham wavefunctions are purely real at $\boldsymbol{k} = 0$.

The simplest scheme to sample the BZ in supercell calculations is to use the Γ point only. When the size of the supercell increases, the wavefunctions calculated correspond to several $\boldsymbol{k}$ points of the underlying perfect bulk lattice, and the perfect lattice $\boldsymbol{k}$ space is evenly sampled. In order to improve the description of, in particular, that of the delocalized bulk-like states and the description of the electron density, it is beneficial to use $\boldsymbol{k}$ points other than the Γ point. Thereby also components with wavelengths larger than the supercell lattice constant are included. This idea leads to the special $\boldsymbol{k}$-point schemes mentioned above. They are widely used to sample the BZ also in defect calculations.

The accuracy of a given $\boldsymbol{k}$-point mesh depends naturally on the supercell size (BZ volume). *Makov* et al. [75] utilized this fact to suggest sampling meshes that would in fact extrapolate the integration result towards larger unit-cell sizes. They proposed a scheme to choose $\boldsymbol{k}$ points for supercell calculations so that the electronic defect–defect interactions are minimized in the total energy.

Defect calculations often focus on total-energy differences, and a tacit assumption is that errors due to BZ sampling largely cancel out when taking the differences (bulk vs. defect, for example) treated with equal-size supercells. This assumption is not a priori warranted and should be checked carefully. It has been demonstrated [76] that even for relatively large supercells the sampling errors can depend on the defect type and charge state and thus do not cancel when taking the difference. *Shim* et al. [76] investigated vacancy and interstitial defects in diamond and silicon, and found that for a given supercell size, the $\boldsymbol{k}$-point sampling errors in the total energy can vary considerably depending on the charge state and defect type.

14 Charged Defects and Electrostatic Corrections

Although the supercell approximation describes accurately and in an economical way the crucial local rearrangement of bonding between atoms and the underlying crystal structure, it also introduces artificial long-range interactions between the periodic images. The most dramatic artefact is the

divergence of the overall electrostatic (Coulomb) energy for charged defects in the periodic superlattice.

This divergence of the total energy of charged defects in the supercell approximation is usually circumvented by the introduction of a fictitious neutralizing background charge, often in the form of a uniform "jellium" distribution. The influence of this neutralizing charge in the total energy of the supercell needs to be included. This is known as the electrostatic or Madelung correction ΔE_{c}.

With the neutralizing background added, in the large-supercell limit the electrostatic interaction of a charged defect with its periodic images in an overall neutralized system becomes, in principle, negligible. However, there is no guarantee that the convergence of the Coulomb energy as a function of the supercell size is particularly fast [77]. In fact, classical electrostatics for localized charges in an overall neutral system predicts an asymptotic L^{-1} dependence, where $L = \sqrt[3]{V}$ and V is the supercell volume. This scaling law is unfortunately slow in converging, and, moreover, its prefactor is in general unknown. Consequently, electrostatic errors are easily introduced into total-energy calculations and they do not necessarily cancel when taking energy differences, as the magnitude depends on the details of the charge distribution within the supercell. Over the years, several attempts to reliably estimate ΔE_{c} have been proposed.

Leslie and *Gillan* [78] and *Payne* and coworkers [79, 80] have developed correction formulae to be applied for electrostatic correction of charged-defect arrays. They considered an array of localized charges immersed in a structureless dielectric and neutralized by jellium compensation. By considering the multipole expansion of the defect charge distribution, the correction formula can be derived in the form

$$\Delta E_{\mathrm{c}} \approx -\frac{q^2\alpha}{2\varepsilon L} - \frac{2\pi q Q}{3\varepsilon L^3} + O(L^{-5})\,, \tag{27}$$

where q is the charge of the defect, α the Madelung constant of the superlattice, and ε the static dielectric constant of the host material. For cubic geometries, α is 2.837 3, 2.888 3 and 2.885 for SC, BCC and FCC supercells, respectively. The first term corresponds to a point-charge array and a compensating background in a uniform dielectric.

The second term in the right-hand side of (27) arises from the shape-dependent interaction charge distribution inside the supercell with the neutralizing background. The parameter Q is the second radial moment of the defect charge density

$$Q = \int d^3\boldsymbol{r} r^2 [\rho_d(\boldsymbol{r}) - \rho_b(\boldsymbol{r})]\,. \tag{28}$$

Kantorovich [81] re-examined the method for arbitrary supercell shapes, and suggested a modified formula ignoring dipole–dipole interactions.

There is obvious ambiguity in this correction scheme. First, the static dielectric constant is introduced as an external parameter and is not consistently defined. Secondly, the correction scheme is derived assuming that the charge perturbation introduced by the defect is well localized within the supercell. This is strictly speaking not valid, as in classical theory the polarization and thus the aperiodic screening charge distribution extend over the whole macroscopic crystal. Thus, as described in detail by *Lento* et al. [82], Q depends on the size of the supercell and (27) does not lead to a well-defined correction value. Nevertheless, (27), often known as the Makov–Payne formula, is popular in estimations of electrostatic corrections.

Segev and *Wei* [83] showed that for charges with a Gaussian distribution with a certain width σ interacting with the background, the Madelung correction should vanish as σ becomes large. Moreover, situations may arise where local symmetry-breaking distortions in fact induce a net dipole in the supercell. Dipole contributions cannot be discarded in such a situation, as they do not cancel in the total-energy differences between perfect and defected supercells.

Shim et al. [84] have systematically studied the behavior of the electrostatic correction for supercells of different sizes for vacancy and interstitial defects in diamond. For negatively charged vacancies and positively charged interstitials, the formation energies show a clear dependence on the supercell size and are in qualitative agreement with the Makov–Payne trend. For positively charged vacancies and negatively charged interstitials, the electrostatic corrections are weak. An analysis of the spatial charge density distributions reveals that these large variations in electrostatic terms with defect type originate from differences in the screening of the defect-localized charge. A strongly localized charge is close to the point-ion model, whereas delocalized defect states spread out and result in weak Madelung corrections. To convince oneself of the proper elimination of the spurious electrostatic energy, one should, in the general case, carry out a proper scaling analysis of the L^{-1} behavior and extrapolate to the infinite-L limit.

Another approach is based on fitting a multicenter Gaussian distribution to the defect charge density, and calculating and subtracting explicitly the electrostatic interactions between the Gaussian distributions and the background. This approach has so far only been used for charged molecules in vacuum [85], but it may be possible to generalize to defect supercells with aperiodic densities.

A different route to the electrostatic correction for charged defects was taken by *Schultz* [86], based again on the linearity of the Poisson equation but avoiding the neutralizing homogeneous background. An aperiodic defect charge distribution $\rho_{\mathrm{LM}}(\boldsymbol{r})$ is constructed so that it matches the electrostatic moments of the system up to a given order. When this charge density is separated from the supercell charge distribution, the remaining periodic distribution has zero net charge and is momentless. Thus, its electrostatic energy can be accurately calculated within the periodic boundary conditions. The

Coulomb energy of $\rho_{\mathrm{LM}}(\boldsymbol{r})$ is calculated using "cluster boundary conditions", with the surrounding polarizable, defect-free crystal replaced by perfect, non-polarizable bulk crystal. There is no analytical formula similar to (27) for this local-moment counter-charge (LMCC) method. Its main content is to use periodic boundary conditions to solve the Kohn–Sham equations but not for the Poisson equation.

One can then calculate the electrostatic energy of a single defect-containing supercell surrounded by cells with the perfect crystal charge density. In order to correctly align the arbitrarily referenced potentials, *Schultz* [87] has suggested a method where a common reference potential can be calculated for all types of defects and also for the valence-band edge of the perfect crystal.

The LMCC method should, in principle, lead to a rigorous $1/L$ convergence of the supercell energy, with a prefactor proportional to $(1-\varepsilon^{-1})q^2$. This method does not need a compensating background, and there are no interactions between the defect charge and its periodic images. It also provides a way to define a reference energy independent of the defect charge state for aligning the band edges. The implementation of the LMCC method would thus seem very much worth the effort for calculations of charged defects.

15 Energy-Level References and Valence-Band Alignment

As is obvious from (13), another important parameter for supercell-based total-energy calculations is the position of the valence-band maximum (VBM), E_{v}, which is usually taken as the reference energy for the electron chemical potential. The position of the VBM of the defect-containing supercell is different from that of the defect-free supercell, and this difference depends, in general, on the charge state of the defect. Consequently, a valence-band alignment (VBA) correction is usually applied by matching the two values. The magnitude of this correction is

$$\Delta E_{\mathrm{VBA}} = \left\langle V^{\mathrm{eff}}_{\mathrm{bulk}} \right\rangle - \left\langle V^{\mathrm{eff}}_{\mathrm{defect}} \right\rangle , \tag{29}$$

where V^{eff} is the effective Kohn–Sham potential and the brackets denote averaging over the bulk and defect supercells, respectively.

Other potential-energy references are naturally possible. One possibility is to align an electronic core or semicore level energy (in all-electron calculations), or define all energies with respect to the so-called crystal zero, the potential energy at the surface of a neutral Wigner–Seitz cell. This choice is particularly simple and useful when using the atomic-sphere approximation (ASA), for example in the context of the linear-muffin-tin-orbital (LMTO) method.

16 Examples: The Monovacancy and Substitutional Copper in Silicon

The literature on supercell calculations of defects in semiconductors is vast and not even a partial review is attempted here. Instead, we discuss in more detail a recent systematic case study, where the practical implementations of two popular approaches were critically examined. The study involves the prototypical (but far from trivial) point defects in Si, the monovacancy V_{Si} and substitutional copper $\mathrm{Cu_{Si}}$.

The fundamental nature of V_{Si} means that it is a much-studied defect. It has four unpaired electrons, and in a simple LCAO picture they occupy four orbitals in the atoms surrounding the vacancy in the diamond structure [66]. These induce electronic states in the gap, which are modified by symmetry-lowering structural distortions driven by the Jahn–Teller effect.

A comprehensive study of the convergence of the formation energy of the neutral silicon vacancy as a function of the supercell size has been given by *Puska* et al. [18] (see also [88]). The result is shown in Fig. 5. It turns out that very large supercells are necessary for proper convergence of the vacancy formation energy, and that the relaxation pattern around the vacancy undergoes a qualitative change (from outward to inward relaxation) as the supercell size grows. This is also shown by calculations for large finite clusters [19]. The total-energy hypersurface is flat, especially for negatively charged defects, which makes the unambiguous determination of the point symmetry difficult. This is also reflected in the difficulty of determining the ionization levels accurately. This will now be discussed for both V_{Si} and $\mathrm{Cu_{Si}}$.

$\mathrm{Cu_{Si}}$ has a similar structure to V_{Si}: the vacancy is perturbed by the Cu $3d$ states that lie in the valence band. The Jahn–Teller theorem predicts that $\mathrm{Cu_{Si}}$ should undergo a symmetry-lowering distortion with magnitude constrained by the presence of the metal atom. Spin-orbit coupling effects may also reduce the amount of distortion by lifting the orbital degeneracy.

16.1 Experiments

The greater structural freedom of V_{Si} introduces another possible (and often encountered) complication. The total energy gained by pairing electrons in dangling bonds, associated with a structural distortion, can outweigh their mutual Coulomb repulsion. This is the famous "negative" effective-U phenomenon, which leads to unexpected ordering of the ionization levels in the bandgap. In the case of V_{Si} the pairing for the neutral defect is much stronger than for the singly positive +1 state, which shifts the ionization (donor) level $E_{\mathrm{d}}(0/+1)$ below that for the second donor $E_{\mathrm{d}}(+1/+2)$. Thus the singly positive silicon vacancy is a metastable state, a fact confirmed with DLTS and EPR measurements by *Watkins* and *Troxell* [89]. The energy position measured in these experiments gives for the double donor level the value $E_{\mathrm{d}}(+1/+2) \approx E_{\mathrm{v}} + 0.13\,\mathrm{eV}$, while EPR studies show that $\mathrm{V_{Si}^{+}}$

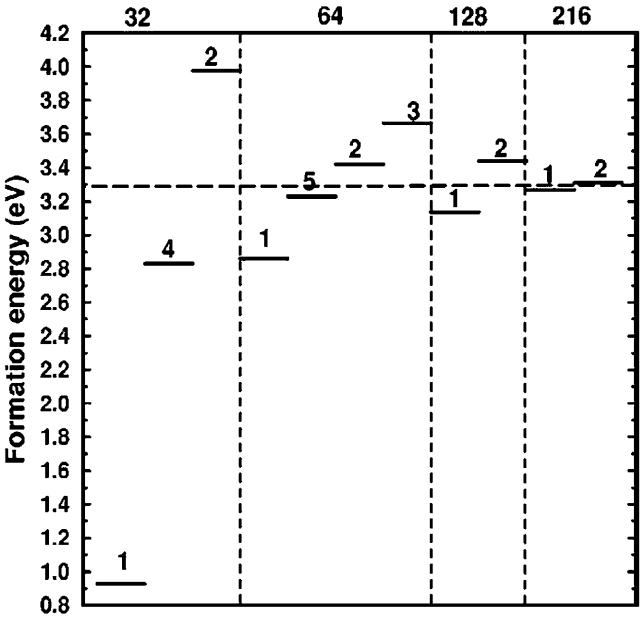

Fig. 5. Formation energy of the neutral vacancy in Si. The supercell size (number of atoms) is given on the top of each panel, and the $\boldsymbol{k}$-point set used for Brillouin-zone sampling is indicated as $1 = \Gamma$, $2 = 2^3$, $3 = 3^3$, $4 = (1/4, 1/4, 1/4)$ and $5 = \Gamma + L$. From [18]

ionized by photoexcitation decays to the neutral state from a donor level at $E_\mathrm{d}(0/+1) \approx E_\mathrm{v} + 0.05\,\mathrm{eV}$. EPR studies together with stress-alignment experiments also demonstrate that the Jahn–Teller effect causes structural distortions in accord with a straightforward one-electron model. This model predicts that the neutral and positively charged defects have D_{2d} symmetry, while the negative charge states have C_{2v} symmetry. The energies of the acceptor levels $E_\mathrm{d}(-1/0)$ and $E_\mathrm{d}(-2/-1)$ are experimentally unknown.

For $\mathrm{Cu_{Si}}$, there are accurate experimental values [90, 91] for its ionization levels and thus it provides a good testing ground for theory and computation, even if experimental structural information is lacking. The defect has a single donor level at $E_\mathrm{d}(0/+1) \approx E_\mathrm{v} + 0.207\,\mathrm{eV}$, an acceptor level $E_\mathrm{d}(-1/0)$ at $\approx E_\mathrm{v} + 0.478\,\mathrm{eV} \approx E_\mathrm{c} - 0.69\,\mathrm{eV}$, and a double acceptor at $E_\mathrm{d}(-2/-1)$ at $E_\mathrm{c} - 0.167\,\mathrm{eV}$.

The Jahn–Teller distortions for V_Si and $\mathrm{Cu_{Si}}$ (and more generally substitutional transition-metal impurities in Si) possess two components. The main one is tetragonal in character and gives the defects D_{2d} symmetry. The direction of this distortion may occur in two senses, and this determines the relative energetic ordering of the resulting electronic states. The two senses correspond to two different shapes for the bounding box with {100} faces that contains the four atoms surrounding the center: one is short and broad

while the other is long and thin. They correspond to different relative lengths for the six Si–Si distances between these four atoms. For the D_{2d} point group, four of the distances belong to one equivalent set and two belong to the second equivalent set. One type of distortion (type A) has the pair longer than the four, and for the other (type B) it is the other way round. Type A splits the t_2 state into a singlet a_1 state above the doublet e state, while type B does the opposite. Transition-metal impurities relax in the A pattern, and monovacancies in Si relax in the opposite sense, B.

When the system contains sufficient electrons to occupy the e state, a weaker trigonal distortion is expected. It lowers the symmetry to C_{2v}. The equivalent pair of Si–Si lengths in the D_{2d} case become unequal. This splits the e state into orbitals of b_1 and b_2 character. Spin-orbit coupling may also affect the splitting of these states for transition-metal impurities [92].

16.2 Calculations

In a recent study, *Latham* et al. [93] carried out a systematic quantitative study of the electrical levels and associated structural relaxations of V_{Si} and $\mathrm{Cu}_{\mathrm{Si}}$. To conduct the study, two different computer programs, AIMPRO [94] and VASP [95, 96] were used. The former uses localized Gaussian orbitals as the basis set for the Kohn–Sham wavefunctions, while the charge density is expanded in plane waves. VASP uses plane waves throughout. In AIMPRO, the core electrons of atoms are represented by pseudopotentials constructed according to the Hartwigsen–Goedecker–Hütter (HGH) norm-conserving scheme [45]. The VASP package includes pseudopotentials based on the Vanderbilt ultrasoft [46] construction or, alternatively, the projected-augmented-wave (PAW) method [48]. The Cu $3d$ electrons are included explicitly in the HGH and PAW schemes, while they are not in the Vanderbilt pseudopotential.

In both methods, the exchange-correlation energy is evaluated using the LSDA formula described by *Perdew* and *Wang* [97]. Unit cells of 216 atoms are used in all cases, and the supercell band structure is sampled by using the Monkhorst–Pack scheme with $2^3\boldsymbol{k}$ points, folded according to the symmetry of the system and shifted to avoid the Γ point. This enables the total energy convergence of better than 10^{-4} eV per supercell, with the plane-wave kinetic energy cutoffs chosen to meet this requirement (in AIMPRO the Coulomb energy is evaluated also using the plane-wave expansion). The actual value of the cutoff depends on the atoms present and the chosen pseudopotential. For charged defects, a compensating background charge is introduced as discussed in Sect. 13, and a finite-size scaling is performed.

For bulk Si, both methods reproduce the lattice parameter a and the bulk modulus B_0 well (see Table 1). The calculation of the bandgap E_{g} is a deficiency of LSDA-DFT formalism, as discussed above. The two methods give similar values, $E_{\mathrm{g}}(\mathrm{AIMPRO}) = 0.71$ while $E_{\mathrm{g}}(\mathrm{VASP}) = 0.60$ eV. The

VASP results do not depend on whether the core electrons are treated by the ultrasoft pseudopotential or PAW.

Table 1. Calculated and measured values for the lattice parameter a and the bulk modulus B_0 of Si

Method	a (Å)	B_0 (GPa)
AIMPRO+HGH	5.395	97.0
VASP+VUS	5.390	87.7
VASP+PAW	5.403	72.8
Experiment	5.431	97.9

All the defect electron levels considered here are sufficiently deep so that the energies of donors can be given with respect to E_c and acceptors with respect to E_v. As reference (marker) energies Latham et al. choose addition/removal energies of ideal Si supercells with the same size (216 atoms) as used for the defect calculations. This means that the energy levels are referenced to the nearest ideal-crystal band edge (i.e., the marker method with respect to the two band edges). None of the studied defects are so shallow that their energy would be between the theoretical and true bandgap. If this were the case, it would be necessary to use the opposite ideal-crystal band edge as the reference potential.

A summary of the calculated electrical levels, compared with experimental data, is presented in Table 2.

Table 2. Calculated and measured electrical levels [93] for V_{Si} and Cu_{Si}

	Transition state	AIMPRO	VASP VUS	VASP PAW	[15]	Expt. [66, 90, 91]
V_{Si}	0/+2	$E_v + 0.0$	$E_v + 0.06$		$E_v + 0.15$	$E_v + 0.09$
V_{Si}	+1/+2	$E_v + 0.05$	$E_v + 0.11$		$E_v + 0.19$	$E_v + 0.13$
V_{Si}	0/+1	$E_v - 0.04$	$E_v + 0.02$		$E_v + 0.11$	$E_v + 0.05$
V_{Si}	−1/0	$E_c - 0.31$	$E_c - 0.34$		$E_c - 0.57$	exists
V_{Si}	−2/−1	$E_c - 0.43$	$E_c - 0.23$		$E_c - 0.40$	exists
V_{Si}	−2/0	$E_c - 0.37$	$E_c - 0.29$		$E_c - 0.49$	
Cu_{Si}	0/+1	$E_v + 0.17$	$E_v + 0.07$	$E_v + 0.17$		$E_v + 0.21$
Cu_{Si}	−1/0	$E_c - 0.50$	$E_c - 0.45$	$E_c - 0.34$		$E_c - 0.69$
Cu_{Si}	−2/−1	$E_c - 0.24$	$E_c - 0.18$	$E_c - 0.26$		$E_c - 0.17$

When both program packages are set to use similar thresholds in terms of energies and forces to decide structural optimization, the outcome shows some differences. The AIMPRO package finds that the direction of the main tetragonal distortion (D_{2d} symmetry) is in the expected direction, type B

for V_{Si} and type A for $\mathrm{Cu_{Si}}$. The VASP package also finds type B D_{2d} for neutral V_{Si}. VASP yields type A D_{2d} distortion for $\mathrm{Cu_{Si}}$ when the Vanderbilt pseudopotential is used, but no significant deviation from the perfect T_d symmetry when PAW is used.

Both methods give equal energies for the $(0/+1)$ level of $\mathrm{Cu_{Si}}$. Regardless of constraint choice, no clear evidence is seen for the expected trigonal component of the distortion of $\mathrm{Cu_{Si}}$. The situation is more complex for V_{Si}. While the optimized structure in positive and neutral charge states has D_{2d} symmetry as expected, the unconstrained lowest-energy structure is a D_{3d} symmetry "split-vacancy" configuration. The split-vacancy configuration is one where a Si atom is located between two unoccupied lattice sites, and is 0.04 eV lower in energy than the ideal T_d monovacancy. This result for negatively charged vacancies was also found by *Puska* et al. [18].

According to Latham et al., the energy differences between the C_{2v} and C_{3v} configurations of V_{Si} in the $q=-1$ and $q=-2$ charge states are 0.02 eV and 0.20 eV, respectively. This means that V_{Si} in the C_{3v} symmetry is calculated to have a negative-U behavior also for the acceptor levels. Little is presently known from experiment about the negative charge states other than that they apparently exist.

The calculated energy for the $(-1/0)$ acceptor level of $\mathrm{Cu_{Si}}$ near midgap is somewhat deeper than the measured value, while the second acceptor is shallower. The Jahn–Teller effect moves these levels apart, moving each by 0.1 eV. Thus, the calculations seem to slightly underestimate the magnitude of the Jahn–Teller effect. It is reasonable to suppose that a similar pattern will be found for the acceptor states of V_{Si} if they can be measured, and the expected C_{2v} model without a negative-U effect would prevail.

17 Summary and Conclusions

Calculations based on the density-functional theory and the supercell method enable quantitative estimates for the formation energies, diffusion barriers, structural parameters and electrical levels of defects in semiconductors. Moreover, vibrational entropies and free energies can be estimated from the calculated force fields, which also enable first-principles molecular dynamics simulations. It is also possible to feed the first-principles results for migration and reaction barriers into kinetic Monte Carlo simulations based on master equations. This opens up the possibility for multiscale simulations of annealing kinetics and related phenomena. Quantitatively accurate total-energy calculations open the way to first-principles thermodynamics and kinetics.

However, reaching the desired quantitative accuracy in the total-energy calculations is not always straightforward, and requires careful consideration of several possible sources of error. In particular, the positioning of defect-related electronic levels in the fundamental semiconducting gap can be quite problematic. The sources of errors may be divided into three main categories.

One is the underlying theory, especially the treatment of electronic exchange and correlation. Local and semilocal approximations do not properly describe the discontinuity of the corresponding density functional, which contributes to the underestimation of the fundamental gap. The second source of error are the finite-size effects inherent in the supercell method, which need proper scaling analysis as a function of the cell size. These errors include the spurious defect–defect interaction, dispersion of defect-related electronic states, incomplete sampling of the reciprocal cell, and the constrained relaxation of atoms around the defect. The third source of errors are the approximations required to construct a specific implementation, including the pseudopotential generation, basis-set construction, and numerical accuracy of the algorithms.

The power of the density-functional methods is considerable in revealing and rationalizing the systematic trends in electronic and structural properties of defects in semiconductors. These include the nature (acceptor or donor) of the deep levels, their spin structure, point symmetry, and energetics. They can also predict the localized vibrational modes associated with defects, and provide the starting point (for example, the ground-state electron density) for quantitative calculations of many experimentally observed characterizing signals (such as positron-annihilation parameters). Total-energy calculations also reveal the details of the potential-energy hypersurfaces for defects moving in the lattice, which makes it possible to study the long-time kinetics by using Monte Carlo techniques.

Quantitatively accurate supercell calculations for defects in semiconductors are notoriously difficult and demand considerable computational resources. While many confusing and seemingly contradictory results have been published in the scientific literature, the available computational methods have steadily matured. With the increase of computer power available for such calculations, it is now possible to perform well-benchmarked calculations with considerable predictive power.

Acknowledgements

I have benefited from several useful discussions with Chris Latham and Maria Ganchenkova. This work has been supported by the Academy of Finland through the Center of Excellence Grant (2000–2005).

References

[1] J. Bourgoin, M. Lannoo: *Point Defects in Semiconductors II. Experimental Aspects*, vol. 35, Springer Series in Solid State Sciences (Springer, Berlin, Heidelberg 1983) 29
[2] R. M. Nieminen: J. Phys.: Condens. Matter **14**, 2859 (2002) 30
[3] W. Kohn, L. J. Sham: Phys. Rev. **136**, B864 (1964) 30
[4] R. Martin: *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge 2004) 30

[5] A. E. Mattsson, P. A. Schultz, M. P. Desjarlais, T. R. Mattsson, K. Leung: Modelling Simul. Mater. Sci. Eng. **13**, R1 (2005) 32
[6] M. J. Puska, R. M. Nieminen: Rev. Mod. Phys. **66**, 841 (1994) 32
[7] C. G. van de Walle, P. E. Blöchl: Phys. Rev. B **47**, 4244 (1993) 32
[8] C. J. Pickard, F. Mauri: Phys. Rev. B **63**, 245101 (2001) 32
[9] C. J. Pickard, F. Mauri: Phys. Rev. Lett. **88**, 086403 (2002) 32
[10] R. P. Messmer, G. D. Watkins: Phys. Rev. Lett. **25**, 656 (1970) 33
[11] S. Ögut, H. Kim, J. R. Chelikowsky: Phys. Rev. B **56**, R11353 (1997) 33
[12] G. Baraff, M. Schlüter: Phys. Rev. B **19**, 4965 (1979) 33
[13] O. Gunnarsson, O. Jepsen, O. K. Andersen: Phys. Rev. **27**, 7144 (1983) 33
[14] M. J. Puska, O. Jepsen, O. Gunnarsson, R. M. Nieminen: Phys. Rev. B **34**, 2695 (1986) 33
[15] S. K. Estreicher, D. S. Marinyck: Phys. Rev. Lett. **56**, 1511 (1986) 33, 59
[16] S. B. Zhang, J. E. Northrup: Phys. Rev. Lett. **67**, 2339 (1991) 34, 49
[17] H. Dreyssé (Ed.): *Electronic Structure and Physical Properties of Solids. The Uses of the LMTO Method*, Lecture Notes in Physics (Springer, Berlin, Heidelberg 2000) 35
[18] M. J. Puska, S. Pöykkö, M. Pesola, R. M. Nieminen: Phys. Rev. B **58**, 1318 (1998) 35, 36, 56, 57, 60
[19] S. Ögut, J. R. Chelikowsky: Phys. Rev. B **64**, 245206 (2001) 35, 56
[20] K. Laaksonen, H. P. Komsa, E. Arola, T. T. Rantala, R. M. Nieminen: Phys. Rev. B in press 37
[21] M. Städele, J. A. Majewski, P. Vogl, A. Görling: Phys. Rev. Lett. **79**, 2089 (1997) 37
[22] M. Städele, M. Moukara, J. A. Majewski, P. Vogl, A. Görling: Phys. Rev. B **59**, 10031 (1999) 37
[23] R. J. Magyar, A. Flezsar, E. K. U. Gross: Phys. Rev. B **69**, 045111 (2004) 37
[24] D. Vogel, P. Krüger, J. Pollmann: Phys. Rev. B **54**, 5495 (1996) 37
[25] S. Sharma, J. K. Dewhurst, C. Ambrosch-Draxl: Phys. Rev. Lett. **95**, 136402 (2005) 37
[26] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, M. Levy: Phys. Rev. B **53**, 3764 (1996) 38
[27] J. Lento, R. M. Nieminen: J. Phys.: Condens. Matter **15**, 4387 (2003) 38
[28] M. S. Hybertsen, S. G. Louie: Phys. Rev. B **34**, 5390 (1986) 38
[29] S. Ismail-Beigi, S. G. Louie: Phys. Rev. Lett. **95**, 156401 (2005) 38
[30] J. Laegsgaard, K. Stokbro: Phys. Rev. Lett. **86**, 2834 (2001) 38
[31] M. d'Aveszac, M. Calandra, F. Mauri: Phys. Rev. B **71**, 205210 (2005) 39
[32] J. Perdew, A. Zunger: Phys. Rev. B **23**, 5048 (1981) 39
[33] A. Svane, O. Gunnarsson: Phys. Rev. B **37**, 9919 (1988) 39
[34] L. Hedin: Phys. Rev. **139**, A796 (1965) 39
[35] G. Onida, L. Reining, A. Rubio: Rev. Mod. Phys. **74**, 601 (2002) 39
[36] M. Rohlfing, S. G. Louie: Phys. Rev. B **62**, 4927 (2000) 39
[37] A. Janotti, C. G. van de Walle: Appl. Phys. Lett. **87**, 122102 (2005) 40
[38] S. Lang, A. Zunger: Phys. Rev. B **72**, 035215 (2005) 40
[39] V. I. Anisimov, F. Aryasetiawan, A. I. Liechtenstein: J. Phys.: Condens. Matter **9**, 767 (1997) 40
[40] W. K. Leung, R. J. Needs, G. Rajagopal, S. Itoh, S. Ihara: Phys. Rev. Lett. **83**, 2351 (1999) 40

[41] R. Q. Hood, P. R. C. Kent, R. J. Needs, P. R. Briddon: Phys. Rev. Lett. **91**, 076403 (2003) 40
[42] C. G. Van de Walle, J. Neugebauer: J. Appl. Phys. **95**, 3851 (2004) 40
[43] D. R. Hamann: Phys. Rev. B **40**, 2980 (1989) 41
[44] N. Troullier, J. L. Martins: Phys. Rev. B **43**, 1993 (1991) 41
[45] C. Hartwigsen, S. Goedecker, J. Hütter: Phys. Rev. B **58**, 3641 (1998) 41, 58
[46] D. Vanderbilt: Phys. Rev. B **41**, 7892 (1990) 41, 58
[47] L. Kleinman, D. M. Bylander: Phys. Rev. Lett. **48**, 1425 (1982) 41
[48] P. E. Blöchl: Phys. Rev. B **50**, 17953 (1994) 41, 58
[49] G. Kresse, D. Joubert: Phys. Rev. B **59**, 1758 (1999) 41
[50] S. G. Louie, S. Froyen, M. L. Cohen: Phys. Rev. B **26**, 1738 (1982) 41
[51] D. Porezag, M. R. Pederson, A. V. Liu: Phys. Rev. B **60**, 14132 (1999) 41
[52] D. Vogel, P. Krüger, J. Pollmann: Phys. Rev. B **55**, 12836 (1997) 41
[53] M. Fuchs, M. Bockstedte, E. Pehlke, M. Scheffler: Phys. Rev. B **57**, 2134 (1998) 42
[54] P. Ordejon, E. Artacho, J. M. Soler: Phys. Rev. B **53**, R10441 (1995) 42
[55] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejón, D. Sanchez-Portal: J. Phys.: Condens. Matter **14**, 2745 (2002) 42
[56] O. Treutler, R. Ahlrichs: J. Chem. Phys. **102**, 346 (1995) 42
[57] T. L. Beck: Rev. Mod. Phys. **72**, 1041 (2000) 43
[58] M. Heiskanen, T. Torsti, M. J. Puska, R. M. Nieminen: Phys. Rev. B **63**, 245106 (2001) 43
[59] S. Nosé: J. Chem. Phys. **81**, 511 (1984) 43
[60] G. H. Hoover: Phys. Rev. A **31**, 1695 (1985) 43
[61] M. Parrinello, A. Rahman: J. Appl. Phys. **52**, 7182 (1980) 44
[62] E. Runge, E. K. U. Gross: Phys. Rev. Lett. **52**, 997 (1984) 44
[63] M. Petersilka, U. J. Gossmann, E. K. U. Gross: Phys. Rev. Lett. **76**, 1212 (1996) 44
[64] G. Vignale, W. Kohn: Phys. Rev. Lett. **77**, 2037 (1996) 44
[65] H. A. Jahn, E. Teller: Proc. Roy. Soc. A **161**, 220 (1937) 45
[66] G. D. Watkins: *The Lattice Vacancy in Silicon* (Gordon and Breach, Yverdon 1992) 45, 56, 59
[67] G. Baraff, E. O. Kane, M. Schlüter: Phys. Rev. Lett. **43**, 956 (1979) 46
[68] G. D. Watkins: Physica B **117–118**, 9 (1983) 46
[69] L. Torpo, R. M. Nieminen, S. Pöykkö, K. Laasonen: Appl. Phys. Lett. **74**, 221 (1999) 46
[70] S. Baroni, S. de Gironcoli, A. Dal Corso: Rev. Mod. Phys. **73**, 515 (2001) 47
[71] J. M. Pruneda, S. K. Estreicher, J. Junquera, J. Ferrer, P. Ordejon: Phys. Rev. B **65**, 075210 (2002) 47
[72] J. Coutinho, V. J. B. Torres, R. Jones, P. R. Briddon: Phys. Rev. B **67**, 035205 (2003) 50
[73] D. J. Chadi, M. L. Cohen: Phys. Rev. B **8**, 5747 (1973) 52
[74] H. J. Monkhorst, J. D. Pack: Phys. Rev. B **13**, 5188 (1976) 52
[75] G. Makov, R. Shah, M. C. Payne: Phys. Rev. B **53**, 15513 (1996) 52
[76] J. Shim, E.-K. Lee, Y. J. Lee, R. M. Nieminen: Phys. Rev. B **71**, 035206 (2005) 52
[77] S. W. de Leeuw, J. W. Perram, E. R. Smith: Proc. Roy. Soc. London, Ser. A **373**, 27 (1980) 53
[78] M. Leslie, M. J. Gillan: J. Phys. C **18**, 973 (1985) 53

[79] G. Makov, M. C. Payne: Phys. Rev. B **51**, 4014 (1995) 53
[80] M. R. Jarvis, I. D. White, R. W. Godby, M. C. Payne: Phys. Rev. B **56**, 14972 (1997) 53
[81] L. N. Kantorovich: Phys. Rev. B **60**, 15476 (1999) 53
[82] J. Lento, J.-L. Mozos, R. M. Nieminen: J. Phys.: Condens. Matter **14**, 2637 (2002) 54
[83] D. Segev, S. H. Wei: Phys. Rev. Lett. **91**, 126406 (2003) 54
[84] J. Shim, E.-K. Lee, Y. J. Lee, R. M. Nieminen: Phys. Rev. B **71**, 245204 (2005) 54
[85] P. Blöchl: J. Chem. Phys. **103**, 7482 (1995) 54
[86] P. A. Schultz: Phys. Rev. B **60**, 1551 (1999) 54
[87] P. A. Schultz: Phys. Rev. Lett. **84**, 1942 (2000) 55
[88] D. A. Drabold, J. D. Dow, P. A. Fedders, A. E. Carlsson, O. F. Sankey: Phys. Rev. B **42**, 5345 (1990) 56
[89] G. D. Watkins, J. R. Troxell: Phys. Rev. Lett. **44**, 593 (1980) 56
[90] S. Knack, J. Weber, H. Lemke: Physica B **273–274**, 387 (1999) 57, 59
[91] S. Knack, J. Weber, H. Lemke, H. Riemann: Phys. Rev. B **65**, 165203 (2002) 57, 59
[92] G. D. Watkins, P. M. Williams: Phys. Rev. B **52**, 16575 (1995) 58
[93] C. D. Latham, M. Ganchenkova, R. M. Nieminen, S. Nicolaysen, M. Alatalo, S. Öberg, P. R. Briddon: Phys. Scr. **T126**, 61 (2006) 58, 59
[94] J. Coutinho, R. Jones, P. R. Briddon, S. Öberg: Phys. Rev. B **62**, 10824 (2000) 58
[95] G. Kresse, J. Furthmüller: Comp. Mater. Sci. **6**, 15 (1996) 58
[96] G. Kresse, J. Furthmüller: Phys. Rev. B **54**, 11169 (1996) 58
[97] J. P. Perdew, Y. Wang: Phys. Rev. B **46**, 12947 (1992) 58

Index

Marker-Method Calculations for Electrical Levels Using Gaussian-Orbital Basis Sets

J.P. Goss, M.J. Shaw, and P.R. Briddon

School of Natural Science, University of Newcastle, Newcastle upon Tyne, NE1 7RU
{J.P.Goss,M.J.Shaw,Patrick.Briddon}@newcastle.ac.uk

Abstract. The introduction of defect-related states in the bandgap of semiconductors can be both advantageous and deleterious to conduction, and it is therefore of great importance to have quantitative computational methods for determining the location of electrical levels. In particular, where the defect levels are deep in the bandgap, the states involved are typically highly localized and the application of real-space, localized basis sets have clear advantages. In this chapter the use of such basis sets both for cluster and supercell geometries is discussed. Agreement with experiment is often hampered by problems such as the underestimate of bandgaps when using density-functional theory. We show that these can be somewhat mitigated by the use of "markers", either experimental or theoretical, to largely eliminate such systematic errors.

1 Introduction

It is well known that the electrical and optical properties of a crystalline material are influenced by impurities and other defects. In particular, conductivity at a given temperature is related to the free-carrier concentration, that in turn depends on the depth of the donor or acceptor level of the dominant dopant. Therefore one requires dopants with the smallest ionization energy, a value that is limited to the effective-mass level of the host material [1]. Conversely, electron and hole traps deep in the bandgap are detrimental to the conductivity because they reduce the free-carrier concentration and the resultant *charged* defects are scattering centers that reduce carrier mobility. Deep levels are also nonradiative recombination centers affecting optical properties.

The importance of the electrical characteristics of defects has driven a huge effort to develop reliable computational methods that can both explain experimental observations and predict new dopants with the desired properties. Currently, most calculations in this field are performed using density-functional theory (DFT) [2, 3], chiefly simulating crystalline materials using periodic boundary conditions (PBCs), and plane wave (PW) basis sets[1] to represent the wavefunctions of the electrons.

[1] Plane waves represent functions in a Fourier-transformed phase space and are of the form $\exp(\mathrm{i}\boldsymbol{k}.\boldsymbol{r})$, where $\boldsymbol{k}$ is the wavevector that is related to an energy by $E = \hbar^2 k^2/2m$.

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors, Topics Appl. Physics **104**, 69–94 (2007)

With PBCs, a section of host material containing the dopant or defect is constructed and this "supercell" is repeated periodically throughout space. Such a construction represents a crystal of defects, and thus the use of PWs is a natural choice. If the defect and its periodic images are sufficiently spatially separated they do not interact significantly and the calculated properties represent those of isolated defects. This means that in order to obtain reliable data for defects (especially shallow dopants with large wavefunction extents) large supercells must be used. However, the computational cost increases rapidly with system size and in practice most of today's calculations contain up to a few hundred atoms.

The DFT-PBC approach has at least two major problems when calculating electrical properties. The first is that within most commonly used local-density (LDA) or generalized-gradient (GGA) approximations, the bandgap is underestimated, sometimes catastrophically. For example, crystalline Ge, in reality a semiconductor with a bandgap 0.74 eV [4], is found to be a semimetal [5]! Secondly, in order to obtain information about the capacity for defects to adopt different charge states one needs to know the properties of *charged systems*. Strictly this is not possible with PBCs, but charged systems can be *approximated* by the use of a uniform background charge that exactly cancels the charge on the defect. Even then, this represents a space-filling array of charges that have an associated electrostatic energy that does not correspond to the properties of isolated defects.

These problems are both mitigated by the use of atomic clusters. Here, defects are imbedded into large sections of the host material, but instead of repeating it periodically, the surface of the material is passivated (usually by hydrogen) effectively forming a large molecule. The cluster geometry imposes an artificial confining potential increasing the bandgap and localizing electronic states. However, although this effect counteracts the inherent underestimate of the bandgap, the surface–defect interactions are of a similar nature to the defect–defect interactions in supercells.

Although there are several sources of error in the calculations, many of them are somewhat systematic in nature, and hence relative locations of electrical levels may be more reliable than absolute values. This has led to the use of so-called marker methods [6–8], where one references the system of interest to one that is well understood: for example, the donor level of P in diamond is known, whereas those of As and Sb are not. Theory shows that they are shallower than phosphorus [9] and may therefore represent an improvement in the production of n-type diamond. (Of course, such calculations do not reflect on the relative solubilities of these dopants.)

Finally, let us introduce the role of the basis functions used to describe the electrons in our problem. The use of PWs does not represent a fundamental problem, but in order to treat many systems of interest where the wavefunctions and charge density vary rapidly in space, the use of PWs often necessitates additional effort (such as high-energy cutoffs) or the use of so-called ultrasoft pseudopotentials where one systematically removes the as-

pects of the electrons that are difficult to represent efficiently with PWs due to their localized nature. It is true that PWs may be viewed as a near-ideal basis set for effective-mass-like shallow levels, since such donor and acceptor states are built out of the band-edge Bloch states. However, for deep levels the wavefunctions are often highly localized and can be characterized rather more simply by the use of molecular orbitals.

The DFT package AIMPRO (Ab-Initio Modeling PROgram) uses a real-space basis made up from Gaussian functions [10, 11]. The real-space basis also allows the same electronic-structure methods to be applied to atomic clusters and molecules. In this Chapter, we first present a detailed description of the basis sets used in our calculations. In Sect. 3 we describe the computational methods for calculating electrical levels, and in particular the marker method and present a review a range examples of its use in different group-IV materials in Sect. 4.

2 Computational Method

Our intention is to detail the calculation of electrical levels and not review the entire formalism used in the calculations. However, briefly, AIMPRO is an electronic-structure program using LDA- and GGA-DFT. Atoms are treated using pseudopotentials [12] to remove the core electrons, and the distribution of the electrons solved self-consistently for a given set of atoms. AIMPRO can be used with a range of pseudopotential types, including the seminonlocal forms [13–15], and the dual-space separable pseudopotentials [16]. In addition to the efficiencies offered by their separability, the latter pseudopotentials offer extended norm conservation, accounting for a large number of occupied and unoccupied atomic levels. Currently, by default, AIMPRO uses the Perdew–Wang functional [17] for PBC calculations and a Padé approximation to this functional [18] for cluster calculations, although a number of alternative functionals are available. Historically, cluster calculations used an extremely fast parameterized analytic form for the exchange-correlation functional as described in detail previously [11]. GGA calculations are performed using a White–Bird [19] implementation of the GGA functional [20].

AIMPRO can be used both in periodic and cluster modes and the atomic structures optimized either using static minimization methods or molecular dynamics. From these calculations a range of observables can be derived, such as the electrical levels introduced above, vibrational-mode frequencies and symmetries, electronic-structures, hyperfine and zero-field splitting tensors, electron energy loss spectra, migration barriers, and so on. However, in contrast to the majority of PBC calculations, AIMPRO represents the wavefunctions using a range of real-space, Gaussian-orbital functions.

2.1 Gaussian Basis Set

In order to construct and solve the Hamiltonian for our systems we expand our wavefunctions in terms of a basis set. Many different basis types are used for first-principles calculations, and each has associated advantages and disadvantages. There is no firm consensus on the optimum basis set to use, and since the choice of basis set plays a significant role in determining the structure and nature of the calculations, the particular bases used and issues concerning them are worthy of some comment here. In the discussion that follows we shall consider only the arguments surrounding the selection of basis sets for pseudopotential calculations: clearly all-electron calculations, which must describe the rapid fluctuations of the wavefunctions in the ionic core region place additional demands on the basis set.

As alluded to above, of the many basis sets that are routinely used, perhaps the most significant distinction lies between the expansion in terms of PWs, and the use of basis functions that are localized functions in real space. `AIMPRO` uses the latter type, specifically a set of Cartesian Gaussian orbitals (CGOs) centered on atoms and possibly other positions. These basis functions are products of a simple Gaussian and a polynomial in the position vector relative to the center of the Gaussian:

$$\phi_{i,n_1n_2n_3}(\boldsymbol{r}) = x^{n_1}y^{n_2}z^{n_3}e^{-\alpha_i r^2}, \tag{1}$$

where α_i reflects the spatial extent of the function; larger (smaller) values correspond to more localized (diffuse). The term $x^{n_1}y^{n_2}z^{n_3}$ defines the angular dependence, and relates to the spherical-harmonic associated with the angular-momentum and magnetic quantum numbers, l and m. For example, a p_x-orbital has $n_1 = 1$, $n_2 = n_3 = 0$. These polynomial combinations enable the basis to describe atomic-like states of arbitrary angular momentum.

The basis-set is defined by the combination of Gaussian exponents and angular-momentum to which the combinations of polynomial functions extend. The advantages and disadvantages of PW and these real-space basis sets are already documented in the literature [11], however, it is useful to consider some of the key differences between them.

A key advantage of CGOs is that they can be *nonuniformly distributed in space*, lending a more flexible basis to regions that require it (for example in the region of a defect in a crystal). Such an approach cannot be taken with PW bases: the region with greatest complexity defines the density of PWs everywhere. For systems containing a small region in which the wavefunctions vary rapidly a high-energy cutoff might be required, and this dense PW basis set must exist throughout the whole crystal. This leads to a very large Hamiltonian and slow calculation. A particular example of this problem is where surfaces are treated using a PBC via the inclusion of regions of vacuum.

An advantage of a PW representation is that the basis functions are necessarily orthogonal. Consequently, the number of PWs can be increased without

limit while retaining the stability of the calculation. CGOs are not orthogonal and the use of too large a basis set can result in numerical instability developing. The stability of the basis becomes a significant issue when one attempts to address the question of how one should choose the basis for a particular problem. For a PW calculation it is possible to ensure that one has arrived at a converged result by systematically increasing the PW cutoff. For a CGO basis set it is not possible to perform a systematic test to ensure that a convergent basis has been arrived at. Indeed, the choice of exponents for the functions in the basis set is a nontrivial one. We shall discuss the details of the basis set optimization procedure in more detail below.

We now consider the CGOs in more detail. The task is to obtain a set of exponents for which the pseudowavefunctions are accurately described without rendering the calculations unstable. In order to achieve this, the variational principle is employed: when one compares any two basis sets, that which allows the wavefunctions to describe the true wavefunction most accurately results in the lowest total energy. The minimization of the total energy with respect to the exponents of the CGOs therefore results in the optimum basis set for a given number of functions. This process may then be repeated for different numbers of functions to assess whether the basis is large enough. Again, the number of CGOs included cannot be increased without limit as the calculations become unstable. However, it is typically possible to obtain a convergent set of exponents prior to the onset of instability. Of course, an increased number of exponents increases the size of the Hamiltonian, and in turn the computational cost. Typically, the optimum number of basis functions is obtained by balancing the absolute quality of the basis and the associated accuracy of the results against the computational load.

In addition to the number and value of the exponents, α_i, the polynomial functions included with a given Gaussian (s, p, d, f and so on) must also be chosen. In general, a basis set contains a range of values of angular momentum and each CGO is treated independently. The angular momentum has a dramatic impact upon the number of functions in the basis set: for example, angular momentum up to s, p and d results in 1, 4 and 10 functions per exponent, respectively[2].

As an example, four exponents per atom each of which include s and p, gives rise to 16 functions per atom. Increasing the basis to include d-functions increases the basis size to 40, and since diagonalization of the Hamiltonian scales as the cube of the basis size, the time to obtain a total energy would increase by more than an order of magnitude.

[2] It should be noted that the linear combinations of six functions with $n_1 + n_2 + n_3 = 2$ in (1) produce the five d-type functions (xy, xz, yz, $2z^2 - x^2 - y^2$ and $x^2 - y^2$), and an s-type function ($x^2 + y^2 + z^2$). Our standard practice in using the uncontracted basis sets discussed here is to include this additional s-type function, and hence the $n_1 = n_2 = n_3 = 0$, $n_1 + n_2 + n_3 = 1$ and $n_1 + n_2 + n_3 = 2$ terms yield ten rather than nine functions.

It is therefore paramount that basis functions of higher angular momentum are used selectively where the physical demands of the system merit them. Indeed, using a small number of high angular momentum functions is common practice in Hartree–Fock calculations where the basis sets are relatively small due to the extreme computational cost of this method, and are built up using the angular momentum of the atomic species with the addition of a single, higher angular momentum orbital as a polarization function.

For most calculations all of the CGOs are centered on atoms, and move with the atoms during structural optimization. It is also possible to center basis functions at sites where there are no atoms. For example, historically, centering basis functions at appropriate positions on the bonds between adjacent atoms in the crystal was used to improve the description of the angular variations due to the formation of bonding orbitals. Locating basis functions at bond centers allows for polarization to be incorporated without increasing the maximum angular momentum, reducing the computational cost, but introducing a poorly defined set of sites: during relaxation, structural changes may change the atoms between which bonding is present.

However, increases in the speed of computational resources, together with improved algorithms has meant that bond-centered orbitals have largely been superseded by the use of basis sets containing higher angular momentum. The higher angular momentum components provide the necessary angular fluctuations without the need for the additional complexity of bond-centered basis functions. Another situation in which localized basis functions may be centered away from atoms is the use of "ghost" atoms when treating surfaces. In this case, basis functions are placed in the vacuum region as though atoms were there (although of course, no atomic potential is included). This approach can provide additional basis functions to help describe the evanescent wavefunction variations close to the surface.

For the majority of calculations the basis functions are atom-centered CGOs. A set of basis functions is therefore associated with each atom. The size of the basis set, and the particular exponents in it, will be optimized for each species of atom. Additionally, it is sometimes appropriate to treat different atoms of the same species with different basis sets within the same calculation (for example, surface atoms may require a larger basis than bulk atoms in a surface calculation).

In the basis sets specified so far, all of the CGOs are independent basis functions, with coefficients free to change during the calculation, and each contributes to the overall dimension of the Hamiltonian. It is possible, however, to exploit the physical properties of the system to take linear combinations CGOs to form more complex basis functions that provide a comparable accuracy but with a reduced number of independent parameters. These are referred to as contracted basis sets and are developed in the following way.

A small reduction in the basis sets involving d-type (and higher angular momentum) functions can be obtained by taking linear combinations of the functions in (1). Formally,

$$\psi_{i,nlm} = \sum c_{nlm,n_1n_2n_3}\phi_{i,n_1n_2n_3} \,. \tag{2}$$

In the case containing up to quadratic prefactors there are ten functions ϕ, from which we may choose to produce a set of nine functions, ψ, which transform with angular momentum $l = 0, 1, 2$.

A more radical approach to reduce the size of a basis set uses combinations of these $\psi_{i,nlm}$:

$$\Psi_{nlm} = \sum C_i^{nl}\psi_{i,nlm} \,, \tag{3}$$

where the number of functions of the form of Ψ is dramatically less than the related set of ψ. Particular examples of these contracted bases appear below in Sect. 2.3.

2.2 Choice of Exponents

As mentioned above, the values of α_i may be chosen by consideration of the total energy. Typically, this optimization process is performed for a prototypical example of the system to be studied. For example, in the case of studies of defects in silicon, the basis associated with the Si atoms is chosen to best represent a bulk Si cell. A problem of optimizing the values of the exponents is a multidimensional minimization problem, and may be tackled by one of the many standard minimization procedures available. `AIMPRO` allows for the optimization of the exponents by a downhill simplex or a conjugate gradient minimization of the total energy.

Given that a typical element will require of the order of four different exponents in its basis set, the free optimization of the exponents is a four-dimensional minimization problem. Calculations have shown that the four-dimensional energy surface generally is very flat and undulating, resulting in many local minima, and causing difficulties in producing a *unique* choice of exponent. This is mitigated by the use of "even-tempered" basis sets in which the n exponents are constrained to be in a geometric series, and although it is generally true that an unrestricted basis set will lower the total energy, it is found that this is often not a significant effect.

For contracted basis sets, as with the full Gaussian basis sets, fixed parameters (in this case both coefficients and exponents) may be obtained with reference to an appropriate prototypical system. Each combined contracted basis function adds just one to the dimension of the Hamiltonian, and yet contains variations over a number of different length scales that reflect the properties of the physical system used in the fitting procedure. Once the contraction coefficients have been optimized, which is typically performed

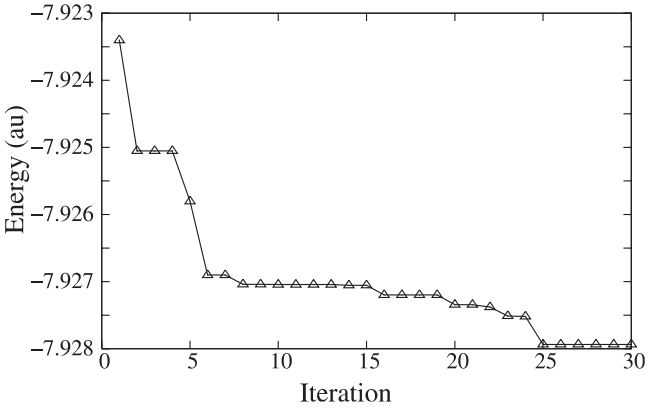

Fig. 1. Total energy of Si bulk unit cell during basis-exponent optimization. The energy is shown as a function of the number of simplex iterations

using a simplex optimization within `AIMPRO`, they remain fixed during the subsequent self-consistent field calculations.

The optimization of the exponents is an important step prior to commencing a real calculation. Clearly, a poorly chosen basis may impact on the accuracy of the final calculation, by affecting the accuracy of the wavefunction description. However, as we shall demonstrate in the example below, the physical properties of the material (for example lattice constant and bulk modulus) are not sensitive to the fine tuning of the basis. Only when the basis set has been carelessly chosen is there a significant impact upon the physical properties.

2.3 Case Study: Bulk Silicon

Let us consider the case of a range of basis sets for bulk silicon. Here, we shall consider the optimization of exponents in a four-exponent basis set, with angular momenta up to d on the first and second exponent, and up to p on the third and fourth exponent (where the first α is smallest). We refer to this as a *ddpp* basis. As an initial guess for the exponents we shall use the values $\{0.1, 0.3, 1.5, 4.0\}$. The range of these values results from a simple consideration of the length scales of the CGOs, the lowest exponent (longest-ranging Gaussian) allowing our basis to represent functions extending into the space between adjacent atoms, the highest exponent (shortest-range Gaussian) allowing us to describe the short-ranged variations of the valence pseudowavefunctions. (Note: the basis sets are described here for pseudopotential calculations – if all-electron calculations are to be performed, very much larger bases would be required, the highest exponents of which would be far greater than those needed for pseudopotentials.)

We show in Fig. 1 the total energy as a function of the number of steps in a simplex optimization of this basis. Note how the reduction in energy is relatively steep in the early stages followed by a long tail, reflecting how a modest attempt to obtain exponents that reflect the properties of the system is sufficient to obtain a reasonably well-converged total energy.

Several of the partially optimized bases were used to compute the lattice constant and bulk modulus of the Si by fitting to the Birch–Murnaghan equation of state. These data are presented in Table 1.

Table 1. Total energies (Hartrees) at energy minimum, lattice constants (au) and bulk moduli (GPa) for Si basis sets at various stages of optimization. For comparison, the experimental lattice constant and bulk modulus of silicon are 10.263 au and 97.9 GPa, respectively [21, 22]

Iterations	Total energy	Lattice constant	Bulk modulus
1	−7.9234	10.24	91
4	−7.9250	10.23	91
6	−7.9269	10.21	94
15	−7.9271	10.21	95
24	−7.9276	10.21	93
30	−7.9280	10.21	94
100	−7.9280	10.20	96

Since the LDA is generally expected to obtain lattice constants to around 1 % accuracy, and bulk moduli to around 5 % to 10 % it can be seen from Table 1 that the variation in the structural parameters with basis is relatively small. In particular, the initial, somewhat arbitrary parameters yield reasonable results, and after only around ten iterations do the changes in lattice constant and bulk modulus become minimal on this scale.

Table 2. Total energies (Hartrees) at energy minimum, lattice constants (au) and bulk moduli (GPa) for different optimized Si basis sets

Basis	No. functions per atom	Total energy	Lattice constant	Bulk modulus
dddd	40	−7.9335	10.17	95
ddpp	28	−7.9280	10.20	96
pdpp	22	−7.9268	10.21	95
pppp	12	−7.8974	10.35	85

We must also consider the role of the angular momentum in the CGOs. The size of basis set increases with higher angular momentum, slowing the calculation but providing a more flexible and accurate basis set. There is a

compromise to make between accuracy and computational cost. Typically, for materials like diamond and silicon *pdpp* or *ddpp* bases are found to be sufficient. However, in calculations of defects it is often the case that heavier bases are used for the atoms at the defect itself, such as *dddd*. This exploits the localized nature of the CGOs, improving the accuracy of the basis in a particular region with little effect on the overall accuracy or computational cost of the calculation.

Table 2 details the effect on the structural properties of silicon for various basis sets. The additional cost of the extra functions in the *dddd* basis is unnecessary since the results are effectively the same as the computationally cheaper *pdpp* basis. However, the *pppp* basis set is seen to be too small and the omission of any higher angular momentum functions prevents a sufficiently accurate description of the directional bonds found in covalent semiconductors.

As with the plain CGO basis, a number of different contracted basis sets can be defined according to the number of different combinations of the functions, and the angular momentum to which they extend. To label these contracted basis sets we draw on the conventional nomenclature of quantum chemistry, and specify, for example, a $4G$ basis to be one in which four CGOs (the "G") of different exponents are combined into fixed functions. Separate combinations are formed for n and l representing the occupied atomic valence states of the element in question. For example, for the s and p orbitals in silicon this results in a total of four independent functions, one s- and three p-polynomial combinations. The same contraction coefficients are used for different components of a given function (for example p_x, p_y and p_z), although they are of course included as independent basis functions. This level of contraction is known as a minimal basis set as they cannot change shape during the calculation.

To improve the flexibility of the basis set, a second set of contracted coefficients may be defined resulting in an additional set of four basis functions (a total of eight basis functions, two s-types and 6 p-type). In our nomenclature this would be referred[3] to as $44G$. Such basis sets are greatly restricted due to the fact that the highest angular momentum that can be represented is limited to the angular momentum of the occupied atomic states. Although these can describe the isolated atom well, higher angular momentum components are needed to describe directional bonds. To account for these we may introduce additional "polarization" functions of higher angular momentum CGOs (for example d functions in silicon) with a single free exponent. The inclusion of polarization functions in a basis set is indicated by a $*$ in the notation, for example $44G^*$. The addition of the polarization functions increases the basis

[3] This is similar to the "31" part of the quantum chemists standard 6-31G basis set. The only difference is that in our two contractions are combinations of all four underlying Gaussians; the 6-31G basis has 3 functions contracted together and one function left uncontracted, a more restrictive prescription.

from 8 to 13 functions per atom, which remains considerably smaller than the *ddpp* plain Gaussian basis.

We have extensively studied the effect of reducing the degrees of freedom in our basis sets by moving to increasingly small contracted sets. In Table 3, the minimum total energy, lattice parameter and bulk modulus are presented for contracted basis sets, and compared with the uncontracted *ddpp* basis.

Table 3. Total energies (Hartrees) at energy minimum, lattice constants (au) and bulk moduli (GPa) for contracted basis sets $44G^*$, $4G^*$, and $4G$, as defined in the text

Basis	No. functions per atom	Total energy	Lattice constant	Bulk modulus
ddpp	28	−7.9280	10.20	96
$44G^*$	13	−7.9269	10.19	97
$4G^*$	9	−7.9226	10.21	103
$4G$	4	−7.8854	10.39	91

A $44G^*$ basis for materials such as Si and diamond is found to be as effective as a 28-function *ddpp* basis. In other words, the $44G^*$ basis provides a convergent description of our wavefunctions for our typical host elements, at a much reduced cost in terms of the Hamiltonian. These contracted basis sets are therefore used in many of our calculations, giving an optimum performance with regard to computational cost. Cheaper bases still, such as the $4G$ basis are found to result in wider variations of lattice constant and bulk modulus, as well as giving poor electronic band structures, particularly in the conduction band.

Although the $4G$ bases are not suitable for routine calculations, there are a number of potential applications for these basis sets. One possibility is that these extremely cheap bases (just four functions per atom in Si) could be used for bulk regions in very large unit cell calculations. Fuller bases could be used to describe the essential regions of the material (for example the defect region and adjacent atoms) with the $4G$ bases filling the remainder of the bulk-like unit cell. For such schemes to work it is necessary to match the optimum lattice constant of the two bases to avoid unphysical internal strain at the boundary between the two different bases. Although we have implemented this approach and tested the idea with encouraging results, it has yet to be used in a real application.

2.4 Charge-Density Expansions

In AIMPRO the charge density, $n(\boldsymbol{r})$, is also fitted to a set of basis functions. With the PBC $n(\boldsymbol{r})$ is expanded in PWs, so that the approximate charge density

$$\tilde{n}(\boldsymbol{r}) = \sum_{G} n(G) \mathrm{e}^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}} , \tag{4}$$

where all PWs with $G^2/2 < E_{\text{cut-off}}$ are included, with the energy cutoff chosen according to the atomic species present. The number of PWs is sufficient to make $\tilde{n}$ a very accurate representation of the charge density $n(\boldsymbol{r})$ given by the underlying Gaussian basis set. This does not incur significant penalties in either speed or memory as the expansion is performed only for one function, $n(\boldsymbol{r})$, rather than each occupied band as is the case in conventional fully PW basis set calculations. As a consequence, the energy cutoff can be very large: typical values are 300 Ry and 80 Ry for carbon and silicon, respectively.

In a cluster calculation, the charge density is expanded in uncontracted Gaussians (1) typically up to d-type functions. More details regarding the charge-density basis functions in cluster calculations can be found elsewhere [11].

3 Electrical Levels

We now turn our attention to the calculation of electrical levels. Since the marker method, which forms the main part of this section, can be understood in terms of the more commonly used formation-energy method, we briefly describe this and its associated problems.

Before we do so, however, it is important to define what is meant by an *electrical level*. The electrical levels of a dopant (or any active defect) correspond to a thermodynamic property of the system, and relate to the chemical potential, μ_{e} of the electrons (sometimes erroneously referred to as a Fermi level, which is strictly only a zero-temperature quantity). Where μ_{e} lies below, say, the donor level of an impurity, energy is released by moving an electron from the defect to the electron reservoir at μ_{e}, and so this transfer of charge proceeds. Conversely, if μ_{e} is higher in the bandgap than the donor level of the defect it would *cost* energy to move the electron from the donor to the electron reservoir, and therefore it does not happen. It should therefore be noted that these levels relate to a *change of charge state* and cannot be obtained from the electronic-structure of any one of the charge states involved alone.

3.1 Formation Energy

The charge-state- (q) dependent, zero-temperature formation energy of a system of atoms and electrons (X) is usually written [23]:

$$E^{\mathrm{f}}(X,q) = E^{\mathrm{t}}(X,q) - \left(\sum_{\text{atoms}} \mu_i\right) + q\left\{E_{\mathrm{v}}(X,q) + \mu_{\mathrm{e}}\right\} + \xi(X,q)\,, \qquad (5)$$

where E^{t} is the total energy calculated for the system such as that obtained using `AIMPRO`, μ_i and μ_{e} are the atomic and electron chemical potentials, $E_{\mathrm{v}}(X,q)$ is the energy of the valence-band top and $\xi(X,q)$ is a term that takes into account artifacts introduced by the computational framework. For PBCs $\xi(X,q)$ reflects defect–defect interactions, whereas for clusters it relates to defect–surface interactions including quantum confinement. Usually, the relevant $E^{\mathrm{t}}(X,q)$ is that of the lowest-energy configuration of X in charge state q, but not always, as noted below.

For PBCs ξ may take the form of a power series in moments of the electron distribution, which can be viewed as terms arising from an array of monopoles, dipoles, quadrupoles and so on. The magnitudes of each term may be taken from a simple model, such as that of *Makov* and *Payne* [24] where only the terms in a Madelung energy and a quadrupole term are typically retained. The material is taken into account simply using the static dielectric constant, ε. Note, the Madelung term scales as q^2 and hence grows rapidly with q so that, for example, for a cubic supercell of diamond of side length $2a_0$, the Madelung correction for $q = \pm 3e$ is already of the order of the bandgap! This simple approach has received considerable criticism over recent years, so other, often computationally demanding, approaches have been adopted [25–28].

Notwithstanding the details of $\xi(X,q)$, one can use $E^{\mathrm{f}}(X,q)$ to estimate the electrical levels of X. The approach is to determine the thermodynamically most stable charge state for all relevant values of μ_{e}, and the critical values at which charge states change are the electrical levels. For example, the acceptor level is given by the value of μ_{e} that satisfies $E^{\mathrm{f}}(X,0) = E^{\mathrm{f}}(X,-1)$:

$$\mu_{\mathrm{e}} = \left[E^{\mathrm{t}}(X,-1) - E^{\mathrm{t}}(X,0)\right] + \left[\xi(X,-1) - \xi(X,0)\right] - E_{\mathrm{v}}(X,-1)\,, \qquad (6)$$

as illustrated in Fig. 2. The diagram also illustrates the difficulties in interpreting levels that lie in the energy range between the theoretical and experimental bandgaps. Note, usually the electrical levels represent the transition between the lowest-energy structural configurations of each charge state, but not always. For example, there may be a considerable energy barrier between conformations of a set of atoms, so that the electrical levels relate to the change of charge state of a specific arrangement of atoms. An example of this is that of boron-vacancy complexes in silicon, where the equilibrium number of host sites between the impurity and the lattice vacancy is charge state

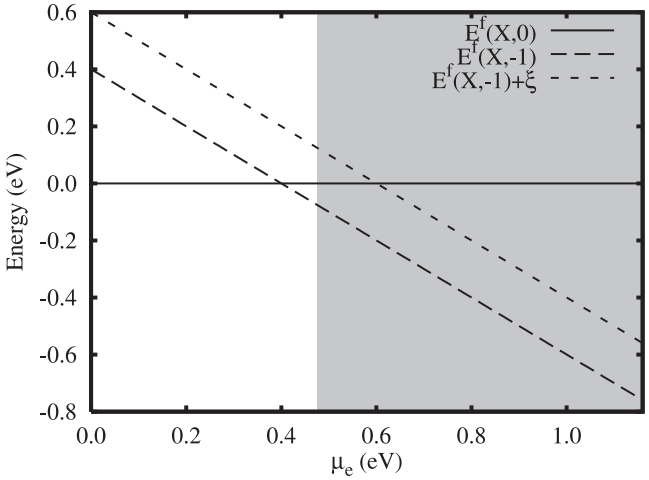

Fig. 2. Illustration of the use of the formation energy to obtain an acceptor level for the system X in silicon. Above the acceptor level X^- is thermodynamically more stable that X^0, whereas below the reverse is true. The shaded area indicates the region between the theoretical and experimental bandgaps. The inclusion of ξ in the charged formation energy pushes the acceptor level above the theoretical value for E_c, but it remains deep in the experimental bandgap

dependent [29]. Such *metastability* must be taken into account when trying to assign calculated electrical transitions to those measured experimentally.

This approach has perhaps three chief problems, the first of which is the poorly defined correction term, ξ which we have already touched upon.

The second is that μ_e is defined relative to the valence-band top in the defective system, which may be difficult to establish with any precision. If one does not wish simply to use the approximation $E_v(X, -1) = E_v$ (bulk, 0) then one might use corrections obtained either by examination of the lowest occupied Kohn–Sham level [30], or of the electrostatic potentials in bulk and defective systems [31] (see Sect. 3.2).

The third, and often most critical problem that may affect the location of electrical levels is the underestimate of the bandgap. For instance, the bandgap of hexagonal ZnO is around 3.4 eV [4], but simple application of the LDA- or GGA-DFT leads to bandgaps around 1 eV. Recently, hydrogen has been shown to be a shallow donor in this material [32, 33], but charge-dependent formation energies yield donor levels around $E_v + 2$ eV [34, 35].

However, as previously suggested, for a given material it is often found that many of the problems introduce *systematic* errors. Therefore, one chooses alternative reference states that mitigate the charged systems, bandgap error and other interactions introduced by the geometry: this is the marker method.

3.2 Calculation of Electrical Levels Using the Marker Method

We may avoid the problems highlighted above by referencing, say, the acceptor level of one system to that of another. Equation (6) for two different systems, X and Y, where the acceptor level $\mu_{\mathrm{e}}(Y)$ is *known* combine to yield the expression:

$$\begin{aligned}\mu_{\mathrm{e}}(X) = \mu_{\mathrm{e}}(Y) &+ \left[E^{\mathrm{t}}(X,-1) - E^{\mathrm{t}}(Y,-1)\right] - \left[E^{\mathrm{t}}(X,0) - E^{\mathrm{t}}(Y,0)\right] \\ &- \left[E_{\mathrm{v}}(X,-1) - E_{\mathrm{v}}(Y,-1)\right] \\ &+ \left[\{\xi(X,-1) - \xi(Y,-1)\} - \{\xi(X,0) - \xi(Y,0)\}\right] . \end{aligned} \tag{7}$$

If terms in ξ cancel and $E_{\mathrm{v}}(X,-1) = E_{\mathrm{v}}(Y,-1)$ this simply reduces to:

$$\mu_{\mathrm{e}}(X) = \mu_{\mathrm{e}}(Y) + \left[E^{\mathrm{t}}(X,-1) - E^{\mathrm{t}}(Y,-1)\right] - \left[E^{\mathrm{t}}(X,0) - E^{\mathrm{t}}(Y,0)\right] . \tag{8}$$

This defines an unknown acceptor level, $\mu_{\mathrm{e}}(X)$, with respect to the *known marker* at $\mu_{\mathrm{e}}(Y)$ with reference only to total energies. Note, (7) and (8) explicitly show that the calculations involve differences of energies between systems with the same charge.

However, we have made at least two rather bold assumptions, and it is necessary to explore where and why they are likely to be valid.

We first examine the notion that $E_{\mathrm{v}}(X,q) = E_{\mathrm{v}}(Y,q)$. Since one can add a constant potential to any system (i.e., the zero of the potential energy scale is often poorly defined in PBC calculations), the valence-band states may be rigidly offset from those of a defect-free system, and one can estimate the potential difference between X and Y by finding the average electrostatic potential in bulk-like regions of these systems [31]. For example, for substitutional impurities in a diamond-structure material, one might characterize the difference in the background potentials by finding the difference in the total electrostatic potentials at the T-interstitial sites far from the impurities. Alternatively, the potential difference between two systems can be estimated by the difference in the energies of the lowest occupied levels in X and Y [30] under the assumption that these characterize the average potentials in these systems. In our experience, such corrections may be up to a few tenths of an eV, but are usually small [31].

We now turn to the assumption that the ξ terms cancel. The suggested corrections for PBCs of *Makov* and *Payne* [24] assume an array of point charges, but it is clear from several studies that this approach tends to give a rather poor estimate. However, it is instructive to consider $\xi(X,q)$ as arising from a series of multipole interactions. For chemically similar systems, the larger terms in the series would naturally be close in magnitude. Therefore, the best accuracy using the marker method is expected when comparing structurally and chemically similar systems. This remains true even when considering transitions between highly charged states, which is important because of the q^2 dependence of the simple Makov–Payne type of correction.

However, there are two important areas where the marker method may be difficult to apply. The first is where the available markers are far in energy and

chemical nature from that of the system of interest, such as using a shallow donor as a reference for a very deep donor. Under these circumstances there is unlikely to be very complete cancellation in the ξ, and hence the error bars in the calculation become more significant. Furthermore, problems associated with the underestimate of the bandgap will also become more important. The second is that in some materials there may be *no* experimental data. For example, to our knowledge there are no unambiguous observations of double donors or acceptors in diamond.

Where there are no appropriate markers, one might use a bulk system as the reference, Y, in (7) and (8). Such an approach was detailed for boron in silicon [7] and defects in diamond [8]. Here, the first donor and acceptor levels of a bulk system are, by definition, E_{v} and E_{c}, respectively. However, one should exercise some caution in using a bulk system as a marker; deep levels and band edges are typically very different in character. Furthermore, the levels of shallow defects are referenced to markers at the other extreme of the bandgap, making the underestimate of the bandgap a significant factor.

Finally, in the preceding discussion we have largely emphasized the use of the marker method for PBCs, but the principles are the same for cluster calculations. Indeed, the use of clusters for narrow-gap materials has two important advantages over the supercell approach. The first is that the lack of periodicity means that there is no dispersion in the electronic-structure. Secondly, the confinement of an atomic cluster tends to oppose the typical underestimation of the bandgap. For materials such as germanium for which the underestimation is acute, this is a very significant, qualitative effect, as reflected in the results presented below.

In the final section we review a number of examples of the application of the marker method using `AIMPRO` in both supercell and cluster modes.

4 Application to Defects in Group-IV Materials

4.1 Chalcogen–Hydrogen Donors in Silicon

`AIMPRO` was recently used to analyze the properties of a range of chalcogens and their complexes with hydrogen in silicon. In this study [31] many of the issues raised above were explicitly examined, including the role of supercell size, Brillouin-zone sampling, basis and the average electrostatic potentials.

One significant factor in using chalcogens and their complexes with hydrogen in this study is that they have been characterized extensively by experiment, with spin densities, vibrational modes and for the substitutional chalcogens, electrical levels being available. It was therefore possible to show that the calculations were able to reproduce important aspects of the defects of interest other than the electrical characteristics, thereby validating the conclusions drawn. In particular, Mulliken bond-population analysis indicated that the donor wavefunctions were being faithfully reproduced in

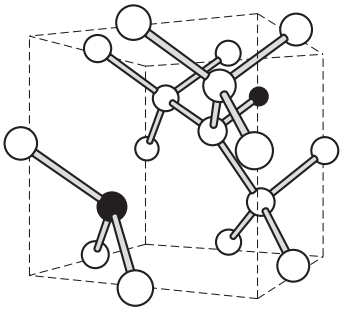

Fig. 3. Schematics of the partially passivated chalcogen–H complex in silicon. *White* and *black circles* represent host and impurity atoms, respectively, H being represented by the *smaller black circle*. *Dashed lines* indicated the cubic axes

these calculations, and that the characteristics were similar for the different chalcogen species, indicating that the marker method should be appropriate. Additionally, for the chalcogens the defect-level dispersions are very similar suggesting that this effect also would cancel out using an empirical marker.

The chalcogens S, Se, and Te lie onsite, each having two donor levels, as listed in Table 4. The calculated donor levels listed were obtained using the formation energies with a Madelung correction and the marker method using bulk silicon and S as references.

Table 4. Electrical levels relative to E_c (eV) of chalcogen and chalcogen–hydrogen complexes in silicon (see [31]). Experimental data from [36]

	Experiment		Formation energy		Bulk marker		Sulfur marker	
Defect	(0/+)	(+/2+)	(0/+)	(+/2+)	(0/+)	(+/2+)	(0/+)	(+/2+)
S	0.29	0.59	0.78	1.25	0.42	0.48	0.29	0.59
Se	0.29	0.54	0.79	1.23	0.40	0.45	0.28	0.55
Te	0.20	0.36	0.75	1.22	0.27	0.28	0.25	0.38
S–H	–	–	0.41	1.78	0.13	1.18	0.01	1.28
Se–H	–	–	0.37	1.77	0.10	1.18	−0.02	1.28
Te–H	–	–	0.27	1.77	0.00	1.18	−0.12	1.28

The shallow nature predicted for the X–H complexes, the structure of which is shown schematically in Fig. 3, was viewed to be consistent with the fact that they had not been detected using deep-level transient spectroscopy [37, 38], but had been detected via their vibrational modes [39].

Three important conclusions can be drawn from this study. The first is that the formation-energy approach yields qualitatively erroneous results: S, Se, and Te have no double-donor level in the bandgap, and the single-

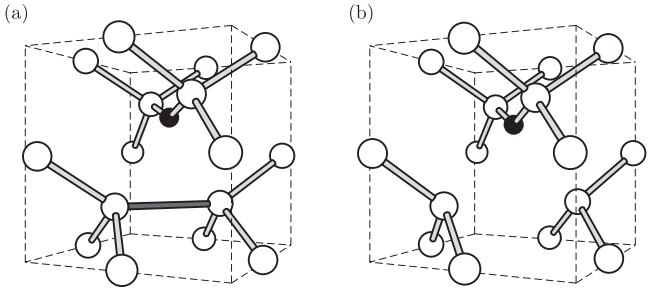

Fig. 4. Schematics of the VO center in Si and Ge. *White* and *black circles* represent host and oxygen atoms, respectively. The *dark bond* in (**a**) indicates the reconstruction present in the neutral charge state. (**b**) shows the unreconstructed center characteristic of the negative charge state

donor levels are far too deep. This can be traced to the underestimate in the bandgap and the nature of the correction terms, ξ. The second is that, although the single-donor levels are reasonably reproduced using the bulk-marker method, the second donor level is very poor, due to an underestimate of the correlation energy. Finally, the positive result is that, at least for classes of similar defects, the empirical marker method yields very good agreement with experiment.

4.2 VO Centers in Silicon and Germanium

The properties of oxygen in silicon and germanium have been studied extensively as a consequence of the incorporation of O during growth and the properties that it lends to the materials. Of all the various roles and structures in which oxygen is involved, one of the most primitive is oxygen substituting for a host atom. Oxygen is a relatively small atom so that this center is generally referred to as a complex of (interstitial) oxygen with a lattice vacancy (VO), the structure of which is shown schematically in Fig. 4. VO centers in the *neutral* charge state can be understood as follows. The removal of a host atom creates four "dangling bonds". Divalent oxygen passivates two of these and the remaining two dangling-bonds then combine together in a reconstruction, resulting in a defect that is fully chemically coordinated.

However, these centers can trap additional electrons by breaking the relatively weak reconstruction, and the resultant dangling bonds lead to states in the bandgap. Experimentally, the VO center in silicon has an acceptor level at $E_{\mathrm{v}} + 1$ eV [40]. However, the DFT calculations using supercells and formation energies have resulted in a rather poor agreement, predicting an acceptor level at $E_{\mathrm{v}} + 0.4$ eV [41]. However, recalling that in these calculations the bandgap of bulk silicon is underestimated by more than 50 % of the experimental value, this donor level is approximately the correct depth below

the *theoretical* E_c, which might be viewed as agreeing with experiment, and it is not clear how to interpret this calculation.

The marker method using atomic clusters [6] where the bandgap underestimate is mitigated and using the acceptor level of interstitial carbon ($E_c - 0.10$ eV [42]) as a marker yielded an acceptor level for VO in silicon at $E_c - 0.13$ eV, in close agreement with the experimental level. Indeed, the small error is a testament to how robust the method is since, although they are close by in energy, the character of the acceptor wavefunctions for the VO and carbon interstitials are not closely related.

The importance of the bandgap error is even greater when considering Ge. Recent calculations have also adopted the cluster configuration of `AIMPRO` to analyze a range of defects in this material [43]. In particular, VO experimentally has acceptor and double-acceptor levels at $E_v + 0.27$ eV and $E_v + 0.49$ eV, respectively [44]. Using `AIMPRO` and atomic clusters that have approximately the experimental bandgap of 0.7 eV the two acceptor levels of VO have been calculated relative to the acceptor and double-acceptor levels of substitutional Zn at $E_v + 40$ meV and $E_v + 100$ meV, respectively [4]. The calculated second acceptor level is in excellent agreement with the experimental result, lying at $E_v + 0.47$ eV. However, the first acceptor level calculated at around $E_v + 0.42$ eV is in less good agreement, but still within 0.2 eV. The reason for the larger deviation in the single-acceptor level is unclear, but even such an error bar is useful in the context of the more traditional supercell approaches that suggest this material has no bandgap at all.

4.3 Shallow and Deep Levels in Diamond

In contrast to Si and Ge, diamond, with an indirect bandgap of around 5.5 eV [4], is a good example of an insulator, not obviously to be associated with the semiconductor industry. However, p- and n-type diamond can be produced via doping with boron and phosphorus, respectively, with heavily boron-doped materials becoming metallic and opaque.

However, although phosphorus acts as a donor, the activation energy is very high at around 0.6 eV, so the room-temperature ionization fraction, and hence the number of free electrons, is rather small. This has led to an effort to deduce which, if any, other system may result in a shallow donor level.

As well as the electrical properties of diamond, the optical transparency, especially in the infrared, of the pure material has led to applications in high-specification optics. However, as-grown diamond may contain deep-level defects that absorb and luminesce, and an understanding of these centers is also obviously of relevance.

Regrettably, the number of definitively characterized donor and acceptor levels in diamond is relative small, and it is unsurprising that computational methods have been used in an attempt to assess both shallow and deep centers. We first address examples of potential shallow donors. Since P is the

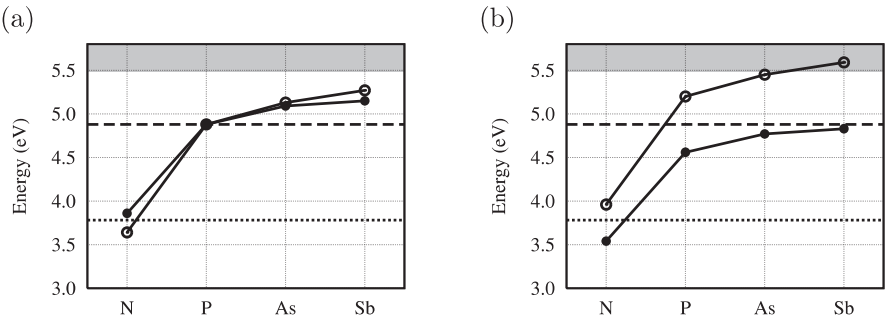

Fig. 5. Location of donor levels in the bandgap relative to E_v for pnictogen donors in diamond. (**a**) and (**b**) shows values calculated using phosphorus and the bulk supercell being used as a marker, respectively. *Empty* and *filled circles* show 64 and 216 atom cells, respectively. The *dashed* and *dotted lines* show the experimental donor levels for P and N, and the *shaded area* shows the experimental conduction band

best-established shallow donor in diamond, it seems a good idea to establish if, as in silicon, the other pnictogens may have useful donor properties. Figure 5 plots the calculated donor levels using `AIMPRO` as described in [9].

Figure 5a shows the use of the experimental donor level [45] for P at $E_c - 0.6$ eV as a marker for supercells containing 64 or 216 atoms (cubes of side length $2a_0$ and $3a_0$, respectively). The trends are the same for both curves and both supercells reproduce the experimental donor level of nitrogen.

Figure 5b shows the same data, but referenced to bulk diamond. Now, the marker is the valence-band top, so it is far (around 5 eV!) from the anticipated location of the pnictogen donor levels. This exemplifies the need for appropriate markers for the method to have a quantitative role. However, irrespective of the marker, the important result in this application is that shallower donors of a simple nature, i.e., As and Sb, are expected to lead to improved n-type diamond.

Recently, we also used supercell calculations to compare the use of the more traditional formation energy approach with the marker method for a range of centers, and in particular for those for which electrical levels are known experimentally [8]. Table 5 lists some of these results.

The overall agreement between the marker method and experiment is conspicuously better than using the formation-energy approach, indicating that the marker method may usefully be applied to wide-gap materials for both deep and shallow electrical levels. For the small set of data presented in Table 5 one could adopt any acceptor and donor as an empirical marker and get very similar accuracy.

Table 5. Electrical levels (eV) for various centers in diamond from experiment, compared to those calculated using the formation-energy method and the marker method using a bulk cell as a marker. For the formation-energy method, we quote the levels above E_v and below the theoretical and experimental locations of E_c ($E_c^{th} = E_v + 4.2$ eV and $E_c^{expt} = E_v + 5.5$ eV). Asterisks indicate the marker-method results where the bandgap energy of 5.5 eV has been used to reference to the opposite band edge. Subscript s indicate substitutional impurities and v represents a lattice vacancy

		Formation energy			Bulk marker	
Defect	Experiment	E_v	E_c^{th}	E_c^{expt}	E_v	E_c
Acceptors						
B_s	$E_v + 0.37$ [46]	0.2	4.0	5.3	0.5*	5.0
Ni_s	$E_c - 2.49$ [47, 48]	3.3	0.9	2.2	3.0*	2.5
V–N_s	$E_c - 2.583$ [49]	1.8	2.4	3.7	2.2*	3.3
Donors						
N_s	$E_c - 1.7$ [50]	3.0	1.2	2.5	4.0	1.5*
P_s	$E_c - 0.6$ [45]	4.2	0.0	1.3	5.2	0.3*
N_s–N_s	$E_c - 3.8$ [51]	0.8	3.6	4.7	1.8	3.7*

5 Summary

The characterization of the electrical properties of defects and dopants using standard methodologies such as LDA- and GGA-DFT is a demanding problem. The most common approach of calculating charge-dependent formation energies has well-publicized frailties associated with periodic boundary conditions for charged systems, but these may be somewhat reduced by employing schemes where the electrical levels of two or more systems are compared. Additionally, when comparing total energies for chemically and structurally similar systems in the PBC approach, the dispersion of the defect levels in the first Brillouin-zone and multipole interaction, which are present even for neutral systems, are likely to be comparable and cancel to a large degree.

In addition, for electrical levels in narrow-gap materials, such as germanium, the use of carefully constructed atomic clusters also significantly reduces the underestimation of bandgaps in these methods, allowing a fully quantitative analysis of the levels in these materials. We have also shown that the marker method is effective in wide-gap materials where excellent agreement may be obtained over the full range of the bandgap, provided that one has a suitable marker available.

Since the marker method is applicable when looking at similar systems, it may be of particular use when studying dilute alloys such as SiGe. Here, a known marker in silicon (for a Si-rich alloy) could be used to determine the effect of having a nearby Ge atom, such as that of VO centers in such alloys [52].

The use of bulk materials as markers is a qualified success in the sense that single donor and acceptor levels are in many cases in good agreement with experiment. However, multiple charges introduce larger errors and, for example with the chalcogen species in silicon, the second donor levels are in much less good agreement.

We conclude that where experimental data allow, the most accurate quantitative method for predicting electrical levels in crystalline semiconductors and insulators is by comparison of a chemically and structurally similar defect for which the electrical levels are known. Other marker species, including the bulk supercells, are likely to introduce larger errors, but in our experience even then the results are rather favorable.

Acknowledgements

The authors would like to thank J. Coutinho, R. Jones and colleagues for useful discussions and access to data presented in this Chapter.

References

[1] G. Burns: *Solid State Physics*, International ed. (Academic, Orlando, Florida 1985) 69
[2] P. Hohenberg, W. Kohn: Phys. Rev. **136**, 864 (1964) 69
[3] W. Kohn, L. J. Sham: Phys. Rev. **140**, A1133 (1965) 69
[4] S. M. Sze: *Physics of Semiconductor Devices*, 2nd ed. (Wiley-Interscience, New York 1981) 70, 82, 87
[5] G. E. Engel, W. E. Pickett: Phys. Rev. B **54**, 8420 (1996) 70
[6] A. Resende, R. Jones, S. Öberg, P. R. Briddon: Phys. Rev. Lett. **82**, 2111 (1999) 70, 87
[7] J.-W. Jeong, A. Oshiyama: Phys. Rev. B **64**, 235204 (2001) 70, 84
[8] J. P. Goss, P. R. Briddon, S. J. Sque, R. Jones: Diamond Relat. Mater. **13**, 684 (2004) 70, 84, 88
[9] S. J. Sque, R. Jones, J. P. Goss, P. R. Briddon: Phys. Rev. Lett. **92**, 017402 (2004) 70, 88
[10] R. Jones, P. R. Briddon: Identification of defects in semiconductors, in M. Stavola (Ed.): *Semicond. Semimet.*, vol. 51A (Academic, Boston 1998) Chap. 6, p. 287 71
[11] P. R. Briddon, R. Jones: Phys. Stat. Sol. B **217**, 131 (2000) 71, 72, 80
[12] W. E. Pickett: Comput. Phys. Rep. **9**, 115 (1989) 71
[13] D. R. Hamann, M. Schlüter, C. Chiang: Phys. Rev. Lett. **43**, 1494 (1979) 71
[14] G. B. Bachelet, D. R. Hamann, M. Schlüter: Phys. Rev. B **26**, 4199 (1982) 71
[15] N. Troullier, J. L. Martins: Phys. Rev. B **43**, 1993 (1991) 71
[16] C. Hartwigsen, S. Goedecker, J. Hutter: Phys. Rev. B **58**, 3641 (1998) 71
[17] J. P. Perdew, Y. Wang: Phys. Rev. B **45**, 13244 (1992) 71
[18] S. Goedecker, M. Teter, J. Hutter: Phys. Rev. B **54**, 1703 (1996) 71
[19] J. A. White, D. M. Bird: Phys. Rev. B **50**, R4954 (1994) 71
[20] J. P. Perdew, K. Burke, M. Ernzerhof: Phys. Rev. Lett. **77**, 3865 (1996) 71

[21] E. R. Cohen, B. N. Taylor: J. Res. Nat. Bur. Standards **92**, 85 (1987) 77
[22] H. J. McSkimin, P. Andreatch, Jr.: J. Appl. Phys. **35**, 2161 (1964) 77
[23] S. B. Zhang, J. E. Northrup: Phys. Rev. Lett. **67**, 2339 (1991) 81
[24] G. Makov, M. C. Payne: Phys. Rev. B **51**, 4014 (1995) 81, 83
[25] H. Nozaki, S. Itoh: Phys. Rev. E **62**, 1390 (2000) 81
[26] U. Gerstmann, P. Deák, R. Rurali, B. Aradi, T. Frauenheim, H. Overhof: Physica B **340–342**, 190 (2003) 81
[27] C. W. M. Castleton, S. Mirbt: Phys. Rev. B **70**, 195202 (2004) 81
[28] J. Shim, E.-K. Lee, Y. J. Lee, R. M. Nieminen: Phys. Rev. B **71**, 035206 (2005) 81
[29] J. Adey, R. Jones, D. W. Palmer, P. R. Briddon, S. Öberg: Phys. Rev. B **71**, 165211 (2005) 82
[30] P. Deák, B. Aradi, A. Gali, U. Gerstmann: Phys. Stat. Sol. B **325**, 139 (2003) 82, 83
[31] J. Coutinho, V. J. B. Torres, R. Jones, P. R. Briddon: Phys. Rev. B **67**, 035205 (2003) 82, 83, 84, 85
[32] S. F. J. Cox, E. A. Davis, S. P. Cottrell, P. J. C. King, J. S. Lord, J. M. Gil, H. V. Alberto, R. C. Vilão, J. Piroto Duarte, N. Ayres de Campos, A. Weidinger, R. L. Lichti, S. J. C. Irvine: Phys. Rev. Lett. **86**, 2601 (2001) 82
[33] D. M. Hofmann, A. Hofstaetter, F. Leiter, H. Zhou, F. Henecker, B. K. Meyer, S. B. Orlinskii, J. Schmidt, P. G. Baranov: Phys. Rev. Lett. **88**, 45504 (2002) 82
[34] C. G. Van de Walle: Phys. Rev. Lett. **85**, 1012 (2000) 82
[35] M. G. Wardle, J. P. Goss, P. R. Briddon: Phys. Rev. B **71**, 155205 (2005) 82
[36] H. G. Grimmeiss, E. Janzém: in S. T. Pantelides (Ed.): *Deep Centers in Semiconductors*, 2nd ed. (Gordon and Breach, Switzerland 1996) p. 97 85
[37] G. Pensl, G. Roos, C. Holm, E. Sirtl, N. M. Johnson: Appl. Phys. Lett. **51**, 451 (1987) 85
[38] G. Roos, G. Pensl, N. M. Johnson, C. Holm: J. Appl. Phys. **67**, 1897 (1990) 85
[39] R. E. Peale, K. Muro, A. J. Sievers: Mater. Sci. Forum **65–66**, 151 (1990) 85
[40] G. D. Watkins, J. W. Corbett: Phys. Rev. **121**, 1001 (1961) 86
[41] M. Pesola, J. von Boehm, T. Mattila, R. M. Nieminen: Phys. Rev. B **60**, 11449 (1999) 86
[42] L. W. Song, G. D. Watkins: Phys. Rev. B **42**, 5759 (1990) 87
[43] R. Jones, A. Carvalho, J. Coutinho, V. J. B. Torres, S. Öberg, P. R. Briddon: Sol. St. Phenom. **108–109**, 697 (2005) 87
[44] V. P. Markevich, I. D. Hawkins, A. R. Peaker, V. V. Litvinov, L. I. Murin, L. Dobaczewski, J. L. Lindström: Appl. Phys. Lett. **81**, 1821 (2002) 87
[45] E. Gheeraert, S. Koizumi, T. Teraji, H. Kanada, M. Nesládek: Phys. Stat. Sol. A **174**, 39 (1999) 88, 89
[46] P. A. Crowther, P. J. Dean, W. F. Sherman: Phys. Rev. **154**, 772 (1967) 89
[47] D. M. Hofmann, M. Ludwig, P. Christmann, D. Volm, B. K. Meyer, L. Pereira, L. Santos, E. Pereira: Phys. Rev. B **50**, 17618 (1994) 89
[48] R. N. Pereira, W. Gehlhoff, N. A. Sobolev, A. J. Neves, D. Bimberg: J. Phys. Condens. Matter **13**, 8957 (2001) 89
[49] J. W. Steeds, S. J. Charles, J. Davies, I. Griffin: Diamond Relat. Mater. **9**, 397 (2000) 89
[50] R. Farrer: Solid State Commun. **7**, 685 (1969) 89

[51] G. Davies: J. Phys. C **9**, L537 (1976) 89
[52] V. P. Markevich, A. R. Peaker, J. Coutinho, R. Jones, V. J. B. Torres, S. Öberg, P. R. Briddon, L. I. Murin, L. Dobaczewski, N. V. Abrosimov: Phys. Rev. B **69**, 125218 (2004) 89

Index

Dynamical Matrices and Free Energies

Stefan K. Estreicher and Mahdi Sanati

Physics Department, Texas Tech University, Lubbock, TX 79409-1051
stefan.estreicher@ttu.edu

Abstract. The calculation of the entire dynamical matrix of a periodic supercell (containing a defect or not) provides several most useful pieces of information. At first, the eigenvalues of this matrix are all the normal mode frequencies of the cell, including the local, pseudolocal, and resonant modes associated with the defect under study. The eigenvalues can also be used to construct phonon densities of state which in turn allow the calculation of (Helmholtz) free energies, vibrational entropies, and specific heats. The eigenvectors of the dynamical matrix can be used to prepare a system in thermal equilibriumat a desired temperature. This allows constant-temperature MD simulations to be peformed without thermalization or thermostat. Applications to the calculation of vibrational lifetimes and decay channels are discussed. Finally, the vibrational, rotational, and charge-carrier contributions to the free energy are described. Configurational entropies are calculated in realistic systems.

1 Introduction

Vibrational spectroscopy provides essential experimental data about defects in semiconductors, not only because of the microscopic nature of the information it provides but also because many of the quantities measured can be calculated from first principles. Fourier transform infrared absorption (FTIR) and Raman spectroscopies often produce sharp lines associated with local vibrational modes (LVMs) of impurities lighter than the hosts atoms. These lines are often above the highest normal mode of the crystal, the Γ phonon but, when the phonon density of states has a gap (as in the case of GaN [1]), they can be between the acoustic and optic modes as well. Isotope substitutions provide element identification as well as information about how many atoms of a given species are part of the defect. Uniaxial stress experiments give the symmetry of the defect through the splitting of the IR or Raman lines. Annealing studies provide various activation energies. Examples of such studies and a discussion of the techniques are found in [2].

By themselves, the measured LVMs, isotope shifts, symmetry, and activation energies are often insufficient to identify a unique structure for the defect. However, they are precious to theorists who use first-principles techniques to calculate the structures (symmetry), LVMs and their isotope shifts, binding, migration and/or reorientation energies, and thus prove or disprove

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 95–114 (2007)

the assignment of an IR or Raman spectrum to a specific defect. The interplay between vibrational spectroscopy and first-principles theory has led to the unambiguous identification of many defects and has provided realistic tests of the accuracy of theory.

For many years, the calculation of defect-related vibrational modes has been limited to those LVMs that are IR or Raman active. The stretch mode of a Si–H bond, in a Si vacancy for example, can be predicted quite accurately by calculating the total energy of a supercell containing this defect in its equilibrium configuration as well as for several (at least 2) positions of H along the Si–H bond. The 1-dimensional potential along this axis is fitted to a polynomial and the frequency and zero-point energy of this mode are calculated. The same can also be done for wag modes. There are many examples of such calculations [3–5].

This method works well, but a lot of additional and very useful information can be obtained when the entire dynamical matrix of the supercell is calculated. For a system of N atoms and therefore $3N$ normal modes, the dimension of this matrix is $3N \times 3N$ and its calculation is computationally expensive. The *eigenvalues* of the dynamical matrix are all the normal mode frequencies of the supercell: acoustic and optic phonons as well as defect-related modes. These can be LVMs, located above the Γ phonon, resonant and/or pseudolocal vibrational modes (pLVMs). Resonant modes occur when the strain (associated with a defect) stretches or compresses host–atom bonds, resulting in new vibrational frequencies near the Γ phonon. pLVMs are localized impurity-related modes buried in the phonon continuum. Such modes are sometimes visible experimentally as phonon sidebands in photoluminescence (PL) spectra [6, 7]. Two copper-related centers in Si have recently been theoretically identified [8, 9] thanks to such pLVMs. Further, the knowledge of all the normal mode frequencies allows the construction of phonon densities of state, which are needed to calculate (Helmholtz) free energies.

The *eigenvectors* of the dynamical matrix make it possible to find (and identify the symmetry of) all the localized modes associated with any atom or group of atoms in the supercell. The eigenvectors can also be used to prepare the supercell in thermal equilibrium and any temperature, thus allowing constant-temperature molecular dynamics (MD) simulations to be performed without the need for thermalization or even a thermostat. This in turn makes it possible to calculate vibrational lifetimes and decay channels of specific modes as a function of temperature. For all these reasons, the calculation of dynamical matrices is well worth its cost.

Our results are obtained from self-consistent, first-principles theory based on local density-functional theory in 64-host-atom periodic supercells. The calculations are performed with the SIESTA code [10, 11]. The exchange-correlation potential is that of *Ceperley–Alder* [12] as parameterized by *Perdew* and *Zunger* [13]. Norm-conserving pseudopotentials in the *Kleinman–Bylander* form [14] are used to remove the core regions from the calculations. The basis sets for the valence states are linear combinations of numerical

atomic orbitals of the *Sankey* type [15–17], generalized to be arbitrarily complete with the inclusion of multiple-zeta orbitals and polarization states [10]. We use double-zeta (two sets of s and p orbitals) for first- and second-row atoms (H through Ne) and add a set of polarization functions (e.g., one set of d orbitals) for third-row atoms and below. The charge density is projected on a real-space grid with an equivalent cutoff of 150 Ry to calculate the exchange-correlation and Hartree potentials. A $2 \times 2 \times 2$ Monkhorst–Pack k-point sampling [18] is used to optimize the structures.

An overview of dynamical matrices is given in Sect. 2. Examples of LVMs and pLVMs are given in Sect. 3. The calculation of vibrational lifetimes is given in Sect. 4. Vibrational free energies and specific heats are discussed in Sect. 5, and the properties of defects at finite temperatures in Sect. 6.

2 Dynamical Matrices

The calculation of dynamical matrices is a well-known topic discussed in many textbooks, such as [19]. Further, the calculation of the force-constant matrix F is implemented in many (if not all) software packages that have MD capabilities. The brief summary below is therefore only necessary in order to define the notation to be used in this Chapter and make a few comments.

We consider a solid represented by periodic supercells of N atoms of mass m_α (with $\alpha = 1, 2, \ldots, N$). The nuclei oscillate around their equilibrium positions. The dynamical matrix is related to the force-constant matrix by

$$D_{\alpha\beta,ij} = \frac{F_{\alpha\beta,ij}}{\sqrt{m_\alpha m_\beta}}\,, \tag{1}$$

where $i, j = x, y, z$ are Cartesian indices.

The calculation of $F_{\alpha\beta,ij}$ is computationally intensive for the type of systems required in the study of defects in semiconductors. Today's typical supercells contain of the order of 100 atoms, and there is little doubt that several hundred (or even thousand) atoms will become the norm in the near future. Once F is known, changing the masses of selected atoms (that is, playing with isotopes) is trivial. Techniques to calculate F are well known, ranging from the direct "frozen phonon" approach implemented in many MD software packages to MD-based methods based on correlation functions [20, 21], or linear-response theory [22–25]. Note that dynamical matrices can also be computed using order-N methods [26].

The eigenvalues of the dynamical matrix are the normal mode frequencies of the system ω_s, and the corresponding eigenvectors $e^s_{\alpha,i}$ are orthonormal:

$$\sum_{\alpha,i} e^s_{\alpha,i} e^{s'}_{\alpha,i} = \delta_{ss'} \quad \text{and} \quad \sum_s e^s_{\alpha,i} e^s_{\beta j} = \delta_{\alpha\beta}\delta_{ij}\,. \tag{2}$$

In the harmonic approximation, the normal-mode coordinates q_s are

$$q_s(t) = A_s \cos(\omega_s t + \varphi_s) = \sum_{\alpha,i} \sqrt{m_\alpha} u_{\alpha,i} e^s_{\alpha,i} \,, \tag{3}$$

where the amplitudes A_s are temperature dependent and the $u_{\alpha,i}$ are Cartesian displacements of the nucleus α away from equilibrium.

This relationship can be used to prepare a system in equilibrium at a temperature T, then perform MD runs at constant temperature without the need for thermalization or even a thermostat [27]. Indeed, the unknown amplitudes of the normal modes can be obtained by requiring that the kinetic energy of each mode is, on the average, $k_\mathrm{B}T/2$, that is

$$\left\langle \frac{1}{2} \left(\frac{\partial q_s}{\partial t} \right)^2 \right\rangle = \left\langle \tfrac{1}{2} \omega_s^2 A_s^2 \sin^2(\omega_s t + \varphi_s) \right\rangle = \tfrac{1}{4} \omega_s^2 A_s^2 = \tfrac{1}{2} k_\mathrm{B} T \,. \tag{4}$$

Thus, the average amplitude of the normal mode s is $\langle A_s \rangle = \sqrt{2k_\mathrm{B}T}/\omega_s$.

Note that assigning the amplitudes $A_s = \langle A_s \rangle$ to each mode s implies that the total energy of each mode is exactly $k_\mathrm{B}T$. It is better to pick a random distribution

$$\zeta_s = \int_0^{E_s} \frac{1}{k_\mathrm{B}T} e^{-E/k_\mathrm{B}T} \, \mathrm{d}E \tag{5}$$

with $0 < \zeta_s < 1$. This leads to $A_s = \sqrt{-2k_\mathrm{B}T \ln(1-\zeta_s)}/\omega_s$. Thus, in the harmonic approximation, the Cartesian coordinates and corresponding velocities needed to prepare the system in equilibrium at the temperature T are

$$u_{\alpha i} = \sqrt{\frac{2k_\mathrm{B}T}{m_\alpha}} \sum_s \frac{1}{\omega_s} \sqrt{-\ln(1-\zeta_s)} \cos(\omega_s t + \varphi_s) e^s_{\alpha i} \,, \tag{6}$$

and

$$\frac{\partial u_{\alpha i}}{\partial t} = -\sqrt{\frac{2k_\mathrm{B}T}{m_\alpha}} \sum_s \sqrt{-\ln(1-\zeta_s)} \sin(\omega_s t + \varphi_s) e^s_{\alpha i} \,. \tag{7}$$

The initial phases $0 \le \varphi_s < 2\pi$ are random, as each mode has a random amount of kinetic and potential energy at the time $t = 0$. A similar transformation has been used by *Gavartin* and *Stoneham* [28] to calculate the energy dissipation in quantum dots. However, we discuss it here in the context of establishing initial conditions with the appropriate random distribution of energies and phases. This is used in Sect. 4 to calculate vibrational lifetimes and decay channels from first principles.

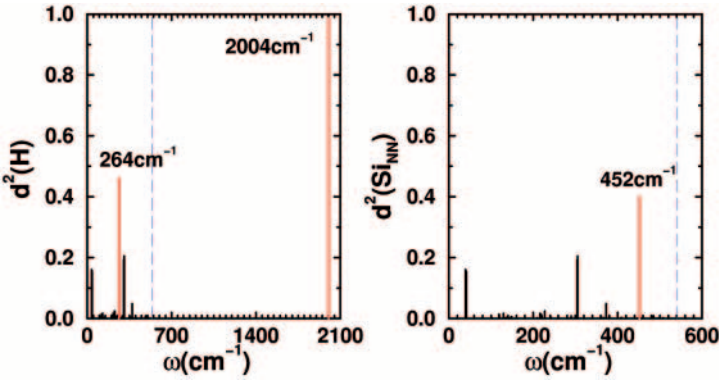

Fig. 1. LVMs and pLVMs associated with H_{bc}^{+} in the Si_{64} supercell. The plots of $d^2(\alpha) = \sum_i |e^s_{\alpha,i}|^2$, where α is H (*left*), or its two Si nearest neighbors (*right*) show the vibrational modes localized on the impurity and its two Si neighbors. The *vertical dashed lines* show the calculated Γ phonon

3 Local and Pseudolocal Modes

The simplest and most immediate consequence of the knowledge of the dynamical matrix is the identification of all the localized modes of a defect, and their symmetry. Indeed, a plot of $\sum_i |e^s_{\alpha,i}|^2$ vs. s (that is, vs. the normal mode frequencies ω) for a specific atom α (or set of atoms) provides quantitative information about the localization of the modes associated with the given atom(s). For example, Fig. 1 shows the localized modes associated with bond-centered (bc) hydrogen in crystalline Si (H_{bc}^{+}): the asymmetric stretch (IR active) of H is a LVM at 2004 cm^{-1} (observed [29] at 1998 cm^{-1}), the two degenerate wag modes (not observed) are pLVMs, far below the Γ phonon, at 264 cm^{-1}, and the symmetric stretch is a pLVM at 452 cm^{-1} (also not observed). The latter does not involve any H motion at all, but shows up when $\sum_i |e^s_{\alpha,i}|^2$ includes the two Si nearest neighbors to H.

Figure 2 shows the eigenvectors of the dynamical matrix in one of the two degenerate wag modes at 264 cm^{-1}. The H displacement in this mode accounts for about 70 % the displacements of all the atoms in the cell. Note that the eigenvectors only give the *relative* amplitudes of the atomic displacements. Their absolute values depend on the temperature, and are exaggerated in the figure.

Thus, plotting $\sum_i |e^s_{\alpha,i}|^2$ quantifies the localization of all the local modes associated with an atom or group of atoms. The corresponding eigenvectors allow the identification of the symmetry of the mode, which is needed to establish whether a mode is IR active or not, and whether it can produce phonon sidebands in PL spectra or not [6]. This is precisely how two low-frequency copper-related defects have recently been identified [8, 9]. A very

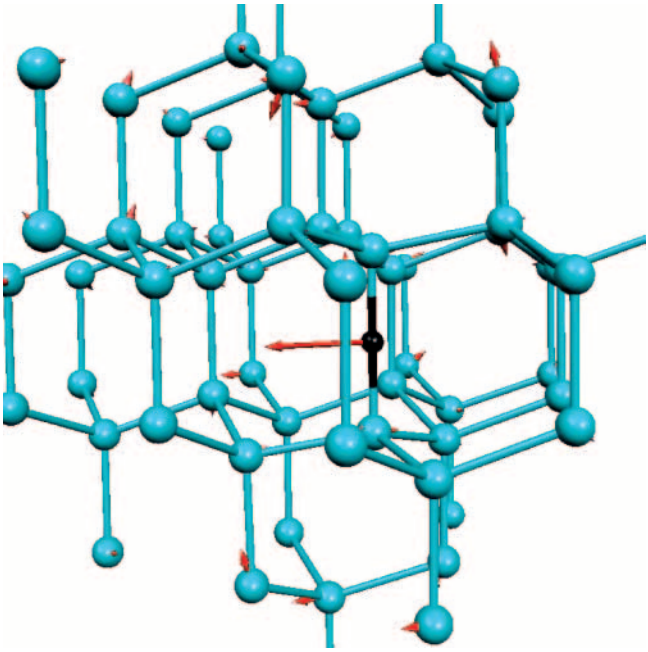

Fig. 2. Eigenvectors of the dynamical matrix in one of the 264 cm^{-1} wag modes of H_{bc}^{+}. The H atom is the *small black ball*, the *gray balls* are all Si. The size of the arrows is exaggerated to make the smaller oscillation amplitude more visible. The *arrows* only show the *relative* amplitudes of the motion of the atoms in this mode

useful software package[1] readily reads dynamical matrices and graphically displays the eigenvectors for any chosen normal mode.

4 Vibrational Lifetimes and Decay Channels

The vibrational lifetimes of the LVMs of light impurities in Si have recently been measured by transient bleaching spectroscopy [30, 31]. Surprisingly, the lifetimes of nearly identical H-related LVMs differ by some two orders of magnitude. For example, at low temperatures, the measured lifetimes of the 2072 cm^{-1} mode of the divacancy–dihydrogen complex ($VH \cdot HV$), the 1998 cm^{-1} mode of H_{bc}^{+} [30], and the 2062 cm^{-1} mode of the H_2^* pair [31] are 295 ps, 8 ps, and 4 ps, respectively. Since the decay of all these modes must involve at least four phonons (six phonon processes are proposed [29, 31]), it is not clear why the lifetimes of LVMs with almost identical frequencies vary by two orders of magnitude or why any of them would be short-lived at all.

[1] For information about the Molekel visualization package, see `www.cscs.ch/molekel/`.

The calculations of vibrational lifetimes begin with the dynamical matrices of the supercells containing the defects. The eigenvectors of this matrix determine the equilibrium background temperature of the cell at the time $t = 0$, as discussed in Sect. 2. This is used to prepare the cell in equilibrium at the same sample temperatures as in the experimental work. The initial excitation of the LVM of interest is assigned to be its zero-point energy plus one phonon, that is $3\hbar\omega/2$ (kinetic energy at $t = 0$). This mimics the laser excitation of the specific LVM. Then, constant-temperature (classical) MD simulations are performed, with a time step of 0.3 fs in the case of H. Since ab initio MD simulations are limited to real times of a few times 10 ps, it is critical to be able to begin the MD runs at a relatively elevated background temperature, because the measured lifetimes are much shorter at higher temperatures. The calculations are then repeated at lower temperatures to monitor the increase in the lifetime until such low temperatures that the calculations become computationally prohibitive. Note that at very low temperatures, all the normal modes of the system have amplitudes consistent with their zero-point energy. In classical MD simulations, these amplitudes go to zero, the modes become harmonic and the lifetimes very long [27].

At every time step, the $3N$ Cartesian coordinates of the atoms are written as linear combinations of the $3N$ normal modes of the supercell. Thus, the energies of all the modes can be plotted as a function of time. This approach allows not only the decay of the LVM of interest to be calculated, but also the identification of the receiving modes, all of it as a function of time for various background temperatures of the cell. The pLVMs of the defect play a critical role in the decay process.

We illustrate this approach with the case of the asymmetric stretch of H_{bc}^{+} [30]. The result of the run performed at a background temperature of 75 K is shown in Fig. 3. The calculated decay of the LVM produces a good exponential fit leading to a calculated lifetime of 7.8 ps at 75 K, a value that agrees very well with experiment [30]. An analysis of the time dependence of the energies of the normal modes of the cell shows that both wag modes and the symmetric stretch of the defect (at 262 cm^{-1} and 452 cm^{-1}, respectively) play important roles in the decay.

5 Vibrational Free Energies and Specific Heats

Although the calculation of potential-energy surfaces is playing a most useful role in our understanding of the behavior of defects in semiconductors, the real-world involves nonzero temperatures. Samples undergo various thermal anneals, they are implanted and exposed to light, and most devices function at or above room temperature. The physics and chemistry of defects is obviously temperature dependent, as one observes processes such as diffusion and association or dissociation at various temperatures.

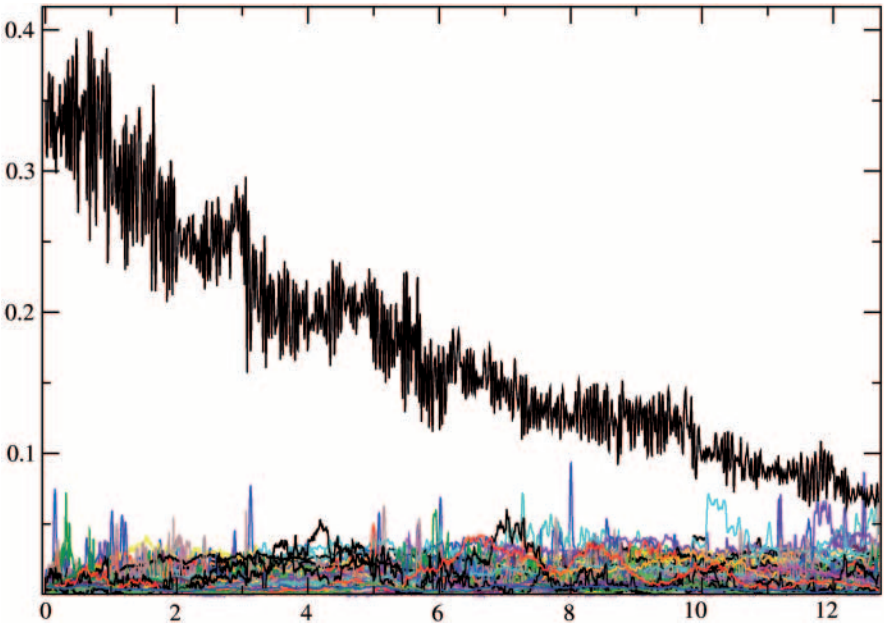

Fig. 3. Plot of the energy (eV) of the asymmetric stretch mode of H_{bc}^{+} at 2004 cm^{-1} in the Si_{64} supercell, at the background temperature of 75 K vs. time (ps). The energies of 194 other modes of the cell are plotted as well. The first step of the decay involves the pLVMs of the defect, which themselves have very short lifetimes

In all these processes, the Gibbs free energy is the most relevant quantity since virtually all experiments are done at constant pressure. However, we focus here on the Helmholtz free energy. Working at constant volume rather than constant pressure is appropriate up to several hundred degrees Celsius in most semiconductors because their thermal expansion coefficient is very small. In Si for example, this coefficient is $4.68 \times 10^{-6}\,K^{-1}$ at room temperature and the phonon frequencies shift slowly with T. Indeed, the difference between the constant-pressure and constant-volume specific heats [32][2] $C_P - C_V$ is $0.0165\,J/mol \cdot K$ at room temperature, a correction of only 0.08 % to $C_P = 20\,J/mol \cdot K$. Thus, in the case of semiconductors such as Si and in the temperature range where the use of harmonic dynamical matrices is justified, calculating the phonon density of states $g(\omega)$ at $T = 0\,K$ and ignoring the temperature dependence of the lattice constant are reasonable approximations. Note that this ignores the frequency shifs associated with the anharmonicity. This greatly simplifies the calculations. In fact, it renders them possible since calculating temperature-dependent phonon densities of state is a formidable task.

[2] In this paper, "cal/g atom" should read "cal/mol" (or the numbers be divided by the atomic mass of Si).

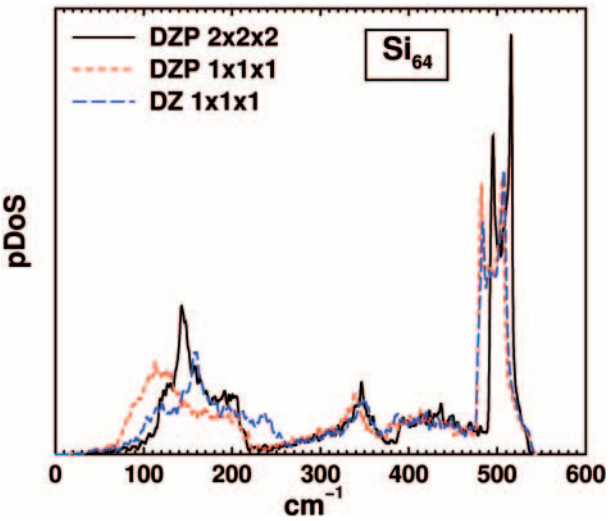

Fig. 4. Phonon densities of state $g(\omega)$ calculated from harmonic dynamical matrices of the Si_{64} supercell evaluated at 90 q points in the BZ of the cell. The *dashed line* (first peak at low frequency) was obtained with a double-zeta basis set and a $1\times1\times1$ Monkhorst–Pack k-point sampling, the *solid line* (second peak) with a double-zeta polarized basis set and a $2\times2\times2$ k-point sampling, and the *long-dash line* with a double-zeta polarized basis set and a $1\times1\times1$ k-point sampling. As expected, the solid line best matches the experimental data [33]

The phonon densities of states of perfect (defect-free) crystals are normally calculated from the dynamical matrix of the primitive unit cell evaluated at thousands of q points in the Brillouin zone (BZ) of the crystal. When studying defects, large periodic supercells must be used, and the BZ of the supercell is distinct from that of the perfect solid. The eigenvalues of the dynamical matrix only provide a small number of normal mode frequencies, and the phonon density of states extrapolated from those few hundred frequencies lead to rather poor $g(\omega)$s. However, evaluating the dynamical matrix at many q points in the BZ of the supercell works very well. Figure 4 shows three phonon densities of state obtained from harmonic dynamical matrices extrapolated at about 90 q points in the BZ of the supercell. The dynamical matrices were calculated with different basis sets and Monkhorst–Pack k-point sampling in the Si_{64} supercell. The best fit to the measured data [33] is obtained with the largest basis set (double-zeta polarized) and k-point sampling ($2\times2\times2$).

In the harmonic approximation, the Helmholtz free energy is given by

$$F_{\mathrm{vib}}(T) = k_{\mathrm{B}}T\int_0^{\infty} \ln\{\sinh(\hbar\omega/2k_{\mathrm{B}}T)\}g(\omega)\,\mathrm{d}\omega\,, \tag{8}$$

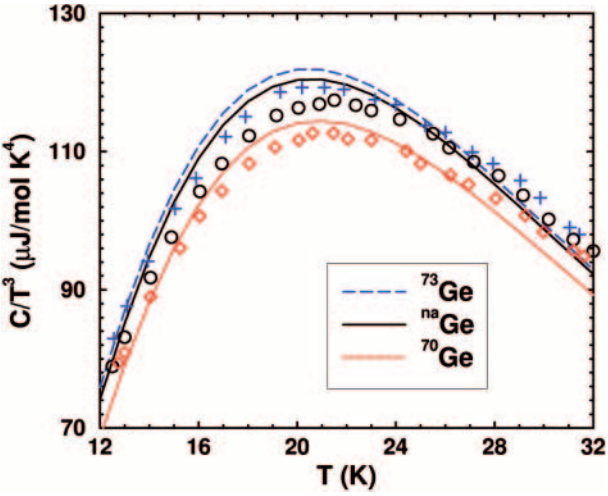

Fig. 5. Calculated and measured peak in the specific heat over T^3 for various isotopically pure samples of Ge. The *dashed line* (expt: *crosses*) is for ^{73}Ge, the *solid line* (expt: *circles*) is for the natural isotopic abundance, and the *dotted line* (expt: *squares*) is for ^{70}Ge

where $k_{\rm B}$ is the Boltzmann constant. In the perfect cell, the integration is carried out up to the Γ phonon. With a defect in the supercell, the integral extends up to the highest normal mode of the cell (perturbed Γ phonon) and becomes a simple sum for the higher LVMs. Note that $F_{\rm vib}(T=0)$ gives the total zero-point energy. Once $F_{\rm vib}$ is calculated, the vibrational entropy and specific heat at constant volume are given by

$$S_{\rm vib} = -\left(\frac{\partial F_{\rm vib}}{\partial T}\right)_{\rm V} \qquad C_{\rm V} = -T\left(\frac{\partial^2 F_{\rm vib}}{\partial T^2}\right)_{\rm V} . \tag{9}$$

The latter can be compared to the measured $C_{\rm p}$ in order to determine up to what temperature the constant-volume and harmonic approximations are appropriate. We have demonstrated that the approximations work quite well up to some 700 K in the case of c-C, Si, Ge, and GaN (see [1, 34, 35]), and that the agreement with even fine features is very good at low temperature. Figure 5 shows the calculated isotope-dependent peak in C/T^3 in Ge. The temperature at which it is predicted to occur and the splitting associated with various isotopes quantitatively reproduce the experimental data [36].

These tests give us confidence that the phonon densities of states calculated in the manner described above are also accurate when defects are present in the same supercell, and therefore that the calculated Helmholtz vibrational free energies are accurate up to several hundred degrees Celsius. Of course, the $F_{\rm vib}$ obtained for a defect in a 64-host-atom cell correspond to a defect concentration of about 1.5 atomic per cent, which is very high. However, the few calculations we have performed with 128- and 216-atom cells

show that cell-size effects are not very significant, probably because $g(\omega)$ is only the weight function used in the integration, and small changes in this function have a minor impact on the result.

6 Theory of Defects at Finite Temperatures

We are always interested in energy *differences*. This could be the energy difference between two configurations of the same defect, the energy difference between a bound complex and its dissociation products (binding energy), etc. At finite temperatures, such a total free-energy difference may contain several contributions:

$$\Delta F = \Delta U + \Delta F_{\rm vib} + \Delta F_{e/h} + \Delta F_{\rm rot} + \cdots - T\Delta S_{\rm config}\,. \tag{10}$$

ΔU is the potential-energy difference obtained (in this Chapter) from first-principles density-functional theory. The vibrational free energy $\Delta F_{\rm vib}$ has been discussed above.

Depending on the defect under study, there may be different concentrations of electrons (in the conduction band) or holes (in the valence band) because different configurations of a given defect have different electrical activities. This has nothing to do with the electrons or holes provided by the background dopants, which can be numerous. Instead, $\Delta F_{e/h}$ refers only to the change in the number of free carriers associated with the two configurations of the defect under study and the contributions of the background (dopant-associated) charge carriers cancel out. Unless one is dealing with unusually high changes in carrier concentrations, this term is very small and can be neglected [37].

Additional contributions to the free energy are associated with energy levels arising from rotational, magnetic, spin or other degrees of freedom specific to a particular defect. These terms can often be calculated directly from the appropriate partition functions [37]. One example is provided by the rotational free energies of interstitial H_2, HD, and D_2 molecules (for a review of interstitial hydrogen molecules in semiconductors, see [38]) in Si. The rotational energies are $E_j = j(j+1)\hbar^2/MR^2$, where M is the nuclear mass and R the internuclear separation. Since H_2 consists of two protons (fermions), there are three *ortho* states (even combinations of spin: $\uparrow\uparrow$, $\downarrow\downarrow$, and $\uparrow\downarrow + \downarrow\uparrow$) for which only odd values of j are allowed, and one *para* state (odd combination of spin: $\uparrow\downarrow - \downarrow\uparrow$) for which only even values of j are allowed. This leads to the familiar set of two Raman lines with 3 : 1 intensity ratio. Since D_2 consists of two bosons, the total wavefunction must be symmetric, and a similar argument leads to two Raman lines with 1 : 2 intensity ratios. Thus, the rotational free energy per molecule is given by

$$F_{\rm rot}(T) = -g_{\rm o}k_{\rm B}T\ln Z_{\rm o} - g_{\rm p}k_{\rm B}T\ln Z_{\rm p}\,, \tag{11}$$

where the partition functions are

$$Z_{\rm o} = \Sigma_{j=1,3,5,\ldots}(2j+1)e^{-j(j+1)\theta/T} \tag{12}$$

$$Z_{\rm p} = \Sigma_{j=0,2,4,\ldots}(2j+1)e^{-j(j+1)\theta/T}\,, \tag{13}$$

and $\{g_{\rm o}, g_{\rm p}\} = \{3/4, 1/4\}$ for H_2 and $\{1/3, 2/3\}$ for D_2. Since HD has no symmetry restrictions, all the values of j are allowed (single Raman line) and $F_{\rm rot}(T) = -k_{\rm B}T \ln Z$ with $Z = \Sigma_j(2j+1)e^{-j(j+1)\theta/T}$. The effective temperature $\theta = \hbar^2/2Ik_{\rm B}$ contains the (classical) moment of inertia I of the molecule, which depends on the host. For free H_2, $\theta = 85.4\,{\rm K}$. In Si, the molecule has a longer bond length than in free space [39] and $\theta = 73.0\,{\rm K}$, 48.7 K and 36.5 K for H_2, HD and D_2, respectively. The values of the rotational free energies at 77 K, 300 K, and 800 K for interstitial H_2, HD, and D_2 in Si are in Table 1. The largest rotational free energy per molecule occurs in the case of HD. When comparing the free energies of the interstitial H_2 molecule with the interstitial H_2^* complex (which consists of a Si–Si replaced by two Si–H bonds: Si–$H_{\rm bc}\cdots$Si–$H_{\rm ab}$ along a trigonal axis) which has no rotational degrees of freedom, the sum $\Delta F_{\rm vib} + \Delta F_{\rm rot}$ clearly favors H_2 at higher temperatures [37]. This prediction is consistent with the fact that samples hydrogenated at high temperatures then rapidly quenched [40] show the presence of only H_2 molecules.

Table 1. Rotational free energy $F_{\rm rot}$ (eV) for interstitial H_2, HD, and D_2 in Si

T (K)	77	300	800
H_2	−0.005	−0.030	−0.129
HD	−0.004	−0.048	−0.194
D_2	−0.004	−0.040	−0.168

The importance of the configurational entropy term depends on the situation. When comparing the two metastable configurations of the CH_2^* complex [37], $\Delta S_{\rm config}$ is exactly zero. When comparing interstitial H_2 and H_2^*, the configurational entropy contribution is not zero but very small. Indeed, interstitial H_2 occupies tetrahedral interstitial (t) sites, while H_2^* is at a bc site. Since there are twice as many bc as t sites, a (very) small $\Delta S_{\rm config}$ results.

The situation is very different when calculating binding free energies. Indeed, for a complex {A, B} with dissociation products A and B, there are often vastly different numbers of configurations for the dissociated species than for the complexes, sometimes leading to large values for $\Delta S_{\rm config}$. The calculation must be done using realistic concentrations of the species involved, and its details depend on the specific situation.

The difference in configurational entropy per complex is $\Delta S_{\rm config} = (k_{\rm B}/[\{{\rm A,B}\}])\ln(\Omega_{\rm pair}/\Omega_{\rm nopair})$, where $[\{{\rm A,B}\}]$ is the number of complexes,

and Ω_{pair} and Ω_{nopair} are the number of configurations with all possible complexes forming and with all complexes dissociated, respectively. We set $\{\text{A}, \text{B}\}$ to be "dissociated" when no B species is within a sphere of radius r_{c} (effective capture radius) of any A. The results are not very sensitive to the actual value of r_{c}. The sites of A, B, and $\{\text{A}, \text{B}\}$ are known and the concentrations [A] and [B] are estimated from experiment. If [A] is larger than [B], the maximum number of complexes is $[\{\text{A}, \text{B}\}] = [\text{B}]$. Although a real sample has traps for the dissociation products A and/or B that are distinct from isolated A and/or B in a perfect crystal, we ignore this additional complication.

We consider here two boron–oxygen complexes in Si. Both of them contain an interstitial oxygen dimer ($\{\text{O}_\text{i}\}_2$) trapped at either substitutional (B_s) or interstitial (B_i) boron. Thus, we consider the binding free energies of the $\{\text{B}_\text{s}, \text{O}_\text{i}, \text{O}_\text{i}\}$ and $\{\text{B}_\text{i}, \text{O}_\text{i}, \text{O}_\text{i}\}$ complexes (both in the +1 charge state). We do not discuss here the reasons why these complexes are important, how they form, and the consequences of complex formation for the sample at hand. These issues are discussed elsewhere [41, 42]. The configurations of the complexes and their dissociation products have been obtained from conjugate gradient geometry optimizations in the appropriate charge states. The binding energies at $T = 0\,\text{K}$ are $\Delta U = 0.54\,\text{eV}$ and $0.61\,\text{eV}$ for $\{\text{B}_\text{s}, \text{O}_\text{i}, \text{O}_\text{i}\}$ and $\{\text{B}_\text{i}, \text{O}_\text{i}, \text{O}_\text{i}\}$, respectively. All the vibrational free energies have been calculated. We are faced now with the calculation of ΔS_{config}.

Since we know the equilibrium sites of all the species involved, we know the number of equivalent orientations. However, we need to assume the concentrations of the various species in order to calculate the total number of configurations. We assume a sample with $N = 5 \times 10^{22}$ substitutional sites, $[\{\text{O}_\text{i}\}_2] = 10^{14}$ oxygen dimers, $[\text{B}_\text{s}] = 10^{19}$ substitutional and $[\text{B}_\text{i}] = 10^{14}$ interstitial boron impurities. The numbers depend on the sample, but these are realistic values that could correspond to an actual experimental situation. In a $1\,\text{cm}^3$ sample, the number of sites for B_s, split-interstitial sites for B_i and staggered or square configurations [42] for $\{\text{O}_\text{i}\}_2$ is 5×10^{22}.

At low temperatures, all the B_ss trap one $\{\text{O}_\text{i}\}_2$. The number of ways one can arrange $[\{\text{O}_\text{i}\}_2]$ dimers among N sites is $N!/[\{\text{O}_\text{i}\}_2]!(N - [\{\text{O}_\text{i}\}_2])!$. Each $\{\text{O}_\text{i}\}_2$ traps at one B_s and each $\{\text{B}_\text{s}, \text{O}_\text{i}, \text{O}_\text{i}\}$ has 12 equivalent orientations. leading to $12^{[\{\text{O}_\text{i}\}_2]}$ possibilities. The remaining $[\text{B}_\text{s}] - [\{\text{O}_\text{i}\}_2]$ borons are distributed among the remaining $N - 12[\{\text{O}_\text{i}\}_2]$ sites. Thus, the number of configurations for $\{\text{B}_\text{s}, \text{O}_\text{i}, \text{O}_\text{i}\}$ complexes is

$$\Omega_{\text{pairs}} = \frac{12^{[\{\text{O}_\text{i}\}_2]} N!(N - 12[\{\text{O}_\text{i}\}_2])!}{[\{\text{O}_\text{i}\}_2]!(N - [\{\text{O}_\text{i}\}_2])!([\text{B}_\text{s}] - [\{\text{O}_\text{i}\}_2])!(N - [\text{B}_\text{i}] - 11[\{\text{O}_\text{i}\}_2])!} \,. \tag{14}$$

At high temperatures, all the $\{\text{B}_\text{s}, \text{O}_\text{i}, \text{O}_\text{i}\}$ complexes are dissociated. We can arrange $[\text{B}_\text{s}]$ borons among N substitutional sites in $N!/[\text{B}_\text{s}]!(N - [\text{B}_\text{s}])!$ ways. If $r_{\text{c}} = 10\,\text{Å}$, no oxygen dimer is within a sphere of radius $10\,\text{Å}$ of any boron, implying that about 150 substitutional sites around each B_s are

not allowed, which means that the number of sites available for the $[\{O_i\}_2]$ oxygen dimers is $N-150[B_s]$. The number of configurations for the $\{O_i\}_2$'s is therefore $(N-150[B_s])!/[\{O_i\}_2]!(N-150[B_s]-[\{O_i\}_2])!$. Thus, the number of configurations for the dissociated complexes is

$$\Omega_{\rm nopairs} = \frac{N!(N-150[B_s])!}{[B_s]!(N-[B_s])![\{O_i\}_2]!(N-150[B_s]-[\{O_i\}_2])!}\,. \tag{15}$$

Using Sterling's formula and an expansion for $\ln(1+\varepsilon)$ with $\varepsilon \ll 1$, we get

$$\Delta S_{\rm config} = k_{\rm B}\left(\ln\frac{12[B_s]}{N} + \frac{278[B_s]}{N} - \frac{[\{O_i\}_2]}{[B_s]}\right). \tag{16}$$

Similar calculations for $\{B_i, O_i, O_i\}$ lead to

$$\Delta S_{\rm config} = k_{\rm B}\left(\ln\frac{24[\{O_i\}_2]}{N} - 1 + 32\frac{[B_i]}{N} + \frac{[\{O_i\}_2]}{N}\right). \tag{17}$$

With the concentrations assumed, this gives $\Delta S_{\rm config} = -0.515\,{\rm meV/K}$ and $-1.538\,{\rm meV/K}$ for $\{B_s, O_i, O_i\}$ and $\{B_i, O_i, O_i\}$, respectively.

The difference between these situations is huge, and the reason for it is quite obvious. Consider an $\{A, B\}$ complex that dissociates into A and B. If A and/or B are abundant (as is the case for B_s), there are many configurations resulting in pairs and relatively few configurations with A away from B. On the other hand, when both A and B are scarce (as is the case for B_i and $\{O_i\}_2$), there are far fewer ways to make pairs and a great number of dissociated configurations. The binding free energies of the $\{B_s, O_i, O_i\}$ and $\{B_i, O_i, O_i\}$ complexes are plotted as a function of temperature in Fig. 6. As shown in [41], the contribution of $\Delta F_{\rm vib}$ is very small and the slope is almost entirely determined by the difference in configurational entropy.

Thus, for a given $\{A, B\}$ complex, the smaller the concentration of A or B, the larger the configurational entropy associated with the dissociated species and the smaller the entropy associated with complex formation. Then, the slope of the binding free energy $E_{\rm b}(T)$ is much steeper. The opposite holds if A and/or B exist in high concentrations. In the example discussed in this Chapter, changing one component of the complex from B_s to B_i changes the relevant concentration from 10^{19} to 10^{14}, and this change of five orders of magnitudes roughly triples $\Delta S_{\rm config}$.

Note that above the temperature T_0 where $E_{\rm b}(T_0) = 0$, the interactions become *repulsive*. The value of T_0 depends on $E_{\rm b}(0)$ and on the slope, that is on $\Delta S_{\rm config}$, which in turn depends on the concentrations of the dissociation products in the sample.

7 Discussion

First-principles calculations of the properties of defects in periodic supercells have become quantitative in many respects. The configurations, energetics,

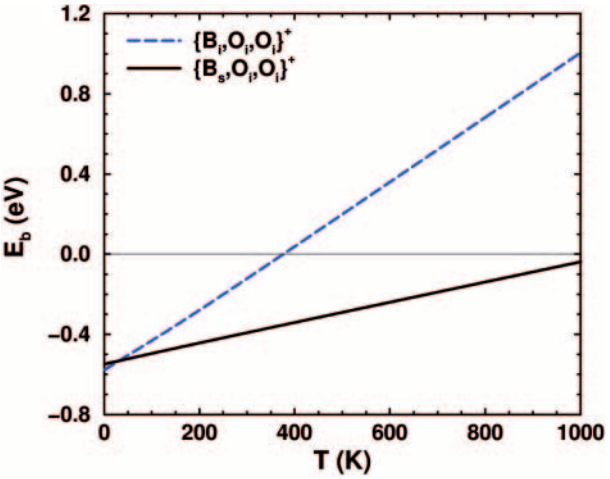

Fig. 6. Binding free energies of the $\{B_s, O_i, O_i\}$ and $\{B_i, O_i, O_i\}$ complexes in Si. The latter complex will not form above room temperature, where the interactions between B_s and $\{O_i\}_2$ are repulsive because of configurational entropy

selected LVMs, spin densities and, to a lesser extent, electrical activities of localized defects can be predicted with very good accuracy. However, the knowledge of the entire dynamical matrix of the system provides much more information.

The eigenvalues of this matrix give all the local, pseudolocal, and resonant defect-related modes, as well as the crystal phonon frequencies. They can be used to obtained high-quality phonon densities of states that in turn allow the calculation of vibrational free energies. Although the latter are limited to the constant-volume (and, in our case, harmonic) approximation, the calculated specific heats show that the results are reliable up to several hundred degrees Celsius.

The eigenvectors of the dynamical matrix allow the localization of local modes to be quantified and their symmetry predicted. The eigenvectors can also be used to prepare a system in thermal equilibrium at any temperature without the need for lengthy thermalizations or even a thermostat. This feature is needed to calculate vibrational lifetimes as a function of temperature.

The calculation of defect energetics at finite temperatures is relatively straightforward once the vibrational free energy is known. Indeed, rotational and other contributions can be obtained (or approximated) analytically. The most tricky, and sometimes most critical, part is the contribution of the configurational entropy. One example has been discussed here in detail. The result depends on the concentrations of the species involved, that is, on the sample.

The binding free energy of an $\{A, B\}$ complex in a crystal varies linearly with temperature, with a slope largely dominated by the difference in

configurational entropy between $\{A, B\}$ and A away from B. If R is a dissociation rate, $R\exp\{-E_b/k_BT\} = R\exp\{\Delta S_{config}/k_B\}\exp\{-E_b(0)/k_BT\}$, and an Arrhenius plot yields a straight line with slope $-E_b(0)/k_B$ and intercept $(\ln R + \Delta S_{config}/k_B)$. Thus, Arrhenius plots of the dissociation of an $\{A, B\}$ complex in samples containing different concentrations of A and/or B should produce parallel lines since the slopes are the same but the intercepts differ. This suggests a way to measure configurational entropies. If we take $R = 10^{11}$ and $\Delta S_{config} = -0.5\,meV/K$ or $-1.0\,meV/K$, the intercepts will be at $25.3 - 5.8 = 19.5$ or $25.3 - 11.6 = 13.7$, a measurable change.

Even though the potential energy surface describing the interactions between A and B has a pronounced minimum when the $\{A, B\}$ complex forms, these interactions become *repulsive* at temperatures $T > T_0$, where the critical temperature T_0 is defined from $E_b(T_0) = 0$. If complex formation begins at a temperature near (but below) T_0, the variations of the slope of $E_b(T)$ with time (that is: as complex formation takes place and the concentrations of the various species change) will cause the interactions to shift from attractive to repulsive, probably resulting in a maximum precipitate size.

Finally, since the difference in configurational entropy depends on the concentrations [A] and [B] in the sample, the binding free energy of a specific complex $\{A, B\}$ at a specific temperature will be different in samples containing different concentrations of A or B.

Many of these consequences of the temperature dependence of binding free energies are unexpected, not to say counterintuitive, because many of us are used to thinking in terms of potential energy alone. Yet, regardless of the value of the configurational entropy or the way it is calculated, the consequences are inescapable.

Acknowledgements

This work is supported in part by the R. A. Welch Foundation and the National Renewable Energy Laboratory.

References

[1] R. K. Kremer, M. Cardona, E. Schmitt, J. Blumm, S. K. Estreicher, M. Sanati, M. Bockowski, I. Grzegory, T. Suski, A. Jezowski: Phys. Rev. B **72**, 075209 (2005) 95, 104
[2] S. J. Pearton, J. W. Corbett, M. Stavola: *Hydrogen in Crystalline Semiconductors* (Springer, Berlin, Heidelberg 1991) 95
[3] S. Limpijumnong, C. G. Van de Walle: Phys. Rev. B **68**, 235203 (2003) 96
[4] A. F. Wright, C. H. Seager, S. M. Myers, D. D. Koleske, A. A. Allerman: J. Appl. Phys. **94**, 2311 (2003) 96
[5] A. Carvalho, R. Jones, J. Coutinho, P. R. Briddon: J. Phys.: Condens. Matter **17**, L155 (2005) 96

[6] G. Davies: Phys. Rep. **176**, 83 (1989) 96, 99
[7] S. Knack: Mater. Sci. Semic. Proc. **7**, 125 (2005) 96
[8] S. K. Estreicher, D. West, J. Goss, S. Knack, J. Weber: Phys. Rev. Lett. **90**, 035504 (2003) 96, 99
[9] S. K. Estreicher, D. West, M. Sanati: Phys. Rev. B **72**, R13532 (2005) 96, 99
[10] D. Sánchez-Portal, P. Ordejón, E. Artacho, J. M. Soler: Int. J. Quant. Chem. **65**, 453 (1997) 96, 97
[11] E. Artacho, D. Sánchez-Portal, P. Ordejón, A. García, J. M. Soler: phys. stat. sol. (b) **215**, 809 (1999) 96
[12] D. M. Ceperley, B. J. Adler: Phys. Rev. Lett. **45**, 566 (1980) 96
[13] S. Perdew, A. Zunger: Phys. Rev. B **32**, 5048 (1981) 96
[14] L. Kleiman, D. M. Bylander: Phys. Rev. Lett. **48**, 1425 (1982) 96
[15] O. F. Sankey, D. J. Niklevski: Phys. Rev. B **40**, 3979 (1989) 97
[16] O. F. Sankey, D. J. Niklevski, D. A. Drabold, J. D. Dow: Phys. Rev. B **41**, 12750 (1990) 97
[17] A. A. Demkov, J. Ortega, O. F. Sankey, M. P. Grumbach: Phys. Rev. B **52**, 1618 (1995) 97
[18] H. J. Monkhorst, J. D. Pack: Phys. Rev. B **13**, 5188 (1976) 97
[19] T. Inui, Y. Tanabe, Y. Onodera: *Group Theory and Its Applications in Physics* (Springer, Berlin, Heidelberg 1990) 97
[20] M. P. Allen, D. J. Tildesley: *Computer Simulations of Liquids* (Clarendon, Oxford 1987) 97
[21] J. L. Gavartin, D. J. Bacon: Comp. Mater. Sci. **10**, 75 (1998) 97
[22] S. Baroni, P. Gioannozzi, A. Testa: Phys. Rev. Lett. **58**, 1861 (1987) 97
[23] A. Fleszar, X. Gonze: Phys. Rev. Lett. **64**, 2961 (1990) 97
[24] X. Gonze, C. Lee: Phys. Rev. B **55**, 10355 (1997) 97
[25] J. M. Pruneda, S. K. Estreicher, J. Junquera, J. Ferrer, P. Ordejón: Phys. Rev. B **65**, 075210 (2002) 97
[26] P. Ordejón, D. A. Drabold, R. M. Martin, S. Itoh: Phys. Rev. Lett. **75**, 1324 (1995) 97
[27] D. West, S. K. Estreicher: Phys. Rev. Lett. **96**, 115504 (2006) 98, 101
[28] J. L. Gavartin, A. M. Stoneham: Phil. Trans. Roy. Soc. Lond. A **361**, 275 (2003) 98
[29] M. Budde, G. Luepke, C. Parks Cheney, N. H. Tolk, L. Feldman: Phys. Rev. Lett. **85**, 1452 (2000) 99, 100
[30] M. Budde, G. Luepke, C. Parks Cheney, N. H. Tolk, L. C. Feldman: Phys. Rev. Lett. **85**, 1452 (2000) 100, 101
[31] G. Lupke, X. Zhang, B. Sun, A. Fraser, N. H. Tolk, L. C. Feldman: Phys. Rev. Lett. **88**, 135501 (2002) 100
[32] P. Flubacher, A. J. Leadbetter, J. A. Morrison: Philos. Mag. **4**, 273 (1959) 102
[33] F. Widulle, T. Ruf, M. Konuma, I. Silier, M. Cardona, W. Kriegseis, V. I. Ozhogin: Solid State Commun. **118**, 1 (2002) 103
[34] M. Cardona, R. K. Kremer, M. Sanati, S. K. Estreicher, T. R. Anthony: Solid State Commun. **133**, 465 (2005) 104
[35] M. Sanati, S. K. Estreicher, M. Cardona: Solid State Commun. **131**, 229 (2004) 104
[36] W. Schnelle, E. Gmelin: J. Phys.: Condens. Matter **13**, 6087 (2001) 104

[37] S. K. Estreicher, M. Sanati, D. West, F. Ruymgaart: Phys. Rev. B **70**, 125209 (2004) 105, 106
[38] S. K. Estreicher: Acta Phys. Polon. A **102**, 403 (2002) 105
[39] S. K. Estreicher, K. Wells, P. A. Fedders, P. Ordejón: J. Phys.: Condens. Matter **13**, 62 (2001) 106
[40] E. E. Chen, M. Stavola, W. B. Fowler: Phys. Rev. B **65**, 245208 (2002) 106
[41] M. Sanati, S. K. Estreicher: Phys. Rev. B **72**, 165206 (2005) 107, 108
[42] J. Adey, R. Jones, D. W. Palmer, P. R. Briddon, S. Öberg: Phys. Rev. Lett. **93**, 055504 (2004) 107

Index

The Calculation of Free-Energies in Semiconductors: Defects, Transitions and Phase Diagrams

E. R. Hernández[1], A. Antonelli[2], L. Colombo[3], and P. Ordejón[1]

[1] Institut de Ciència de Materials de Barcelona (ICMAB–CSIC), Campus de Bellaterra, 08193 Barcelona, Spain
ehe@icmab.es
[2] Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, Unicamp, 13083-970, Campinas, São Paulo, Brazil
[3] SLACS (INFM-CNR) and Department of Physics, University of Cagliari, Cittadella Universitaria, I-09042 Monserrato (Ca), Italy

Abstract. In this chapter we review a series of novel techniques that make possible the efficient calculation of free energies in condensed-matter systems, without resorting to the quasiharmonic approximation. Employing these techniques, it is possible to obtain the free energy of a given system not just at a predefined temperature, but in a whole range of temperatures, from a single simulation. This makes possible the study of phase transitions, as well as the determination of equilibrium concentrations of defects as a function of temperature, as will be illustrated by examples of specific applications. The same techniques, coupled with a scheme to integrate the Clausius–Clapeyron equation, can lead to the efficient determination of phase diagrams, a capability that will be illustrated with the calculation of the phase diagram of silicon.

1 Introduction

The free-energy plays a central role in understanding the thermal properties of materials. From it, other properties of a material may be derived, such as the internal energy, the volume, entropy, etc. The phase behavior of a material is controlled by the values of the free-energy of its different phases, and the concentrations of defects or impurities, as well as their partition among coexisting phases, are defined by their chemical potentials, which are themselves obtained from the free-energy. It is clear, therefore, that it is highly desirable to have efficient computational tools that can evaluate the free-energy of modeled materials in different conditions of temperature and pressure (or volume). These methods should be accurate, and ideally it should be possible to combine them with first-principles electronic structure methods, which provide an accurate picture of the structure, bonding and energetics of materials.

In this Chapter we provide a self-contained description of recent theoretical developments that have contributed to making free-energy calculations

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 115–140 (2007)

more accessible and efficient, and we illustrate their capabilities with a series of applications, mostly in the field of semiconductors, the topic of this book. Our intention in writing this Chapter has been to illustrate the potential of these techniques, some of which are still relatively little known. On the other hand, we have not intended to provide an exhaustive review of free-energy techniques, nor to dwell overmuch on the technical details of implementation. Both topics are covered at length in the excellent book by *Frenkel* and *Smit* [1], and also to some extent in the earlier book by *Allen* and *Tildesley* [2]. Nor has it been our intention to list the many examples of applications of free-energy techniques to problems in materials science, chemical physics, geology or biomolecular systems. Applications in some of these fields are reviewed to some extent in the articles by *Rickman* and *LeSar* [3] and by *Ackland* [4]. A very recent review of some of the techniques that will be discussed here, in particular the reversible scaling technique, is that of *de Koning* and *Reinhardt* [5].

2 The Calculation of Free-Energies

Unlike the total internal energy, which depends only on the positions and velocities of the system at a single point in phase-space, the free-energy depends on all configurations (all state points) in the phase-space volume accessible to the system of interest in its given conditions of temperature and volume (or pressure). This is, in essence, why it is more arduous to calculate the free-energy in atomistic simulations of condensed-matter systems. From statistical mechanics we have that the free-energy at temperature T has the form

$$F = -k_{\mathrm{B}} T \ln \mathcal{Z} \,, \tag{1}$$

where k_{B} is Boltzmann's constant and $\mathcal{Z}$ is the *partition function*. In *canonical ensemble* conditions, i.e., constant number of particles N, constant volume V and constant temperature T, the partition function has the following expression:

$$\mathcal{Z}_{NVT} = \frac{1}{N!\, h^{3N}} \int_V \mathrm{d}\Gamma \, e^{-\beta H(\Gamma)} \,, \tag{2}$$

where Γ represents a point in phase-space, i.e., the $6N$-dimensional space formed by the $3N$ momenta and $3N$ coordinates of the system, $\mathrm{d}\Gamma$ is the corresponding volume element, h is Planck's constant, H is the Hamiltonian of the system, $\beta = (k_{\mathrm{B}} T)^{-1}$, and only configurations of volume V contribute to the integral. By writing an integral we are implicitly assuming that the system under consideration is being described by classical mechanics. In the quantum case the integral would be replaced by a discrete sum over quantum states.

Frequently it is desirable to consider the system under conditions of constant temperature and constant pressure (instead of constant volume), conditions that correspond to the so-called *isothermal-isobaric* or NPT ensemble. In this case, the partition function $\mathcal{Z}$ is generalized to

$$\mathcal{Z}_{\mathrm{NPT}} = \frac{1}{N!\, h^{3N}} \int_0^\infty \mathrm{d}V\, e^{-\beta PV} \left[\int_V \mathrm{d}\Gamma\, e^{-\beta H(\Gamma)} \right] . \tag{3}$$

In this case the free-energy is called the *Gibbs free-energy* (and we will label it G), to distinguish it from the canonical free energy (also called the *Helmholtz free energy*).

The partition function plays the role of a normalization factor in thermodynamical averages. Consider, for example, some property A, which depends on the positions, and/or velocities of the particles in the system. Then, the canonical thermal average of A would be given by

$$\langle A \rangle_{NVT} = \frac{\int_V \mathrm{d}\Gamma\, A(\Gamma)\, e^{-\beta H(\Gamma)}}{\int_V \mathrm{d}\Gamma\, e^{-\beta H(\Gamma)}}, \tag{4}$$

where the denominator can be recognized as the canonical partition function of (2) save for some factors that are canceled when taking the quotient of integrals. Such thermal averages are readily estimated by the standard simulation techniques of (canonical) molecular dynamics (MD) or Monte Carlo (MC), but it is important to note that these techniques directly estimate the quotient of integrals appearing in (4), and they cannot evaluate each of the integrals separately. Also, note that the free-energy is not itself a thermal average [it does not have the form of (4)], and therefore it cannot be obtained by direct MD or MC simulation.

2.1 Thermodynamic Integration and Adiabatic Switching

Equations (2) and (3) clearly illustrate the difficulty of evaluating the free-energy. These integrals are multidimensional, with as many dimensions as degrees of freedom in the system, and except for very simple models such as the ideal gas or the harmonic solid, they cannot be evaluated analytically. Furthermore, their high dimensionality precludes any attempt of evaluation by numerical quadrature methods. Thus it would seem that we are faced with an unsurmountable difficulty, though fortunately this is not the case. A way out of the problem is given by considering the dependence of the Hamiltonian H on some parameter λ, and asking ourselves how does the free-energy change with λ. It is immediately apparent that

$$\frac{\partial F}{\partial \lambda} = \frac{\int_V \mathrm{d}\Gamma \left(\frac{\partial H}{\partial \lambda}\right) e^{-\beta H(\Gamma)}}{\int_V \mathrm{d}\Gamma\, e^{-\beta H(\Gamma)}} = \left\langle \frac{\partial H}{\partial \lambda} \right\rangle . \tag{5}$$

Equation (5) shows that, while the free energy itself is not a thermal average, its derivative with respect to any parameter λ *is* a thermal average,

and therefore it can be evaluated by employing conventional simulation techniques. This observation forms the basis of the technique known as the *thermodynamic integration* or *coupling parameter* [6, 7] method for evaluating free-energies, or rather free-energy differences. Consider a system, for example a solid, described by some potential $U(\boldsymbol{r})$, where $\boldsymbol{r}$ stands for the set of (generalized) coordinates specifying the positions of all degrees of freedom, for which we wish to evaluate the free-energy at some given temperature and volume. Now imagine that there is another system, similar in some sense to the system of interest, for which the free-energy is known at the temperature and volume at which we wish to know it for the system of interest. We noted earlier that one of the systems for which the canonical partition function can be evaluated analytically is the harmonic solid in which each atom is tied to its equilibrium lattice site by means of a harmonic spring. We will refer to the system for which the free-energy is known as the *reference system*, and will assume it is described by a potential $U_{\mathrm{ref}}(\boldsymbol{r})$. Let us now define the λ-dependent Hamiltonian

$$H_\lambda = \frac{1}{2}\sum_i \frac{\boldsymbol{p}_i^2}{m_i} + \lambda U(\boldsymbol{r}) + (1-\lambda)U_{\mathrm{ref}}(\boldsymbol{r})\,. \tag{6}$$

Notice that when $\lambda = 0$ the Hamiltonian corresponds to that of the reference system (in our example the Einstein solid), while when $\lambda = 1$ it is that of the system of interest. A direct application of (5) tells us that the derivative of the free-energy F_λ associated with H_λ at $\lambda \in [0,1]$ is

$$\frac{\partial F_\lambda}{\partial \lambda} = \langle U(\boldsymbol{r}) - U_{\mathrm{ref}}(\boldsymbol{r})\rangle_\lambda\,. \tag{7}$$

Thus, to obtain the free-energy difference between the system of interest and the reference all we need to do is to integrate over λ,

$$\Delta F = F - F_{\mathrm{ref}} = \int_0^1 \mathrm{d}\lambda\, \langle U(\boldsymbol{r}) - U_{\mathrm{ref}}(\boldsymbol{r})\rangle_\lambda\,. \tag{8}$$

Since F_{ref} is known, the sought F follows immediately. As stated above, the λ-dependent thermal averages $\langle U(\mathbf{r}) - U_{\mathrm{ref}}(\boldsymbol{r})\rangle_\lambda$ are readily obtained from standard simulation techniques. In practice, a discrete set of λ values in the interval $[0,1]$ is chosen (usually of the order of 5 to 10), and at each one of them an equilibrium simulation is carried out with Hamiltonian H_λ, from which the average $\langle U(\boldsymbol{r}) - U_{\mathrm{ref}}(\boldsymbol{r})\rangle_\lambda$ is obtained. Then the integral in (8) is evaluated numerically. Note that because one samples directly the potential $U(\boldsymbol{r}) - U_{\mathrm{ref}}(\boldsymbol{r})$ using either MD or MC simulation techniques, no harmonic approximation is involved here, and thus anharmonic effects are automatically taken into account.

For this procedure to work, the chosen reference system must be in some sense *similar* to the target system. By this we mean that as λ is changed,

the system described by Hamiltonian H_λ (6) should not undergo any phase transitions. If that were to occur, some of the work spent in transforming the reference system into the system of interest (or vice versa) would actually be spent in the latent heat of the phase transition, and the resulting free-energy estimate would be inaccurate. Also, with a view to making the numerical integration of (8) accurate, it is desirable that the quantity $U(\boldsymbol{r}) - U_{\mathrm{ref}}(\boldsymbol{r})$ is subject to fluctuations as small as possible, and this is another criterion of similarity between target and reference systems. We have already mentioned that a useful reference for free-energy calculations of solids is the Einstein crystal, but this would not be a good reference for a liquid. Fortunately, there are a number of simple model liquids, such as the Lennard–Jones fluid [8] or the inverse-power fluid [9], which have been extensively studied, and for which the free-energy has been tabulated over a wide range of temperature and pressure (density) conditions, and therefore these models serve the purpose of reference systems for liquids. Of course, for systems in the gas phase the ideal gas is an adequate reference.

Equation (8) reminds us again of the fact that it is harder to obtain free-energies than it is to calculate thermal averages. A thermal average is typically obtained from a single equilibrium (MD or MC) simulation, while to evaluate the free-energy of the same system one needs several simulations. The effort is increased further if the free-energy must be evaluated at other temperatures or volumes (pressures). It can easily be seen that the amount of work necessary to find a coexistence point, let alone map out a phase boundary (where two coexisting phases have the same free-energy), soon becomes a daunting task, if one must resort to the schemes described thus far. Fortunately, starting in the early 1990s, a number of alternative techniques of increased efficiency have been developed, which considerably ameliorate this situation, and thanks to them the task of calculating phase boundaries, and even entire phase diagrams is nowadays more accessible. The first such development was proposed by *Watanabe* and *Reinhardt* [10], who showed that accurate estimations of ΔF could be obtained from a single simulation with Hamiltonian H_λ (6), during which the parameter λ is slowly (i.e., quasiadiabatically) switched from 0 to 1 (or vice versa). Thus, the task of performing several equilibrium simulations at different values of λ to obtain ΔF is reduced to that of performing a single nonequilibrium simulation on a system that quasiadiabatically transmutes from the reference system to the system of interest (or the reverse). During this nonequilibrium simulation one computes the *irreversible work* (reversible in the adiabatic limit) $W_{\mathrm{irr}}(t, \lambda{=}0{\rightarrow}1)$, given by

$$W_{\mathrm{irr}}(t, \lambda{=}0{\rightarrow}1) = \int_0^t \mathrm{d}t'\, \dot{\lambda}\, \{U[\boldsymbol{r}(t')] - U_{\mathrm{ref}}[\boldsymbol{r}(t')]\}\,, \tag{9}$$

where $\dot{\lambda} = \mathrm{d}\lambda/\,\mathrm{d}t'$ is the rate of change of λ, and the time variable corresponds to either real time in an MD simulation, or to *simulation* time in

an MC simulation (each MC sweep corresponding to a unit of simulation time). In the adiabatic limit, $W_{\rm irr}(t=\infty, \lambda = 0 \to 1)$ would be equal to the free-energy difference ΔF, but since in practice strict adiabaticity cannot be attained, $W_{\rm irr}(t, \lambda = 0 \to 1)$ is only an approximation to ΔF, due to entropic dissipation effects (see below). Nevertheless, experience shows that (9) provides satisfactorily accurate results with modest computational efforts, indeed more modest than those required in a full thermodynamic integration calculation. We will refer to the method of Watanabe and Reinhardt as the *adiabatic switching* method.

The adiabatic switching method is based on two key observations. The first one is that, by virtue of Liouville's theorem, any classical trajectory, even one generated by a time-dependent Hamiltonian, preserves the phase-space volume. In other words, the trajectory generated by a time-dependent classical Hamiltonian will evolve on a phase-space hypersurface that, though itself changing shape with time, encloses a fixed amount of phase-space volume. The second observation, due to Hertz [10], applies when the evolution of the Hamiltonian is slow, or adiabatic. Under these conditions, the evolving hypersurface of phase-space on which the trajectory moves corresponds, at each instant t, to a constant energy shell of energy $H(t)$. Therefore, a trajectory generated by an adiabatically evolving Hamiltonian preserves the entropy, defined as $S = k_{\rm B} \ln \Omega$, where Ω is the (invariant) phase-space volume enclosed by the adiabatically evolving phase-space hypersurface. This is the mechanical analog of the observation in thermodynamics that an adiabatic process (i.e., one that proceeds along a succession of equilibrium states) exerted on a system preserves its entropy. The consequence of this is that, since $\Delta S = 0$ when λ in (6) is adiabatically switched along the trajectory generated by H_λ in canonical ensemble conditions, the free-energy change involved in this switching process can be measured by the internal energy change alone. As pointed out above, in the limit of strict adiabaticity, we would have that $\Delta F = W_{\rm irr}(t = \infty, \lambda = 0 \to 1)$, but since the switching process cannot be truly adiabatic there will be some entropic dissipative effects, which will cause $W_{\rm irr}(t, \lambda = 0 \to 1)$ to be an upper bound of ΔF. In fact, one can also run the switching simulation in reverse, i.e., switching λ from 1 to 0, and the resulting irreversible work, $-W_{\rm irr}(t, \lambda = 1 \to 0)$ then provides a lower bound to ΔF. This proves to be a convenient way of determining error bars for the estimated free-energy [5].

2.2 Reversible Scaling

Another important development allows one to obtain the free energy in a quasicontinuous range of temperatures, starting from a reference temperature at which the free-energy is known, from either a previous adiabatic switching or thermodynamic integration calculation. This method, known as *reversible scaling*, was introduced by *de Koning* et al. [11], and is based on the

formal equivalence that exists, from a statistical-mechanics perspective, between scaling the temperature and scaling the potential. Indeed, the canonical partition function of a system described by potential $U(\boldsymbol{r})$ at temperature T is

$$\mathcal{Z}_{NVT} = \frac{1}{N!\,\Lambda^{3N}(T)} \int_V \mathrm{d}\boldsymbol{r}\, e^{-\frac{U(r)}{k_\mathrm{B}T}}, \tag{10}$$

where the factor $\Lambda(T) = (h^2/2\pi m k_\mathrm{B} T)^{1/2}$ results from integrating out the momenta, and we have assumed that all particles are identical. Now, consider a system described by a scaled potential, $\lambda U(\boldsymbol{r})$, at a different temperature T_0. For this system, the partition function would be

$$\begin{aligned}\mathcal{Z}_{NVT_0}(\lambda) &= \frac{1}{N!\,\Lambda^{3N}(T_0)} \int_V \mathrm{d}\boldsymbol{r}\, e^{-\frac{\lambda U(r)}{k_\mathrm{B}T_0}} \\ &= \frac{\lambda^{3N/2}}{N!\,\Lambda^{3N}(T_0/\lambda)} \int_V \mathrm{d}\boldsymbol{r}\, e^{-\frac{U(r)}{k_\mathrm{B}T_0/\lambda}}\,. \end{aligned} \tag{11}$$

Now, it is easy to see that, except for a factor of $\lambda^{3N/2}$, the partition function of the scaled system at temperature T_0/λ and that of the unscaled system at temperature T are identical, if we impose that $T_0 = \lambda T$. From this it is easy to derive the following relation between the free-energies of the scaled and unscaled systems:

$$\frac{F(T)}{T} = \frac{F_\mathrm{s}(T_0,\lambda)}{T_0} + \frac{3}{2} N k_\mathrm{B} \ln\lambda\,. \tag{12}$$

In words, we have related the free-energy of a system at temperature T to that of an appropriately scaled system at temperature T_0 such that $T_0 = \lambda T$. From a formal point of view not much has been done. However, using the technique of adiabatic switching described in Sect. 2.1 it is very simple to calculate $F_\mathrm{s}(T_0,\lambda)$, at fixed temperature T_0 but in a quasicontinuous range of λ values in some interval $\lambda \in [\lambda_\mathrm{min}, \lambda_\mathrm{max}]$, and by virtue of (12), each of these $F_\mathrm{s}(T_0,\lambda)$ values corresponds to the free-energy of the unscaled system in a quasicontinuous range of temperatures $T \in [T_0/\lambda_\mathrm{max}, T_0/\lambda_\mathrm{min}]$. $F_\mathrm{s}(T_0,\lambda)$ is obtained as

$$F_\mathrm{s}(T_0,\lambda) = F_\mathrm{s}(T_0,\lambda_0) + W_\mathrm{irr}^{NVT}(\lambda_0 \to \lambda), \tag{13}$$

where $F_\mathrm{s}(T_0,\lambda_0)$ is the free-energy of the scaled system at the initial value of the scaling parameter λ_0 (usually either λ_min or λ_max) and $W_\mathrm{irr}^{NVT}(\lambda_0 \to \lambda)$ is the work done by switching the scaling parameter from λ_0 to λ, which, using the adiabatic switching technique is estimated as

$$W_\mathrm{irr}^{NVT}(\lambda_0 \to \lambda) = \int_0^t \mathrm{d}t'\, \dot{\lambda}\, U[\boldsymbol{r}(t')]\,. \tag{14}$$

Thus, instead of having to run a separate adiabatic switching or thermodynamic integration calculation to obtain the free-energy of the system of interest for each temperature in a discrete set spanning the desired range of temperatures, one can perform a single adiabatic switching calculation of the scaled system at some appropriately chosen temperature T_0, and simply quasicontinuously vary the scaling parameter λ from $\lambda_{\min}$ to $\lambda_{\max}$ (or vice versa). Note that in (14) each time step leads to a new value of λ and a corresponding value of $W_{\rm irr}^{NVT}(\lambda_0 \to \lambda)$, which, through (12), leads to an estimate of $F(T)$ at temperature $T = T_0/\lambda$.

For the sake of simplicity, we have presented the method in the context of the canonical ensemble, but the same ideas can be applied in the isothermal-isobaric case [12], which is of more interest when it comes to studying phase transitions. In this case one finds that it is necessary not only to scale the potential energy but the pressure as well, so that the scaled system pressure is $P_{\rm s}(\lambda) = \lambda P$, P being the pressure of the unscaled system. Then (12) transforms into

$$\frac{G(T,P)}{T} = \frac{G_{\rm s}[T_0, P_{\rm s}(\lambda)]}{T_0} + \frac{3}{2} N k_{\rm B} \ln \lambda \,, \tag{15}$$

relating the Gibbs free-energies of the scaled and unscaled systems, and $G_{\rm s}(T_0, P_{\rm s}, \lambda) = G_{\rm s}[T_0, P_{\rm s}(\lambda_0), \lambda_0] + W_{\rm irr}^{\rm NPT}(\lambda_0 \to \lambda)$, with

$$W_{\rm irr}^{\rm NPT}(\lambda_0 \to \lambda) = \int_0^t {\rm d}t' \, \dot{\lambda} \left\{ U[\boldsymbol{r}(t')] + \frac{{\rm d}P_{\rm s}(\lambda)}{{\rm d}\lambda} \, V(t') \right\} , \tag{16}$$

which is obtained from an adiabatic switching calculation carried out under isobaric-isothermal conditions.

The reversible scaling technique discussed above provides an efficient procedure for calculating coexistence points between different phases of a given material or substance. Using reversible scaling in the isothermal-isobaric ensemble one can easily obtain the Gibbs free energy of the two different phases, at constant pressure P, in a range of temperatures bounding the values where the coexistence point is expected to be located. The temperature at which the two free energies match is by definition the coexistence temperature of the two phases at pressure P. This strategy is illustrated for the particular case of the melting point of Si in the diamond phase at 0 pressure in Fig. 1. Si was simulated with the semiempirical tight-binding (TB) [13] Hamiltonian due to *Lenosky* and coworkers [14], which describes very well the structural and thermal properties of Si in several of its phases [15].

2.3 Phase Boundaries and Phase Diagrams

If one wishes to map out an entire phase boundary, rather than just a single coexistence point between two phases in equilibrium, one option would be to iterate the procedure illustrated in Fig. 1 to obtain new coexistence temperatures at different pressures. However, an alternative method exists

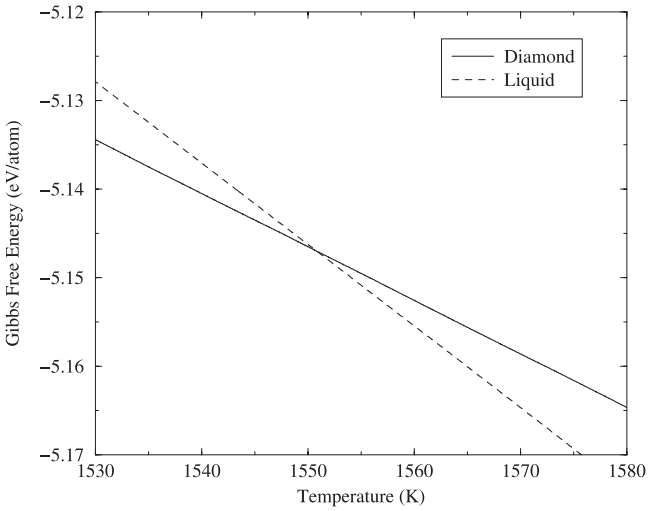

Fig. 1. Gibbs free-energy of the diamond and liquid phases of Si in the neighborhood of the zero-pressure melting point, as obtained from reversible scaling simulations with the *Lenosky* et al. [14] model. The simulations contain 128 atoms in each phase, and four k-points were used to sample the Brillouin zone. It can be seen that the two free-energies become equal at approximately 1551 K, and above this temperature the liquid becomes thermodynamically more stable. More details of these calculations can be found in [15, 16]

that is often more efficient and practical, since it does not require any further calculation of free-energies along the phase boundary. This technique was pioneered by *Kofke* [17, 18], and is known as *Gibbs–Duhem* or *Clausius–Clapeyron integration*. In this method, starting from some previously determined coexistence point along the desired phase boundary, one numerically solves the Clausius–Clapeyron equation:

$$\frac{\mathrm{d}T}{\mathrm{d}P} = T\frac{\Delta V}{\Delta H}\,. \tag{17}$$

This equation relates the slope of the phase boundary, $\mathrm{d}T/\mathrm{d}P$, at the current coexistence point, given by temperature T and pressure P, with the differences of molar volumes ΔV and molar enthalpies ΔH of the coexisting phases. Thus, if ΔV and ΔH are known at the current coexistence point, one can estimate how the equilibrium temperature will change if the pressure changes by some small amount. ΔV and ΔH can be obtained from standard equilibrium simulations using MD or MC. Note that the interface between the coexisting phases need not be taken into account: each phase is independently simulated in a separate simulation box, at identical conditions of temperature and pressure. Once the average volume and enthalpy of each phase is known with sufficient accuracy, new coexistence conditions are derived from (17),

and the process is iterated until the phase boundary has been mapped out. The stability and accuracy of the method have been analyzed in detail by *Kofke* [18]. This method has been extensively employed, and some particular examples of its use will be illustrated in the next section.

A parallelism can be established between Kofke's Clausius–Clapeyron integration method and thermodynamic integration. In the latter, one performs a series of equilibrium simulations at different λ values in the range $[0, 1]$ with the aim of obtaining the free-energy of the system. In Clausius–Clapeyron integration, one performs a series of equilibrium simulations at different values of the independent variable (T or P), and uses (17) to find how the dependent variable (P or T) adapts to the change in the independent variable such that the two phases remain in equilibrium. However, as we saw in Sect. 2.1, the calculation of the free energy can be made more efficiently using adiabatic switching, where instead of several (ca. 5–10) equilibrium simulations, a single nonequilibrium quasiadiabatic simulation is employed. It is therefore natural to ask if a similar gain in efficiency is possible in Clausius–Clapeyron integration, and the answer is affirmative, as *de Koning* et al. [12] have shown. The Clausius–Clapeyron equation (17) is arrived at by demanding that the change in free-energy due to a small change in the independent variable be the same in both phases. For illustrative purposes, let us consider the case in which the temperature T plays the role of independent variable, while the pressure P is the dependent one. As in the reversible scaling method, let us now consider scaled versions of the two phases at constant temperature T_0 and pressure P_s. When $\lambda = 1$ these correspond to some initial coexistence point. The free-energies of the scaled phases will be $G_{\mathrm{S,a}}$ and $G_{\mathrm{S,b}}$, respectively. In the scaled case, now the temperature is fixed, and the role of the independent variable is played by λ. Therefore, as λ is changed, the scaled phases will depart from equilibrium, unless the scaled pressure $P_\mathrm{s}(\lambda)$ evolves in such a way as to make the change in $G_{\mathrm{S,a}}$ equal to that in $G_{\mathrm{S,b}}$. By requiring that this be the case, one arrives at a Clausius–Clapeyron equation for $P_\mathrm{s}(\lambda)$, which reads

$$\frac{\mathrm{d}P_\mathrm{s}(\lambda)}{\mathrm{d}\lambda} = -\frac{\Delta U}{\Delta V}\,, \tag{18}$$

where ΔU is the difference of internal energies between the two phases. If λ is stationary, the values of ΔU and ΔV would be obtained from standard equilibrium simulations, and everything would be exactly as in the case of the Clausius–Clapeyron integration of Kofke, but working with the scaled phases. However, (18) suggests that, as in the reversible scaling method, λ can be varied quasicontinuously (i.e., quasiadiabatically). In these conditions, ΔU and ΔV will adapt to the changing value of λ, and $P_\mathrm{s}(\lambda)$ will evolve according to (18) such that the scaled phases remain in equilibrium. Through the scaling relations $T = T_0/\lambda$, $P = \lambda P_\mathrm{s}(\lambda)$, the equilibrium pressure $P(T)$ at temperature T will be obtained for the unscaled phases. De Koning et al. demonstrated the usefulness of the method by calculating the melting curve

of the Lennard–Jones model for pressures between 0 up to 160 reduced units, obtaining results matching those obtained previously by *Agrawal* and *Kofke* [19] employing the standard (equilibrium) Clausius–Clapeyron integration method. In the next section we will see another application, in which this *dynamic* Clausius–Clapeyron integration method has been recently used to obtain the phase diagram of Si as predicted by a semiempirical TB model.

3 Applications

Let us now discuss some examples of applications of the techniques described in Sect. 2 to the study of thermal properties of defects, phase transitions and phase diagrams. Our focus, given the theme of this book, will be mostly on semiconductors, though occasionally we will mention examples of applications to other types of materials, due to their importance or to their illustrative value. Let us remark that we do not intend to provide an exhaustive review of the literature concerned with thermal properties of semiconducting materials, as this would be beyond the scope of this Chapter. Our aim is rather to illustrate the capabilities of the novel techniques discussed above.

3.1 Thermal Properties of Defects

Most calculations of the free-energy and other thermal properties of defects in semiconductors to be found in the literature rely on the use of the quasiharmonic approximation, and since this topic is going to be covered at length in Chap. 8, we will not dwell much on it here. However, we must mention two studies that employed techniques described in Sect. 2, and that therefore go beyond the quasiharmonic approximation. Both studies illustrate nicely the potential of the techniques discussed in this Chapter, and make evident the need for incorporating the effects of anharmonicity at sufficiently high temperatures.

The first example is the study of *Jääskeläinen* et al. [20], who computed the free-energy of formation of the vacancy and self-interstitial in Si as a function of temperature. A thorough study of self-diffusion requires accurate free-energy calculations (aimed at predicting temperature-dependent equilibrium concentrations) and extensive diffusivity simulations (aimed at computing migration energies and diffusivity prefactors) for all relevant kinds of native defects. In their study, Jääskeläinen and coworkers considered only self-interstitial (I) and vacancy (V) defects.

Once the formation free-energies $F^{\rm f}_{\rm I,V} = E^{\rm f}_{\rm I,V} + TS^{\rm f}_{\rm I,V}$, as well as migration energies $E^{\rm m}_{\rm I,V}$ and diffusivity prefactors $d^0_{\rm I,V}$ are known, the self-diffusion coefficient $D_{SD}(T)$ can be cast in the form

$$D_{\rm SD}(T) = d^0_{\rm I} \exp\left(-\frac{E^{\rm f}_{\rm I} - TS^{\rm f}_{\rm I}}{k_{\rm B}T}\right) \exp\left(-E^{\rm m}_{\rm I}/k_{\rm B}T\right) + d^0_{\rm V} \exp\left(-\frac{E^{\rm f}_{\rm V} - TS^{\rm f}_{\rm V}}{k_{\rm B}T}\right) \exp\left(-E^{\rm m}_{\rm V}/k_{\rm B}T\right) , \quad (19)$$

so that a direct theory vs. experiment comparison is possible.

The thermodynamic integration (TI) method [1] was adopted by *Jääskeläinen* et al. [20] to evaluate $F^{\rm f}_{\rm I,V}$ in c-Si by means of the tight-binding (TB) model provided by *Kwon* et al. [21]. The ensemble average appearing in (8) was performed during constant-volume, constant-temperature molecular dynamics (TBMD) simulations in a 64±1 atom periodically repeated cell. The thermodynamical integration was performed over 16 λ points, while free-energy calculations were performed at four different temperatures, namely 300 K, 500 K, 1000 K and 1400 K. The formation entropy $S^{\rm f}_{\rm I}$ for the self-interstitial defect was found to be almost constant with temperature, the average value being $S^{\rm f}_{\rm I} = 11.2k_{\rm B}$. First-principles calculations [22] predict a value of $S^{\rm f}_{\rm I} \sim 10k_{\rm B}$ when including anharmonic terms through TI calculations as well. The case of a vacancy is more complicated due to its high mobility [23], which effectively adds a sizeable contribution of migration entropy in the TI free-energy calculations. This is confirmed by the fact that the computed value of $S^{\rm f}_{\rm V}$ varied in the range $10.2k_{\rm B}$ to $11.7k_{\rm B}$ in the selected temperature interval, with an average value of $S^{\rm f}_{\rm V} \sim 10.8k_{\rm B}$. In the case of the vacancy, in order to prevent the double counting of the migration entropy, which is already taken into account by the TBMD simulations aimed at measuring $d^0_{\rm V}$, short observation runs were performed, taking care to select only those simulations where V diffusion actually did not take place. The resulting average formation (i.e., configurational + vibrational) entropy was $S^{\rm f}_{\rm V} \sim 8.8k_{\rm B}$.

These TBMD results for $S^{\rm f}_{\rm I}$ and $S^{\rm f}_{\rm V}$ are in good qualitative agreement with first-principles calculations by *Blöchl* and coworkers [22] in the sense that the difference in the entropies of formation for interstitials and vacancies is of the order of $1k_{\rm B}$ to $2k_{\rm B}$ in both studies, with a larger formation entropy for the interstitial.

Diffusivity constants were finally obtained by using the migration prefactors and energies computed by means of the same TBMD functional, following *Tang* et al. [23]. The overall reliability of this TI-TBMD approach is summarized in Fig. 2, where the silicon TBMD total self-diffusion coefficient $D_{SD} = D_I + D_V$ is compared with state-of-the-art experimental data [24,25]. As can be seen, there is extremely good agreement between the different sets of experimental data and the theoretical results. These temper-

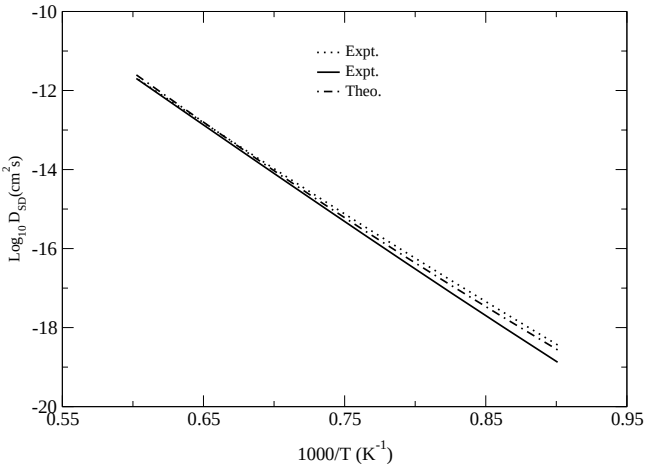

Fig. 2. Temperature dependence of the total self-diffusion coefficient in silicon. The *continuous line* is the experimental results of [24], the *dotted line* is the experimental measurements of [25], and the *dot-dash line* is the theoretical results of [20]. Reprinted with permission from *Jääskeläinen* et al. [20], Copyright (2001) by the American Physical Society

ature-dependent self-diffusion values should be therefore useful in modeling Si bulk processing.

As noted above, Jääskeläinen et al. implicitly assumed that the main contributors to self-diffusion in Si are the vacancy and interstitial defects, thus neglecting the possible contribution of other kinds of defects, such as clusters of interstitials or vacancies. Indeed, both first-principles simulations [26] and tight-binding temperature-accelerated MD simulations have revealed that small clusters of defects can also make substantial contributions to the mobility. In particular, the study of *Cogoni* et al. [27] shows quite conclusively that in all ranges of temperatures the mobility of the interstitial dimer is comparable to that of the single interstitial. At sufficiently high temperatures (above 1400 K) even the trimer diffuses fairly rapidly. However, it should be noted that to better estimate the significance of these clusters of interstitials to self-diffusion in Si, their concentrations, as well as their mobilities have to be taken into account. Since their formation energies are larger than that of the single interstitial, it is likely that in equilibrium conditions their net contribution to self-diffusion will be rather small. On the other hand, the processing of Si samples may incur nonequilibrium defect populations, and under these conditions the contribution of clusters of interstitials to self-diffusion could be important.

Let us now focus on another aspect of the thermal properties of Si. At room temperature, Si is a brittle material, but as the temperature is raised, a dramatic change in its mechanical properties takes place in a very narrow

range of temperatures around 873 K, and Si becomes a ductile material [28]. It is accepted that this change is due to the fact that at this temperature dislocations can be more easily nucleated and emitted at the tip of cracks, blunting the crack tip and making its propagation through the material more difficult. In Si there are two sets of closely packed {111} slip planes, called *shuffle* and *glide*. While dislocations nucleated at the shuffle set have high Peierls stress and low mobility, the dislocations on the glide set can split into partials, having a smaller Peierls stress and higher mobility. It is believed that the brittle–ductile transition in silicon can be related to the change in dominance of one set over the other. Over 10 years ago, *Rice* and collaborators [29–31] proposed a model in which the resistance to nucleation of dislocations at the crack tip can be measured by the so-called unstable stacking energy, which is the lowest energy barrier that needs to be crossed when one half of a crystal slips relative to the other half. Through a combination of density-functional theory calculation with Vineyard's transition state theory, *Kaxiras* and *Duesbery* [32] have shown that an abrupt transition from shuffle to glide dominance happens with temperature. In this approach, the temperature effects were considered without taking into account the real dynamics of the atoms on either side of the slip plane. The calculations by *de Koning* et al. [33] using the adibatic switching method within MD simulations fully included such temperature effects. Figure 3 shows the phase diagram for preferable nucleation of dislocations on shuffle plane versus glide set. One can see from the figure that the inclusion of all anharmonic effects in the calculations by de Koning et al. did not change substantially the diagram previously obtained by Kaxiras and Duesbery.

3.2 Melting of Silicon

Probably the first calculation of the melting point of crystalline Si from simulation was carried out by *Broughton* and *Li* [34], employing the well-known *Stillinger–Weber* [35] potential. Broughton and Li calculated the zero-pressure melting point by obtaining the free-energy of the crystal and liquid phases at different temperatures. The free-energy of the crystal was calculated by thermodynamic integration at different temperatures, taking as a reference system the harmonic solid. For the liquid phase, the free-energy was obtained by first switching off the attractive two-body part of the potential to avoid the occurrence of a possible first-order liquid–vapour transition, and then expanding the system to zero density. They obtained a melting temperature of 1691 K, with an estimated error of ±20 K. This value is in very good agreement with the currently accepted experimental value of 1687 K [36], an agreement that is perhaps not so surprising, as the Stillinger–Weber potential was parametrized on the basis of data from both the crystal and liquid phases. The careful free-energy tabulation of the Stillinger–Weber potential for the crystal and liquid phases of Si achieved by Broughton and Li has al-

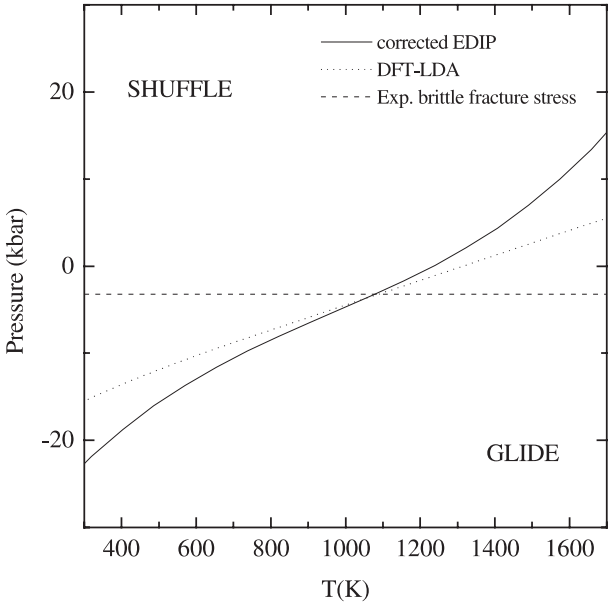

Fig. 3. Coexistence curves separating the preferable nucleation of dislocations on shuffle planes versus glide set. The *full line* is the results obtained from the EDIP potential after applying a correction for the overestimation of γ_{us}, the unstable stacking energy, by this potential. The *dotted line* shows the results obtained with DFT LDA calculations, and the *dashed horizontal line* gives the value of the experimental brittle fracture stress. Figure reprinted with permission from *de Koning* et al. [33]. Copyright (1998) by the American Physical Society

lowed subsequent studies (see below) to use this model as a reference system for free-energy calculations of Si employing higher levels of theory.

In a subsequent study, *Clancy* and *Cook* [37] compared several empirical potential models in their ability to describe the thermal behavior of Si and Ge; in particular they considered the well-known *Tersoff* potential [38–40], and a form of the embedded atom potential modified to describe covalent bonding (modified embedded atom method, MEAM) [41]. In this work the melting point for each model was determined by employing the two-phase coexistence method, i.e., by simulating the solid and liquid in a large cell in direct coexistence. It was found that the Tersoff model for Si led to a melting temperature of 2547 K, well above the experimental value of 1687 K, while the MEAM showed the opposite behavior, underestimating the melting temperature at 1475 K. Both models correctly predict, as does the Stillinger–Weber potential, a slight increase of the density upon melting, though all models underestimate somewhat the crystal and liquid densities. A similar study, comparing several tight-binding models in their ability to reproduce

the liquid structure, and the melting temperature of Si has been recently carried out by *Kaczmarski* et al. [15].

Very soon after the proposal of the adiabatic switching technique by *Watanabe* and *Reinhardt* [10], *Sugino* and *Car* [42] employed this technique, combined with first-principles Car–Parrinello dynamics [43], to study the melting of Si. These authors employed as a reference system the Stillinger–Weber potential for both the crystal and liquid phases, for which the free energy had been previously determined by *Broughton* and *Li* [34] as a function of temperature, as discussed above. Sugino and Car thus calculated the free-energy difference between the reference model and a first-principles description of Si employing density-functional theory (DFT) with the local-density approximation (LDA) for the exchange and correlation energy, combined with the pseudopotential approximation. They predicted a melting temperature of 1350 ± 100 K, some 300 K below the experimental value. This error was attributed to the relative destabilization of the solid phase with respect to the liquid in the LDA, which would result in a reduction of the melting point. Any small error in the difference of free energies translates into a rather large error in the melting temperature, and hence these kind of calculations require very stringent accuracy constraints. More recently, *Alfè* and *Gillan* [44] have revisited the theme of the melting of Si from first-principles simulations. These authors, employing the LDA approximation obtained a melting temperature of 1300 ± 50 K, in good agreement with the value reported earlier by Sugino and Car. They also performed calculations employing a generalized gradient approximation (GGA) to the exchange and correlation energy, obtaining in this case a somewhat improved value of the melting temperature of 1492 ± 50 K, which, although still below the experimental value by nearly 200 K, clearly improves upon the LDA result. This observation confirms the suggestion made by Sugino and Car that the underestimation of the melting point was most likely attributable to errors in the exchange and correlation functional, and not in any other of the approximations involved in the simulations (pseudopotential, basis set, etc.). It is worth noting that Si is a troublesome case, as the melting point separates two phases of different electronic character: while the crystal is semiconducting, the liquid is metallic. Errors due to the approximate nature of the functionals may, in general, be of different size in insulators and metals. Indeed, when systems in which the electronic character of the material does not change upon melting have been studied, DFT methods can predict melting temperatures within a few tens of K of the experimental result. This is the case for Al, which has been studied by *de Wijs* et al. [45], who reported a value of $T_{\mathrm{m}} = 890 \pm 20$ K, compared to an experimental value of 933.47 K.

Both liquid silicon (l-Si) and water present an anomalous behavior of the density and specific heat with temperature. Also, like water, l-Si is a very poor glass former, in the sense that it is extremely difficult to experimentally obtain the amorphous phase by quenching from the melt. In the case of silicon, as a matter of fact, this has never been achieved. Due to these similarities, it has

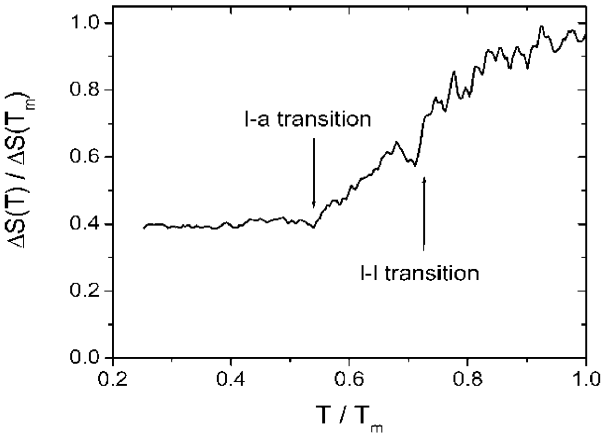

Fig. 4. Kauzmann plot for the environment-dependent interatomic potential (EDIP)-Si. Normalized entropy difference between the disordered and crystalline phases according to the EDIP model as a function of normalized temperature. Reprinted with permission from [46]. Copyright (2004), American Institute of Physics

been proposed that, as in water, supercooled l-Si could present polymorphism, as well as polyamorphism. Experiments have suggested the existence of a first-order-like liquid–liquid (l–l) transition in supercooled l-Si [47, 48]. Also, it has been considered, based on experimental results, that the transition from amorphous silicon (a-Si) to l-Si (l–a transition) would also be a discontinuous transition [47]. From the computational point of view, MD simulations using the Stillinger–Weber potential [49] have determined a l–l transition at about 1060 K, but have not found evidence of a l–a transition. Other computational results, using the EDIP [50] model for silicon, found a first-order-like l–a transition at 1170 K and no evidence of a l–l transition [51]. In a recent study using the reversible scaling method within MC simulations and also using the EDIP model, *Miranda* and *Antonelli* [46] have shown that a l–l transition occurs at 1135 K and a l–a transition takes place at 843 K. Figure 4 shows the normalized excess entropy of the liquid phase with respect to the crystalline phase as a function of the normalized temperature. From the figure one can see that as the temperature is lowered, first a discontinuous transition takes place, and as the temperature is lowered further, a glass-like transition occurs, beyond which the excess entropy becomes constant. This would be in agreement with recent experiments [52] that have indicated that a glass-like transition takes place at about 1000 K. Also, the configurational entropy obtained by Miranda and Antonelli for the amorphous phase is in agreement with that obtained by *Vink* and *Barkema* [53] using a methodology based on graph and information theories.

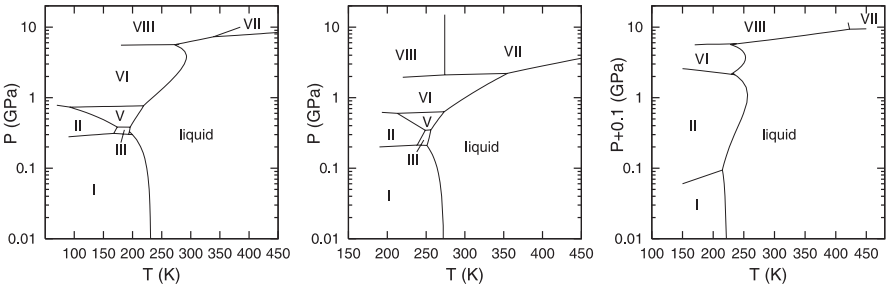

Fig. 5. Phase diagram of water. The *left panel* shows the phase diagram predicted by the TIP4P potential; the *left panel* that obtained from the SPC/E model, and the *central panel* displays the experimental phase diagram. The SPC/E phase diagram has been shifted by 0.1 GPa in order to make phase I visible, as this phase appears at small negative pressures, according to this model. Figure reprinted with permission from *Sanz* et al. [54]. Copyright (2004) by the American Physical Society

3.3 Phase Diagrams

As discussed above, the efficient free-energy techniques exposed in Sect. 2, combined with Kofke's Clausius–Clapeyron integration method, can be employed to calculate phase diagrams of materials entirely from simulation. Two particularly interesting examples of applications of these techniques are the determination of the phase diagram of water by *Sanz* et al. [54], and the phase diagram of carbon by *Ghiringhelli* et al. [55].

The phase diagram of water is very rich, with at least nine stable ice phases documented. Many different empirical potentials aiming at describing this complicated system have been proposed in the literature, and *Sanz* and coworkers [54] have compared two of them, namely the so-called TIP4P model of *Jorgensen* et al. [56], and the SPC/E model of *Berendsen* et al. [57], in their ability to reproduce the experimental phase diagram of water. Sanz et al. used standard thermodynamic integration to locate coexistence points between the different phases, and then obtained the corresponding coexistence lines by using Clausius–Clapeyron integration as proposed by *Kofke* [17, 18]. The theoretical phase diagrams derived from TIP4P and SPC/E models are compared with the experimental water phase diagram in Fig. 5.

As can be seen in Fig. 5, the phase diagram predicted by the TIP4P model agrees fairly well with the experimental one, while that resulting from the SPC/E model deviates noticeably. The TIP4P model correctly predicts ice phases I, II, III, V, VI, VII and VIII to be stable, as indeed they are found to be in real water, while the SPC/E model, on the other hand, predicts wrongly that phases III and V are metastable (i.e., there is no temperature–pressure domain in which these phases are found to have the lowest Gibbs free-energy).

A second example, more within the theme of this book, is provided by the phase diagram of carbon obtained by *Ghiringhelli* et al. [55]. If mod-

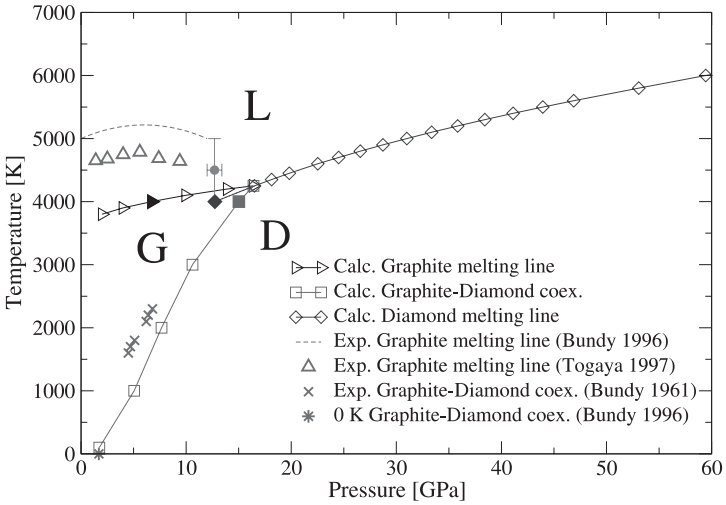

Fig. 6. Phase diagram of carbon as obtained by *Ghiringhelli* and coworkers [55]. This figure shows the phase diagram only up to pressures of 60 GPa, but the *diamond melting line* was followed up to pressures of 400 GPa. The *solid black triangle, blue diamond* and *red square* are the three coexistence points from which later the phase boundaries were obtained by Clausius–Clapeyron integration. The *pink solid circle* is an experimental estimate of the liquid–graphite–diamond triple point. Figure reprinted with kind perission from *Ghiringhelli* et al. [55]. Copyright (2005) by the American Physical Society

eling water accurately is a significant challenge for empirical potentials, as the work of *Sanz* et al. [54] so nicely illustrates, the situation with carbon is by no means easier. The three phases considered in the study of Ghiringhelli and coworkers, namely graphite, diamond and the liquid phase, have markedly different structures and types of bonding, and devising an empirical potential that is capable of reproducing not only these structural and bonding differences, but also the thermodynamic behavior of each phase, is a significant achievement. In this study, the recently developed model of *Los* and *Fasolino* [58] was used. Ghiringhelli et al. proceeded, as in the previously discussed example of water, by first locating coexistence points between the different phases (graphite–diamond, graphite–liquid, and diamond–liquid), and then using Clausius–Clapeyron integration to determine the corresponding phase diagram. A portion of the phase diagram they obtained is shown in Fig. 6, where also some experimentally measured coexistence points are shown for comparison.[1]

The phase diagram of carbon is a challenge not only for simulation, but even more so for experiments. Indeed the melting temperature of graphite

[1] Shortly after this Chapter was sent to the printer, first principles calculations of the phase diagram of carbon were published by *Wang*, *Scandolo* and *Car*, [59].

occurs well above 4000 K. The melting line of diamond has not to date been experimentally characterized. The situation is a little better for the graphite–diamond coexistence line, for which some measurements exist. As can be seen from Fig. 6, the carbon model of *Los* and *Fasolino* [58] predicts a phase diagram that agrees well with the available experimental data. In particular, the slope of the graphite–diamond coexistence curve seems to be well reproduced, although the model predicts this transition to occur at slightly higher pressures than observed experimentally. The graphite melting line occurs at somewhat lower temperatures (ca. 4000 K) than the experimental values, and is predicted to be monotonic with a slight positive pressure derivative, while it appears that the experimental curves show a maximum at around 6 GPa. It should be noticed, however, that the experimental values do not correspond to real measurements, and are in fact indirect estimations based on the ambient-pressure value [55].

A final example that we wish to discuss is that of the phase diagram of Si obtained recently by *Kaczkarski* and coworkers [16]. These authors employed the tight-binding model of *Lenosky* et al. [14], as it is known that no empirical potential currently available predicts correctly the structure of the liquid phase in this system. The range of temperatures and pressures considered by Kaczmarski et al. encompassed four phases, namely the diamond phase, which is the stable one at ambient conditions, the liquid phase, the β-Sn structure, into which the diamond phase transforms when subject to a sufficient amount of compressive pressure, and the Si_{36} clathrate phase, which becomes stable at negative (expansive) pressures. As in the case of C, not much experimental data exists concerning the equilibrium conditions of these phases of Si. It is known, however, that the zero-pressure melting point of the diamond phase occurs at 1687 K, and that its melting line has a negative slope, which results from the fact that the liquid phase is denser than the solid, while having a higher enthalpy. In fact, some potential models and even some tight-binding ones, fail to reproduce this behavior correctly, and would consequently produce a qualitatively wrong melting behavior, even if some of them predict the zero-pressure melting point rather accurately. In a previous study, *Kaczmarski* et al. [15] compared different tight-binding models in their ability not only to accurately reproduce the melting temperature of the Si diamond phase, but also the structural properties of the liquid. The choice of the model due to Lenosky et al. for the study of the phase diagram was based on the results of that comparison.

One difference between the previous examples of water and C and the work of *Kaczmarski* et al. [16] is that in the latter the dynamical version of the Clausius–Clapeyron integration method due to de Koning et al. was used. Also, coexistence points were located from free-energies obtained from the reversible scaling and adiabatic switching techniques. The combined use of these techniques considerably reduced the computational costs compared to what would have been required if standard thermodynamic and Clausius–Clapeyron integration had been used instead.

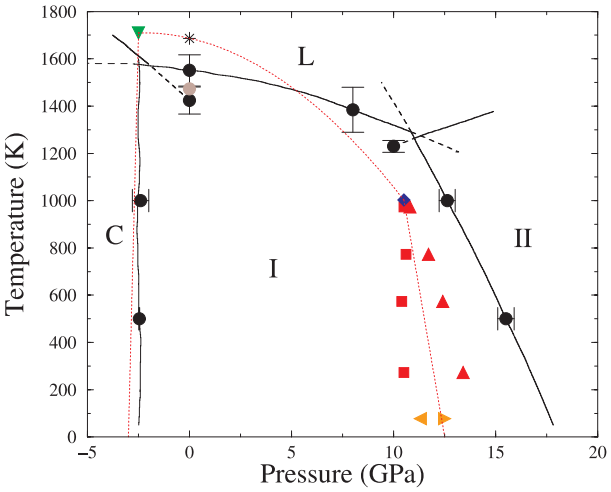

Fig. 7. Silicon phase diagram as predicted by the Lenosky model. The *continuous* and *dashed black curves* are calculated coexistence lines. *Dashed curves* indicate phase boundaries in regions where the separated phases are metastable, while *continuous black curves* separate thermodynamically stable phases; uncertainty bounds estimated at specific points of the phase diagram (marked by *filled circles*) are provided by the error bars. For comparison purposes, a schematic phase diagram summarizing the experimental data is shown with *red dotted lines*, and experimental data at specific temperatures and pressures is shown in the form of *colored symbols*. The *asterisk* corresponds to the zero-pressure melting point of phase **I** (*diamond*), 1687 K [36]; the *brown circle* is the zero-pressure melting point of the (metastable) clathrate phase (**C**), at 1473 K [60]; the *blue diamond* is the diamond–liquid–β-Sn triple point, with estimated coordinates of 1003 ± 20 K and 10.5 ± 0.2 GPa [61]; *red squares* and *triangles* indicate the pressures at which the β-Sn phase (**II**) was first observed and where the diamond phase ceased to be detected, respectively, in the experiments of *Voronin* et al. [61]; *left* and *right pointing orange triangles* give the same information as obtained by *Hu* et al. [62]; finally, the *downward-pointing green triangle* is the estimated diamond–clathrate–liquid triple point, at 1710 K and -2.5 GPa [60]. Figure reprinted from [16]. Copyright (2005) by the American Physical Society

Figure 7 shows the computed phase diagram for Si, compared with available experimental data. It is seen that the overall features of the phase diagram are qualitatively well reproduced by the model. In particular, the negative slope of the melting line of the diamond phase is correct, even if the actual absolute slope value is probably underestimated. The diamond–β-Sn coexistence line is also predicted to have a negative slope, though again the value seems to be smaller than the experimental one. The coexistence line between the diamond and clathrate phases is in good agreement with the only available experimental data on this transition [60].

4 Conclusions and Outlook

In this Chapter we have given an overview of recent theoretical developments in the methodology of free-energy calculations, and we have also provided a selection of applications, most of them in the field of semiconductors, which illustrate current capabilities and provide a glimpse of what the future may bring in this topic. Many outstanding problems still remain, however, and applications in fields such as metallurgy and mineral science, particularly when variable chemical composition is involved, are still extremely challenging. We feel sure, nevertheless, that future theoretical and methodological developments will render such problems more accessible than they are at present.

Acknowledgements

E.R.H. and P.O. thank the Spanish Ministry of Education and Science (MEC) for funding their work through project BFM2003-03372-C03 and the bilateral Italian–Spanish collaborative action HI2003-0337; A. A. acknowledges the support from the Brazilian funding agencies FAPESP, CNPq and FAEP-UNICAMP; L. C. acknowledges financial support under project MIUR-FIRB2001 (contract RBAU01LLX2).

References

[1] D. Frenkel, B. Smit: *Understanding Computer Simulation* (Academic, New York 1996) 116, 126
[2] M. P. Allen, D. J. Tildesley: (Clarendon, Oxford 1987) 116
[3] J. M. Rickman, R. LeSar: Annu. Rev. Mater. Res. **32**, 195 (2002) 116
[4] G. J. Ackland: J. Phys.: Condens. Matter **14**, 2975 (2002) 116
[5] M. de Koning, W. P. Reinhardt: in S. Yip (Ed.): *Handbook of Materials Modeling* (Springer, Berlin, Heidelberg 2005) 116, 120
[6] J. G. Kirkwood: J. Chem. Phys. **3**, 300 (1935) 118
[7] D. Frenkel, A. J. C. Ladd: J. Chem. Phys. **81**, 3188 (1984) 118
[8] J. K. Johnson, J. A. Zollweg, K. E. Gubbins: Mol. Phys. **78**, 591 (1993) 119
[9] D. A. Young, F. J. Rogers: J. Chem. Phys. **81**, 2789 (1984) 119
[10] M. Watanabe, W. P. Reinhardt: Phys. Rev. Lett. **65**, 3301 (1990) 119, 120, 130
[11] M. de Koning, A. Antonelli, S. Yip: Phys. Rev. Lett. **83**, 3973 (1999) 120
[12] M. de Koning, A. Antonelli, S. Yip: J. Chem. Phys. **115**, 11025 (2001) 122, 124
[13] C. M. Goringe, D. R. Bowler, E. R. Hernández: Rep. Prog. Phys. **60**, 1447 (1997) 122
[14] T. J. Lenosky, J. D. Kress, I. Kwon, A. F. Voter, B. Edwards, D. F. Richards, S. Yang, J. B. Adams: Phys. Rev. B **55**, 1528 (1997) 122, 123, 134
[15] M. Kaczmarski, R. Rurali, E. R. Hernández: Phys. Rev. B **69**, 214105 (2004) 122, 123, 130, 134

[16] M. Kaczmarski, O. N. Bedoya-Martínez, E. R. Hernández: Phys. Rev. Lett. **94**, 095701 (2005) 123, 134, 135
[17] D. A. Kofke: Mol. Phys. **78**, 1331 (1993) 123, 132
[18] D. A. Kofke: J. Chem. Phys. **98**, 4149 (1993) 123, 124, 132
[19] R. Agrawal, D. A. Kofke: Mol. Phys. **85**, 43 (1995) 125
[20] A. Jääskeläinen, L. Colombo, R. Nieminen: Phys. Rev. B **64**, 233203 (2001) 125, 126, 127
[21] I. Kwon, R. Biswas, C. Z. Wang, K. M. Ho, C. M. Soukoulis: Phys. Rev. B **49**, 7242 (1994) 126
[22] R. Car, P. Blochl, E. Smargiassi: Mater. Sci. Forum **83–87**, 433 (1992) 126
[23] M. Tang, L. Colombo, J. Zhu, T. Diaz de la Rubia: Phys. Rev. B **55**, 14279 (1997) 126
[24] H. Bracht, E. E. Haller, R. Clark-Phelps: Phys. Rev. Lett. **81**, 393 (1998) 126, 127
[25] T. Y. Tan, U. Gösele: Appl. Phys. A: Solids Surf. **37**, 1 (1985) 126, 127
[26] S. K. Estreicher, M. Garaibeh, P. A. Fedders, P. Ordejón: Phys. Rev. Lett. **86**, 1247 (2001) 127
[27] M. Cogoni, B. P. Uberuaga, A. F. Voter, L. Colombo: Phys. Rev. B **71**, 121203 (2005) 127
[28] J. Samuels, S. G. Roberts: Proc. Roy. Soc. London, Ser. A **421**, 1 (1989) 128
[29] J. R. Rice: J. Mech. Phys. Solids **40**, 239 (1992) 128
[30] J. R. Rice, G. E. Beltz: J. Mech. Phys. Solids **42**, 333 (1994) 128
[31] Y. Sun, G. E. Beltz: Mater. Sci. Eng. A **170**, 67 (1993) 128
[32] E. Kaxiras, M. S. Duesbery: Phys. Rev. Lett. **70**, 3752 (1993) 128
[33] M. de Koning, A. Antonelli, M. Z. Bazant, E. Kaxiras, J. F. Justo: Phys. Rev. B **58**, 12555 (1998) 128, 129
[34] J. Q. Broughton, X. P. Li: Phys. Rev. B **35**, 9120 (1987) 128, 130
[35] F. H. Stillinger, T. A. Weber: Phys. Rev. B **31**, 5262 (1985) 128
[36] R. Hull (Ed.): *Properties of Crystalline Silicon* (Inspec, London 1999) 128, 135
[37] S. J. Cook, P. Clancy: Phys. Rev. B **47**, 7686 (1993) 129
[38] J. Tersoff: Phys. Rev. Lett. **56**, 632 (1986) 129
[39] J. Tersoff: Phys. Rev. B **37**, 6991 (1988) 129
[40] J. Tersoff: Phys. Rev. B **38**, 9902 (1988) 129
[41] M. I. Baskes, J. S. Nelson, A. F. Wright: Phys. Rev. B **40**, 6085 (1989) 129
[42] O. Sugino, R. Car: Phys. Rev. Lett. **74**, 1823 (1995) 130
[43] R. Car, M. Parrinello: Phys. Rev. Lett. **55**, 2471 (1985) 130
[44] D. Alfè, M. J. Gillan: Phys. Rev. B **68**, 205212 (2003) 130
[45] G. A. de Wijs, G. Kresse, M. J. Gillan: Phys. Rev. B **57**, 8223 (1998) 130
[46] C. R. Miranda, A. Antonelli: J. Chem. Phys. **120**, 11672 (2004) 131
[47] E. P. Donovan, F. Spaepen, D. Turnbull, J. M. Poate, D. C. Jacobson: J. Appl. Phys. **57**, 1795 (1985) 131
[48] S. K. Deb, M. Wilding, M. Somayazulu, P. F. McMillan: Nature **414**, 528 (2001) 131
[49] S. Sastry, C. A. Angell: Nature Mater. **2**, 739 (2003) 131
[50] J. F. Justo, M. Z. Bazant, E. Kaxiras, V. V. Bulatov, S. Yip: Phys. Rev. B **58**, 2539 (1998) 131
[51] P. Keblinski, M. Z. Bazant, R. K. Dash, M. M. Treacy: Phys. Rev. B **66**, 064104 (2002) 131

[52] A. Hedler, S. L. Klaumunze, W. Wesch: Nature Mater. **3**, 804 (2004) 131
[53] R. L. C. Vink, G. T. Barkema: Phys. Rev. Lett. **89**, 076405 (2002) 131
[54] E. Sanz, C. Vega, J. L. F. Abascal, L. G. MacDowell: Phys. Rev. Lett. **92**, 255701 (2004) 132, 133
[55] L. M. Ghiringhelli, J. H. Los, E. J. Meijer, A. Fasolino, D. Frenkel: Phys. Rev. Lett. **94**, 145701 (2004) 132, 133, 134
[56] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein: J. Chem. Phys. **79**, 926 (1983) 132
[57] H. J. C. Berendsen, J. R. Grigera, T. P. Straatma: J. Phys. Chem. **91**, 6269 (1987) 132
[58] J. H. Los, A. Fasolino: Phys. Rev. B **68**, 024107 (2003) 133, 134
[59] X. Wang, S. Scandolo, R. Car: Phys. Rev. Lett. **95**, 185701 (2005) 133
[60] P. F. McMillan: Nature Mater. **1**, 19 (2002) 135
[61] G. A. Voronin, C. Pantea, T. W. Zerda, L. Wang, Y. Zhao: Phys. Rev. B **68**, 020102 (2003) 135
[62] J. Z. Hu, L. D. Merkle, C. S. Menoni, I. L. Spain: Phys. Rev. B **34**, 4679 (1986) 135

Index

Quantum Monte Carlo Techniques and Defects in Semiconductors

R. J. Needs

Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom
rn11@cam.ac.uk

Abstract. The continuum variational and diffusion quantum Monte Carlo (VMC and DMC) techniques are stochastic methods for obtaining expectation values of many-body wavefunctions. These methods are capable of achieving very high accuracy, yet the scaling with system size is sufficiently favorable to allow applications to condensed-matter systems. Defects in semiconductors pose a variety of challenges to computational electronic-structure methods. As well as the large system sizes required, one is faced with calculating energy differences between rather disparate interatomic bonding configurations, and calculating excited state energies, sometimes including the energies of multiplets. Applications of VMC and DMC to defects in semiconductors are still in their infancy but, as I will describe, these methods possess features that make them well suited for addressing some important issues in the field. In this chapter I first discuss why we expect VMC and DMC calculations to be valuable in studying defects in semiconductors. I then review the techniques themselves, concentrating on the issues most pertinent to calculations for defects in solids. I describe recent applications to self-interstitial defects in crystalline silicon, the low-energy electronic states of the neutral vacancy in diamond, and the Schottky defect energy in magnesium oxide. The chapter ends with a perspective on the future of such studies.

1 Introduction

Quantum Monte Carlo (QMC) methods in the variational (VMC) and diffusion (DMC) forms are stochastic methods for evaluating expectation values of many-body wavefunctions. (For a review of the VMC and DMC methods and applications to solids, see [1].) One of the main attractions of these methods is that the computational cost scales roughly as the square or cube of the number of particles, which is very favorable compared with quantum-chemical many-body techniques, allowing applications to condensed-matter. Applications of QMC to defects in semiconductors are still in their infancy but, as I will describe, these methods possess features that make them well suited for addressing important issues in the field. In this Chapter I first discuss why we expect VMC and DMC calculations to be valuable in studying defects in semiconductors. I then review the techniques themselves, concentrating on the issues most pertinent to defects in solids. Finally, I describe recent applications to point defects in crystalline silicon, diamond and magnesium oxide, ending with a perspective on the future of such studies.

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 141–164 (2007)

It is important to treat electron-correlation effects accurately when describing the energetics of condensed-matter systems. Defects in semiconductors are not special in this regard, but they exemplify almost all the difficulties that may be encountered. For example, we would like to calculate the formation and migration energies of defects. These quantities are related to the energy differences between structures, which may involve atoms with different coordination numbers and bond lengths. We might also be interested, for example, in the properties of a $3d$ transition metal impurity in a semiconductor, whose electronic properties are difficult to describe accurately. We may also be interested in the excited electronic states of defects, which may involve calculating the energies of multiplets. These are severe tests of electronic-structure methods, which can only be met by techniques that incorporate a sophisticated description of electronic correlation.

QMC methods have the potential to provide very accurate energies and, in principle, there are no restrictions on the types of electronic state that can be treated. The $N^2 - N^3$ scaling of the computational cost with the number of electrons N allows applications to quite large systems, which are certainly sufficient for studies of defects in semiconductors. However, the prefactor is large, and a QMC calculation is generally much more expensive than the corresponding density-functional theory (DFT) one.

The DMC calculations of *Ceperley* and *Alder* [2] provided accurate energies for the homogeneous electron gas, which have been used in constructing parameterized density functionals. Since then, the DMC method has been used in studies of the electrons in molecules, clusters and solids. The power of the DMC method for describing electronic correlation in large systems has been amply demonstrated by recent applications, including studies of point defects in semiconductors [3, 4], the reconstruction of the silicon (001) surface [5] and its interaction with H_2 [6], and the calculation of optical excitation energies [7, 8]. The methodology of QMC calculations is being improved rapidly that, together with advances in computing technologies, will allow them to be applied to more complicated systems. I believe that QMC will play an important role in the future of electronic-structure computations. They will not replace DFT studies, but rather complement them by providing higher accuracy, albeit at considerably higher cost.

2 Quantum Monte Carlo Methods

The VMC method is conceptually very simple, the energy being calculated as the expectation value of the Hamiltonian with an approximate many-body trial wavefunction. In the more sophisticated DMC method the estimate of the ground-state energy is improved by performing an evolution of the wavefunction in imaginary-time.

2.1 The VMC Method

The variational theorem of quantum mechanics states that, for a proper trial wavefunction Ψ_{T}, the variational energy,

$$E_{\mathrm{V}} = \frac{\int \Psi_{\mathrm{T}}^{*}(\boldsymbol{R})\hat{H}\Psi_{\mathrm{T}}(\boldsymbol{R})\,\mathrm{d}\boldsymbol{R}}{\int \Psi_{\mathrm{T}}^{*}(\boldsymbol{R})\Psi_{\mathrm{T}}(\boldsymbol{R})\,\mathrm{d}\boldsymbol{R}}\,, \tag{1}$$

is an upper bound on the exact ground-state energy E_0, i.e., $E_{\mathrm{V}} \geq E_0$. In (1), Ψ_{T} is the many-body wavefunction, the symbol $\boldsymbol{R}$ denotes the $3N$-dimensional vector of the electron coordinates, and $\hat{H}$ is the many-body Hamiltonian.

The VMC method was first used by *McMillan* in a study of liquid ^{4}He [9]. To facilitate the stochastic evaluation, E_{V} is written as

$$E_{\mathrm{V}} = \int p(\boldsymbol{R})E_{\mathrm{L}}(\boldsymbol{R})\,\mathrm{d}\boldsymbol{R}\,, \tag{2}$$

where the probability distribution p is

$$p(\boldsymbol{R}) = \frac{|\Psi_{\mathrm{T}}(\boldsymbol{R})|^2}{\int |\Psi_{\mathrm{T}}(\boldsymbol{R})|^2\,\mathrm{d}\boldsymbol{R}}\,, \tag{3}$$

and the *local energy*, E_{L}, is

$$E_{\mathrm{L}}(\boldsymbol{R}) = \Psi_{\mathrm{T}}^{-1}\hat{H}\Psi_{\mathrm{T}}\,. \tag{4}$$

In VMC one uses a Metropolis algorithm [10] to sample the distribution p, generating M *configurations*, $\boldsymbol{R}_i$, drawn from p. The variational energy is then estimated as

$$E_{\mathrm{V}} \simeq \frac{1}{M}\sum_{i=1}^{M} E_{\mathrm{L}}(\boldsymbol{R}_i)\,. \tag{5}$$

Equation (2) is an *importance sampling* transformation of (1). Equation (2) exhibits the *zero-variance property*; as the trial wavefunction approaches an exact eigenfunction ($\Psi_{\mathrm{T}} \to \phi_i$), the local energy approaches the corresponding eigenenergy, E_i, everywhere in configuration space. As Ψ_{T} is improved, E_{L} becomes a smoother function of $\boldsymbol{R}$ and the number of sampling points, M, required to achieve an accurate estimate of E_{V} is reduced.

VMC is a simple and elegant method. There are no restrictions on the form of trial wavefunction that can be used and it does not suffer from a fermion sign problem. The main problem is, essentially, that you get out what you put in. Even if the underlying physics is well understood it is very difficult to prepare trial wavefunctions of equivalent accuracy for different systems, and therefore the energy differences between them will be biased. We use the VMC method mainly to optimize parameters in trial wavefunctions, see Sect. 2.4, and our main calculations are performed with the more sophisticated DMC method, which is described in the next section.

2.2 The DMC Method

The DMC procedure largely overcomes the problem of bias inherent in VMC. In DMC the operator $\exp(-t\hat{H})$ is used to project out the ground-state from the initial state. This can be viewed as solving the imaginary-time Schrödinger equation,

$$-\frac{\partial}{\partial t}\Phi(\boldsymbol{R},t) = \left(\hat{H} - E_{\mathrm{T}}\right)\Phi(\boldsymbol{R},t)\,, \tag{6}$$

where t is a real variable measuring the progress in imaginary-time and E_{T} is an energy offset. The solution of (6) can be obtained by expanding $\Phi(\boldsymbol{R},t)$ in the eigenstates of the Hamiltonian,

$$\Phi(\boldsymbol{R},t) = \sum_i c_i(t)\phi_i(\boldsymbol{R})\,, \tag{7}$$

which leads to

$$\Phi(\boldsymbol{R},t) = \sum_i \exp[-(E_i - E_{\mathrm{T}})t]c_i(0)\phi_i(\boldsymbol{R})\,. \tag{8}$$

For long times one finds

$$\Phi(\boldsymbol{R},t\to\infty) \simeq \exp[-(E_0 - E_{\mathrm{T}})t]\, c_0(0)\phi_0(\boldsymbol{R})\,, \tag{9}$$

which is proportional to the ground-state wavefunction, ϕ_0. If we interpret the initial state, $\sum_i c_i(0)\phi_i$, as a probability distribution, then the time-evolution process can be thought of as taking a set of points in configuration space distributed according to the initial state and subjecting them to diffusion and branching (branching is also referred to as "birth" and "death" of configurations). The diffusion process arises from the kinetic energy operator and the branching from the potential energy. So far we have assumed that the wavefunction can be interpreted as a probability distribution, but a wavefunction for more than two identical fermions must have positive and negative regions. One can construct algorithms that use two distributions of configurations, one for positive regions and one for negative, but such algorithms suffer from a fermion sign problem that manifests itself by the signal-to-noise ratio decaying exponentially with imaginary-time.

The *fixed-node approximation* [11, 12] allows us to avoid these problems. This method is equivalent to placing infinite potential barriers everywhere on the nodal surface of a trial wavefunction Ψ_{T}. (The nodal surface is the $3N-1$ dimensional surface in configuration space on which the wavefunction is zero and across which it changes sign.) The infinite potential barriers have no effect if the trial nodal surface is correct, but if it is incorrect the energy is always raised. The DMC projection can then be done separately in each nodal pocket, which gives the lowest-energy solution in each pocket. It therefore follows that the DMC energy is always less than or equal to the VMC energy

with the same trial wavefunction, and always greater than or equal to the exact ground-state energy, i.e., $E_{\mathrm{V}} \geq E_{\mathrm{D}} \geq E_0$. In addition, it turns out that all the nodal pockets of the exact fermion ground-state are equivalent [13] and, if Ψ_{T} has this *tiling property*, we do not need to sample all of the pockets.

Such a fixed-node DMC algorithm is extremely inefficient and a vastly superior algorithm can be obtained by introducing an importance sampling transformation [14, 15]. Consider the *mixed distribution*,

$$f(\boldsymbol{R}, t) = \Psi_{\mathrm{T}}(\boldsymbol{R})\Phi(\boldsymbol{R}, t)\,, \tag{10}$$

which has the same sign everywhere if the nodal surface of $\Phi(\boldsymbol{R}, t)$ is constrained to be the same as that of $\Psi_{\mathrm{T}}(\boldsymbol{R})$. Substituting in (6) for Φ we obtain

$$-\frac{\partial f}{\partial t} = -\frac{1}{2}\nabla^2 f + \nabla \cdot [\boldsymbol{v} f] + [E_{\mathrm{L}} - E_{\mathrm{T}}] f\,, \tag{11}$$

where the $3N$-dimensional drift velocity is defined as

$$\boldsymbol{v}(\boldsymbol{R}) = \Psi_{\mathrm{T}}^{-1}(\boldsymbol{R})\nabla\Psi_{\mathrm{T}}(\boldsymbol{R})\,. \tag{12}$$

The three terms on the right-hand side of (11) correspond to diffusion, drift, and branching processes, respectively. The importance sampling transformation has several consequences. First, the density of configurations is increased where Ψ_{T} is large. Second, the rate of branching is now controlled by the local energy that is generally a much smoother function than the potential energy. Finally, the importance sampling actually enforces the fixed-node constraint in (11). In practice one has to use a *short-time approximation* (for a review of the VMC and DMC methods and applications to solids see [1]) for the Green's function diffusion equation that means that occasionally configurations attempt to cross the nodal surface, but such moves are simply rejected.

This algorithm generates the distribution f, but fortunately we can calculate the corresponding DMC energy using the formula

$$E_{\mathrm{D}} = \frac{\int f E_{\mathrm{L}}\, \mathrm{d}\boldsymbol{R}}{\int f\, \mathrm{d}\boldsymbol{R}} \tag{13}$$

$$\simeq \frac{1}{M}\sum_{i=1}^{M} E_{\mathrm{L}}(\boldsymbol{R}_i)\,. \tag{14}$$

The importance-sampled fixed-node fermion DMC algorithm was first used by *Ceperley* and *Alder* [2].

It is not possible to obtain directly from f the proper expectation values of operators that do not commute with the Hamiltonian, such as the charge-density or pair-correlation functions. There are, however, three methods by which such expectation values can be obtained: the approximate (but often very accurate) extrapolation technique [16], the forward-walking technique (explained in *Hammond* et al. [17]) that is formally exact but statistically poorly behaved, and the *reptation* QMC technique of *Baroni* and *Moroni* [18], which is formally exact and well behaved, but quite expensive.

2.3 Trial Wavefunctions

It should be clear that the trial wavefunction is of central importance in VMC and DMC calculations. The trial wavefunction introduces importance sampling and controls both the statistical efficiency and the accuracy obtained. In DMC the accuracy depends only on the nodal surface of the trial wavefunction via the fixed-node approximation, while in VMC it depends on the entire trial wavefunction, so that VMC energies are more sensitive to the quality of the approximate wavefunction than DMC energies.

QMC calculations require a compact trial wavefunction, and most studies of electronic systems have used the Slater–Jastrow form, in which a pair of up- and down-spin determinants is multiplied by a Jastrow correlation factor,

$$\Psi_{\mathrm{T}} = e^{\mathrm{J}(\boldsymbol{R})} D_{\uparrow}(\boldsymbol{R}_{\uparrow}) D_{\downarrow}(\boldsymbol{R}_{\downarrow}) \,, \tag{15}$$

where e^{J} is the Jastrow factor and the D are determinants of single-particle orbitals. The quality of the single-particle orbitals is very important, and they are often obtained from DFT or Hartree–Fock (HF) calculations.

The Jastrow factor is a positive, symmetric, function of the electron coordinates, which is chosen to be small when electrons are close to one another, thereby introducing electron-correlation into the wavefunction. The Jastrow factor is also used to enforce the electron–electron Kato *cusp conditions* [19]. The idea is that when two electrons are coincident the contribution to the local energy from the Coulomb interaction diverges, and therefore the exact wavefunction must have a cusp at this point so that the kinetic energy operator supplies an equal and opposite divergence. It seems very reasonable to enforce these conditions on a trial wavefunction as they are obeyed by the exact wavefunction, and in VMC, and more particularly DMC, calculations they are very important because divergences in the local energy can lead to poor statistical behavior.

Many excellent QMC results have been obtained using Slater–Jastrow wavefunctions, but it may be profitable to consider other forms. One variant is to use more than one pair of determinants,

$$\Psi_{\mathrm{T}} = e^{\mathrm{J}(\boldsymbol{R})} \sum_i \alpha_i D_{i,\uparrow}(\boldsymbol{R}_{\uparrow}) D_{i,\downarrow}(\boldsymbol{R}_{\downarrow}) \,, \tag{16}$$

where the α_i are coefficients. This approach is not suitable for improving the energy per atom of a solid because the number of determinants required would have to increase exponentially with system size. It is, however, useful when considering a point defect with an open-shell electronic configuration in a solid, in which case the electronic states of the multiplet may be described by a small number of determinants. An example of this is described in Sect. 6.3. Another possibility is to include *backflow correlations,* which may be derived by demanding that the wavefunction conserves the local current [20]. This approach has been used successfully in QMC studies of the

electron gas [21], but it has not been explored much for electrons in real atomic systems [22]. Including backflow correlations improves the nodal surface of a wavefunction and leads to a compact form that is suitable for use in QMC calculations, although it is more complicated than the Slater–Jastrow form and the computational cost is increased.

2.4 Optimization of Trial Wavefunctions

Optimizing trial wavefunctions is perhaps the most important technical issue in QMC calculations, and it consumes large amounts of human and computing resources. It is standard to include a number of variable parameters in the Jastrow factor whose values are determined by a stochastic optimization procedure. The determinant coefficients, α_i, in (16) may also be optimized, and in some calculations the orbitals themselves have been optimized.

Developing better wavefunction-optimization techniques is currently a very active field, with many new approaches being tested. Rather than delving into the technical details of this enterprise, I will mention some of the key issues. Ideally the wavefunction should be optimized within DMC, but this is generally much too costly and so optimizations are performed at the VMC level. Two measures of wavefunction quality have generally been considered. The most obvious idea is to minimize the expectation value of the VMC energy, although more commonly the variance of the VMC energy has been minimized. There are good reasons to suppose that better DMC results would be obtained by using energy-minimized trial wavefunctions, rather than variance-minimized ones. First, it seems likely that a better nodal surface would be obtained by minimizing the energy rather than the variance. Second, it turns out that the variance of the DMC energy is approximately proportional to the difference between the VMC and DMC energies, rather than the variance of the VMC energy [23]. Why then has it been more common to minimize the variance of the VMC energy? The answer is simply that it has proved easier to devise algorithms for minimizing the variance that are stable and efficient in the presence of statistical noise [24–26]. Most of the current effort is being devoted to developing energy-minimization schemes, although the applications to defects in semiconductors described in this Chapter were performed using variance minimization.

2.5 QMC Calculations within Periodic Boundary Conditions

QMC calculations for defects in solids may be performed using cluster models or periodic boundary conditions, just as in other techniques. Periodic boundary conditions are preferred because they give smaller finite-size effects. In the standard *supercell* approach a point defect in an otherwise perfect crystal is modeled by taking a large cell containing a single defect and repeating it throughout space. Just as in other techniques, one must ensure that the

supercell is large enough for the interactions between defects in different supercells to be small.

In techniques involving many-electron wavefunctions such as QMC it is not possible to reduce the problem to solving within a primitive unit cell (when a defect is present the primitive unit cell is the supercell). Such a reduction is allowed in single-particle methods because the Hamiltonian is invariant under the translation of a single electronic coordinate by a translation vector of the primitive lattice, but this is not a symmetry of the many-body Hamiltonian. Consequently QMC calculations may be performed only at a single k-point. It is possible to perform QMC calculations at different k-points [27, 28] and average the results [29], but this is not exactly equivalent to a Brillouin-zone integration within single-particle theories. In the applications described in this Chapter a single k-point was used.

Many-body techniques such as QMC also suffer from another source of finite-size effects – the interaction of the electrons with their periodic images. Such an effect is absent in local DFT calculations because the interaction energy is written in terms of the electronic charge density, but the effect is present in HF calculations. These "Coulomb finite-size effects" are often corrected for using extrapolation techniques (see [1] and references therein), and they may also be reduced using the so-called model periodic Coulomb (MPC) interaction [30].

2.6 Using Pseudopotentials in QMC Calculations

The cost of a DMC calculation increases rapidly with the atomic number, Z, as roughly $Z^{5.5}$ [23, 31]. This effectively rules out applications to atoms with Z greater than about 10, and consequently pseudopotentials are commonly used in DMC calculations. The use of nonlocal pseudopotentials within VMC is quite straightforward, but DMC poses an additional problem, because the matrix element of the imaginary-time propagator is not necessarily positive, which leads to a sign problem analogous to the fermion sign problem. To circumvent this difficulty an additional approximation known as the pseudopotential *localization approximation* has been used [32]. If the trial wavefunction is accurate the error introduced by the localization approximation is small and is proportional to $(\Psi_{\rm T} - \phi_0)^2$ [33], although the error term can be of either sign. Recently, *Casula* et al. [34] have introduced a fully variational technique for dealing with nonlocal pseudopotentials within DMC.

Currently it is not possible to generate pseudopotentials entirely within a QMC framework, and therefore they are obtained from other sources. There is evidence that indicates that HF theory provides better pseudopotentials than DFT for use within QMC calculations [35], and recently we have developed smooth relativistic HF pseudopotentials for H to Ba and Lu to Hg that are suitable for use in QMC calculations [36, 37].

3 DMC Calculations for Excited States

It might appear that the DMC method cannot be applied to excited states because the imaginary-time propagation would evolve it towards the ground-state. However, the fixed-node constraint ensures convergence to the lowest-energy state compatible with the nodal surface of the trial wavefunction. If the nodal surface is exactly that of an eigenstate then DMC gives the exact energy of that state. If the nodal surface is approximately that of an excited-state then the DMC energy gives an approximation to the energy of that excited state. There is, however, an important difference from the ground-state case, in that the existence of a variational principle for the energy of an excited-state depends on the symmetry of the trial wavefunction [38]. In practice, DMC works quite well for excited-states [39, 40].

4 Sources of Error in DMC Calculations

Here we summarize the potential sources of errors in DMC calculations.

- Statistical error. The standard error in the mean is proportional to $1/\sqrt{M}$, where M is the number of moves. It therefore costs a factor of 100 in computer time to reduce the error bars by a factor of 10. On the other hand, a random error is much better than a systematic one.
- The fixed-node error. This is the central approximation of the DMC technique, and is often the limiting factor in the accuracy of the results.
- Timestep bias. The short-time approximation leads to a bias in the f distribution and hence in expectation values. This bias can be serious and is often corrected for by using several different timesteps and extrapolating to zero timestep.
- Population control bias. In DMC the f distribution is represented by a finite population of configurations that fluctuates due to the branching procedure. The population is usually controlled by occasionally changing the trial energy, E_{T}, in (11) so that the population returns towards a target size. This procedure results in a *population-control bias*. In practice, the population control bias is normally extremely small, in fact so small that it is difficult to detect.
- Pseudopotential errors including the localization error. Pseudopotentials inevitably introduce errors that can be significant in some cases. The localization error also appears to be quite small, although this has not been tested in many cases.
- Finite-size errors within periodic boundary conditions. It is important to study the finite-size effects carefully, and to correct for them. A number of correction procedures have been devised, mainly involving extrapolation to the thermodynamic limit.

In electronic-structure theory one is almost always interested in the difference in energy between two similar systems. All electronic-structure methods for complex systems rely on large cancelations of the errors in the energy differences between systems. In DMC this helps with all the above sources of error except the statistical errors. Fixed-node errors tend to cancel because the DMC energy is an upper bound, but even though DMC often retrieves 95 % or more of the correlation energy, noncancelation of nodal errors is a severe problem for DMC calculations. As mentioned before, the most straightforward way forward is simply to improve the quality of the trial wavefunctions, but perhaps it might also be possible to *balance* the errors in different systems more successfully than is currently possible.

5 Applications of QMC to the Cohesive Energies of Solids

A number of VMC and DMC studies have been performed of the cohesive energies of solids. The cohesive energy is calculated as the energy difference between the isolated atom and an atom in the solid. This is a severe test of QMC methods because the trial wavefunctions used for the atom and solid must be closely matched in quality. Data for Si [30], Ge [28], C [4], Na [41] and NiO [42] have been collected in Table 1. The local spin density approximation (LSDA) data shows the standard overestimation of the cohesive energy, while the QMC data is in good agreement with experiment. These studies have been important in establishing DMC as an accurate method for calculating the energies of solids.

Table 1. LSDA, VMC and DMC cohesive energies of solids in eV per atom (eV per 2 atoms for NiO), compared with the experimental values. The numbers in brackets indicate standard errors in the last significant figure. All calculated values contain a correction for the zero-point energy of the solid

Method	Si	Ge	C	Na	NiO
LSDA	5.28	4.59	8.61	1.20	10.96
VMC	4.48(1)	3.80(2)	7.36(1)	0.9694(8)	8.57(1)
DMC	4.63(2)	3.85(2)	7.346(6)	1.0221(3)	9.442(2)
Exp.	4.62	3.85	7.37	1.13	9.45

6 Applications of QMC to Defects in Semiconductors

6.1 Using Structures from Simpler Methods

In the applications of QMC to defects in semiconductors reported to date, the structures were obtained from DFT calculations and the energies recalculated within QMC. Surely this leads to a bias in the energy differences that might be considerable? The answer must certainly be yes, it does lead to a bias, but here I argue that it may often be quite small. We can pose the question more clearly as "is it sensible to calculate structures using a simple method of limited accuracy and then recalculate the energy difference between the structures with a sophisticated method of much higher accuracy?".

Consider a system with a single structural coordinate Q, and suppose there are two structures of interest, denoted by 1 and 2, corresponding to local minima in the energy at coordinates Q_1 and Q_2, with energies E_1 and E_2. Around the two energy minima the energy can be expanded as

$$\text{Around } Q_1 \qquad E(Q) = E_1 + \alpha_1 (Q - Q_1)^2 \tag{17}$$

$$\text{Around } Q_2 \qquad E(Q) = E_2 + \alpha_2 (Q - Q_2)^2 \,. \tag{18}$$

Suppose further that we have a simple method for calculating the energy $E(Q)$ that gives a smoothly varying error of the form

$$\Delta E(Q) = \beta Q \,, \tag{19}$$

where β is a constant.

Within the simple method the errors in the coordinates of the equilibrium structures are obtained by minimizing $E + \Delta E$, which gives

$$\Delta Q_1 = -\frac{\beta}{2\alpha_1} \,, \quad \Delta Q_2 = -\frac{\beta}{2\alpha_2} \,, \tag{20}$$

and the error in the energy difference between these structures is

$$\Delta(E_1 - E_2) = \beta(Q_1 - Q_2) - \frac{\beta^2}{4} \left(\frac{1}{\alpha_1} - \frac{1}{\alpha_2} \right) . \tag{21}$$

Now suppose that we use the structures from the simple method, but recalculate the energies within a sophisticated method that gives a negligible error in the energy. The error in the energy difference between the structures is then

$$\Delta(E_1 - E_2) = \frac{\beta^2}{2} \left(\frac{1}{\alpha_1} - \frac{1}{\alpha_2} \right) . \tag{22}$$

Within the simple method the error in the energy difference is first order in β, while for the sophisticated method it is second order. The error in

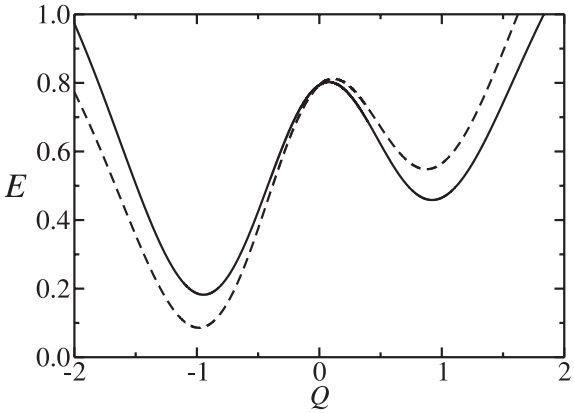

Fig. 1. The energy E versus a structural coordinate Q showing two minima for a model system. *Solid line*: exact energy, *dashed line*: approximate energy from a simple theory that has an error of $0.1Q$. The error in the energy difference between the two minima obtained from the simple theory is 0.186, while the error in the energy difference calculated using the accurate energy curve, but evaluated at the minima obtained from the simple theory, is 0.000 45

the energy difference within the simple method is proportional to $Q_1 - Q_2$, so that the error is large if the structures are very different, while in the sophisticated method the error is independent of $Q_1 - Q_2$. This behavior is illustrated in Fig. 1. We therefore conclude that using the sophisticated method to calculate the energy difference between structures obtained from the simple method leads to only quite small errors. Note also that within this model the quadratic coefficients in the energy, α_1 and α_2, are unaltered by the linear error term, so that the simple method gives accurate values for the vibrational frequencies.

The model is, of course, highly simplified, but it illustrates my contention that it is reasonable to take structures from DFT calculations and recalculate the energy differences between them within QMC.

6.2 Silicon Self-Interstitial Defects

The diffusion of dopant impurity atoms during thermal processing limits how small silicon devices can be made, and understanding this effect requires a knowledge of diffusion on the microscopic scale in situations far from equilibrium. Dopant diffusion is mainly controlled by the presence of intrinsic defects such as self-interstitials and vacancies, and therefore it is important to improve our understanding of these defects.

Both experimental and theoretical techniques have been brought to bear on these defects but, unfortunately, it has not been possible to detect silicon self-interstitials directly and we must rely on theoretical studies to elucidate

their microscopic properties. Electron paramagnetic resonance studies [43] have determined unambiguously that the symmetry of neutral vacancies in silicon is D_{2d}, which is consistent with a Jahn–Teller distortion. Positron-lifetime experiments have given a value for the enthalpy of formation of a neutral vacancy in silicon of $3.6 \pm 0.2\,\mathrm{eV}$ [44].

The self-diffusivity of silicon at high temperatures follows an Arrhenius behavior with an activation energy in the range 4.1 eV to 5.1 eV [45]. Significant difficulties arise with the interpretation of experimental data when the contributions to the self-diffusivity from different mechanisms are considered. The self-diffusivity is usually written as the sum of contributions from independent diffusive mechanisms. The contribution of a particular microscopic mechanism can be written as the product of the diffusivity, D_i, and the concentration, C_i, of the relevant defect, i.e.,

$$D_{\mathrm{SD}} = \sum_i D_i C_i \,. \tag{23}$$

Gösele et al. [46] give estimates of the contributions to the self-diffusivity as

$$D_{\mathrm{I}} C_{\mathrm{I}} = 914 \exp(-4.84/k_{\mathrm{B}}T)\,\mathrm{cm}^2 \cdot \mathrm{s}^{-1}\,, \tag{24}$$

from self-interstitials and

$$D_{\mathrm{V}} C_{\mathrm{V}} = 0.6 \exp(-4.03/k_{\mathrm{B}}T)\,\mathrm{cm}^2 \cdot \mathrm{s}^{-1}\,, \tag{25}$$

from vacancies, where $k_{\mathrm{B}}T$ is in units of eV. It is believed that self-diffusivity in silicon is dominated by vacancies at low temperatures and self-interstitials above 1300 K. The experimental situation regarding the individual values of D_i and C_i is, however, controversial. Indeed, experimental data has been used to support values of the diffusivity of the silicon self-interstitial, D_{I}, which differ by ten orders of magnitude at the temperatures of around 1100 K at which silicon is processed [47].

The main challenge to theory is to identify the important defects and to predict their properties, comparing with experimental data where available. This includes identifying the diffusive mechanisms of the defects, providing accurate values for the D_i and the equilibrium values of C_i, and reproducing the experimental data for the self-diffusivity. This programme represents an enormous challenge, and although many theoretical studies of point defects in silicon have been performed, the goal is still far away.

A number of different structures for silicon self-interstitials have been postulated, including the split-$\langle 110 \rangle$, hexagonal and tetrahedral defects, see Fig. 2. The consensus view from modern DFT calculations is that the split-$\langle 110 \rangle$ and hexagonal structures are the lowest-energy self-interstitials in silicon [48–50], and that the tetrahedral interstitial is unstable. However, even here there has been a surprise. Recent DFT calculations have shown that the hexagonal interstitial is unstable to a distortion out of the hexagonal ring by

0.48 Å, which lowers the formation energy by 0.03 eV [50]. Another surprise has been that, within DFT, it turns out that the lowest-energy point defect in silicon is neither an interstitial nor the vacancy! *Goedecker* et al. [49] found that a fourfold coordinated defect (FFCD) had a formation energy at least 0.5 eV lower than the interstitials and vacancy. It is not clear whether this defect could play a role in self-diffusion, as its possible diffusive mechanisms have not yet been studied.

There are some discrepancies between DFT values reported for the formation energies of the split-⟨110⟩ and hexagonal self-interstitials and for the associated migration energies. However, modern calculations [48–50] indicate that the sum of the DFT formation and migration energies is significantly smaller than the exponent of 4.84 eV deduced from experimental studies, see (24).

An interesting suggestion by *Pandey* and coworkers [51, 52], based on the results of DFT calculations, was that an exchange of neighboring atoms in the perfect lattice atoms would contribute to self-diffusion in silicon. In Pandey's *concerted exchange* mechanism two nearest-neighbor atoms interchange via a complicated 3-dimensional path that allows the atoms to avoid large energy barriers [51, 52]. One of the main planks of Pandey's argument was that the calculated value for the activation energy for this concerted exchange was within the experimental range for self-diffusion of 4.1 eV to 5.1 eV [45]. However, as we have already noted, DFT gives even lower energies for interstitials and vacancies, and therefore Pandey's argument is unsatisfactory.

All of the calculations mentioned so far assumed no thermal motion at all, but it is also important, of course, to study finite-temperature effects. Two DFT molecular dynamics studies have been reported [53, 54], but the energy barriers to interstitial self-diffusion are very small (at least within DFT) and highly accurate calculations are required to obtain reliable results. It is rather easier to calculate the equilibrium concentration of defects including the effects of thermal vibrations, and a recent study [50] using DFT perturbation theory gave results that are broadly in agreement with the rather scattered experimental data.

Given this background, what role can QMC calculations play in unravelling this complicated problem? In the medium term its role must be to give accurate values for the energies of various defect structures. Given that the LSDA and generalized gradient approximation (GGA) self-interstitial defect-formation energies can differ by 0.5 eV it is valuable to quantify the errors in DFT calculations of the defect formation energies. As argued in Sect. 6.1, it should be sufficient to use defect structures obtained from DFT calculations for this purpose.

6.2.1 DFT Calculations on Silicon Self-Interstitials

Leung et al. [3] and *Needs* [48] reported LSDA and PW91-GGA results for various interstitial defect formation energies and the saddle point of Pandey's

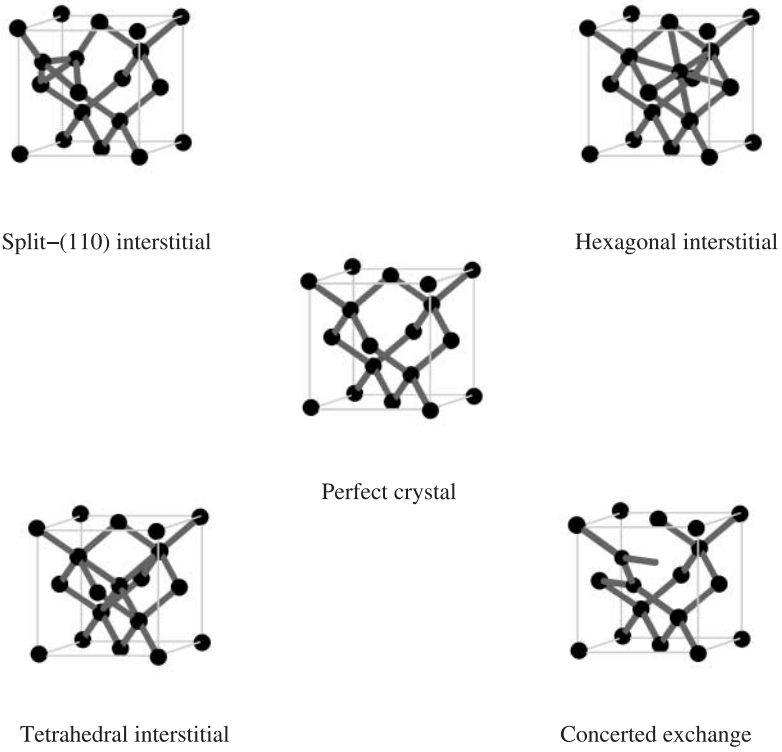

Fig. 2. The structures of the split-⟨110⟩, hexagonal, and tetrahedral interstitial defects, the saddle point of the concerted-exchange mechanism, and the perfect silicon crystal. After *Needs* [48]

concerted-exchange mechanism, see Table 2. This work also indicated that the LSDA and PW91-GGA equilibrium defect structures are very similar. One of the features of the various defect structures is the wide range of interatomic bonding they exhibit. In the split-⟨110⟩ structure the two atoms forming the defect are fourfold coordinated, but two of the surrounding atoms are fivefold coordinated. The hexagonal interstitial is sixfold coordinated and its six neighbors are fivefold coordinated. The tetrahedral interstitial is fourfold coordinated and has four neighbors, which are therefore fivefold coordinated. At the saddle point of Pandey's concerted exchange two bonds are broken so that the exchanging atoms and two other atoms are fivefold coordinated.

The DFT energy barriers to diffusive jumps between the low-energy structures were calculated, which gave a path for split-⟨110⟩-hexagonal diffusion with a barrier of 0.15 eV (LDA) and 0.20 eV (PW91-GGA), and a barrier for hexagonal–hexagonal diffusive jumps of 0.03 eV (LDA) and 0.18 eV (PW91-GGA).

Table 2. LDA, PW91-GGA, and DMC formation energies in eV of the self-interstitial defects and the saddle point of the concerted-exchange mechanism. [a]16-atom supercell, [b]54-atom supercell

Defect	LDA	GGA	DMC[a]	DMC[b]
Split-⟨110⟩	3.31	3.84	4.96(24)	4.96(28)
Hexagonal	3.31	3.80	4.70(24)	4.82(28)
Tetrahedral	3.43	4.07	5.50(24)	5.40(28)
Concerted exchange	4.45	4.80	5.85(23)	5.78(27)

6.2.2 QMC Calculations on Silicon Self-Interstitials

The DMC calculations were performed for 16- and 54-atom *fcc* structures obtained from LDA calculations. The Slater–Jastrow trial wavefunctions were formed from determinants of LDA orbitals and Jastrow factors containing up to 64 parameters, whose optimal values were obtained by minimizing the variance of the energy.

The DMC defect-formation energies are shown in Table 2. Note that the DMC results for the 16- and 54-atom simulation cells are consistent, indicating that the residual finite-size effects are small. The clearest conclusion is that the DMC formation energies are roughly 1 eV larger than the PW91-GGA values and 1.5 eV larger than the LDA values. The hexagonal interstitial has the lowest formation energy within DMC, while the split-⟨110⟩ interstitial is slightly higher in energy, and the tetrahedral interstitial has a considerably higher energy. The saddle point of the concerted exchange has an energy that is too high to explain self-diffusion in silicon.

Calculating the energy barriers to diffusion within DMC is currently prohibitively expensive, but they can be roughly estimated as follows. The tetrahedral interstitial is a saddle point of a possible diffusion path between neighboring hexagonal sites. The DMC formation energy of 5.4 eV for the tetrahedral interstitial is therefore an upper bound to the formation plus migration energy of the hexagonal interstitial. The true formation plus migration energy is expected to be less than this. Within the LDA there is a diffusion path for the hexagonal interstitial with a barrier of only 25 % of the tetrahedral-hexagonal energy difference. Applying the same percentage reduction to the DMC barrier then gives an estimate of the formation plus migration energy of the hexagonal interstitial of 5 eV. This estimate is in good agreement with the experimental activation energy for self-interstitial diffusion of 4.84 eV [46].

This study demonstrated the importance of a proper treatment of electron correlation when calculating defect-formation energies in silicon. DFT predicts formation plus migration energies that are smaller than those deduced from experiment. The larger defect-formation energies found in the DMC calculations indicate a possible resolution of this problem, which might be an important step in improving our understanding of self-diffusion in silicon.

6.3 Neutral Vacancy in Diamond

The vacancy in diamond is a dominant defect associated with radiation damage, and it exhibits a wider variety of physical phenomena than its counterpart in silicon. Self-diffusion in diamond is relevant to growing single-crystal diamond thin films on nondiamond substrates. DFT studies have suggested that self-diffusion in diamond is dominated by vacancies [55] and that interstitials do not play a significant role.

A vacancy may be formed by removing an atom, leaving four dangling-bonds. The simplest model of the electronic structure of a vacancy is a one-electron molecular defect picture in which symmetry-adapted combinations of the four dangling-bond states have a_1 and t_2 symmetry. These levels are, respectively, singly and triply degenerate. In the neutral defect four electrons are placed in the dangling-bond states, giving a $a_1^2 t_2^2$ configuration. This system is unstable to Jahn–Teller distortion and a spontaneous reduction in symmetry occurs in which the t_2 states split into a doubly occupied a_1 and empty e state, and a structure with D_{2d} symmetry results. This model is appropriate for the vacancy in silicon but, because of the large electron–electron interaction effects, not for diamond.

The prominent GR1 optical transition at 1.673 eV [56] is associated with the neutral vacancy. Uniaxial stress perturbations show that the GR1 transition is between a ground state that is orbitally doubly degenerate with symmetry E and an orbitally triply degenerate excited-state of T symmetry [57]. Both the ground and excited-states undergo Jahn–Teller relaxations, but the effects are dynamic during absorption and the vacancy therefore appears to maintain the T_d point group of an atomic site in diamond [58].

Correlation effects among the electrons in the dangling-bonds are very important, but the strong coupling to the solid-state environment must also be included. The electronic states of the defect form a multiplet requiring a multideterminant description [59, 60]. The GR1 optical transition cannot be expressed as a transition between one-electron states, which makes a first-principles approach very difficult. For example, approaches based on one-electron states, such as Kohn–Sham DFT or the *GW* self-energy approach [61, 62], cannot give a full description of the multiplet structure. *Von Barth* has suggested an approximate scheme for obtaining multiplet energies from DFT calculations [63], but unfortunately this scheme does not permit the calculation of the energy of what is believed to be the excited-state of the GR1 band [64, 65]. QMC appears to be an ideal method for this problem, as one can use a multideterminant description of the defect states while maintaining the favorable scaling with system size.

6.3.1 VMC and DMC Calculations on the Neutral Vacancy in Diamond

Hood et al. [4] performed VMC and DMC calculations of the neutral vacancy in diamond using periodic boundary conditions and a simulation cell

containing 53 atoms. The relaxed ionic positions were obtained from an LSDA calculation in which each of the t_2 states was fractionally occupied. The nearest-neighbor atoms of the vacancy were found to relax outwards by 0.1 Å, while the relaxation of the next nearest neighbors was negligible.

The trial wavefunctions were of the Slater–Jastrow type with the single-particle orbitals obtained from an LSDA calculation using a Gaussian basis set. To calculate the multiplet structure of the electronic states of the vacancy, symmetrized multideterminants were used with the configuration in which the a_1 states were fully occupied and the t_2 states were doubly occupied. This gives rise to $^1A_1$, ^1E, $^1T_2$, and $^3T_1$ states, with orbital degeneracies of 1, 2, 3, 3, respectively. The corresponding trial wavefunctions took the form,

$$\Psi_{^1A_1} = e^{\mathrm{J}(\boldsymbol{R})}\left(D^x_\uparrow D^x_\downarrow + D^y_\uparrow D^y_\downarrow + D^z_\uparrow D^z_\downarrow\right) \tag{26}$$

$$\Psi_{^1E} = e^{\mathrm{J}(\boldsymbol{R})}\left(2D^x_\uparrow D^x_\downarrow - D^y_\uparrow D^y_\downarrow - D^z_\uparrow D^z_\downarrow\right) \tag{27}$$

$$\Psi_{^1T_2} = e^{\mathrm{J}(\boldsymbol{R})}\left(D^y_\uparrow D^z_\downarrow - D^z_\uparrow D^y_\downarrow\right) \tag{28}$$

$$\Psi_{^3T_1} = e^{\mathrm{J}(\boldsymbol{R})} D^{yz}_\uparrow D_\downarrow \,. \tag{29}$$

In this notation, the superscripts x, y, z label which of the t_2 orbitals are included in the determinant, in addition to the a_1 and lower-energy orbitals. The values of the parameters in the Jastrow factors were optimized by minimizing the variance of the local energy [24, 25].

Figure 3 shows the DMC energies of the states obtained by *Hood* et al. [4]. The ^1E state was found to be the ground-state, in agreement with experiment. The state with the lowest-energy electronic excitation from the ground-state that is spin and orbitally dipole allowed was the $^1T_2$ state, and these states are identified as giving rise to the GR1 line. The associated DMC transition energy was calculated to be 1.51(34) eV, which lie within the statistical error bars of the experimental value of 1.673 eV.

The vacancy-formation energy was calculated to be 6.98 eV within the LSDA, and 5.96(34) eV within DMC. Both of these values include a correction for the Jahn–Teller relaxation energy of the vacancy, which is estimated to be 0.36 eV [64]. The DMC formation energy indicates that it is approximately one eV more favorable to form a neutral vacancy in diamond than predicted by the LSDA. The vacancy-formation energy in diamond has yet to be measured.

6.4 Schottky Defects in Magnesium Oxide

Alfè and *Gillan* [66] recently used DMC to study the Schottky defect formation energy in MgO. Although MgO is classified as an insulator rather than a semiconductor, these calculations illustrate very nicely the methodology that can be used to calculate the energetics of charged defects within QMC.

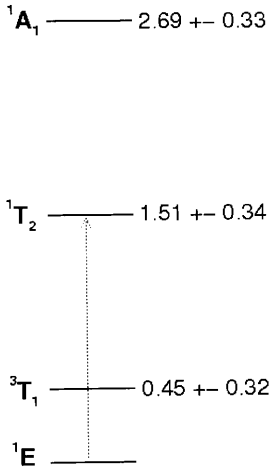

Fig. 3. DMC energy differences from the ground-state in eV with statistical errors for the lowest-symmetry states of the neutral vacancy in diamond. The *dotted arrow* indicates the lowest spin and orbitally dipole allowed transition. After *Hood* et al. [4]

Oxide materials are, of course, of great importance as insulating layers in semiconductor devices.

The Schottky defect energy of MgO is defined to be the energy change when a Mg^{2+} and an O^{2-} ion are removed from a perfect MgO crystal, the ions being replaced to form an additional unit cell of perfect crystal. The Schottky defect energy, E_S, governs the concentration of vacancies present in thermal equilibrium.

The methodology used for calculating E_S follows that used in DFT calculations. One way to model the Schottky defect would be to remove an Mg atom and an O atom from a large supercell, but in electronic-structure calculations it is normally preferable to study supercells containing only a single defect, because otherwise the interactions between the defects would be large.

Consider a crystal containing N cation and N anion sites. Let $E_N(\nu^+, \nu^-)$ denoted the total energy of a crystal containing ν^+ cation vacancies and ν^- anion vacancies. The Schottky defect energy is then

$$E_S = E_N(1,0) + E_N(0,1) - \frac{2(N-1)}{N} E_N(0,0)\,. \tag{30}$$

The systems with a single Mg^{2+} vacancy and a single O^{2-} vacancy were modeled by supercells subject to periodic boundary conditions. These supercells would each carry a net charge, in which case the electrostatic energy of the system cannot be defined, and therefore an appropriate uniform background charge was added to each supercell to make them charge neutral. In the limit of large N the presence of the uniform background charges makes no difference to E_S. In reality, a finite value of N must be used, which introduces finite-size errors. The finite-size error due to the interaction between the periodic images of the defects was corrected for by a standard method [67].

The DMC calculations used $N = 27$, i.e., 53 atoms for the defective crystals and 54 for the perfect crystals, and the relaxed atomic positions were

obtained from LSDA calculations. Slater–Jastrow wavefunctions were used, with orbitals obtained from plane-wave LSDA calculations, and optimized Jastrow factors. For efficient evaluation within the QMC calculations the orbitals were re-expanded in a basis of B-splines or "blips", which are functions that are strongly localized in real space [68].

The DMC calculations gave $E_S = 7.5 \pm 0.53\,\mathrm{eV}$, while equivalent LSDA calculations using the same pseudopotentialand cell size gave $E_S = 6.99\,\mathrm{eV}$, and a highly converged LSDA calculation gave $E_S = 6.76\,\mathrm{eV}$. It is very difficult to measure the Schottky defect energy, but it is estimated to lie within the range 4 eV to 7 eV.

7 Conclusions

QMC techniques provide a unified treatment of both the ground- and excited-state energies of correlated electron systems. They are therefore widely applicable and hold great promise as a computational method to enhance our understanding of defects in solids.

Considering that QMC is very costly, there seems little point in using it unless the results are *reliably* more accurate than those obtained from less costly methods such as DFT. The pursuit of higher accuracy in QMC calculations is, in my opinion, the most important practical issue facing today's practitioners. Very high accuracy has already been demonstrated for small systems, but the situation for larger systems is less clear. I believe that many careful studies are required before we can confidently assert that QMC techniques reliably achieve highly accurate results for complicated systems such as defects in semiconductors.

QMC techniques are currently in a phase of rapid development. The most challenging problem in fermion QMC is the infamous *fermion sign problem*. Solving the sign problem is certainly extremely difficult and perhaps impossible [69]. It is thought that exact fermion methods that avoid the sign problem will be exponentially slow on a classical computer. However, it is clear that QMC methods *can* deliver highly accurate results provided the trial wavefunctions are accurate enough. It is important to appreciate that trial wavefunctions can be improved using optimization techniques, whereas improving DFT results requires the development of better density functionals, which seems likely to be a much harder problem. Another important issue is that currently it is not possible to calculate accurate atomic forces within QMC for large systems, and therefore it is not possible to relax defect structures. So far this problem has been avoided by using relaxed structures obtained from DFT calculations. The problem of calculating accurate forces within QMC for large systems is unlikely to be insurmountable, and I believe that a satisfactory solution will be developed in due course. Although QMC methods have a long way to go before they are routinely applied to defects

in semiconductors I hope that I have persuaded the reader that such a goal is both desirable and possible.

I have discussed applications of the DMC method to self-interstitials in silicon, the neutral vacancy in diamond and Schottky defects in magnesium oxide. These studies have demonstrated the feasibility of DMC studies of defects in semiconductors, and have already produced interesting results that challenge those from other methods.

Acknowledgements

I would like to thank all of my collaborators who have contributed so much to our QMC project. Much of this work has been supported by the Engineering and Physical Sciences Research Council (EPSRC), UK.

References

[1] W. M. C. Foulkes, L. Mitas, R. J. Needs, G. Rajagopal: Rev. Mod. Phys. **73**, 33 (2001) 141, 145, 148
[2] D. M. Ceperley, B. J. Alder: Phys. Rev. Lett. **45**, 566 (1980) 142, 145
[3] W.-K. Leung, R. J. Needs, G. Rajagopal, S. Itoh, S. Ihara: Phys. Rev. Lett. **83**, 2351 (1999) 142, 154
[4] R. Q. Hood, P. R. C. Kent, R. J. Needs, P. R. Briddon: Phys. Rev. Lett. **91**, 076403 (2003) 142, 150, 157, 158, 159
[5] S. B. Healy, C. Filippi, P. Kratzer, E. Penev, M. Scheffler: Phys. Rev. Lett. **87**, 016105 (2001) 142
[6] C. Filippi, S. B. Healy, P. Kratzer, E. Pehlke, M. Scheffler: Phys. Rev. Lett. **89**, 166102 (2002) 142
[7] A. J. Williamson, J. C. Grossman, R. Q. Hood, A. Puzder, G. Galli: Phys. Rev. Lett. **89**, 196803 (2002) 142
[8] N. D. Drummond, A. J. Williamson, R. J. Needs, G. Galli: Phys. Rev. Lett. **95**, 096801 (2005) 142
[9] W. L. McMillan: Phys. Rev. A **138**, 442 (1965) 143
[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller: J. Chem. Phys. **21**, 1087 (1953) 143
[11] J. B. Anderson: J. Chem. Phys. **63**, 1499 (1975) 144
[12] J. B. Anderson: J. Chem. Phys. **65**, 4121 (1976) 144
[13] D. M. Ceperley: J. Stat. Phys. **63**, 1237 (1991) 145
[14] R. C. Grimm, R. G. Storer: J. Comput. Phys. **7**, 134 (1971) 145
[15] M. H. Kalos, D. Levesque, L. Verlet: Phys. Rev. A **9**, 257 (1974) 145
[16] D. M. Ceperley, M. H. Kalos: *Monte Carlo Methods in Statistical Physics* (Springer, Berlin, Heidelberg 1979) pp. 145–194 145
[17] B. L. Hammond, W. A. Lester, Jr., P. J. Reynolds: *Monte Carlo Methods in ab initio Quantum Chemistry* (World Scientific, Singapore 1994) 145
[18] S. Baroni, S. Moroni: Phys. Rev. Lett. **82**, 4745 (1999) 145
[19] T. Kato: Comm. Pure Appl. Math. **10**, 151 (1957) 146
[20] R. P. Feynman, M. Cohen: Phys. Rev. **102**, 1189 (1956) 146

[21] Y. Kwon, D. M. Ceperley, R. M. Martin: Phys. Rev. B **58**, 6800 (1998) 147
[22] M. Holzmann, D. M. Ceperley, C. Pierleoni, K. Esler: Phys. Rev. E **68**, 046707 (2003) 147
[23] D. M. Ceperley: J. Stat. Phys. **43**, 815 (1986) 147, 148
[24] C. J. Umrigar, K. G. Wilson, J. W. Wilkins: Phys. Rev. Lett. **60**, 1719 (1988) 147, 158
[25] P. R. C. Kent, R. J. Needs, G. Rajagopal: Phys. Rev. B **59**, 12344 (1999) 147, 158
[26] N. D. Drummond, R. J. Needs: Phys. Rev. B **72**, 085124 (2005) 147
[27] G. Rajagopal, R. J. Needs, S. Kenny, W. M. C. Foulkes, A. James: Phys. Rev. Lett. **73**, 1959 (1994) 148
[28] G. Rajagopal, R. J. Needs, A. James, S. Kenny, W. M. C. Foulkes: Phys. Rev. B **51**, 10591 (1995) 148, 150
[29] C. Lin, F. H. Zong, D. M. Ceperley: Phys. Rev. E **64**, 016702 (2001) 148
[30] P. R. C. Kent, R. Q. Hood, A. J. Williamson, R. J. Needs, W. M. C. Foulkes, G. Rajagopal: Phys. Rev. B **59**, 1917 (1999) 148, 150
[31] A. Ma, N. D. Drummond, M. D. Towler, R. J. Needs: Phys. Rev. E **71**, 066704 (2005) 148
[32] M. M. Hurley, P. A. Christiansen: J. Chem. Phys. **86**, 1069 (1987) 148
[33] L. Mitas, E. L. Shirley, D. M. Ceperley: J. Chem. Phys. **95**, 3467 (1991) 148
[34] M. Casula, C. Filippi, S. Sorella: unpublished 148
[35] C. W. Greeff, W. A. Lester, Jr.: J. Chem. Phys. **109**, 1607 (1998) 148
[36] J. R. Trail, R. J. Needs: J. Chem. Phys. **122**, 014112 (2005) 148
[37] J. R. Trail, R. J. Needs: J. Chem. Phys. **122**, 224322 (2005) 148
[38] W. M. C. Foulkes, R. Q. Hood, R. J. Needs: Phys. Rev. B **60**, 4558 (1999) 149
[39] A. J. Williamson, R. Q. Hood, R. J. Needs, G. Rajagopal: Phys. Rev. B **57**, 12140 (1998) 149
[40] M. D. Towler, R. Q. Hood, R. J. Needs: Phys. Rev. B **62**, 2330 (2000) 149
[41] R. Maezono, M. D. Towler, Y. Lee, R. J. Needs: Phys. Rev. B **68**, 165103 (2003) 150
[42] R. J. Needs, M. D. Towler: Int. J. Mod. Phys. B **17**, 5425 (2003) 150
[43] G. D. Watkins: *Defects and Their Structure in Non-Metallic Solids* (Plenum, New York 1976) p. 2003 153
[44] S. Dannefaer, P. Mascher, D. Kerr: Phys. Rev. Lett. **56**, 2195 (1986) 153
[45] W. Frank, U. Gösele, H. Mehrer, A. Seeger: *Diffusion in Crystalline Solids* (Academic, Orlando 1985) p. 64 153, 154
[46] U. Gösele, A. Plössl, T. Y. Tan: *Process Physics and Modeling in Semiconductor Technology* (Electrochemical Society, Pennington 1996) p. 309 153, 156
[47] D. Eaglesham: Phys. World **8**, 41 (1995) 153
[48] R. J. Needs: J. Phys.: Condens. Matter **11**, 10437 (1999) 153, 154, 155
[49] S. Goedecker, T. Deutsch, L. Billard: Phys. Rev. Lett. **88**, 235501 (2002) 153, 154
[50] O. Al-Mushadani, R. J. Needs: Phys. Rev. B **68**, 235205 (2003) 153, 154
[51] K. C. Pandey: Phys. Rev. Lett. **57**, 2287 (1986) 154
[52] E. Kaxiras, K. C. Pandey: Phys. Rev. B **47**, 1659 (1993) 154
[53] P. E. Blöchl, E. Smargiassi, R. Car, D. B. Laks, W. Andreoni, S. T. Pantelides: Phys. Rev. Lett. **70**, 2435 (1993) 154

[54] S. J. Clark, G. J. Ackland: Phys. Rev. B **56**, 47 (1997) 154
[55] J. Bernholc, A. Antonelli, T. M. Del Sole, Y. Bar-Yam, S. T. Pantelides: Phys. Rev. Lett. **61**, 2689 (1988) 157
[56] G. Davies, C. P. Foy: J. Phys. C **13**, 4127 (1980) 157
[57] C. D. Clark, J. Walker: Proc. Roy. Soc. London Ser. A **234**, 241 (1973) 157
[58] G. Davies: Rep. Prog. Phys. **44**, 787 (1981) 157
[59] C. A. Coulson, M. J. Kearsley: Proc. Roy. Soc. London Ser. A **241**, 433 (1957) 157
[60] M. Lannoo, J. Bourgoin: *Point Defects in Semiconductors I: Theoretical Aspects* (Springer, Berlin, Heidelberg 1981) pp. 141–145 157
[61] L. Hedin: Phys. Rev. **139**, A796 (1965) 157
[62] G. Onida, L. Reining, A. Rubio: Rev. Mod. Phys. **74**, 601 (2002) 157
[63] U. von Barth: Phys. Rev. A **20**, 1693 (1979) 157
[64] S. J. Breuer, P. R. Briddon: Phys. Rev. B **51**, 6984 (1995) 157, 158
[65] A. Zywietz, J. Furthmüller, F. Bechstedt: Phys. Rev. B **62**, 6854 (2000) 157
[66] D. Alfè, M. J. Gillan: Phys. Rev. B **71**, 220101 (2005) 158
[67] M. Leslie, M. J. Gillan: J. Phys. C **18**, 973 (1985) 159
[68] D. Alfè, M. J. Gillan: Phys. Rev. B **70**, 161101 (2004) 160
[69] M. Troyer, U.-J. Wiese: Phys. Rev. Lett. **94**, 170201 (2005) 160

Index

Quasiparticle Calculations for Point Defects at Semiconductor Surfaces

Arno Schindlmayr[1,2] and Matthias Scheffler[2]

[1] Institut für Festkörperforschung, Forschungszentrum Jülich, 52425 Jülich, Germany
A.Schindlmayr@fz-juelich.de
[2] Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, 14195 Berlin-Dahlem, Germany
scheffler@fhi-berlin.mpg.de

Abstract. We present a quantitative parameter-free method for calculating defect states and charge-transition levels of point defects in semiconductors. It combines the strength of density-functional theory for ground-state total energies with quasiparticle corrections to the excitation spectrum obtained from many-body perturbation theory. The latter is implemented within the G_0W_0 approximation, in which the electronic self-energy is constructed non-self-consistently from the Green's function of the underlying Kohn–Sham system. The method is general and applicable to arbitrary bulk or surface defects. As an example we consider anion vacancies at the (110) surfaces of III–V semiconductors. Relative to the Kohn–Sham eigenvalues in the local-density approximation, the quasiparticle corrections open the fundamental band gap and raise the position of defect states inside the gap. As a consequence, the charge-transition levels are also pushed to higher energies, leading to close agreement with the available experimental data.

1 Introduction

The electric properties of semiconductors, and hence their applicability in electronic devices, are to a large degree governed by defects that are either intrinsic or incidentally or intentionally introduced impurities. Considerable efforts, therefore, focus on determining the factors that lead to the formation of point defects and their influence on a material's electric properties.

The progress in the understanding of the atomic structure of point defects at cleaved III–V semiconductor surfaces, which serve as an illustration in this work, was recently reviewed by *Ebert* [1]. The possibility to image individual defects using scanning tunneling microscopy (STM) with atomic resolution, in particular, has yielded a wealth of data, but as STM provides a somewhat distorted picture of the electronic states close to the Fermi energy, these results cannot (and should not) be identified directly with the atomic geometry. Electronic-structure calculations have hence turned out to be an indispensable tool for the interpretation of the experimental findings. A deeper discussion of this point together with an example for the practical structure determination of a semiconductor surface is given in [2]. For point

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 165–192 (2007)

defects at the (110) surfaces of III–V semiconductors several calculations were reported [3–8]. They are based on density-functional theory [9], where the exchange-correlation energy is typically treated in the local-density approximation (LDA) [10] or generalized gradient approximations (GGA) [11]. The agreement with experimental STM data appears to be very good. For example, the enhanced contrast of the empty p_z orbitals of the two Ga atoms nearest to an anion vacancy in p-type GaAs(110) observed under positive bias and initially interpreted as an outward relaxation [12] could thus be understood to result, instead, from a downward local band bending accompanied by an *inward* relaxation [3]. The band bending itself is caused by the positive charge of the defect. Another controversy centers on the lateral relaxation of the positively charged anion vacancy. While STM images show a density of states preserving the mirror symmetry of the surface at the defect site [12], early theoretical studies of the lattice geometry based on total-energy minimization produced conflicting evidence for [3] and against [4, 5] a possible breaking of the mirror symmetry. Well-converged electronic-structure calculations later confirmed that the distortion is indeed asymmetric [6–8] and that the apparently symmetric STM image results from the thermally activated flip motion between two degenerate asymmetric configurations.

In contrast, theoretical predictions of the electronic properties of point defects have been less successful and still show significant quantitative deviations from experimental results. The principal quantities of interest are the location of defect states in the fundamental band gap as well as the charge-transition levels. We will carefully distinguish in this chapter between these two quantities: "defect states" or "defect levels" on the one hand and "charge-transition levels" on the other. The former are part of the electronic structure and can, in principle, be probed by photoemission spectroscopy, although standard spectroscopic techniques are often not applicable due to the low density of the surface defects. The Franck–Condon principle is typically well justified, because the rearrangement of the atoms happens on a much slower timescale than the photoemission process. Nevertheless, the coupling of the electrons to the lattice may be visible in the linewidths and lineshapes. The defect levels thus contain the full *electronic* relaxation in response to the created hole in direct photoemission or the injected extra electron in inverse photoemission, but no atomic relaxation. Although investigations of electronic properties frequently rely on the Kohn–Sham eigenvalues from density-functional theory, these only provide a first approximation to the true band structure, and quantitative deviations from experimental results must be expected. In particular, for many materials the eigenvalue gap both in the LDA and the GGA underestimates the fundamental band gap significantly. Likewise, the position of defect states in the gap cannot be determined without systematic errors. With these words of warning we note, however, that the Kohn–Sham eigenvalues constitute a good and well-justified starting point for calculating band structures and defect states [13]. Therefore, a

perturbative approach starting from the Kohn–Sham eigenvalue spectrum is indeed appropriate.

While the measurement of the defect levels (see the discussion on the page previous of this Chapter) probes the geometry before the electron is added or removed, the charge-transition levels are thermodynamic quantities and specify the values of the Fermi energy where the stable charge state of the defect changes. Therefore, the charge-transition levels are affected noticeably by the atomic relaxation taking place upon the addition or removal of an electron. A quantitative analysis must hence be able to accurately compare the formation energies of competing configurations with different numbers of electrons. Density-functional theory at the level of the LDA or GGA is capable of giving a good account of the atomic geometries for many materials. A critical feature is the nonlinearity of the exchange-correlation functional. The pseudopotential approximation, which effectively removes the inner shells from the calculation by modeling their interaction with the valence electrons in terms of a modified potential, linearizes the core–valence interaction and thus does not treat this contribution correctly. In particular, freezing the d electrons in the core of a pseudopotential leads to poor lattice constants and a distorted electronic structure for some III–V semiconductors, such as GaN, where the Ga $3d$ states resonate strongly with the N $2s$ states [14]. For GaAs and InP this is a lesser problem, because the cation d states are energetically well below the anion $2s$ states and thus are relatively inert. As a result, the LDA and the GGA yield only small deviations from the experimental lattice constants and provide a good starting point for quasiparticle band-structure calculations. However, when competing configurations with a different number of electrons are compared, the relevant energy differences lack the required quantitative accuracy. As we will show in more detail below, the reason is that these jellium-based approximations of the exchange-correlation energy ignore important features of the exact functional, such as the discontinuity of the exchange-correlation potential with respect to a change in the particle number. As a consequence, previous calculations of charge-transition levels based on the LDA exhibit systematic errors, for example for anion vacancies at InP(110) [6].

As an alternative approach to the electronic structure of point defects, we employ techniques adapted from many-body perturbation theory that we have found to be very fruitful in the past [15]. Exchange and correlation effects are here described by a nonlocal and frequency-dependent self-energy operator. The solutions of the ensuing nonlinear eigenvalue equations have a rigorous interpretation as excitation energies and can be identified with the electronic band structure. We discuss the calculation of defect states as well as charge-transition levels within this framework and show that the results improve significantly upon earlier values obtained from the Kohn–Sham scheme in the LDA. As an example we consider anion vacancies at GaAs(110) and InP(110), but the method is general and can also be applied to other defects at surfaces as well as in the bulk. The (110) surfaces are not only

the natural cleavage planes of III–V semiconductors, but they have several characteristics that make them particularly interesting for defect studies. As no surface states exist inside the fundamental band gap [16, 17], the Fermi energy of a system that is clean and free from intrinsic defects is not pinned but controlled by the doping of the crystal. Only imperfections, such as anion vacancies or antisite defects, can introduce gap states and pin the Fermi energy at the surface. STM can probe filled and empty surface states by reversing the bias voltage [12], and both the GaAs(110) and the InP(110) surface are well characterized experimentally.

This Chapter is organized as follows. We start in Sect. 2 by reviewing the computational methods. In Sect. 3 we then explain the physics of the defect-free GaAs(110) and InP(110) surfaces. The calculation of defect levels is discussed in Sect. 4 and that of charge-transition levels in Sect. 5, together with a comparison with the available theoretical and experimental data. Finally, Sect. 6 summarizes our conclusions. Unless explicitly indicated otherwise, we use Hartree atomic units.

2 Computational Methods

A quantitative analysis of the electronic properties of point defects requires computational schemes that describe not only the ground state but also the excitation spectrum. While density-functional theory with state-of-the-art exchange-correlation functionals can be used to determine ground-state atomic geometries, many-body perturbation theory is the method of choice for excited states. In this work we take the Kohn–Sham eigenvalues in the LDA as a first estimate and then apply the G_0W_0 approximation for the electronic self-energy as a perturbative correction. The latter provides a good account of the discontinuity as well as other shortcomings of the LDA. For this reason we first review both schemes, emphasizing their strengths as well as limitations.

2.1 Density-Functional Theory

Density-functional theory is based on the Hohenberg–Kohn theorem [9], which observes that the total energy $E_{N,0}$ of a system of N interacting electrons in an external potential $V_{\text{ext}}(\boldsymbol{r})$ is uniquely determined by the ground-state electron density $n_N(\boldsymbol{r})$. While the Hohenberg–Kohn theorem itself makes no statement about the mathematical form of this functional, it has inspired algorithms that exploit the reduced number of degrees of freedom compared to a treatment based on many-particle wavefunctions. In the Kohn–Sham scheme [10], which underlies all practical implementations, the

density is constructed from the orbitals of an auxiliary noninteracting system according to

$$n_N(\boldsymbol{r}) = 2\sum_{j=1}^{\infty} f_{N,j}\,|\varphi_{N,j}(\boldsymbol{r})|^2\ . \tag{1}$$

The occupation numbers $f_{N,j}$ are given by the Fermi distribution; at zero temperature they equal one for states below the Fermi energy and zero for states above. The factor 2 stems from the spin summation. Here we only consider nonmagnetic systems and thus assume two degenerate spin channels throughout, although the formalism can easily be generalized if necessary. The energy functional is decomposed as

$$E_{N,0} \equiv E[n_N] = T_\mathrm{s}[n_N] + \int V_\mathrm{ext}(\boldsymbol{r}) n_N(\boldsymbol{r})\,\mathrm{d}^3 r + E_\mathrm{H}[n_N] + E_\mathrm{xc}[n_N]\,, \tag{2}$$

where $T_\mathrm{s}[n_N]$ is the kinetic energy of the auxiliary noninteracting system and $E_\mathrm{H}[n_N]$ the Hartree energy. The last term incorporates all remaining exchange and correlation contributions and is not known exactly. In practical implementations it must be approximated, for example by the LDA, which replaces the exchange-correlation energy $E_\mathrm{xc}[n_N]$ by that of a homogeneous electron gas with the same local density [10]. A variational analysis finally shows that the total energy is minimized if the single-particle orbitals satisfy

$$\left[-\tfrac{1}{2}\nabla^2 + V_\mathrm{ext}(\boldsymbol{r}) + V_\mathrm{H}([n_N];\boldsymbol{r}) + V_\mathrm{xc}([n_N];\boldsymbol{r})\right]\varphi_{N,j}(\boldsymbol{r}) = \varepsilon^\mathrm{KS}_{N,j}\varphi_{N,j}(\boldsymbol{r})\,. \tag{3}$$

The Hartree potential $V_\mathrm{H}([n_N];\boldsymbol{r})$ and the exchange-correlation potential $V_\mathrm{xc}([n_N];\boldsymbol{r})$ are defined as functional derivatives of the corresponding energy terms with respect to the density. The eigenvalues $\varepsilon^\mathrm{KS}_{N,j}$ are Lagrange parameters that enforce the normalization of the orbitals.

We use the Kohn–Sham scheme to determine the equilibrium geometry of clean surfaces and surface defects by relaxing the atomic coordinates and allowing the system to explore its energetically most favorable configuration. Although the Kohn–Sham eigenvalues differ from the true excitation energies and constitute only an approximation to the quasiparticle band structure, they are often numerically close in practice. This follows from the transition-state theorem of *Slater* [18] and *Janak* [19] and allows a correction within perturbation theory. Here, we are interested in the position of the defect state, which may be occupied or unoccupied, relative to the surface valence-band maximum. As the defect state is separated from the valence-band maximum by a finite energy difference, the location obtained from the Kohn–Sham eigenvalue spectrum contains two sources of errors: In addition to the chosen approximation for the exchange-correlation functional, there is another systematic error that is due to fundamental limitations of the Kohn–Sham scheme and would also be present if the exact functional was employed. In order to understand the latter, we now briefly sketch its origin.

One rigorous result is that the eigenvalue of the highest occupied Kohn–Sham state matches the corresponding quasiparticle energy [20], which is in turn equal to the ionization potential that marks the threshold for photoemission, i.e., $\varepsilon^{\mathrm{KS}}_{N,N} = E_{N,0} - E_{N-1,0}$. If the highest occupied state corresponds to the valence-band maximum, then the unoccupied defect state equals the electron affinity $\varepsilon^{\mathrm{KS}}_{N+1,N+1} = E_{N+1,0} - E_{N,0}$, because it is the next to be populated by one extra electron added to the system. Unfortunately, this is not the same as the eigenvalue $\varepsilon^{\mathrm{KS}}_{N,N+1}$ of the first unoccupied state obtained from (3). The difference is due to the fact that the exchange-correlation potential of an insulator $V_{\mathrm{xc}}([n_{N+1}];\boldsymbol{r}) = V_{\mathrm{xc}}([n_N];\boldsymbol{r}) + \Delta_{\mathrm{xc}} + O(1/N)$ with $\Delta_{\mathrm{xc}} > 0$ changes discontinuously upon addition of an extra electron [21, 22]. A similar argument can be made if the defect state is occupied. As a consequence, the Kohn–Sham eigenvalue gaps differ systematically from the gaps in the true quasiparticle band structure. The magnitude of the discontinuity is still a matter of controversy but is believed to be significant. For pure *sp*-bonded semiconductors like GaAs the LDA underestimates the experimental band gap by about 50 %. The GGA, designed only to improve the total energy, yields a very similar eigenvalue spectrum as the LDA when applied to the same atomic geometry. An entirely different construction that permits a more systematic treatment of exchange and correlation is the optimized effective potential method [23]. When evaluated to first order in the coupling constant, this approach yields the exact exchange potential, which can be used in band-structure calculations [24]. Remarkably, for many semiconductors the resulting Kohn–Sham eigenvalue gaps are very close to the true quasiparticle band gaps [25]. The significance of this observation is under debate, however, because there are indications that it may not uphold if correlation is treated on the same footing. Preliminary results suggest that the eigenvalue gaps are again close to the LDA values if correlation is included within the random-phase approximation [26, 27], but there is currently too little data to make a definite statement, and all existing calculations at this level also contain additional simplifications, for example a shape approximation for the potential in the linearized muffin-tin orbital method in [27].

The reasoning above suggests that density-functional theory is still, in principle, applicable for calculating the difference in total energy between ground-state configurations with different electron numbers, which is needed to determine the energetically most favorable charge state of a point defect. However, all jellium-based functionals lack the derivative discontinuity Δ_{xc} of the exact exchange-correlation potential. This neglect reduces the electron affinity $E_{N+1,0} - E_{N,0}$ both in the LDA and the GGA if the additional electron occupies a state separated from the valence-band maximum by a finite energy difference. For systems in contact with an electron reservoir, such as defects in solids, it hence lowers the threshold for an increase of the electron population. This is consistent with the observation that the calculated charge-transition levels for materials like InP are significantly smaller than the available values deduced from experimental measurements [6]. The *exact exchange* potential,

an implicit functional of the density defined in terms of the Kohn–Sham orbitals, includes a derivative discontinuity [13], but the latter exceeds the experimental band gap significantly and must hence be partially canceled by a correlation contribution with similar magnitude and opposite sign [25]. As the total energies are by construction close to the corresponding Hartree–Fock values, the electron affinities $E_{N+1,0} - E_{N,0}$ are grossly overestimated, and no reliable charge-transition levels can be obtained in this way.

Our implementation of density-functional theory employs the plane-wave pseudopotential method in combination with the LDA. We use the parametrization by *Perdew* and *Zunger* [28], which is in turn based on the quantum Monte-Carlo data of *Ceperley* and *Alder* for the homogeneous electron gas [29]. The norm-conserving pseudopotentials are of the fully separable Kleinman–Bylander form [30]. We choose d as the local component for all pseudopotentials except for In, where p is used instead. The Kohn–Sham wavefunctions are expanded in plane waves with a cutoff energy of 15 Ry. Our calculations are performed with the FHImd code [31, 32]. The bulk lattice constants obtained in this way, 5.55 Å for GaAs and 5.81 Å for InP in the absence of zero-point vibrations, are in good agreement with previously published data [33] and slightly smaller than the experimental values at room temperature by 1.8 % and 1.1 %, respectively [34]. We use the theoretical lattice constants in order to prevent errors resulting from a nonequilibrium unit-cell volume during the relaxation of the surface geometries.

2.2 Many-Body Perturbation Theory

Many-body perturbation theory [35] provides powerful techniques to analyze the electronic structure in the gap region, because the framework is designed specifically to give access to excited states. Quasiparticle excitations created by the addition or removal of one electron are obtained from the one-particle Green's function

$$G(\boldsymbol{r},\boldsymbol{r}';t-t') = -\mathrm{i}\langle\Psi_{N,0}|\mathcal{T}\{\hat{\psi}(\boldsymbol{r},t)\hat{\psi}^{\dagger}(\boldsymbol{r}',t')\}|\Psi_{N,0}\rangle\,, \tag{4}$$

where $|\Psi_{N,0}\rangle$ denotes the ground-state wavefunction of the interacting electron system in second quantization, $\hat{\psi}^{\dagger}(\boldsymbol{r}',t') = \exp(\mathrm{i}\hat{H}t')\hat{\psi}^{\dagger}(\boldsymbol{r}')\exp(-\mathrm{i}\hat{H}t')$ and $\hat{\psi}(\boldsymbol{r},t) = \exp(\mathrm{i}\hat{H}t)\hat{\psi}(\boldsymbol{r})\exp(-\mathrm{i}\hat{H}t)$ are the electron creation and annihilation operators in the Heisenberg picture, respectively, and the symbol $\mathcal{T}$ sorts the subsequent list of operators according to ascending time arguments from right to left with a change of sign for every pair permutation. The Green's function can be interpreted as a propagator: For $t > t'$ it describes a process in which an extra electron is added to the system at time t'. The resulting wavefunction is, in general, no eigenstate of the Hamiltonian $\hat{H}$ but a linear combination of many eigenstates $|\Psi_{N+1,j}\rangle$. Between the times t' and t each projection evolves with its own characteristic phase, and a Fourier analysis of this oscillatory behavior immediately yields the energy spectrum

$\varepsilon_{N,j} = E_{N+1,j} - E_{N,0}$ of the accessible excited states. Likewise, for $t' > t$ the Green's function describes the propagation of an extra hole between t and t', yielding the energies $\varepsilon_{N,j} = E_{N,0} - E_{N-1,j}$. In mathematical terms, we insert a complete set of eigenstates between the field operators in (4) and Fourier transform to the frequency axis. The resulting expression

$$G(\boldsymbol{r},\boldsymbol{r}';\omega) = \lim_{\eta\to+0}\sum_{j=1}^{\infty} f_{N,j}\frac{\psi_{N,j}(\boldsymbol{r})\psi_{N,j}^*(\boldsymbol{r}')}{\omega-\varepsilon_{N,j}-\mathrm{i}\eta} + \lim_{\eta\to+0}\sum_{j=1}^{\infty}(1-f_{N,j})\frac{\psi_{N,j}(\boldsymbol{r})\psi_{N,j}^*(\boldsymbol{r}')}{\omega-\varepsilon_{N,j}+\mathrm{i}\eta} \tag{5}$$

shows that the poles of the Green's function correspond directly to the quasiparticle energies $\varepsilon_{N,j}$. Significantly, these not only yield the true band structure, but the highest occupied state also equals the exact ionization potential $E_{N,0} - E_{N-1,0}$ and the lowest unoccupied state the exact electron affinity $E_{N+1,0} - E_{N,0}$. Therefore, many-body perturbation theory provides a convenient way to analyze both defect states and the energetics of charge transitions. The wavefunctions $\psi_{N,j}(\boldsymbol{r}) = \langle\Psi_{N-1,j}|\hat{\psi}(\boldsymbol{r})|\Psi_{N,0}\rangle$ for occupied and $\psi_{N,j}(\boldsymbol{r}) = \langle\Psi_{N,0}|\hat{\psi}(\boldsymbol{r})|\Psi_{N+1,j}\rangle$ for unoccupied states are obtained from the quasiparticle equations

$$\left[-\tfrac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\boldsymbol{r}) + V_{\mathrm{H}}(\boldsymbol{r})\right]\psi_{N,j}(\boldsymbol{r}) + \int \Sigma_{\mathrm{xc}}(\boldsymbol{r},\boldsymbol{r}',\varepsilon_{N,j})\psi_{N,j}(\boldsymbol{r}')\,\mathrm{d}^3r' = \varepsilon_{N,j}\psi_{N,j}(\boldsymbol{r})\,. \tag{6}$$

The self-energy $\Sigma_{\mathrm{xc}}(\boldsymbol{r},\boldsymbol{r}',\varepsilon)$ incorporates all contributions from exchange and correlation processes. In contrast to the exchange-correlation potential of density-functional theory, it is nonlocal, energy dependent and has a finite imaginary part, which is proportional to the damping rate resulting from electron–electron scattering. Together with other relevant decay channels, such as scattering from phonons or impurities, this mechanism is responsible for a finite lifetime of the excitations.

For real materials the self-energy can only be treated approximately. Like the majority of practical applications, we use the G_0W_0 approximation [36]

$$\Sigma_{\mathrm{xc}}(\boldsymbol{r},\boldsymbol{r}';t-t') = \lim_{\eta\to+0}\mathrm{i}G_0(\boldsymbol{r},\boldsymbol{r}';t-t')W_0(\boldsymbol{r},\boldsymbol{r}';t-t'+\eta)\,. \tag{7}$$

The Fourier transform of $G_0(\boldsymbol{r},\boldsymbol{r}';t-t')$ on the frequency axis is constructed in analogy to (5) from the eigenstates of an appropriate mean-field system, in our case from the Kohn–Sham orbitals $\varphi_{N,j}(\boldsymbol{r})$ and eigenvalues $\varepsilon_{N,j}^{\mathrm{KS}}$. The dynamically screened Coulomb interaction $W_0(\boldsymbol{r},\boldsymbol{r}';t-t')$ can be modeled in different ways. Many implementations use plasmon-pole models, in which the frequency dependence is described by an analytic function whose parameters are determined by a combination of known sum rules and asymptotic limits [37,38]. This simplification has the advantage that it facilitates an analytic

treatment, but it ignores details of the dynamic screening processes in a material and thus constitutes a potential source of errors. Here we take the full frequency dependence of the dielectric function into account by employing the random-phase approximation

$$W_0(\boldsymbol{k},\omega) = v(\boldsymbol{k}) + v(\boldsymbol{k})P_0(\boldsymbol{k},\omega)W_0(\boldsymbol{k},\omega) \tag{8}$$

with the bare Coulomb potential $v(\boldsymbol{k}) = 4\pi/|\boldsymbol{k}|^2$ and the polarizability

$$P_0(\boldsymbol{r},\boldsymbol{r}';t-t') = -2\mathrm{i}G_0(\boldsymbol{r},\boldsymbol{r}';t-t')G_0(\boldsymbol{r}',\boldsymbol{r};t'-t)\,. \tag{9}$$

The inclusion of dynamic screening derives from the concept of quasiparticles, which comprise an electron or hole together with its surrounding polarization cloud. The composite is called a quasiparticle because it behaves in many ways like a single entity. The polarization cloud is created by the repulsive Coulomb potential and reduces the effective charge of the quasiparticle compared to that of the bare particle at its center. The G_0W_0 expression (7) constitutes the leading term in the expansion of the self-energy and is of first order in the dynamically screened interaction. Finally, we exploit the formal similarity between (6) and the Kohn–Sham equations (3) by evaluating the quasiparticle energies within first-order perturbation theory as

$$\varepsilon_{N,j} = \varepsilon^{\mathrm{KS}}_{N,j} + \langle\varphi_{N,j}|\Sigma_{\mathrm{xc}}(\varepsilon_{N,j}) - V_{\mathrm{xc}}[n_N]|\varphi_{N,j}\rangle\,. \tag{10}$$

The treatment within first-order perturbation theory is justified if the eigenvalues of the underlying Kohn–Sham system are already sufficiently close to the expected quasiparticle band structure. This is guaranteed by the transition-state theorem [18, 19]. Besides, the orbitals $\varphi_{N,j}(\boldsymbol{r})$ are usually a good approximation to the true quasiparticle wavefunctions. For the homogeneous electron gas both are plane waves and thus coincide exactly. Numerical calculations for bulk semiconductors indicate an overlap close to unity between the Kohn–Sham orbitals in the LDA and the quasiparticle wavefunctions obtained from (6) with the G_0W_0 approximation for states near the band edges [39]. Larger effects have been observed for surfaces, especially for image states, because the latter are located outside the surface in a region where the LDA potential is qualitatively wrong [40, 41]. Changes in the wavefunctions of other surface states can also be identified but have only a minor influence on the quasiparticle energies in the gap region. For GaAs(110) this has been confirmed explicitly [42].

In principle, the equations (4) to (9) could be solved self-consistently by successively updating the self-energy with the quasiparticle orbitals derived from it. This approach is appealing on formal grounds because it makes the results independent of the original mean-field approximation. In addition, the self-consistent Green's function satisfies certain sum rules, including particle-number conservation [43]. In practice, however, this procedure produces a poor excitation spectrum. The reason lies in the mathematical structure of

Hedin's equations, which describe the recipe for constructing the self-energy from the Green's function by means of functional-derivative techniques [36]. The G_0W_0 approximation is the result of a single iteration. Further iterations would not only ensure self-consistency in the Green's function but would also introduce higher-order self-energy terms, so-called vertex corrections. If the latter are neglected, then the spectral features deteriorate, as was first demonstrated for the homogeneous electron gas [44]. For bulk semiconductors self-consistency appears to lead to a gross overestimation of the fundamental band gap [45] in combination with poor spectral weights and linewidths. We hence follow the established procedure for practical applications and terminate the cycle at the G_0W_0 approximation. Of course, the final results depend on the input Green's function in this case.

Another point that has been raised in the same context concerns possible errors resulting from the pseudopotential approach, which has traditionally dominated applications of the G_0W_0 approximation. Numerical deviations must be expected because the matrix elements of the self-energy in (10) are influenced by the pseudoization of the wavefunctions – that is the neglect of core states plus appropriate smoothening of the valence states in the core region – and the treatment of the core–valence interaction in pseudopotential calculations. Indeed, a number of early all-electron implementations of the G_0W_0 approximation, based on the linearized augmented plane-wave or the linearized muffin-tin orbital method, found quasiparticle band gaps for prototype semiconductors that were significantly smaller than previously reported values and blamed the discrepancy on the pseudopotential approximation [46,47]. However, this claim was later refuted by all-electron calculations using plane waves [48,49]. The issue is currently still under debate, but there is mounting evidence that a large part of the observed discrepancy is due to insufficient convergence of the early calculations in combination with deficiencies of the linearized basis sets for the description of high-lying unoccupied states [48,50]. If these factors are properly accounted for, then the deviation is significantly reduced, and the all-electron results are again in good agreement with experimental measurements. A certain discrepancy with respect to pseudopotential calculations still remains, although the difference is small compared to that from the Kohn–Sham eigenvalues. While some errors of the pseudopotential approach can be partially suppressed, for example through a better description of the core–valence interaction [51], many others are inherent. In particular, the pseudopotential construction also entails an incorrect description of high-lying states. A careful analysis of the quantitative implications of the pseudopotential approximation is the subject of active research.

The good quantitative agreement between the G_0W_0 approximation and experimental band structures has been demonstrated for a wide range of bulk materials, including III–V semiconductors [52–54]. Due to the high computational cost in present implementations, which stems from the evaluation of the nonlocality of the propagators, their frequency dependence, and the

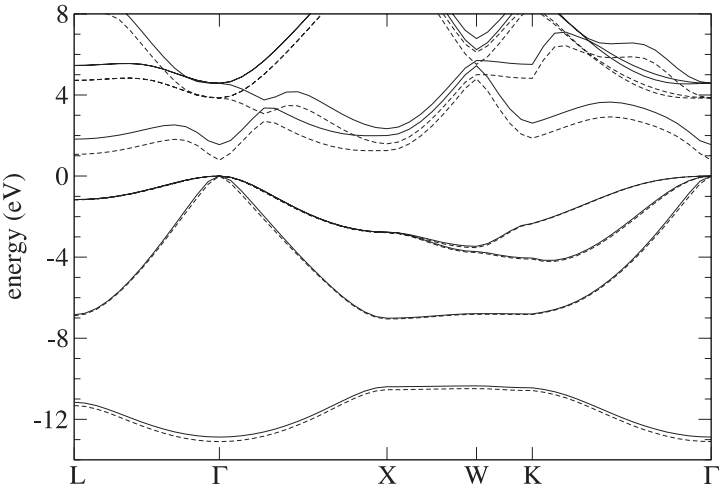

Fig. 1. Bulk band structure of GaAs. The G_0W_0 approximation (*straight lines*) opens the too small Kohn–Sham eigenvalue gap in the LDA (*dashed lines*) and is in very good agreement with the experimental band gap. The calculation is carried out at the theoretical lattice constant 5.55 Å, and the valence-band maximum is set to zero in both schemes

need to include a large number of unoccupied states, there have been relatively few applications to more complex systems so far, however. Furthermore, these often contain additional simplifications: the two available studies of the quasiparticle band structure of GaAs(110) both employed plasmon-pole models instead of the more accurate random-phase approximation [55, 56], and the only published results for InP(110) were obtained within an even more restrictive tight-binding formalism [57].

For a quasiparticle band-structure calculation the matrix elements of the self-energy in (6) must be evaluated in the frequency domain for states with a given wavevector $\boldsymbol{k}$. Therefore, most practical implementations of the G_0W_0 approximation choose a reciprocal-space representation for all propagators. The cell-periodic part of the wavefunctions is often expanded in plane waves [52, 55], although localized basis sets like Gaussian orbitals have also been used [54]. The disadvantage of the representation in reciprocal space is that the products (7) and (9) turn into numerically expensive multidimensional convolutions. Therefore, we employ a representation in real space and imaginary time [58, 59], in which the self-energy and the polarizability can be calculated by simple multiplications. The projection on wavevectors and imaginary frequencies used to solve the Dyson-type equation (8) can be done efficiently by exploiting fast Fourier transforms. The imaginary time and frequency arguments are chosen because the functions are smoother on these axes and can be sampled with fewer grid points, although they contain

exactly the same information. The physical self-energy on the real frequency axis is eventually recovered by an analytic continuation.

As an illustration of the G_0W_0 approximation we show the calculated bulk band structure of GaAs in Fig. 1. The band gap is direct and located at Γ in the center of the Brillouin zone. While the LDA underestimates the fundamental band gap and yields a Kohn–Sham eigenvalue gap of only 0.78 eV at a lattice constant of 5.55 Å, the subsequent addition of the self-energy correction opens the band gap to 1.55 eV, which is in very good agreement with the experimental value of 1.52 eV [34]. As we measure all energies relative to the valence-band maximum, we set the latter to zero and align the two sets of curves at this point. The principal effect of the G_0W_0 approximation is a rigid upward shift of the conduction bands, although the dispersion is also slightly modified, as can be seen by the reduced bandwidth at the bottom of the valence band. The band structure of InP looks very similar. In this case we obtain a Kohn–Sham eigenvalue gap of 0.76 eV at the theoretical lattice constant 5.81 Å and a quasiparticle band gap of 1.52 eV that is again close to the experimental value 1.42 eV [34]. Incidentally, the band gap depends sensitively on the lattice constant. For GaAs it decreases at a rate of $-4.07\,\mathrm{eV/Å}$ in the LDA and $-4.59\,\mathrm{eV/Å}$ in the G_0W_0 approximation when the lattice constant increases. For InP the values are $-3.13\,\mathrm{eV/Å}$ and $-3.68\,\mathrm{eV/Å}$.

3 Electronic Structure of Defect-Free Surfaces

Before focusing on defect states we first briefly discuss the electronic structure of the defect-free GaAs(110) and InP(110) surfaces. The nonpolar (110) surface, illustrated in Fig. 2, is the natural cleavage plane of the zincblende lattice, because it cuts the smallest number of bonds per unit area. Both the cations and anions in the terminating layer are threefold coordinated, and each possess one dangling bond extending into the vacuum. Due to the different electron affinities of the two species, charge is transfered from the dangling bonds of the cations to those of the anions. This charge transfer is the driving force for a structural relaxation, which consists of an outward movement of the anion atoms and a corresponding inward movement of the cation atoms [33]. As a result, the orbitals of the latter rehybridize from an sp^3 towards an energetically more favorable sp^2 bonding situation in a nearly planar environment; the empty p_z-like orbitals perpendicular to this plane are pushed to higher energies and form an unoccupied surface band. At the same time, the three bonds between the anions and their neighboring group-III atoms are rearranged at almost right angles and become more p-like in character; the nonbonding electron pairs in the fourth orbital pointing away from the surface are in turn lowered in energy and give rise to an occupied surface band. The relaxation preserves the $C_{1\mathrm{h}}$ point-group symmetry with a single mirror plane perpendicular to the $[\bar{1}10]$ direction.

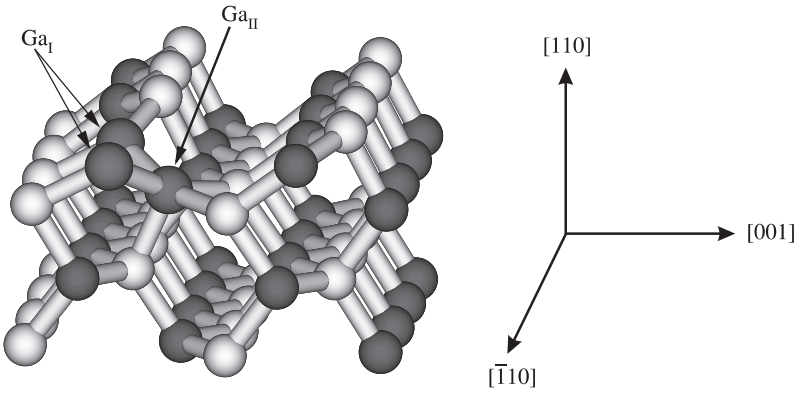

Fig. 2. Geometry of the anion vacancy at the GaAs(110) surface. The As atoms are shown in *light* and the Ga atoms in *dark gray*. The three Ga atoms nearest to the vacancy are indicated by *arrows*

In Fig. 3 we show the calculated band structure of the clean and defect-free GaAs(110) surface at the theoretical lattice constant 5.55 Å, modeled as a slab element placed in a supercell with periodic boundary conditions in all directions. The slab consists of six atomic layers, of which the top three layers are allowed to relax, while the three base layers are kept fixed at their ideal bulk positions. The dangling bonds at the bottom of the slab are passivated by pseudoatoms with noninteger nuclear charges of 0.75 and 1.25 for anion and cation termination, respectively. The slabs are separated by a vacuum region equivalent to four atomic layers. The gray-shaded regions in the figure mark the projection of the G_0W_0 bulk bands onto the two-dimensional surface Brillouin zone, shown in the inset. The dashed lines indicate the occupied and unoccupied surface bands obtained from the LDA, the straight lines are the corresponding G_0W_0 results. For the calculation of the ground-state density and the Kohn–Sham eigenvalues we used four Monkhorst–Pack $\boldsymbol{k}$-points [60] in the irreducible part of the Brillouin zone, while the self-energy was evaluated at the four high-symmetry points $\overline{\Gamma}$, $\overline{\mathrm{X}}'$, $\overline{\mathrm{M}}$, and $\overline{\mathrm{X}}$. In our implementation the $\boldsymbol{k}$-point set enters merely as the reciprocal grid of the real-space mesh used to describe the nonlocality of $G_0(\boldsymbol{r}, \boldsymbol{r}'; t - t')$ and $\Sigma_{\mathrm{xc}}(\boldsymbol{r}, \boldsymbol{r}'; t - t')$ [58]; the four selected $\boldsymbol{k}$-points correspond to a real-space mesh that extends over four surface unit cells. This is sufficient, because the correlation length is of the order of the interatomic distance [61]. The position of the Kohn–Sham surface bands relative to the corresponding bulk bands was determined by aligning the electrostatic potential in the central part of the slab with that of the bulk. We then applied the self-energy correction independently to surface and bulk bands and again chose the valence-band maximum as the energy zero. From the figure it is evident that the G_0W_0 approximation has only a small effect on the dispersion of the surface bands.

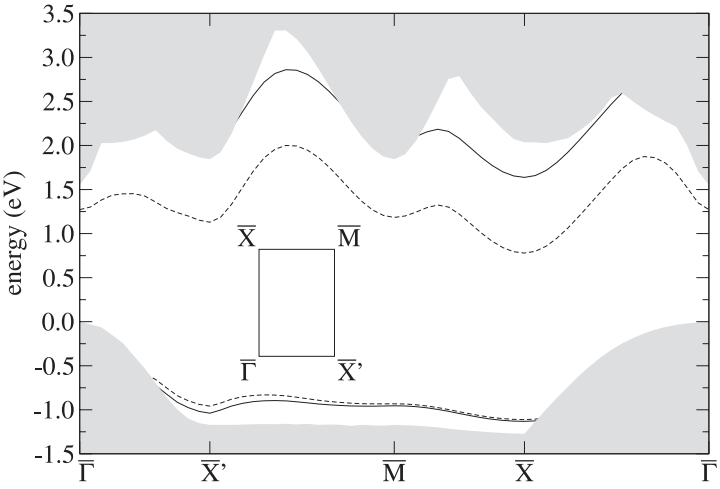

Fig. 3. Surface band structure of GaAs(110) in the LDA (*dashed lines*) and the G_0W_0 approximation (*straight lines*) at the theoretical lattice constant 5.55 Å. The *gray-shaded regions* mark the projected G_0W_0 bulk bands. The *inset* indicates the two-dimensional surface Brillouin zone

While the position of the occupied surface band relative to the valence-band edge at $\overline{\Gamma}$ remains almost unchanged, the upward shift of the unoccupied surface band (0.86 eV) slightly exceeds that of the bulk conduction bands (0.77 eV). The larger impact of the self-energy correction on the surface gap was noticed before [55, 56]; our result for the gap correction lies between the two previously published values. Small deviations of about 0.1 eV are due to differences in the implementations, for example the reliance on plasmon-pole models in [55, 56] compared to the random-phase approximation for the dynamically screened Coulomb interaction in this work. Both the occupied and the unoccupied surface bands are in close proximity to the projected bulk bands and, in fact, overlap with them in large parts of the Brillouin zone. As neither extends into the fundamental gap between the bulk valence and conduction band edges at $\overline{\Gamma}$, they do not pin the surface Fermi level.

Although prevalent in electronic-structure calculations for surfaces, the supercell approach is a drastic alteration of the system's geometry whose influence must be carefully monitored, because the occurrence of electric multipole moments may lead to artificial long-range interactions between the periodic slabs. Static dipoles, if present, can be eliminated in density-functional theory [62]. With this correction the limit of isolated slabs is quickly reached. Dynamic dipoles are always created in dielectric media, however, and contribute to the polarizability in the G_0W_0 approximation. Actually, there are two contributions that must be distinguished. The first is the dynamic polarization between the slabs, which gives rise to an additional slowly varying potential that reduces the band gap. This effect can be understood and even

quantitatively modeled in terms of classical image charges [63]; for the geometry used in this work it amounts to about 0.1 eV. The model thus allows an a-posteriori extrapolation to the limit of an isolated slab. The second contribution is the finite width of the slab, which increases the gap due to quantization effects. At present there is no obvious cure for this problem within the G_0W_0 approximation, and as the two effects counteract each other, we have not applied any partial correction to eliminate the dynamic polarization between the slabs either. In principle, both problems could be avoided by studying systems comprised of semi-infinite matter and vacuum regions. Within density-functional theory several methods have indeed been proposed for this purpose [64–67]. Their efficiency relies on the fact that a perturbation breaking the translational symmetry, such as a surface or defect, modifies the effective potential only in its immediate vicinity. In the G_0W_0 formalism this cannot be exploited to the same degree, because all propagators are explicitly nonlocal, and a much larger simulation cell must hence be taken into account. So far, only one G_0W_0 calculation for an effectively one-dimensional system, a semi-infinite jellium surface, has been reported [68].

In preparation for later applications to larger supercells, we repeated the G_0W_0 calculation with lower cutoff energies and found that the self-energy correction to the surface gap remained stable up to 10 Ry. The reason for the rapid convergence, which is well known and can be exploited to reduce the computational expense considerably, is that the kinetic energy and the electrostatic Hartree potential, the two largest contributions to the quasiparticle energies, are already included in the Kohn–Sham eigenvalues in (10); the matrix elements of the self-energy are less sensitive to the number of plane waves. Besides, we obtained essentially the same surface gap when reducing the width of the slab to four layers. For this geometry and a cutoff energy of 10 Ry, we performed test calculations in which we included up to 1049 unoccupied bands in the Green's function. With 379 bands the results are already converged within 0.02 eV, sufficient for our purpose.

4 Defect States

The geometry of the anion vacancy is illustrated in Fig. 2 for the GaAs(110) surface. The removal of the As atom leaves each of the three Ga atoms surrounding the vacancy with a dangling bond. As a consequence, the two Ga_I atoms in the first layer move downwards while the Ga_{II} atom in the second layer moves upwards and forms two new bonds with the Ga_I atoms across the void. Its coordination number thus increases from four to five, while the threefold coordination of the Ga_I atoms remains unchanged. The relaxation preserves the C_{1h} point-group symmetry, except in the positive charge state where an asymmetric distortion that pushes the unoccupied defect level in the band gap to higher energies is more favorable [6, 8]. We find that the distortion lowers the total energy by 0.17 eV for GaAs and 0.11 eV for InP.

The anion vacancy gives rise to three electronic states, all localized at the Ga_I–Ga_{II} bond pair. They are labeled as 1a′, 1a″ and 2a′, where a′ denotes states that are even with respect to the mirror plane and a″ denotes states that are odd. Although the asymmetric relaxation in the positive charge state destroys this symmetry and deforms the orbitals slightly, it leaves the order of the states intact, and we continue to use the same notation for simplicity. The 1a′ state is located several eV below the valence-band maximum and thus always filled with two electrons, while the 2a′ state is too high in energy to become populated. Only the 1a″ state falls inside the fundamental band gap. Depending on the doping, it can be occupied either by zero, one or two electrons, which corresponds to the positive, neutral or negative charge state, respectively. It is important to note that the charge state influences the defect geometry, as the Ga_I–Ga_{II} bonds contract with increasing electron occupancy, reflecting the bonding character of the 1a″ state. In the following we examine the position of this defect level for a given charge state; the question of which charge state is preferred under specific conditions, such as doping, is answered in the next section.

For the density-functional calculations we choose a supercell consisting of 2×4 surface unit cells and six atomic layers. For charged systems we include a uniform charge density with opposite sign in order to compensate the extra electron or hole and restore overall neutrality in the supercell. Instead of a well-defined defect state, the supercell periodicity gives rise to an artificial dispersion as illustrated in Fig. 4. At the surface of each slab the defects form a rectangular grid whose lattice parameters a_x and a_y equal the dimensions of the supercell. Since the 1a″ state is odd with respect to the mirror plane, one can regard it as a p orbital that exhibits π-type bonding along the [001] and σ-type bonding along the $[\bar{1}10]$ direction. We hence consider a tight-binding model

$$\begin{aligned}\varepsilon_{\boldsymbol{k}}^{\mathrm{KS}} = {} & \varepsilon_{1\mathrm{a}''}^{\mathrm{KS}} + 2V_{1\pi}^{\mathrm{KS}} \cos(k_x a_x) + 2V_{1\sigma}^{\mathrm{KS}} \cos(k_y a_y) \\ & + V_2^{\mathrm{KS}} \cos(k_x a_x) \cos(k_x a_x) + 2V_{3\pi}^{\mathrm{KS}} \cos(2k_x a_x) + 2V_{3\sigma}^{\mathrm{KS}} \cos(2k_y a_y)\end{aligned} \tag{11}$$

with parameters fitted to the calculated Kohn–Sham band, where $\varepsilon_{1\mathrm{a}''}^{\mathrm{KS}}$ equals the eigenvalue of a single defect and the other parameters have the meaning of hopping integrals. The above expression includes interactions up to third-nearest neighbors and reproduces the dispersion with a correlation coefficient close to 0.999 9 for all systems under consideration.

As the G_0W_0 formalism involves nonlocal propagators, the amount of data that must be processed grows rapidly with the system size. In order to limit the computational expense we use a smaller (2×2) supercell and four atomic layers instead of the (2×4) cell to determine the self-energy correction of the 1a″ state. The stronger defect–defect interaction along the $[\bar{1}10]$ direction increases the dispersion but does not change it qualitatively. As the presence of the defect does not modify the range of the nonlocal propagators appre-

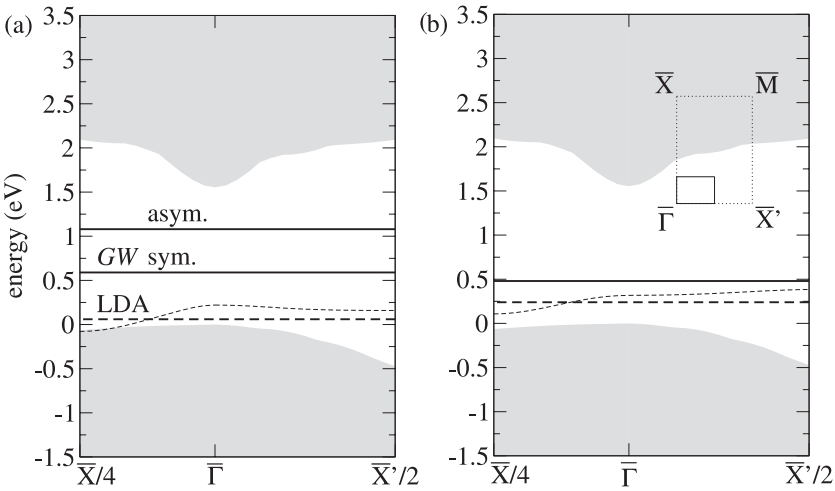

Fig. 4. Calculated 1a″ defect level of the As vacancy at GaAs(110) in the (**a**) positive and (**b**) negative charge state (*dashed line*: LDA; *solid line*: G_0W_0). The artificial dispersion in the LDA is due to the supercell periodicity; the *horizontal lines* mark the actual defect level $\varepsilon^{\mathrm{KS}}_{1a''}$ and the corresponding G_0W_0 result. As the calculations in (**a**) refer to the constrained symmetric relaxation, the additional shift due to the asymmetric distortion is shown separately. The *inset* in (**b**) indicates the downfolded Brillouin zone for the (2×4) supercell

ciably, we use one $\boldsymbol{k}$-point and 1500 unoccupied bands for the construction of the Green's function G_0, which corresponds to the same Brillouin-zone sampling as the four $\boldsymbol{k}$-points and 379 bands that we found satisfactory for the (1×1) unit cell of the defect-free surface. For the positively charged As vacancy at GaAs(110) we evaluated the self-energy correction in the entire Brillouin zone [15]. By relating the calculated quasiparticle dispersion to a tight-binding model equivalent to (11), we found very similar values for the parameters $V_{3\pi}$ and $V_{3\sigma}$ as in the LDA, which implies that the influence of the third-nearest neighbors on the self-energy correction is negligible. This observation can be exploited as follows. At $\boldsymbol{k}' = (2\pi/4)(1/a_x, 1/a_y, 0)$ the contributions from the first- and second-nearest neighbors vanish, so that the Kohn–Sham eigenvalue is given by

$$\varepsilon^{\mathrm{KS}}_{\boldsymbol{k}'} = \varepsilon^{\mathrm{KS}}_{1a''} - 2V^{\mathrm{KS}}_{3\pi} - 2V^{\mathrm{KS}}_{3\sigma}\,, \tag{12}$$

and the corresponding quasiparticle energy by

$$\varepsilon_{\boldsymbol{k}'} = \varepsilon_{1a''} - 2V_{3\pi} - 2V_{3\sigma}\,. \tag{13}$$

If the surviving hopping integrals on the right-hand sides coincide, then the identity $\varepsilon_{1a''} - \varepsilon^{\mathrm{KS}}_{1a''} = \varepsilon_{\boldsymbol{k}'} - \varepsilon^{\mathrm{KS}}_{\boldsymbol{k}'}$ holds. On the other hand, the self-energy correction is defined through the relation

$$\varepsilon_{\boldsymbol{k}'} - \varepsilon^{\mathrm{KS}}_{\boldsymbol{k}'} = \langle \varphi^{\mathrm{KS}}_{\boldsymbol{k}'} | \Sigma(\varepsilon^{\mathrm{KS}}_{\boldsymbol{k}'}) - V_{\mathrm{xc}} | \varphi^{\mathrm{KS}}_{\boldsymbol{k}'} \rangle \,. \tag{14}$$

Instead of a scan over the whole Brillouin zone, we can hence determine the self-energy correction for an isolated defect from a single calculation at $\boldsymbol{k}'$. The quasiparticle results are obtained by adding this correction to the Kohn–Sham value $\varepsilon^{\mathrm{KS}}_{1a''}$ of the larger (2×4) supercell, where the position of the latter relative to the valence-band maximum can be established more accurately. We estimate that the uncertainty of the final quasiparticle energies, which results from the discrete $\boldsymbol{k}$-point sampling, the approximate treatment of the core–valence interaction in the pseudopotential approach, the finite size of the supercell, and other convergence factors amounts to 0.1 eV to 0.2 eV.

We performed explicit G_0W_0 calculations for the positive and negative charge states, in the first case using a constrained symmetric relaxation. The results for GaAs(110) are displayed in Fig. 4. For the neutral anion vacancy the supercell contains an odd number of electrons; in combination with the requirement of spin degeneracy this leads to fractional occupation numbers in each spin channel. At present we cannot treat such systems. For the positive charge state with the proper asymmetric distortion the defect level in the G_0W_0 approximation is not calculated directly but deduced as follows. As the lowest unoccupied state, the 1a″ defect level equals the electron affinity, i.e., $\varepsilon_{1a''} = E^{\mathrm{vac}}(0, Q^{\mathrm{asym}}_+) - E^{\mathrm{vac}}(+, Q^{\mathrm{asym}}_+)$, where $E^{\mathrm{vac}}(q, Q)$ denotes the total energy of a surface featuring an anion vacancy with the actual electron population $q \in \{+, 0, -\}$ and a geometric structure optimized for the charge state $Q \in \{Q^{\mathrm{asym}}_+, Q^{\mathrm{sym}}_+, Q_0, Q_-\}$. This expression can be rewritten as

$$\begin{aligned}\varepsilon_{1a''} = {} & \left[E^{\mathrm{vac}}(0, Q^{\mathrm{asym}}_+) - E^{\mathrm{vac}}(0, Q^{\mathrm{sym}}_+)\right] \\ & + \left[E^{\mathrm{vac}}(0, Q^{\mathrm{sym}}_+) - E^{\mathrm{vac}}(+, Q^{\mathrm{sym}}_+)\right] \\ & + \left[E^{\mathrm{vac}}(+, Q^{\mathrm{sym}}_+) - E^{\mathrm{vac}}(+, Q^{\mathrm{asym}}_+)\right]\end{aligned} \tag{15}$$

by adding and subtracting intermediate configurations. The term in the second line on the right-hand side equals the quasiparticle energy for the corresponding symmetric relaxation, which can be calculated with less computational cost by exploiting the C_{1h} point-group symmetry. The other two terms are simple total-energy differences between the symmetric and the asymmetric geometry for a constant number of electrons; both are positive and can be obtained from density-functional theory.

The calculated defect levels are summarized in Table 1 for GaAs(110) and Table 2 for InP(110). Our LDA results are consistent with most of the previously published calculations [3, 5, 7]. A notable exception are the values by *Zhang* and *Zunger* [3] for the negative charge state of the As vacancy at GaAs(110) and for the constrained symmetric geometry of the positive charge

Table 1. Calculated defect levels for the As vacancy at GaAs(110) in eV relative to the valence-band maximum. For the positive charge state the first column refers to the asymmetric and the second to the constrained symmetric relaxation. The quasiparticle band gap of 1.55 eV in this work, calculated at the theoretical lattice constant 5.55 Å, is close to the experimental value of 1.52 eV

Charge state	(+1)		(0)	(−1)
LDA (this work)	0.70	(0.06)	0.13	0.24
LDA [3]	0.73	(0.41)		0.5
LDA [5]		(0.06)	0.23	0.24
G_0W_0 (this work)	1.08	(0.59)		0.48

Table 2. Calculated defect levels for the P vacancy at InP(110) in eV relative to the valence-band maximum. For the positive charge state the first column refers to the asymmetric and the second to the constrained symmetric relaxation. The quasiparticle band gap of 1.52 eV in this work, calculated at the theoretical lattice constant 5.81 Å, slightly exceeds the experimental value of 1.42 eV

Charge state	(+1)		(0)	(−1)
LDA (this work)	0.89	(0.39)	0.49	0.60
LDA [7]		(0.326)	0.479	0.580
G_0W_0 (this work)	1.36	(0.91)		1.01

state in Table 1. After a full relaxation of the asymmetric distortion in the latter case, the discrepancy vanishes, however. The agreement with the results of *Kim* and *Chelikowsky* [5] for the same system is very good, except for a difference of 0.1 eV for the neutral charge state. The origin of these deviations cannot be traced, because both groups of authors give very little information about their computational details, and the size of the corresponding Kohn–Sham eigenvalue gap is not stated. For all configurations with symmetric geometries the defect level moves steadily upwards with increasing electron population. The asymmetric distortion has a large effect on the unoccupied defect level and pushes it to significantly higher energies. Concomitant with the opening of the fundamental band gap, the G_0W_0 approximation adds a positive quasiparticle correction to all defect levels and predicts larger values than the LDA for all charge states. The size of the self-energy shift depends both on the charge state and the geometry, although the differences are of the same order as the error bar of the calculation. We note that our numerical values differ slightly from those reported earlier in [15]. The discrepancy is due to an improved description of the anisotropic screening in the slab geometry in this work but lies within the estimated overall error bar of the calculation.

For the P vacancy at InP(110) we obtain a similar picture, although the defect levels are at higher energies both in the LDA and the G_0W_0 approximation.

Unfortunately, due to experimental difficulties, no direct measurements of the defect states by photoemission are available. In this situation one can only try to extract values from indirect methods like surface photovoltage imaging with STM. One study using this technique claimed a value of $0.62 \pm 0.04\,\mathrm{eV}$ for the As vacancy in the positive charge state, based on the known position of the sample's Fermi energy 0.09 eV above the valence-band maximum and a local band bending of 0.53 eV [69]. An STM measurement on a different sample found a band bending of only 0.1 eV, however [12]. The origin of this discrepancy is unclear but points to a strong influence of experimental conditions and/or sample quality. Because of this uncertainty, no experimental values are included in the tables.

5 Charge-Transition Levels

The occurrence of different charge states is a direct consequence of the position of the defect level inside the fundamental band gap: in p-doped materials the defect state lies above the Fermi energy and is hence depopulated, whereas it is fully occupied in n-doped materials with a higher Fermi energy close to the conduction-band edge. Indeed, a charge state of $(+1)$ has been confirmed experimentally for anion vacancies at p-GaAs(110) [70] and p-InP(110) [71] and a charge state of (-1) at n-GaAs(110) [72] and n-InP(110) [73]. In a theoretical treatment the stable charge state can be identified by comparing the different formation energies

$$E^{\mathrm{form}}(q, \mu_{\mathrm{A}}, \varepsilon_{\mathrm{F}}) = E^{\mathrm{vac}}(q, Q_q) + \mu_{\mathrm{A}} + q\varepsilon_{\mathrm{F}} - E^{\mathrm{surf}}\,, \tag{16}$$

where we use the same notation for the total energy of the vacancy as in the previous section. The chemical potential μ_{A} of the anion atoms is controlled by the partial pressure and temperature, ε_{F} denotes the Fermi level and E^{surf} is the total energy of the defect-free surface. The qualitative behavior of the formation energies is illustrated in Fig. 5: due to their different slopes the stable charge state with the lowest formation energy changes from positive to negative as the Fermi energy varies between the valence-band maximum and the conduction-band minimum. In between there may be a region where the neutral vacancy is stable. The charge-transition levels are defined as the values of the Fermi energy where the curves intersect and the stable charge state changes. They are given explicitly by $\varepsilon^{+/0} = E^{\mathrm{vac}}(0, Q_0) - E^{\mathrm{vac}}(+, Q_+)$ for the transition from $q = +1$ to $q = 0$ and $\varepsilon^{0/-} = E^{\mathrm{vac}}(-, Q_-) - E^{\mathrm{vac}}(0, Q_0)$ for the transition from $q = 0$ to $q = -1$.

All quantities in (16) are ground-state energies and can, in principle, be calculated within density-functional theory. As explained above, however, the parametrizations most commonly used in practical implementations lack

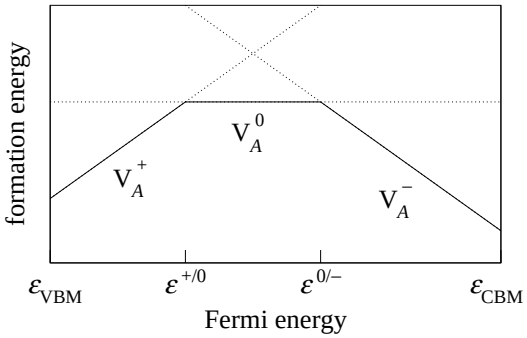

Fig. 5. Qualitative behavior of the formation energies for anion vacancies in the positive (V_A^+), neutral (V_A^0), and negative (V_A^-) charge state. The Fermi energy is limited by the valence-band maximum and the conduction-band minimum. The charge-transition levels $\varepsilon^{+/0}$ and $\varepsilon^{0/-}$ mark the values of the Fermi energy where the stable charge state with the lowest formation energy changes

essential properties of the exact exchange-correlation functional, so that the charge-transition levels obtained in this way suffer from systematic errors. For a more accurate quantitative description we use the same trick as in (15) and rewrite $\varepsilon^{+/0}$ as

$$\varepsilon^{+/0} = [E^{\mathrm{vac}}(0, Q_0) - E^{\mathrm{vac}}(0, Q_+)] + [E^{\mathrm{vac}}(0, Q_+) - E^{\mathrm{vac}}(+, Q_+)] \quad (17)$$

by adding and subtracting the total energy $E^{\mathrm{vac}}(0, Q_+)$ of a configuration with the atomic geometry of the positively charged vacancy but with one extra electron. In this way the charge-transition level is naturally decomposed into two distinct contributions. The first term on the right-hand side is purely structural and describes the relaxation energy of the neutral system from the atomic structure optimized for the positive charge state to its own equilibrium geometry. It is always negative and can be calculated within density-functional theory. We obtain $-0.59\,\mathrm{eV}$ for GaAs(110) and $-0.54\,\mathrm{eV}$ for InP(110) when taking the asymmetric distortion into account. The second term on the right-hand side is purely electronic and equals the electron affinity of the positively charged vacancy, which in a many-body framework corresponds to the lowest unoccupied state, i.e., the empty $1a''$ defect level in the band gap. This was already calculated within the G_0W_0 approximation in the preceding section and can be taken from Tables 1 and 2. In the same spirit we rewrite $\varepsilon^{0/-}$ as

$$\varepsilon^{0/-} = [E^{\mathrm{vac}}(-, Q_-) - E^{\mathrm{vac}}(0, Q_-)] + [E^{\mathrm{vac}}(0, Q_-) - E^{\mathrm{vac}}(0, Q_0)] \; . \quad (18)$$

The first term now equals the ionization potential of the negatively charged vacancy, corresponding to the highest occupied quasiparticle state. Again this is the $1a''$ defect level, which in this case is filled with two electrons, and the G_0W_0 results can be taken from the tables in the previous section.

The second term is the energy difference between the electrically neutral system in the atomic structure optimized for the positive charge state and its own relaxed geometry. This contribution is always positive and can be obtained from density-functional theory. We obtain 0.12 eV for GaAs(110) and 0.08 eV for InP(110). The structural component is much smaller than for $\varepsilon^{+/0}$ because there is no symmetry-breaking distortion in this case, only a minor reduction of the Ga–Ga bond length across the vacancy.

Incidentally, the charge-transition levels within the LDA, which are usually obtained according to the definitions $\varepsilon^{+/0} = E^{\mathrm{vac}}(0, Q_0) - E^{\mathrm{vac}}(+, Q_+)$ and $\varepsilon^{0/-} = E^{\mathrm{vac}}(-, Q_-) - E^{\mathrm{vac}}(0, Q_0)$ by evaluating the total-energy differences directly, can also be decomposed into structural and electronic energy contributions. The latter are *not* given by the Kohn–Sham eigenvalues in Tables 1 and 2, however. Instead, they must be calculated with the help of the transition-state theorem [18] from intermediate configurations with half-integer occupation numbers of 1/2 and 3/2 electrons, respectively.

Table 3. Calculated charge-transition levels for the As vacancy at GaAs(110) in eV relative to the valence-band maximum. For $\varepsilon^{+/0}$ the first column refers to the asymmetric and the second to the constrained symmetric relaxation

Transition level	$\varepsilon^{+/0}$		$\varepsilon^{0/-}$
LDA (this work)	0.24	(0.07)	0.15
LDA [3]	0.32		0.4
LDA [5]		(0.10)	0.24
G_0W_0 (this work)	0.49	(0.32)	0.60

In Table 3 we summarize the results for the As vacancy at GaAs(110). In contrast to earlier studies [3, 5] that found a small energy window in which the neutral charge state is stable, our own calculation at the level of the LDA predicts $\varepsilon^{+/0} > \varepsilon^{0/-}$ if the correct asymmetric distortion is taken into account and hence a direct transition from the positive to the negative charge state, but the small energy difference is within the uncertainty of the calculation. The G_0W_0 approximation, on the other hand, reverses this ordering and simultaneously moves all charge-transition levels to higher energies. The values for the constrained symmetric relaxation are merely shown for the purpose of comparison with earlier work. The deviations from earlier LDA studies, especially by *Zhang* and *Zunger* [3], are related to differences of similar magnitude in the defect states, which were already mentioned above.

The results for the P vacancy at InP(110) are given in Table 4. For this material an indirect experimental measurement of $\varepsilon^{+/0}$, obtained with a combination of scanning tunneling microscopy and photoelectron spectroscopy, is available [6]. Consistent with previous studies [6, 7], we find that the LDA significantly underestimates the experimentally deduced value. The larger

Table 4. Calculated charge-transition levels for the P vacancy at InP(110)

Transition level	$\varepsilon^{+/0}$		$\varepsilon^{0/-}$
LDA (this work)	0.47	(0.39)	0.54
LDA [6]	0.52	(0.45)	
LDA [7]	0.388		0.576
G_0W_0 (this work)	0.82	(0.79)	1.09
Expt. [6]	0.75 ± 0.1		

G_0W_0 result, however, lies within the error bar of the experimental measurement. We take this as a confirmation of our approach and an indicator that the other defect states and charge-transition levels calculated within the same framework are also meaningful. Nevertheless, further calculations for different systems are necessary to establish the general validity of this scheme.

6 Summary

We have presented a parameter-free method for calculating defect states and charge-transition levels of point defects in semiconductors. Compared to previous studies that extracted these quantities directly from the self-consistent iteration of the Kohn–Sham equations, it apparently corrects important errors that are inherent in all jellium-based exchange-correlation functionals and employs a separation of structural and electronic energy contributions. While the former are accurately obtainable in density-functional theory, we use many-body perturbation theory and the G_0W_0 approximation for the self-energy to calculate the latter with proper quasiparticle corrections. The scheme is general and can be applied to arbitrary bulk or surface defects. As an example we examined the electronic structure of anion vacancies at the (110) surfaces of III–V semiconductors. For the As vacancy at GaAs(110) our calculation indicates that all three charge states including the neutral configuration are stable, in contrast to the LDA that predicts a direct transition from the positive to the negative charge state. Due to a general lack of experimental data, a direct comparison between theoretical and experimental values is only possible for the charge-transition level $\varepsilon^{+/0}$ of the P vacancy at InP(110). In this case our calculation is in good agreement with the experimentally deduced result and constitutes a clear improvement over previous LDA treatments. Nevertheless, besides an improvement of experimental techniques, further developments in theoretical and computational procedures are highly desirable. Our method only opens the door to novel approaches; with present implementations the results are derived at a cost that may be infeasible for more complex systems, and the numerical uncertainty of 0.1 eV to 0.2 eV is often of the same order as the relevant energy differences. The

study of excitons, which, e.g., play an important role at GaAs(110) [74], further requires an extension of the mathematical framework beyond single quasiparticles.

Acknowledgements

We thank Magnus Hedström, Günther Schwarz, and Jörg Neugebauer for fruitful collaborations in the course of this work and Philipp Ebert for useful discussions. This work was funded in part by the EU through the Nanophase Research Training Network (Contract No. HPRN-CT-2000-00167) and the Nanoquanta Network of Excellence (Contract No. NMP-4-CT-2004-500198).

References

[1] P. Ebert: Curr. Opin. Solid State Mater. Sci. **5**, 211 (2001) 165
[2] V. P. LaBella, H. Yang, D. W. Bullock, P. M. Thibado, P. Kratzer, M. Scheffler: Phys. Rev. Lett. **83**, 2989 (1999) 165
[3] S. B. Zhang, A. Zunger: Phys. Rev. Lett. **77**, 119 (1996) 166, 182, 183, 186
[4] H. Kim, J. R. Chelikowsky: Phys. Rev. Lett. **77**, 1063 (1996) 166
[5] H. Kim, J. R. Chelikowsky: Surf. Sci. **409**, 435 (1998) 166, 182, 183, 186
[6] P. Ebert, K. Urban, L. Aballe, C. H. Chen, K. Horn, G. Schwarz, J. Neugebauer, M. Scheffler: Phys. Rev. Lett. **84**, 5816 (2000) 166, 167, 170, 179, 186, 187
[7] M. C. Qian, M. Göthelid, B. Johansson, S. Mirbt: Phys. Rev. B **66**, 155326 (2002) 166, 182, 183, 186, 187
[8] M. C. Qian, M. Göthelid, B. Johansson, S. Mirbt: Phys. Rev. B **67**, 035308 (2003) 166, 179
[9] P. Hohenberg, W. Kohn: Phys. Rev. **136**, B864 (1964) 166, 168
[10] W. Kohn, L. J. Sham: Phys. Rev. **140**, A1133 (1965) 166, 168, 169
[11] J. P. Perdew, K. Burke, M. Ernzerhof: Phys. Rev. Lett. **77**, 3865 (1996) 166
[12] G. Lengel, R. Wilkins, G. Brown, M. Weimer: Phys. Rev. Lett. **72**, 836 (1994) 166, 168, 184
[13] A. Görling: Phys. Rev. A **54**, 3912 (1996) 166, 171
[14] V. Fiorentini, M. Methfessel, M. Scheffler: Phys. Rev. B **47**, 13353 (1993) 167
[15] M. Hedström, A. Schindlmayr, M. Scheffler: phys. stat. sol. (b) **234**, 346 (2002) 167, 181, 183
[16] L. Sorba, V. Hinkel, H. U. Middelmann, K. Horn: Phys. Rev. B **36**, 8075 (1987) 168
[17] B. Reihl, T. Riesterer, M. Tschudy, P. Perfetti: Phys. Rev. B **38**, 13456 (1988) 168
[18] J. C. Slater: Adv. Quant. Chem. **6**, 1 (1972) 169, 173, 186
[19] J. F. Janak: Phys. Rev. B **18**, 7165 (1978) 169, 173
[20] C.-O. Almbladh, U. von Barth: Phys. Rev. B **31**, 3231 (1985) 170
[21] J. P. Perdew, M. Levy: Phys. Rev. Lett. **51**, 1884 (1983) 170
[22] L. J. Sham, M. Schlüter: Phys. Rev. Lett. **51**, 1888 (1983) 170
[23] A. Görling, M. Levy: Phys. Rev. A **50**, 196 (1994) 170

[24] A. Görling: Phys. Rev. B **53**, 7024 (1996) 170
[25] M. Städele, J. A. Majewski, P. Vogl, A. Görling: Phys. Rev. Lett. **79**, 2089 (1997) 170, 171
[26] R. W. Godby, M. Schlüter, L. J. Sham: Phys. Rev. B **37**, 10159 (1988) 170
[27] T. Kotani: J. Phys.: Condens. Matter **10**, 9241 (1998) 170
[28] J. P. Perdew, A. Zunger: Phys. Rev. B **23**, 5048 (1981) 171
[29] D. M. Ceperley, B. J. Alder: Phys. Rev. Lett. **45**, 566 (1980) 171
[30] L. Kleinman, D. M. Bylander: Phys. Rev. Lett. **48**, 1425 (1982) 171
[31] M. Bockstedte, A. Kley, J. Neugebauer, M. Scheffler: Comput. Phys. Commun. **107**, 187 (1997) 171
[32] M. Fuchs, M. Scheffler: Comput. Phys. Commun. **119**, 67 (1999) 171
[33] J. L. A. Alves, J. Hebenstreit, M. Scheffler: Phys. Rev. B **44**, 6188 (1991) 171, 176
[34] O. Madelung, U. Rössler, M. Schulz (Eds.): *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology – New Series*, vol. III/41A (Springer, Berlin, Heidelberg 2001) 171, 176
[35] G. D. Mahan: *Many-Particle Physics*, 3rd ed. (Springer, Berlin, Heidelberg 2000) 171
[36] L. Hedin: Phys. Rev. A **139**, 796 (1965) 172, 174
[37] W. von der Linden, P. Horsch: Phys. Rev. B **37**, 8351 (1988) 172
[38] G. E. Engel, B. Farid: Phys. Rev. B **47**, 15931 (1993) 172
[39] M. S. Hybertsen, S. G. Louie: Phys. Rev. B **34**, 5390 (1986) 173
[40] I. D. White, R. W. Godby, M. M. Rieger, R. J. Needs: Phys. Rev. Lett. **80**, 4265 (1998) 173
[41] M. Rohlfing, N.-P. Wang, P. Krüger, J. Pollmann: Phys. Rev. Lett. **91**, 256802 (2003) 173
[42] O. Pulci, F. Bechstedt, G. Onida, R. Del Sole, L. Reining: Phys. Rev. B **60**, 16758 (1999) 173
[43] A. Schindlmayr, P. García-González, R. W. Godby: Phys. Rev. B **64**, 235106 (2001) 173
[44] B. Holm, U. von Barth: Phys. Rev. B **57**, 2108 (1998) 174
[45] W.-D. Schöne, A. G. Eguiluz: Phys. Rev. Lett. **81**, 1662 (1998) 174
[46] W. Ku, A. G. Eguiluz: Phys. Rev. Lett. **89**, 126401 (2002) 174
[47] T. Kotani, M. van Schilfgaarde: Solid State Commun. **121**, 461 (2002) 174
[48] M. L. Tiago, S. Ismail-Beigi, S. G. Louie: Phys. Rev. B **69**, 125212 (2004) 174
[49] K. Delaney, P. García-González, A. Rubio, P. Rinke, R. W. Godby: Phys. Rev. Lett. **93**, 249701 (2004) 174
[50] C. Friedrich, A. Schindlmayr, S. Blügel, T. Kotani: Phys. Rev. B **74**, 045104 (2006) 174
[51] E. L. Shirley, X. Zhu, S. G. Louie: Phys. Rev. B **56**, 6648 (1997) 174
[52] R. W. Godby, M. Schlüter, L. J. Sham: Phys. Rev. B **35**, 4170 (1987) 174, 175
[53] X. Zhu, S. G. Louie: Phys. Rev. B **43**, 14142 (1991) 174
[54] M. Rohlfing, P. Krüger, J. Pollmann: Phys. Rev. B **48**, 17791 (1993) 174, 175
[55] X. Zhu, S. B. Zhang, S. G. Louie, M. L. Cohen: Phys. Rev. Lett. **63**, 2112 (1989) 175, 178
[56] O. Pulci, G. Onida, R. Del Sole, L. Reining: Phys. Rev. Lett. **81**, 5374 (1998) 175, 178
[57] S. J. Jenkins, G. P. Srivastava, J. C. Inkson: Surf. Sci. **254**, 776 (1996) 175

[58] M. M. Rieger, L. Steinbeck, I. D. White, H. N. Rojas, R. W. Godby: Comput. Phys. Commun. **117**, 211 (1999) 175, 177
[59] L. Steinbeck, A. Rubio, L. Reining, M. Torrent, I. D. White, R. W. Godby: Comput. Phys. Commun. **125**, 105 (2000) 175
[60] H. J. Monkhorst, J. D. Pack: Phys. Rev. B **13**, 5188 (1976) 177
[61] A. Schindlmayr: Phys. Rev. B **62**, 12573 (2000) 177
[62] J. Neugebauer, M. Scheffler: Phys. Rev. B **46**, 16067 (1992) 178
[63] P. Eggert, C. Freysoldt, P. Rinke, A. Schindlmayr, M. Scheffler: Verhandl. DPG (VI) **40**, 2/491 (2005) 179
[64] N. D. Lang, A. R. Williams: Phys. Rev. B **18**, 616 (1978) 179
[65] P. J. Braspenning, R. Zeller, A. Lodder, P. H. Dederichs: Phys. Rev. B **29**, 703 (1984) 179
[66] M. Scheffler, C. Droste, A. Fleszar, F. Máca, G. Wachutka, G. Barzel: Physica B **172**, 143 (1991) 179
[67] J. Bormet, J. Neugebauer, M. Scheffler: Phys. Rev. B **49**, 17242 (1994) 179
[68] G. Fratesi, G. P. Brivio, L. G. Molinari: Phys. Rev. B **69**, 245113 (2004) 179
[69] S. Aloni, I. Nevo, G. Haase: Phys. Rev. B **60**, R2165 (1999) 184
[70] K. J. Chao, A. R. Smith, C. K. Shih: Phys. Rev. B **53**, 6935 (1996) 184
[71] P. Ebert, X. Chen, M. Heinrich, M. Simon, K. Urban, M. G. Lagally: Phys. Rev. Lett. **76**, 2089 (1996) 184
[72] C. Domke, P. Ebert, K. Urban: Phys. Rev. B **57**, 4482 (1998) 184
[73] P. Ebert, K. Urban, M. G. Lagally: Phys. Rev. Lett. **72**, 840 (1994) 184
[74] O. Pankratov, M. Scheffler: Phys. Rev. Lett. **75**, 701 (1995) 188

Index

Multiscale Modeling of Defects in Semiconductors: A Novel Molecular-Dynamics Scheme

Gábor Csányi[1], Gianpietro Moras[2], James R. Kermode[1], Michael C. Payne[1], Alison Mainwood[2], and Alessandro De Vita[2,3]

[1] Cavendish Laboratory, University of Cambridge, Madingley Road, CB3 0HE, United Kingdom
{gc121,jrk33,mcp1}@cam.ac.uk
[2] Department of Physics, King's College London, Strand, London, United Kingdom
{gianpietro.moras,alison.mainwood,alessandro.de_vita}@kcl.ac.uk
[3] DEMOCRITOS National Simulation Center and CENMAT-UTS, Trieste, Italy

Abstract. Now that the modeling of simple semiconductor systems has become reliable, accurate and routine, attention is focusing on larger scale, more complex simulations. Many of these necessarily involve multiscale aspects and can only be tackled by addressing the different length scales simultaneously. We discuss some of the types of problems that require multiscale approaches. Finally we describe the LOTF (learn-on-the-fly) hybrid scheme with a series of examples to show its versatility and power.

1 Introduction

Over the last twenty years, the improvement in computational modeling of materials problems has been remarkable. This is due to the significant increase in the capacity and speed of computers matched (and arguably surpassed) by the ingenuity of those who write the computational codes. It would be invidious to do other than refer the reader to other Chapters in this volume to support these assertions. However, these advances are also challenged by the complexities of systems that need to be modeled.

There has been a great deal written about different scales of modeling, usually where the results of calculations on one scale are used to determine the parameters that are used to model the material at a larger scale. Many of the embedding approaches are designed in this way. Additionally, where dynamic processes are being modeled, it may be useful to treat different timescales by different means. A combination of molecular-dynamics and quantum-mechanical modeling using a first-principles technique for example, can model only a few hundred atoms for a period of a few picoseconds. Using more approximate methods, one may be able to extend the number of atoms to hundreds or thousands, or the time period by a few orders of magnitude. In order to extend the time period to the several seconds for some biological processes,

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors, Topics Appl. Physics **104**, 193–212 (2007)

or years or aeons for geological ones, the activation energies and reaction pathways from the static or picosecond MD calculations have to be inserted into Monte Carlo models or analytical rate equations and their variants.

In this Chapter, we suggest a selection of problems where the complexity of such an approach makes it difficult or impossible to treat every length scale with a separate calculation, largely because the physical processes on the various length scales are *strongly coupled* to each other. We then go on to describe a method that allows some of these complex problems to be tackled.

2 A Hybrid View

Radiation or Implantation Damage

An important technique for doping semiconductors is to implant the dopant with an energy (or range of energies) that determines the location of the doped area or layer. However, in this process damage in the form of vacancies and interstitials and their complexes are formed, some of which can be electrically active. Annealing may heal most of the damage (e.g., [1]), but invariably some defects remain and can be detected by photoluminescence and positron-annihilation spectroscopy [2]. Intrinsic defects have very low activation energies for diffusion, so they may migrate long distances before they are trapped by dopants, or other radiation-induced defects. The same problems arise when the radiation damage of silicon particle detectors is modeled. A very comprehensive empirical analysis of the processes involved in the annealing of damage was undertaken by *Huhtinen* [3], but he did not attempt to understand the microscopic processes. However, his analysis showed that some long-range migration of radiation damage products and displaced impurities was occurring. It is well within the current state-of-the-art to model the dynamics of the initial impact of a particle or ion with a silicon atom in a small finite cluster (or periodic supercell), and the subsequent displacement of that atom. If the damage trail formed by the particle, the knock-on atom and any further damage products can all be contained in the finite cluster or supercell, then a reasonable model of the full process can be built up. More commonly, though, one or more of the damage products may exit the cluster in a direction not easily predicted from the initial conditions, or the intrinsic defects may migrate longer distances than the dimensions of the largest clusters. The problem is then to construct a method to follow the important species such that their interactions are treated with the required level of accuracy, while keeping the whole problem to within acceptable computational limits.

Point-Defect Diffusion

A very similar problem is the annealing of point defects in semiconductors, where an additional complication is the chemical interaction between the im-

purity and the defects in the host lattice. As an example, atomic hydrogen in silicon has a stable site in a bond-centered position, which is only stable when the bond is greatly elongated – a self-trapped site [4]. It also has a metastable site at the tetrahedral site, where the migration barrier is very low. To model the migration, a full molecular-dynamics model must be used, but it must also be capable of following the long-range migration of the hydrogen through the tetrahedral-site routes. For hydrogen, a proper model using quantum mechanics for the proton as well as for the electrons is necessary, and has been developed in a limited way [5], but the principles illustrated by this example apply to the diffusion of many other impurities and defects. Classical interatomic potentials are available for some of the simpler chemical systems, but they tend to be constructed for bonding situations that are close to equilibrium, and have rather more questionable validity where the local environment of the host–defect complex is greatly distorted.

Both the relative stability and the rate of migration of point defects can be altered by the presence of strain fields and strain gradients that are ubiquitous in semiconductor systems. While the point defects typically alter the local atomic structure only on a scale of a nanometer, the strain fields of epitaxial layers extend much further due to lattice mismatch. Dislocations also give rise to a slow-decaying strain field. This means that while the quantum-mechanical treatment of new bonding arrangements near each individual defect can be accomplished using a few hundred atoms, the environment in which the diffusion and possible interaction of the defects takes place needs to be represented by tens of thousands of atoms or more.

Dislocation Motion

The strength of real materials is dominated by the behavior of its dislocations. Around the core of a dislocation, there is a 1-D region in which the bonding of the host is distorted. Once there is a kink in the dislocation, by which means it moves, the distortions become much more pronounced, and this part of the crystal must be modeled by particularly accurate techniques, see Fig. 1 for a schematic view. Moreover, the movement of the kink means that this highly distorted region also moves, which typically involves bond breaking and forming. More complexity arises when dislocations interact with each other or with grain boundaries, and they may also be sources (or sinks) for point defects.

Grain Boundaries

Nanocrystalline silicon and other semiconductors are becoming more and more technologically important. Similarly, the crystal quality of many of the newer wide-bandgap semiconductors used in electronics (the nitrides, silicon carbide, diamond) is not nearly as perfect as silicon – it is no longer acceptable to ignore dislocations and grain boundaries when trying to understand the

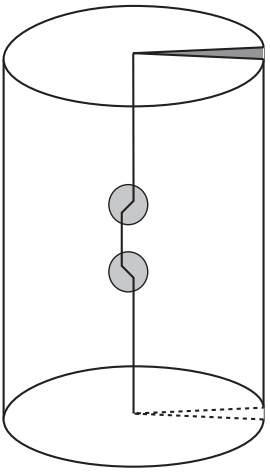

Fig. 1. Schematic view of a screw dislocation with a double kink. In covalent materials, dislocations move by the formation of double kinks. *Gray circles* approximately indicate the region in which active rebonding takes place

degradation or limitations of their electronic properties. The exact role that grain boundaries play in plasticity is far from being completely understood. In particular, the grain boundaries need to be investigated as sources and sinks for dislocations. Significant departure from the equilibrium crystalline structure is present. This takes the form of a mixture of locally disordered bonding and longer-range elastic distortion – a situation that needs a multiscale description. Furthermore, the grain boundaries can act as traps for dopants, electrons and holes, adding to the local chemical complexity. Once again, although the longer-range interactions are well described by empirical potentials, the description of the local chemistry requires quantum-mechanical accuracy.

Fracture

Fracture is perhaps the textbook example of a multiscale problem. The importance of understanding both the catastrophic failure of brittle materials and the conditions under which normally ductile materials become brittle cannot be overemphasized. Classical continuum models of the stress field do an excellent job in predicting the enhancement of stress near the crack tip. However, the divergence of components of the stress tensor at the tip imply that elasticity theory must break down in this region since it is not applicable when distortions go beyond the harmonic range. Indeed, during failure the interatomic bonds are stretched well into the anharmonic region and ultimately are broken. In covalent systems this leaves dangling bonds, whose energetics can only be described by a fully quantum-mechanical treatment. Moreover,

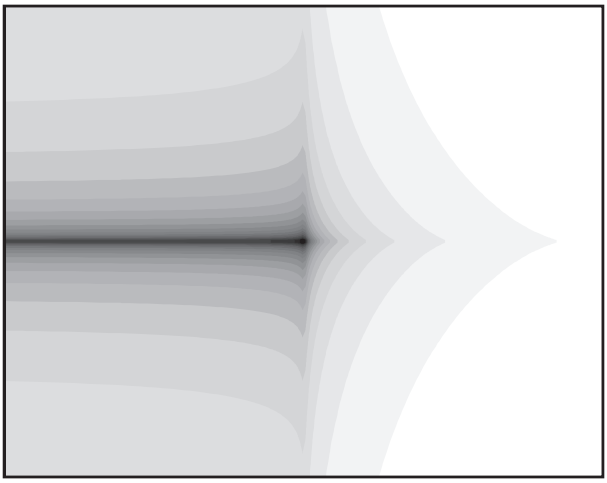

Fig. 2. Displacement field of a propagating crack under uniaxial tension in the opening mode (analytical continuum solution)

the influence of impurities and corrosive agents on failure processes has never been explored theoretically at this level of detail, yet there is evidence for the critical role they play, e.g., in fracture of silicon notches in a wet environment [6, 7] and in the controlled fabrication of thin Si films by hydrogen implantation [8–10].

The key aspect of brittle fracture that makes it amenable to hybrid modeling is that the propagating crack is atomically sharp [11]. Thus, although the crack surface that is opening up is two-dimensional, the active region, where bonds are being broken, is a one-dimensional line, perpendicular to the direction of propagation. The strongly coupled multiscale aspect of crack propagation comes about because the opening crack gives rise to a stress field that has a singularity (in the continuum approximation), diverging as $1/\sqrt{r}$ where r is the distance ahead of the crack tip, and in turn, it is this large stress that breaks the bonds and thus advances the crack forward. The extension of the stress field is nevertheless very large. If we wish to represent all the atoms that contribute significantly to elastic relaxation as the stress field advances, we need to include several tens of thousands of atoms. Only then can we hope to be able to quantitatively predict critical loading levels.

3 Hybrid Simulation

The general pretext in a hybrid simulation is that we are going to use two different physical models to describe the system. It is understood that one of these is accurate enough to describe the physics of the activated regions properly, which, in practice, means we will use some sort of quantum-mechanical

description, e.g., a tight-binding (TB) formalism, or indeed density-functional theory (DFT); henceforth this first model will be referred to as the *QM model*. Using such sophisticated modeling comes at a price, in the very real terms of computing costs. With the currently available hardware, it is typically possible to simulate thousands of atoms using TB, and hundreds of atoms using DFT for tens of picoseconds. These are evidently modest capabilities if we are faced with any of the above problems, and hence we need to employ a second model, to be used in parts of the system that are not activated at a given instant in time. This second model should capture correctly the basic topology of bonding and the response to small deformations, while at the same time be relatively inexpensive to compute. Empirical atomistic models fit the bill perfectly. It should be noted that the recent progress in the field of empirical "ball and spring"-type modeling has been in the direction of refining the analytical forms in an attempt to extend their range of applicability [12–15]. However, once we have decided to employ a hybrid technology there is little benefit in complicating this second, *classical model*. It is more important that it is as robust as possible and uses only a small number of free parameters since in any particular situation where the classical model would fail we will use the QM model.

In theory, one could consider not just two models, but a whole hierarchy of ever more coarse-grained descriptions of physical systems. The next one up from classical atomistic models would be a continuum finite element model. In our experience, the cost ratio between the QM and the classical models is so extreme, that for all intents and purposes the classical atoms can be regarded as "free" in most applications. If we are at liberty to consider millions of classical atoms, the necessity of a continuum description is less pressing. The standard technical tools of classical molecular-dynamics can thus be used on this large, atomistically modeled region to impose the correct elastic or thermal boundary conditions of the problem under study. The boundary regions between the QM and classical zones need, on the contrary, a special treatment and a separate discussion.

Boundary Problems

It is not hard to recognize that the biggest challenge in designing a hybrid simulation scheme is to overcome the problems associated with the artificial boundary that separates regions of the system that are described by the two different models. In fact there are two distinct issues that have to be addressed. The first one generally goes under the name of *termination*, and it concerns the accurate computation of forces on atoms near the boundary. If we attempt the most naive partitioning and simply omit the classical atoms from the quantum-mechanical calculation and vice versa, the atoms close to the boundary will behave like those near an open surface. This is clearly wrong. To get accurate forces on these atoms, we should "trick" the system into believing that these surface atoms are really bulk atoms, while avoiding

implicit inclusion of all the atoms from the other region (which is the point of the hybrid simulation in the first place).

The next simplest solution is to use so-called "termination atoms", which is standard practice in the hybrid simulation schemes of biological systems (called "QM/MM" models [16]). Here, atoms from the region we are presently not considering are also taken away, but wherever covalent bonds are broken by this removal, a monovalent atom (typically a hydrogen) is placed to saturate the dangling bond. For the classical model, this procedure can be seen to be sufficient. The classical description of covalent bonding is very near-sighted, i.e., the energy of the "last" classical atom can be essentially perfectly recovered using the terminator-atom technique. Moreover, since the classical total energy is usually expressed as a simple sum of atomic energies, the terminator atom can simply be excluded from this sum. It has to be noted that if we are dealing with a strongly ionic material, even our classical model will not be so near-sighted, and the problems that are usually associated with the termination of the quantum region in standard QM/MM treatments apply to the classical part as well.

Terminating the quantum-mechanical subsystem is more tricky. Quantum mechanics is not a nearest-neighbor model, so even if our free surface is terminated by monovalent terminator atoms, the atoms close to the boundary will feel an artificial environment. Secondly, it is more or less impossible in any QM scheme to exclude the terminator atoms from the total energy in a consistent fashion, so we have just traded artificial surface atoms for artificial terminator atoms – not an altogether satisfactory situation. These two problems are solved by the same trick: we stop worrying about the total energy and concentrate on obtaining accurate *forces* on the atoms in the QM region, which is achieved by allowing a thicker *termination region*, rather than just terminator atoms. In other words, we include a shell of nominally classical atoms (a thickness of about a nanometer is sufficient in silicon) in the QM calculation, which ensures that the forces that we compute for the QM atoms are accurate. Now in contrast to the total energy, the forces are local quantities, so it is trivial to exclude the forces on atoms in the "contaminated" termination region from the subsequent calculations: from the QM calculation, we only keep the forces on atoms in the original QM region, not the termination region.

Mechanical Matching

And thus we squarely arrive at the second major issue associated with the boundary. Suppose we have computed forces with the desired method in each region, we now have to propagate the system forward in time along its trajectory. On the two sides of the boundary, we originally resolved to compute with different models, but these yield different, and incompatible trajectories for the atoms on the two sides. The set of forces we just computed (from two different models) are not the derivatives of some total energy and do

not necessarily add up to zero for an isolated system. If we simply plugged them into an integrator formula (e.g., velocity, Verlet [17]), the trajectory may be unstable. The solution to this problem forms the basis of our recently developed method [18–20] called "learn on the fly" (LOTF): instead of using the set of forces directly, we consider a simplistic universal "ball and spring" model, but allow every spring to be different, and also to change with time. At every time step, the set of forces we computed using the classical and quantum models are used as targets in an optimization of the spring constants, which are tuned until the forces derived from the universal model match their targets. The trajectory integrator is then used on the universal model, whose forces are now consistent across the whole system. The universal model in effect "interpolates" across the boundary, closely matching the target on both sides, while maintaining global consistency.

In practice, a number of simplifications immediately arise. The target forces do not have to be computed at every time step, as the universal model can be considered an instantaneous interpolator not just in space, but also in time and so it can be used with unchanged spring constants for a few time steps without incurring a significant deviation from the true hybrid trajectory. More importantly however, the scheme is greatly simplified if the universal model is chosen to be the classical model itself! Thus an alternative and complementary point of view is that we take the classical model and remove the constraints from its parameters, allowing them to be different for each atom. We then tune them in the QM region to incorporate the quantum-mechanically accurate force information.

Another important consequence of this scheme is that we are at liberty to move the QM region in space without upsetting the stability of the simulation. The universal model will adapt to the target forces at each optimization, regardless of where and how we obtained the set of target forces. The use of a thick transition region to pad the quantum-mechanical calculation means that as we shift the quantum region in space together with its padding the Hamiltonian in the nominal quantum region changes smoothly.

4 The LOTF Scheme

The Universal Potential

We now describe in detail the LOTF scheme that is based on the above ideas. The choice of the universal force model that we will use to interpolate the quantum and classical models is crucial. There are two obvious choices. One, already alluded to above, is to use the classical model itself. This means that there would only be two force models in our scheme, a classical model with variable parameters and a quantum model. Most of the applications in the following sections have been carried out using this approach. An alternative and more flexible choice is to pick a universal force model that offers a good

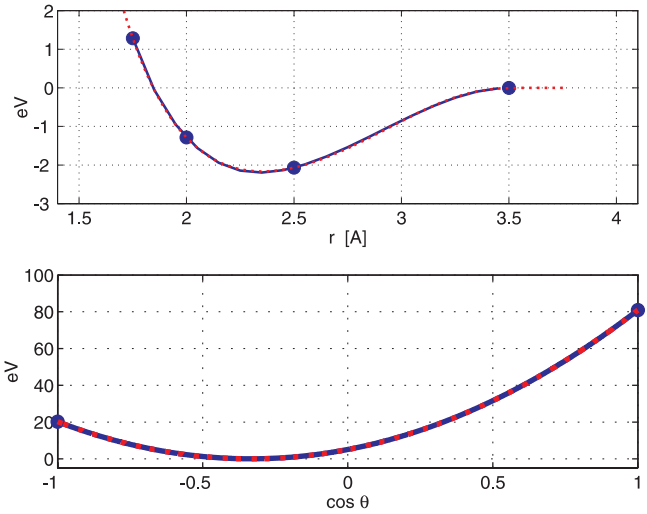

Fig. 3. Examples of cubic spline potentials for the two-body (*top*) and three-body terms (*bottom*). The *large dots* represent spline knots, the *red dotted lines* are the corresponding functions of the Stillinger–Weber potential [22]

compromise between expressive power and robustness. In particular, we will be changing the model parameters on the fly, so we would like the potential form to change smoothly as we change these parameters. In our search for the parameter set optimized at any given point in the simulation, we will make use of gradient search techniques, so the derivative of the potential *with respect to its parameters* should be straightforward and easy to evaluate. Keeping the bond lengths and bond angles as fundamental coordinates, a cubic spline functional form fits the above criteria [21].

For the two-body term, we can take a spline with four knots (e.g., at $r = \{1.5, 2.0, 2.5, 3.5\}$ Å for Si, for other elements scaled appropriately by the atomic radius). The spline function is completely determined by its values at the knot points and its derivatives at the two endpoints. We fix both the value and the derivative at the outer endpoint to be zero, and at the inner endpoint to match the corresponding values of the quantum model for a dimer. This leaves two free parameters, the value of the spline at the two inner knots. Keeping these two values negative guarantees the existence of a single minimum in the spline function between the endpoints.

For the three-body term, we could keep the traditional quadratic form in the cosine of the bond angle θ, but with a variable curvature and position of the minimum. To allow more flexibility, and in particular, an asymmetric angular dependence, we can again take a cubic spline of $\cos\theta$, with two knots at -1 and 1, the free parameters are then the values and derivatives at the endpoints. As long as we restrict the derivatives to have the appropriate sign,

we can guarantee that the spline will have a single minimum in $(-1, 1)$ for any set of values of the parameters.

Figure 3 shows a pictorial illustration of the splines. Since the spline functions are linear functions of the values at the knot points, our free parameters, evaluating the splines and their derivatives at fixed atomic positions for different parameter values is trivial and fast.

Parameters

Since we want to optimize potential parameters associated with atoms close to the quantum region, the choice of initial values is important. Unless the elastic constants of our classical and quantum models match closely, acoustic waves will partially reflect from the artificial quantum/classical boundary. Therefore, the starting values of classical parameters have to be set to match the elastic constants of our quantum model. During the dynamics, as the quantum region moves, we have two choices for atoms that are no longer in the quantum region: either we reset the classical parameters to their initial values, or we leave them with the last-fitted parameters. Which is best depends on the application. In the case of defect diffusion, once the defect leaves a particular area, it makes sense to reset the parameters. In other cases, if the physical effects in the quantum region have resulted in some permanent topological change, e.g., the opening of a new surface, possibly followed by surface reconstruction, we might want to keep the last-fitted parameters, as they could give a better description of the new topology than the original parameters. It seems difficult to formulate an optimal strategy for the general case.

Algorithm

The following sequence of steps constitutes the hybrid scheme, a graphical illustration is shown in Fig. 4.

1. **Initialize** universal force-model parameters.
2. **Extrapolate** the atomic trajectory for n steps using fixed model parameters.
3. **Identify** atoms that need quantum treatment. This is done using geometric and topological criteria, e.g., atoms are selected if they are under- or overcoordinated. For each new application, a suitable selection criterion has to be designed that captures the relevant processes. The user-defined goal of the calculation may determine the exact recipe used here. Indeed in calculations involving surface-chemical reactions in a slab geometry, we may choose to treat quantum-mechanically only one surface of the slab, thus effectively halving the accurately rendered area but doubling the simulation time for a given computer budget.

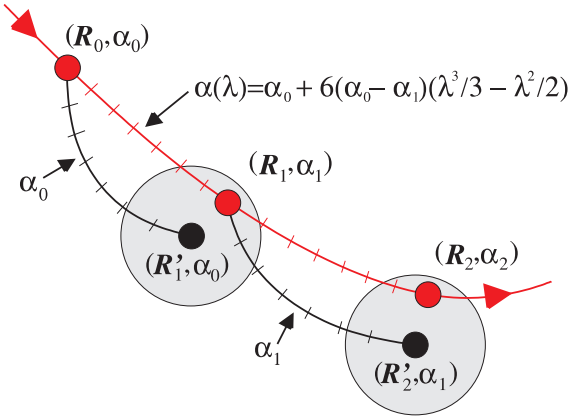

Fig. 4. Predictor–corrector-style parameter fitting. The *black* and *red lines* represent the predictor and corrector part of the trajectory, respectively. The *gray circles* denote the assumed domain of validity of the newly fitted parameters at each fit point (*black dots*)

4. **Quantum-mechanical** calculations are carried out to obtain accurate forces on the selected quantum atoms.
5. **Optimize** the force-model parameters to match the target forces both in the quantum and classical regions.
6. **Interpolate** the trajectory over the previous n steps between the old and new parameters. This achieves a smooth evolution of parameters in time.
7. **Back to 2.**

5 Applications

Point Defects

We now present a series of applications of the LOTF scheme in silicon. In all cases, the classical model is that of *Stillinger–Weber* [22] and, unless stated otherwise, the quantum model is empirical tight-binding with various parametrizations [23, 24]. The first application is more of a quantitative test of the algorithm, rather than a real application that necessitates a hybrid methodology. We consider the diffusion of point defects in crystalline Si, and show that the LOTF scheme, using a spherical QM region around the moving defect, recovers the same diffusion coefficient as that calculated from a fully quantum-mechanical trajectory, while both are significantly different from the case of a purely classical trajectory. Figure 5 shows the diffusivity as a function of temperature for a vacancy in silicon calculated in three different ways. The LOTF simulations have been performed with two different QM

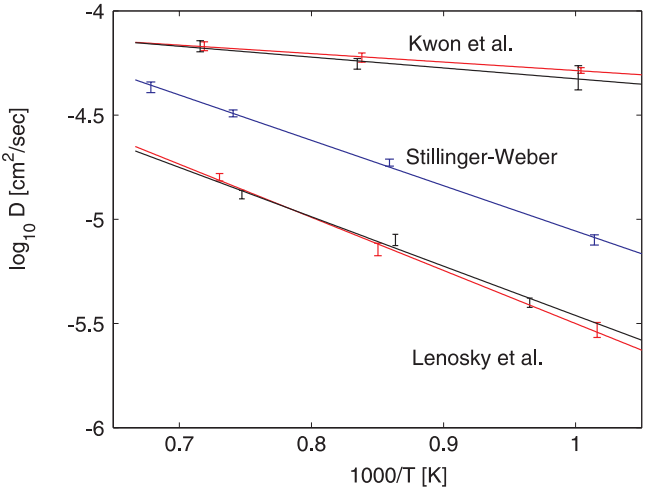

Fig. 5. Vacancy diffusion rate in Si as a function of inverse temperature, as computed using fully quantum (*black*), classical (*blue*) and hybrid (*red*) models. The unit cell had 63 atoms

packages (different orthogonal tight-binding parametrizations), which – not surprisingly – give different values. Critically however, the LOTF simulation reproduces the correct curve, corresponding to whichever QM package was used to get the target forces. Further details can be found in [19] and [20].

Figure 6 shows the mean square displacement of a hydrogen atom, diffusing in crystalline silicon at 1000 K. This test shows up an interesting aspect of the LOTF simulation. Because the algorithm retunes the classical model continuously, it effectively uses a time-dependent Hamiltonian, and thus we can no longer define a total energy that is a constant of motion. A slow (local) heating or cooling of the systems may occur during microcanonic simulations, or in constant-temperature simulations using an inappropriately tuned thermostat, in spite of the forces being all the time accurate within a few per cent. The problem is, however, lifted in the canonic ensemble if a suitable thermostat is used to stabilize the dynamics. In the present case, with two different atomic species and a large mass ratio ($\sim$ 28), a simple velocity-rescaling thermostat is unable to maintain the correct kinetic temperature for both the hydrogen and the silicon atoms. This problem can be corrected with the use of a more sophisticated thermostat. Figure 7 shows the temperature evolution of the same system, where instead of a real QM engine, LOTF was used to track the classical forces, but the force on the H atoms was scaled by 0.9. The same behavior is observed as shown in Fig. 6: without a thermostat, the system is unstable. With a simple velocity-rescaling thermostat, the kinetic temperature of the hydrogen is different from the system (in this case, much larger), while using a Langevin thermostat [25] eliminates this discrepancy. The Langevin thermostat adds a dissipative and a fluctuat-

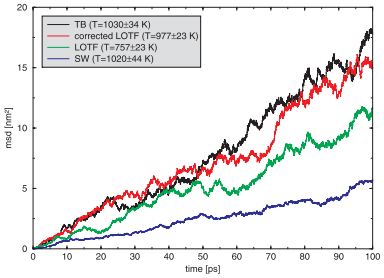

Fig. 6. Mean square displacement of a H atom diffusing in Si. Quantum and classical runs are represented by *black*, *blue lines*, respectively. The *green curve* corresponds to the hybrid simulation with a Nosé–Hoover thermostat, which is unable to maintain, by itself, the correct kinetic temperature of the H atom. Once this is corrected for results match the diffusion rate of the fully quantum-mechanical simulation (*red line*)

ing term to the forces and these are precisely balanced to achieve the desired temperature. In contrast to other schemes, the Langevin thermostat does not rely on efficient coupling of phonons of the hydrogen and the bulk, but couples to each particle directly and separately.

Dislocation Glide

To demonstrate a simple application of the hybrid scheme in a system that definitely needs quantum-mechanical accuracy yet is already too large for an ordinary simulation, we consider the glide of partial dislocations. Silicon partial dislocations move by forming kinks that zip along the dislocation line. The most prevalent partial dislocation in silicon is the 30° partial, and the kink that most easily forms and migrates on it is the left kink [26].

Figure 8 shows a pair of such partials, oppositely directed, so that the whole unit cell is periodic, and the partials enclose a stacking fault. The unit cell is also skewed slightly, thus forcing the existence of a left kink on each partial. In a molecular-dynamics simulation at high temperature (900 K to 1100 K), the partials move toward each other, and eventually annihilate, leaving behind a number of scattered point defects (mostly fourfold-coordinated

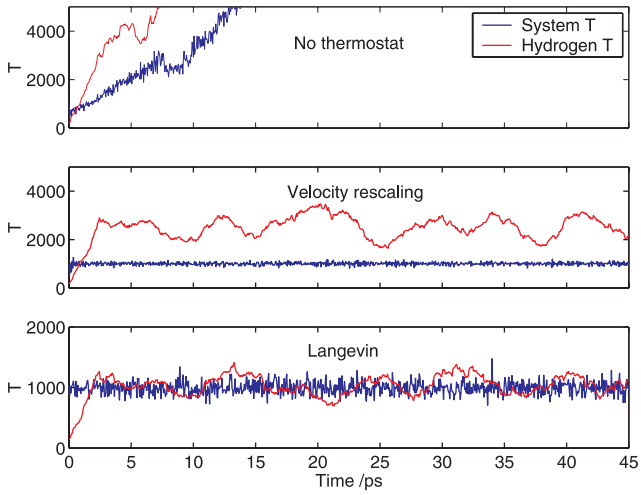

Fig. 7. Temperature evolution of a hydrogen atom in a 64-atom silicon cell using different thermostats with the hybrid scheme. The *top panel* shows that in such a small system, there is a significant energy drift due to the time-dependent Hamiltonian. The *middle panel* shows that a simple velocity-rescaling thermostat that, like the Nosé–Hoover, couples to the entire system via one extra degree of freedom to the entire system, is unable to maintain the correct kinetic temperature for the light H atom. The *bottom panel* shows that a more sophisticated Langevin thermostat, which couples to each atom individually, maintains the correct temperature for all species of atoms

defects [27]). There are a number of salient features of the LOTF simulation that are worth pointing out and that are not observed in a purely classical simulation. First, the glide is an order of magnitude faster in the hybrid simulation. This does not just indicate a lower energy barrier for kink migration, but the configurations that are most prevalent in the hybrid simulation are very different from those in the classical case. The most striking example is the equilibrium state of the left kink itself, shown in Fig. 9. Although the equilibrium configurations in the quantum and classical models agree at zero temperature, at high temperature, the quantum-mechanical free-energy minimum corresponds to a kink configuration with a square and an ejected antiphase defect. During the classical simulation, the kink spent a large proportion of the time trapped in a metastable state that had the effect of pinning the partial.

A unique advantage of a modular hybrid scheme with a black-box quantum engine is that it is easy to investigate the effect of chemical defects. Swapping the empirical tight-binding quantum engine for density-functional tight-binding (DFTB)[28] enables us to simulate the interaction of the partial with dopants without having to worry about constructing a classical model for a new type of atom. If we start the simulation after placing a boron atom

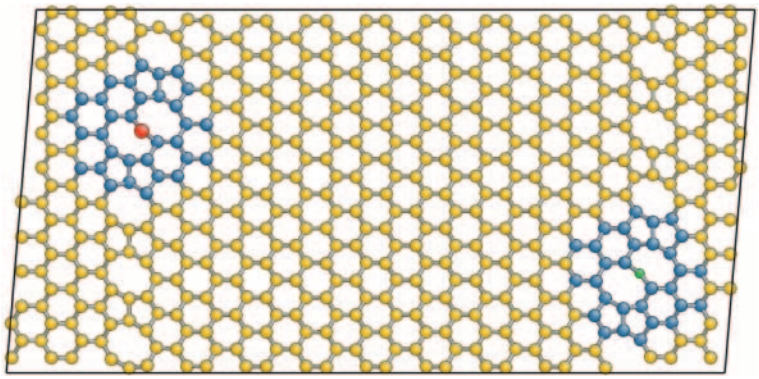

Fig. 8. A pair of oppositely directed 30° partial dislocations enclosing a stacking fault. The complete unit cell consists of 4536 atoms, but only the (111) plane containing the stacking fault is shown. The red (Si) and green (B impurity) atoms are undercoordinated, so they and the surrounding blue (Si) atoms are treated quantum-mechanically

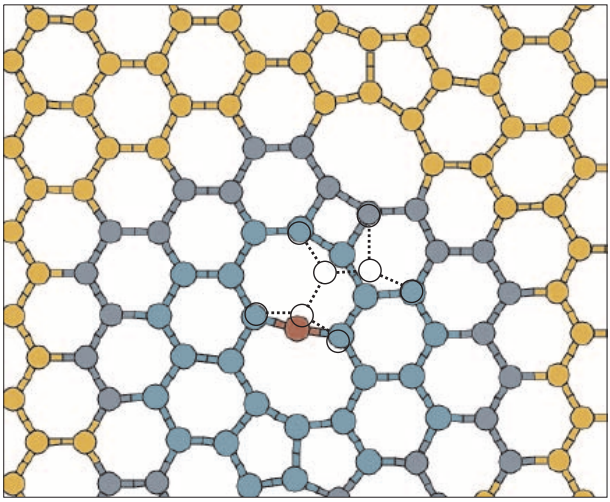

Fig. 9. Detail of the dislocation kink, showing the configuration that corresponds to equilibrium at zero temperature (*dotted lines*), and to what is prevalent at 900 K in the hybrid simulation

at the position shown in green in Fig. 8, the right-hand partial is pinned and remains completely stationary, while the left-hand partial advances as before.

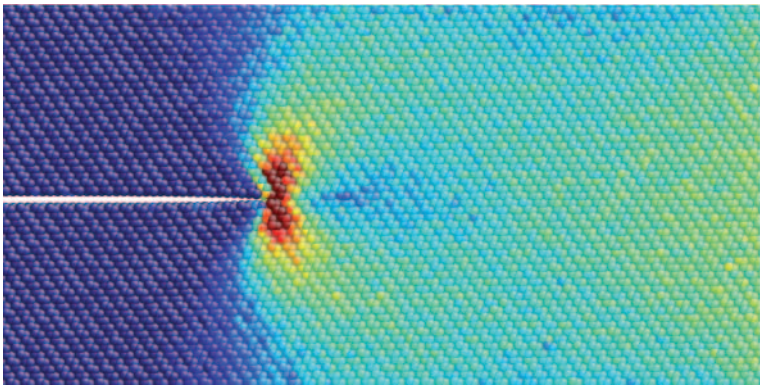

Fig. 10. Von-Mises strain field around a stationary crack tip in the opening mode in silicon

Brittle Fracture

In recent years, the problem of brittle fracture of covalent materials has been a prototypical problem addressed using hybrid methodology [29–31]. Figure 10 depicts the von-Mises strain around a stationary crack tip, as calculated using the Stillinger–Weber potential by optimizing the positions of the interior atoms, while holding the top and bottom rows of atoms fixed, corresponding to a given load in the opening mode. The stress concentrates near the tip and when the load reaches a critical level, the most strained bond gives way and the crack propagates forward. While analytical values for the required loading for crack propagation are easy to compute in the continuum approximation, it is well known that the discreteness of the atomic lattice gives rise to a barrier to this process, called *lattice trapping* [32–34]. The height of this barrier depends sensitively on the atomistic model employed and classical models typically overestimate this barrier, resulting in a much higher critical loading. The barrier associated with the Stillinger–Weber potential in particular is so high that, when the crack finally does propagate, so much elastic energy is released that the opening surfaces roughen, the crack tip blunts, emits dislocations, resulting in a reduction of the local stress field and an arrest to crack propagation: in other words, the ductile behavior is incorrectly predicted. An example of such "ductile" crack propagation is shown in Fig. 11. In contrast, real cracks in silicon at low temperature are atomically flat and propagate continuously above the critical load. This is reproduced by the hybrid simulation, a snapshot of which is shown in Fig. 12. This simulation at 300 K shows the 5–7–5 Pandey reconstruction of the opening (111) surfaces. The motion of the crack tip is tracked with the quantum region by noting which atoms have changed their number of nearest neighbors since the start of the simulation, and treating all atoms quantum-mechanically within 5 Å of these.

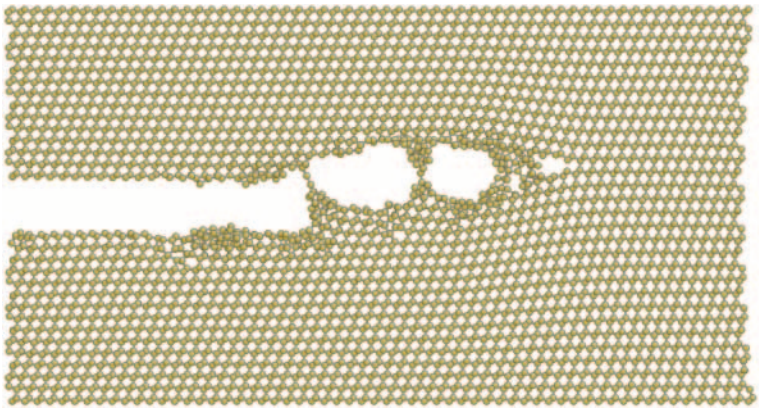

Fig. 11. Snapshot of "ductile" crack propagation using the Stillinger–Weber model. Defects are produced in the crack-opening region and dissipate energy; the onset of cracking is not sharp

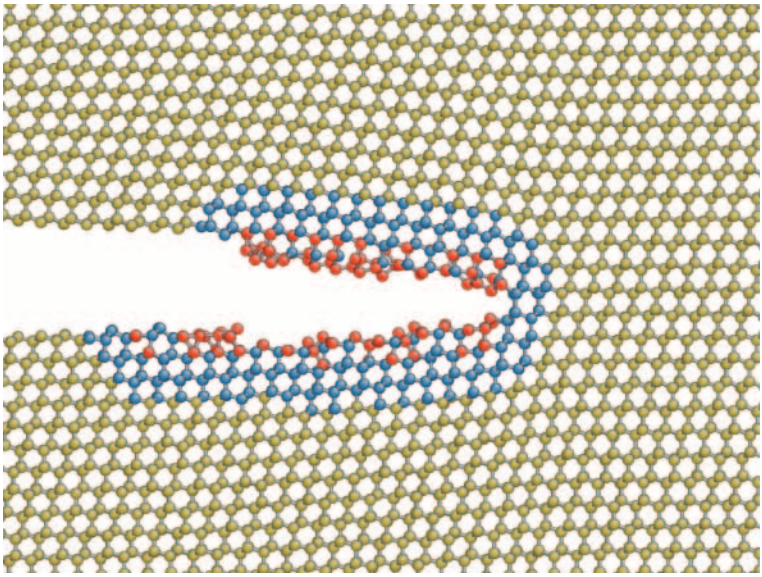

Fig. 12. Snapshot from brittle crack propagation, using the hybrid method at 300 K. Red atoms have been flagged as "active" because they have changed their neighbor count since the start of the simulation, and together with the blue atoms, are treated quantum-mechanically. In contrast to the ductile case, the crack propagates continuously above critical loading

6 Summary

To conclude this overview, we reiterate that a large class of problems in the area of semiconductor mechanical properties are inherently multiscale, and that significant advances in computer simulation will only be made by addressing the different length scales directly and simultaneously. We have introduced a promising new scheme that scales well to large system sizes and that deals with the intricacies of a hybrid simulation with fully controllable approximations. The scheme is very general, and can be extended to deal with arbitrary classical force-field potentials, one foreseen extension being, in particular, in the direction of modeling long-range (classical electrostatic) interactions.

References

[1] L. Pelaz, L. A. Marques, J. Barbolla: J. Appl. Phys. **96**, 5947 (2004) 194
[2] M. Bruzzi, et al.: Nucl. Instr. Meth. A **514**, 189 (2005) 194
[3] M. Huhtinen: Nucl. Instr. Meth. A **491**, 194 (2002) 194
[4] S. K. Estreicher: Mater. Sci. Eng. Rep. **14**, 319 (1995) 195
[5] A. Kerridge, A. H. Harker, A. M. Stoneham: J. Phys. Condens. Matter **16**, 8743 (2004) 195
[6] C. L. Muhlstein, E. A. Stach, R. O. Ritchie: Acta Mater. **50**, 3579 (2002) 197
[7] H. Kahn, R. Ballarini, J. J. Bellante, A. H. Heuer: Science **298**, 1215 (2002) 197
[8] J. Du, W. H. Ko, D. J. Young: Sens. Actuators A **112**, 116 (2004) 197
[9] P. Nguyen, I. Cayrefourcq, K. K. Bourdelle, A. Bougassol, E. Guiot, N. B. Mohamed, N. Sousbie, T. Akatsu: J. Appl. Phys. **97**, 083527 (2005) 197
[10] J. Weber, T. Fischer, E. Hieckmann, M. Hiller, E. V. Lavrov: J. Phys.: Condens. Matter **17**, S2303 (2005) 197
[11] B. Lawn: *Fracture of Brittle Solids* (Cambridge University Press, Cambridge 1993) 197
[12] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard: J. Phys. Chem. A **105**, 9396 (2001) 198
[13] S. J. Stuart, A. B. Tutein, J. A. Harrison: J. Chem. Phys. **112**, 6472 (2000) 198
[14] D. G. Pettifor, I. I. Oleinik, D. Nguyen-Manh, V. Vitek: Comp. Mater. Sci. **23**, 33 (2002) 198
[15] D. W. Brenner: phys. stat. sol. B **217**, 23 (2000) 198
[16] *J. Molecular Structure – Theochem, Special Issue*, vol. 632 (Elsevier Science BV 2003) Combined QM/MM Calculations in Chemistry and Biochemistry 199
[17] W. C. Swope, H. C. Andersen, P. H. Berens, K. R. Wilson: J. Chem. Phys. **76**, 637 (1982) 200
[18] A. De Vita, R. Car: MRS Proc. **491**, 473 (1998) 200
[19] G. Csányi, T. Albaret, M. C. Payne, A. De Vita: Phys. Rev. Lett. **93**, 175503 (2004) 200, 204

[20] G. Csányi, T. Albaret, G. Moras, M. C. Payne, A. De Vita: J. Phys. Condens. Matter **17**, R691 (2005) 200, 204
[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery: *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, Cambridge 1992) 201
[22] F. Stillinger, T. Weber: Phys. Rev. B **31**, 5262 (1985) 201, 203
[23] I. Kwon, R. Biswas, C. Z. Wang, K. M. Ho, C. M. Soukoulis: Phys. Rev. B **49**, 7242 (1994) 203
[24] T. J. Lenosky, J. D. Kress, L. A. Collins, I. Kwon: Phys. Rev. B **55**, 1528 (1997) 203
[25] D. Quigley, M. I. J. Probert: J. Chem. Phys. **120**, 11432 (2004) 204
[26] V. V. Bulatov, J. F. Justo, W. Cai, S. Yip, A. S. Argon, T. Lenosky, M. de Koning, T. D. de la Rubia: Philos. Mag. A **81**, 1257 (2001) 205
[27] S. Goedecker, T. Deutsch, L. Billard: Phys. Rev. Lett. **88**, 235501 (2002) 206
[28] T. Frauenheim, G. Seifert, M. Elstner, Z. Hajnal, G. Jungnickel, D. Porezag, S. Suhai, R. Scholz: phys. stat. sol. B **217**, 41 (2000) 206
[29] N. Bernstein: Europhys. Lett. **55**, 52 (2001) 208
[30] F. F. Abraham, J. Q. Broughton, N. Bernstein, E. Kaxiras: Comp. Phys. **12**, 538 (1998) 208
[31] S. Ogata, E. Lidorikis, F. Shimojo, A. Nakano, P. Vashishta, R. K. Kalia: Comp. Phys. Comm. **138**, 143 (2001) 208
[32] R. Pérez, P. Gumbsch: Phys. Rev. Lett. **84**, 5347 (2000) 208
[33] P. Gumbsch: Brittle fracture and the breaking of atomic bonds, in *Materials Science for the 21st Century*, vol. A (The Society of Materials Science, JSMS, Japan) pp. 50–58 208
[34] N. Bernstein, D. W. Hess: Phys. Rev. Lett. **91**, 025501 (2003) 208

Index

Empirical Molecular Dynamics: Possibilities, Requirements, and Limitations

Kurt Scheerschmidt

Max Planck Institute of Microstructure Physics, Weinberg 2, D-06120 Halle, Germany
schee@mpi-halle.de

Abstract. Classical molecular dynamics enables atomistic structure simulations of nanoscopic systems to be made. The method is extremely powerful in solving the Newtonian equations of motion to predict static and dynamic properties of extended particle systems. However, to yield macroscopically relevant and predictive results, suitable interatomic potentials are necessary, developed on ab-initio-based approximations. The fundamental requirements for performing classical molecular dynamics are presented as well as the relation to statistical methods and particle mechanics, suitable integration and embedding techniques, and the analysis of the trajectories. The applicability of the technique is demonstrated by calculating quantum-dot relaxations and interaction processes at wafer-bonded interfaces.

1 Introduction: Why Empirical Molecular Dynamics?

Classical molecular dynamics (MD) enable atomistic structure simulations of nanoscopic systems and are, in principle, a simple tool to approach the many-particle problem. For given interatomic or intermolecular forces one has to integrate the Newtonian equations of motion assuming suitable boundary conditions for the box containing the model structure. There are at least two advantages of this technique. The molecular dynamics is deterministic and provides the complete microscopic trajectories, i.e., the full static and dynamic information of all particles is available, from which a large number of thermodynamic and mechanically relevant properties of the models can be calculated. Further, one can perform simulations that are macroscopically relevant with the present computational power of even desktop computers. With reasonable computational effort models of nanoscale dimension can be treated for several million particles and up to microseconds of real time. Thus, empirical MD has two main fields of application: The search for the global energetic minima by relaxing nanoscopic structures and the calculation of dynamical parameters by analyzing the lattice dynamics.

Computer simulations are performed on models simulating the reality by using approximations, reduction, localization, linearization, etc., the validity of which has to be critically evaluated for each problem considered. As sketched in Fig. 1, increasing the time and length scales of the models (which is necessary to increase the macroscopic validity and robustness of the calculations) requires an increasing number of approximations and thus leads

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 213–244 (2007)

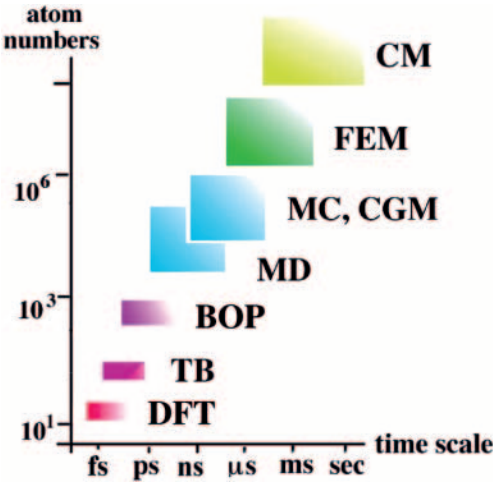

Fig. 1. Length and timescales in various modeling methods: DFT = density-functional theory, TB = tight-binding approximations, BOP = bond-order potentials, MD = empirical molecular dynamics, MC, CGM = Monte-Carlo/conjugate-gradient techniques, FEM = finite element methods, CM = continuum mechanics

to a reduction in the ability to predict some physical properties. Density-functional theory (DFT) and its approximations (e.g., local-density LDA, generalized gradient GGA), and the different kinds of tight-binding treatments (TB) up to the bond-order potential (BOP) approximations start with the Born–Oppenheimer (BO) approximation, thus decoupling the ionic and electronic degrees of freedom. Additional approximations such as pseudopotentials, gradient corrections to the exchange-correlation potential, and incomplete basis sets for the single-particle states are required for DFT calculations. The specifics of the various methods are discussed in various Chapters in the present book and in a number of review articles on DFT, TB, linear scaling techniques and programs such as SIESTA and CASTEP [1–9]. In first-principle MD [10], e.g., using DFT or DFT-TB, the electronic system is treated as parameter free and the resulting Hellmann–Feynman forces are the glue of the ionic interactions. Such simulations are computationally too expensive for large systems.

Figure 1 shows that the empirical MD closes the gap between first-principles, macroscopic, and continuum techniques (FEM = finite element methods, CM = continuum mechanics). The latter neglect the underlying interatomic interactions but allow the description of defects, defect interactions, diffusion, growth processes, etc. Stochastic Monte Carlo techniques (MC) enable the further increase of the timescales, also at the first-principles level, but without access to the dynamics. On the other hand, static energy minimization (e.g., conjugate gradient methods, CGM) enable a drastic increase in the model size. However, this does not necessarily provide the global mini-

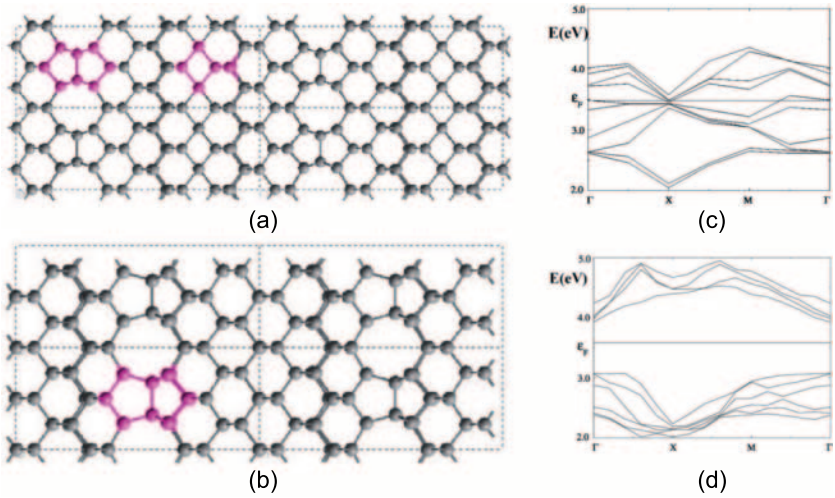

Fig. 2. Interface structures at 90° twist boundaries ((**a**): $Pmm(m)$-layer, (**b**): $\overline{4}2m$-*dreidl*, cf. text in Sect. 4.2), predicted by empirical MD and revealed by ab-nitio DFT simulations yielding semimetallic/isolating behavior as a function of the interface bonding state as demonstrated by the corresponding DFT band structures (**c**) and (**d**), respectively

mum of the potential energy surface. The largest model dimensions (number of atoms) accessible to empirical MD (or MC, CGM) has approximately the same extension as the smallest devices in microelectronics and micromechanics. Thus, today's atomistic modeling approaches the size of actual nanoscale systems.

The assumption of the existence, validity, and accuracy of known empirical interatomic potentials or force fields in analytic form always ignores the underlying electronic origin of the forces, i.e., the quantum structure of the interactions (cf. Sect. 2.4). Therefore, it is important to have better approximations, such as the bond-order potential (BOP), which is developed from tight-binding approximations and discussed briefly in Sect. 3.2. Other possibilities to include electronic properties of the interaction consist in continuously refitting an empirical potential during the calculation, called *learning on the fly* (see Chap. 9 and [11]). Separated subsystems can be treated at an ab-initio level. Suitable handshaking methods can be designed to bridge the embedded subsystem with its surrounding, which is treated semiclassically (cf. [12–15]). However, well-constructed potentials describing sufficiently accurate physical properties also give physical insights and enable a thorough understanding of the underlying processes [2].

Figure 2 shows a simple example demonstrating the difference in the approximation levels. Using empirical MD the correct structural relaxation of special interfaces created by wafer bonding can be treated, as described later

in Sect. 4.2. Two configurations exist, a metastable one (Fig. 2a) and the global minimum structure shown in Fig. 2b, called *dreidl* [16]. The electronic properties demand ab-initio level simulations, done here using a smaller periodic subunit of the interface and applying DFT techniques. It is shown that the higher-level approximation reproduces the structure predicted by the semiempirical techniques, whereas the correct energies and electronic properties (the band structure and the semimetallic or isolating behavior of the interfaces as shown in Figs. 2c,d) can only be described using the DFT formalism.

A second kind of embedding problem occurs because even millions of particles only describe a small part of reality that, even for smaller pieces of matter, is characterized by Avogadro's number (6×10^{23} mol^{-1}). Since isolated systems introduce strong surface effects, each model has to be embedded in suitable surroundings. For a discussion of various types of boundary conditions, see Sect. 3.1, Chap. 9 and [13–15].

The fundamental requirements of classical MD simulations, the relation to statistical methods and particle mechanics, suitable integration and embedding techniques and the analysis of the trajectories are presented in Sect. 2. The enhancements of potentials (bond-order potentials) and boundary conditions (elastic embedding) are discussed in Sect. 3. Selected application of semiempirical MD (relaxation of quantum dots and wafer-bonded interfaces) are given in Sect. 4 together with some examples from the literature. They strongly depend on the approximations assumed in the simulations.

2 Empirical Molecular Dynamics: Basic Concepts

The main steps for applying empirical molecular dynamics consist of the integration of the basic equations (cf. Sect. 2.1) using a suitable interaction potential and embedding the model in a suitable surrounding (cf. Sect. 2.4 and Sect. 2.3, respectively). Textbooks of classical molecular dynamics, e.g., [17–23], describe the technical and numerical details, and provide a good insight into possible applications and the physical properties, which may be predictable. Here, only the main ideas of empirical MD simulations, viz. the basic equations, methods of numerical treatment and the analysis of the trajectories are discussed.

2.1 Newtonian Equations and Numerical Integration

The basic equations of motion solved in empirical MD are the Newtonian equations for N particles ($i = 1, \ldots, N$) characterized by their masses m_i, their coordinates $\boldsymbol{r}_i$, and the forces acting on each particle $\boldsymbol{f}_i$. These may include external forces $\boldsymbol{F}_i$ and interatomic interactions $\boldsymbol{f}_i = \sum_{j \neq i} \boldsymbol{f}_{ij}$:

$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i = -\frac{\partial \mathcal{V}(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)}{\partial \boldsymbol{r}_i} \,. \tag{1}$$

The assumption that a potential $\mathcal{V}$ exists presupposes conservative (nondissipative) interactions and needs some more general considerations, cf. Sect. 2.2. If one assumes pairwise central potentials, $\mathcal{V}(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N) = \sum'_{i,j} \mathcal{V}(r_{ij})$ with $r_{ij} = |\boldsymbol{r}_j - \boldsymbol{r}_i|$ (the dash means that the sum is restricted to $i < j$), it follows that the virial of the external and internal forces is $\mathcal{M} = \sum'_{i,j} \boldsymbol{r}_{ij} \boldsymbol{f}_{ij} = \sum_i \boldsymbol{r}_i \boldsymbol{F}_i$. If the system is isolated (microcanonical ensemble), the conservation of total energy $\mathcal{E} = \mathcal{K} + \mathcal{V}$ (with kinetic energy $\mathcal{K} = \sum_i m_i \dot{\boldsymbol{r}}_i{}^2$) is guaranteed:

$$\dot{\mathcal{E}} = \dot{\mathcal{K}} + \dot{\mathcal{V}} = \sum_i m_i \dot{\boldsymbol{r}}_i \cdot \ddot{\boldsymbol{r}}_i - \sum_i \dot{\boldsymbol{r}}_i \boldsymbol{f}_i = 0\,. \tag{2}$$

Given an initial configuration of particles and suitable boundary conditions (cf. Sect. 2.3), the differential equations can be integrated using one of the standard methods, e.g., Runge–Kutta techniques, predictor–corrector methods, Verlet or Gear algorithm. The numerical integration is equivalent to a Taylor expansion of the particle positions $\boldsymbol{r}_i(t + \delta t)$ at a later time in terms of atomic positions $\boldsymbol{r}_i$ and velocities $\boldsymbol{v}_i$. The forces $\boldsymbol{f}_i$ are the time derivatives of $\boldsymbol{r}_i$ with an increasing order at previous time steps:

$$\boldsymbol{r}_i(t + \delta t) = \boldsymbol{r}_i(t) + a_1 \delta t \boldsymbol{v}_i(t) + a_2 (\delta t)^2 \boldsymbol{f}_i(t) + \ldots + \mathcal{O}(\delta t^n)\,. \tag{3}$$

In addition to good potentials and system restrictions (cf. Sect. 2.3 and Sect. 2.4, respectively) an efficient and stable integration procedure is required to accurately propagate the system. The conservation of energy during the simulation is an important criterion. An increase in order allows larger time increments δt to be used, if the evaluation of the higher derivatives is not too time consuming, which happens with more accurate potentials, like the BOP (Sect. 3.2). The efficiency and the accuracy of the integration can be controlled by choosing suitable series-expansion coefficients $a_j, j = 1, 2, ..., n-1$ and the order n of the method or by mixing the derivatives of $\boldsymbol{r}_i$ at different times in (3) to enhance the procedure. However, the increment δt, of the order of fs, must be at least so small that the fastest particle oscillations are sufficiently sampled.

Better or faster MD calculations may be performed using special acceleration techniques, the three most important methods being:

1) Localization: By using the linked-cell algorithm and/or neighbor lists, it is assumed that for sufficiently rapidly decaying potentials (faster than Coulomb $1/r$ for $r \to \infty$) only a small number of particles have a direct and significant interaction. A cutoff r_c is defined and the interaction potential is assumed to be zero for $r > r_\mathrm{c}$. A transition region is fitted using splines or other suitable functions. Then the system is divided into cells. Their minimum dimension is given by $2r_\mathrm{c}$ to avoid self-interaction of the particles (*minimum image convention*). Only the interactions of the atoms within a cell and its 26 neighboring cells are considered, which reduces the simulation time drastically from the $\propto N^2$ behavior for all particle interactions to linear behavior $\propto 729 c_N N$, where c_N is the average number of particles per cell.

The problem is to find a suitable r_c for a smooth transition and screening behavior that includes a sufficient number of next-neighbor atoms and cells. One also needs a suitable criterion to update the neighbor lists and reorder the cells whenever particles leave a cell during the propagation of the system.

The Coulomb potential may also be screened. However, summations over the long-range $1/r$ potential, representing infinite point-charge distributions, are only conditionally convergent. Thus, it is better to apply the Ewald method (originally developed to calculate cohesive energies and Madelung constants [24]) and its extensions [25, 26] based on successive charge neutralization by including next-neighbor shells around the origin.

2) Parallelization: Using replicas, or dividing the structure into several parts, which are then distributed to different processors, thus allowing parallelization [27, 28]. The replica technique needs suitable criteria for dividing the system into small parts with minimum interaction. More importantly, one must bear in mind that the replicas are not independent over long times. A careful control of the time interval after which the communication between the different parts is required in order to achieve the same results as nonparallelized simulations. This issue is the bottleneck of the technique.

3) Time stretching: Such techniques as hyperdynamics, temperature acceleration, basin-constrained dynamics, on-the-fly Monte-Carlo, and others are subsumed briefly here, because their common idea consists in replacing the true time evolution by a shorter one increasing the potential minima, transition frequencies, system temperature, etc. (see [28]).

2.2 Particle Mechanics and Nonequilibrium Systems

In classical mechanics a system is characterized either by its Lagrangian $\mathcal{L}(q,\dot{q}) = \mathcal{K}(q,\dot{q}) - \mathcal{V}(q)$ or its Hamiltonian related to the Lagrangian by a Legendre transform $\mathcal{H}(q,p) = \sum \dot{\boldsymbol{q}}_i \boldsymbol{p}_i - \mathcal{L} = \mathcal{K} + \mathcal{V}(q)$, where $\boldsymbol{q}_i, \boldsymbol{p}_i$ are generalized coordinates and momenta (conjugate coordinates), respectively, which have to be independent or unrestricted. One derives the generalized momenta from the derivatives of the Lagrangian $\boldsymbol{p}_i = \frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{q}}_i}$, whereas the derivatives of the Hamiltonian

$$\dot{\boldsymbol{q}}_i = \frac{\partial \mathcal{H}}{\partial \boldsymbol{p}_i}, \quad \dot{\boldsymbol{p}}_i = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{q}_i} \tag{4}$$

reproduce Newton's law of motion (1).

Hamilton's principle of least action enables a simple generalization of the mechanics of many-particle systems. The integral over the Lagrangian function has to be an extremum. Variational methods yields, applying the extremal principle:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\boldsymbol{q}}_i} - \frac{\partial \mathcal{L}}{\partial \boldsymbol{q}_i} = 0\,. \tag{5}$$

The advantage of the general formulation (5) is that it allows a simple extension to nonconservative systems by including an explicit time dependence or to systems with constraints, such as fixed bond lengths in subsystems, friction of particles, outer forces, etc. If there are holonomic constraints describing relations between the coordinates $f_l(\boldsymbol{q}_k) = 0$, one gets the generalized additional coordinates $a_{lk} = \partial f_l/\partial \boldsymbol{q}_k$ and $a_l = \partial f_l/\partial t$ creating an additional term on the right-hand side of (5) in the form $\sum_l \lambda_l a_{lk}$ with the set of Lagrange multipliers λ_l. This formalism is the basis for the enhancement of the boundary conditions in Sect. 3.1 and the handshaking methods mentioned above.

In addition, the Lagrangian and Hamiltonian formalisms correlate classical dynamics to statistical thermodynamics and to quantum theory, and allow the evaluation of properties from the trajectories (cf. Sect. 2.5). The set of time-dependent coordinates $\boldsymbol{q}$ (the configuration space) and time-dependent momenta $\boldsymbol{p}$ (momentum space) together is called the phase space $\Gamma = (\boldsymbol{q}, \boldsymbol{p}) = (\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_n, \boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n)$. Presenting Γ in a 6-dimensional hyperspace yields N trajectories, one for each molecule, and allows the study of the behavior of the system using statistical methods. It is called the μ-space (following Ehrenfest) or the Boltzmann molecule phase space. The whole Γ-set as one trajectory in a $6N$-dimensional hyperspace presents the Gibbs phase space, with the advantage to include better interactions and restrictions. However, the system must now be described as a virtual assembly for statistical relations. The phase space Γ contains the complete information on the microscopic state of the many-particle system. Several basic quantities may be derived such as the phase-space flow or the "velocity field" $\dot{\Gamma} = (\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}})$. Applying the relations (4) yields the Liouville equation

$$\nabla_\Gamma \Gamma = \sum \left(\frac{\partial \dot{\boldsymbol{q}}_i}{\partial \boldsymbol{q}_i} + \frac{\partial \dot{\boldsymbol{p}}_i}{\partial \boldsymbol{p}_i} \right) = 0 \,, \tag{6}$$

showing that the flow $\dot{\Gamma}$ behaves like an incompressible liquid, i.e., although the μ-trajectories are independent, their related phase-space volume is constant.

An equivalent formulation of the Liouville statement (6) is the equation of continuity for $\dot{\Gamma}$, $\dot{\rho}$, and the local change of the density $\rho(\boldsymbol{q}, \boldsymbol{p}, t)$, similar to the Heisenberg equation for observables in quantum theory. The density $\rho(\boldsymbol{q}, \boldsymbol{p}, t)$ describes the probability of finding the system within the region $\boldsymbol{p}, \boldsymbol{q}$ and $\boldsymbol{p} + \mathrm{d}\boldsymbol{p}, \boldsymbol{q} + \mathrm{d}\boldsymbol{q}$ of the phase-space volume $\mathrm{d}\boldsymbol{p}\mathrm{d}\boldsymbol{q}$. Because the systems must be somewhere in the phase space, the density can be normalized in the entire phase space. The integral over the whole phase space of the non-normalized density yields the partition function. Integration over $(N - k)$ particles yields the k-particle phase-space density $\rho^{(k)}(\boldsymbol{q}, \boldsymbol{p}, t) = \int \rho^{(N)}(\boldsymbol{q}, \boldsymbol{p}, t) \mathrm{d}\boldsymbol{p}^{(N-k)} \mathrm{d}\boldsymbol{q}^{(N-k)}$, and the integration over the whole momentum space the corresponding k-particle distribution function $n^{(k)}(\boldsymbol{q}) = \int \rho^{(N)}(\boldsymbol{q}, \boldsymbol{p}, t) \mathrm{d}\boldsymbol{q}^{(N-k)} \mathrm{d}\boldsymbol{p}$. Very important for the trajectory analy-

sis (Sect. 2.5) are the cases $k = 1, 2$ defining the (radial) pair-distribution functions $g(\boldsymbol{q}_i, \boldsymbol{q}_j)$:

$$n^{(1)}(\boldsymbol{q}_i)n^{(1)}(\boldsymbol{q}_j)g(\boldsymbol{q}_i, \boldsymbol{q}_j) = n^{(2)}(\boldsymbol{q}_i, \boldsymbol{q}_j). \tag{7}$$

The phase-space trajectories allow the classification of the dynamics of many-particle systems. They may be Poincare-recurrent (the same phase-space configuration occurs repeatedly), Hamiltonian (nondissipative), conservative (no explicit time dependence for the Hamiltonian), or integrable (number of constants of motion equals the number of degrees of freedom, resulting in stable periodic or quasiperiodic systems). Nonintegrable systems may be ergodic or mixing, etc., i.e., the trajectory densely covers different hypersurfaces in the phase space, and allow the characterization of different kinds of instabilities of the system.

2.3 Boundary Conditions and System Control

As mentioned above, even for systems considered large in MD simulations, the number of atoms is small compared to real systems and therefore dominated by surface effects. They are caused by interactions at free surfaces or with the box boundaries. In order to reduce nonrealistic surface effects, periodic boundary conditions are applied, or the box containing the model (supercell) has to be enlarged so that the influence of the boundaries may be neglected. For further discussion and enhancements including elastic embedding, see Sect. 3.1. Periodic boundary conditions means that the supercell is repeated periodically in all space directions with identical image frames or mirror cells. In contrast to the case of fixed boundaries, under periodic boundary conditions the particles can move across the boundaries. If this happens, the positions $\boldsymbol{r}$ have to be replaced by $\boldsymbol{r} - \boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is a translation vector to the image frame to which the particle is moved. Particle number, total mass, total energy, and momenta are conserved in periodic boundary conditions, but the angular momenta are changed and only an average of them is conserved. It should be mentioned here that periodic boundary conditions repeat defects, which in small supercells create high defect concentrations. Antiperiodic and other special boundary conditions may be chosen, where additional shifts of the positions and momenta along the box borders allow correction of a disturbed periodic continuation and a description of reflective walls at free surfaces to be obtained.

According to the boundary conditions and system restrictions, the virtual Gibbs entities are either isolated microcanonical ensembles with constant volume, total energy, and particle numbers (NVE ensembles) or systems that exchange and interact with the environment. Closed ensembles (e.g., canonical NVT or isothermic-isobar NPT systems) have only an energy exchange with a thermostat, whereas open systems have in addition particle exchange with the environment (e.g., grand-canonical $T\mu V$, where μ is the chemical potential).

Simple tools exist to equilibrate systems at a well-defined temperature. For example, velocity rescaling is numerically equivalent to a simple extension of the original microcanonical system with nonholonomic constraints in the Lagrangian formalism and is called the Berendsen thermostat. A similar simple extension may be used for pressure rescaling. It is called the Berendsen barostat [29]. The simple velocity rescaling works by conserving the Maxwell distribution and yielding maximum entropy. However, such nonisolated ensembles are better described using constraints in the Lagrangian similar to those discussed in Sect. 2.2. The extension of $\mathcal{L}$ in (5) by adding terms $\mathcal{L}_{\text{constr}}$ introduces additional generalized variables having extra equations of motion and also additional force terms in the Newtonian equations (1). Some of the most important methods for $\mathcal{L}_{\text{constr}}$ are given below without further comments:

- Nosé–bath and generalized Nosé–Hoover thermostat [30–32]: $\mathcal{L}_{\text{constr}} = M\dot{s}^2/2 - n_f T \ln(s)$ with the fictive or the whole mass $M = \sum m_i$, the degrees of freedom $n_f = 3N + 1$, and the new generalized variable s playing the role of an entropy.
- Andersen isothermic-isobaric system control (NPH ensemble) $\mathcal{L}_{\text{constr}} = \dot{V}^2/2 + PV$ introduces volume and pressure as generalized variables [33].
- Generalized stresses according to *Parrinello* and *Rahman* [34] $NTL\sigma$-ensembles: the Lagrangian is extended by $\mathcal{L}_{\text{constr}} = -1/2Tr\boldsymbol{\Sigma}\boldsymbol{G}$ with a generalized symmetric tensor $\boldsymbol{\Sigma}$ and the metric tensor $\boldsymbol{G}$ of the crystal structure. A further specialization, e.g., for constant strain rates $NTL_x\sigma_{yy}\sigma_{zz}$-ensemble [35], is in principle a mixture of the isothermic-isobaric system with the generalized stress constraint.
- Brownian fluctuations, transport and flow processes, density and other gradients, Langevin damping $\propto \boldsymbol{v}(t)$, etc. demand nonequilibrium MD [36, 37], which is mostly done by including the perturbation as a suitable virtual field $\mathcal{F}(\boldsymbol{p}, \boldsymbol{q}, t)$ into the equations of motion via the Lagrange formalism.

2.4 Many-Body Empirical Potentials and Force Fields

Empirical potentials and force fields exist with a wide variety of forms, and also different classification schemes are used according to their structure, applicability, or physical meaning. It makes no sense to describe a large number of potentials or many details in the present review, therefore only some of the existing and most used potentials are briefly listed below (a good overview can be found in [38–48]. A classification by *Finnis* [2] (and many references therein) describes recent developments and discusses in an excellent way the justification of the potentials by first-principle approximations, which is important for the physical reliability and the insight into the electronic structure of potentials. Interatomic forces are accurate only if the influence of the local environment according to the electronic structure is included. The most

important property of a potential is transferability, that is the applicability of the potential to varied bonding environments. An important additional criterion is its ability to predict a wide range of properties without refitting the parameters. The number of fit parameters decreases as the sophistication of the force fields increases. According to [2] one has:

Pair potentials – Valid for *s-p* bonded metals and mostly approximated by a sum of pairwise potentials. They may be derived as the response to a perturbation in jellium, which can be visualized in the pseudoatom picture as an ionic core and a screening cloud of electrons.
Ionic potentials – To use Coulomb interactions directly for ionic structures, the problem of screening (Ewald summation, Madelung constant) has to be considered as mentioned above. The interactions were originally described by Born and others as the rigid-ion approximation. Starting from the Hohenberg–Kohn–Sham formulation of the DFT-LDA, shell or deformable ion models may be developed beyond the rigid-ion approximation, where the additional shell terms [49] look like electron-density differences in noble gases.
Tight-binding models – Different derivations and approximations for TB-related potentials exist and are nowadays applicable to semiconductors, transition metals, alloys and ionic systems. The analytic bond-order potential is such an approximation and will be discussed in more detail in Sect. 3.2.
Hybrid schemes – Combinations of pair potentials with TB approximations are known as generalized pseudopotentials, effective media theories (EMT [50]), environmental-dependent ionic potentials (EDIP [51]), or embedded atom models (EAM [52]); for details see [2].

From the empirical point of view the simplest form of potentials may be considered to be a Taylor-series expansion of the potential energy with respect to 2-, 3-,. . ., n-body atomic interactions. Pair potentials have a short-range repulsive part, and a long-range attractive part, e.g., of the Morse or Lennard–Jones (LJ) type, mostly "12–6", i.e., $ar^{-12} - br^{-6}$. LJ potentials are successfully applied to noble gases, biological calculations, or to model long-range van der Waals interactions (e.g. [53]). For quasicrystals [54] an LJ potential was constructed with two minima in a golden number distance relation. However, simple pair potentials are restricted in their validity to very simple structures or to small deviations from the equilibrium. Therefore many-body interactions are added and fitted for special purposes, e.g., the MD of molecules and molecule interactions [55,56]. Separable 3-body interactions are widely used: Stillinger–Weber [57] (SW), Biswas–Hamann [58], and Takai–Halicioglu–Tiller [59] potentials. The SW is perhaps the best-known 3-body-type potential. It includes anharmonic effects necessary to reproduce the thermal lattice expansion of Si and Ge [60]. Hybrid force fields are sometimes used to include the interaction of different types of atoms, such as Born–

Mayer–Huggins, Rahmann–Stillinger–Lemberg terms and others applicable to silicate glasses and interdiffusing metal ions or water molecules [61–65].

As mentioned above, other force fields are developed from first-principle approximations that combine sufficient simplicity with high rigor. They are not based upon an expansion involving N-body interactions (cluster potentials). The more or less empirical forms of TB potentials and effective medium force fields are the modified embedded atom model (MEAM, [66] for cubic structures and references therein for other structures), the Finnis–Sinclair (FS) [67] and the Tersoff-type (TS) potentials [68–70]. The TS potential is an empirical bond-order potential with the functional form:

$$V(\boldsymbol{r}_{ij}) = ae^{-\lambda \boldsymbol{r}_{ij}} - b_{ij}e^{-\mu \boldsymbol{r}_{ij}} \,. \tag{8}$$

The bonds are weighted by the bond order $b_{ij} = F(\boldsymbol{r}_{ik}, \boldsymbol{r}_{jk}, \gamma_{ijk})$ including all next neighbors $k \neq i, j$, which gives the attractive interaction the form of an embedded many-body term. The different parameterizations (TI, TII, TIII) of the Tersoff potential have been intensively tested. Other parameterizations exist [71]. They involve other environmental functions and first-principle derivations [72, 73], as well as extensions to include further interactions, H in Si [74], C, Ge [75], C–Si–H [76, 77], AlAs, GaAs, InAs, etc. [78]. Finally, a refitted MEAM potential with SW terms is available for Si [79]. Multipole expansions replaced by spherical harmonics [80] are an alternative to TS potentials.

A comparative study of empirical potentials shows advantages and disadvantages in the range of validity, physical transparency, fitting and accuracy as well as applicability [81]. Restrictions exist for all empirical potential types, even if special environmental dependencies are constructed to enhance the elastic properties near defects. In addition, not all potentials are applicable to long-range interactions, and the electronic structure and the nature of the covalent bonds can only be described indirectly. Thus, it is of importance to find physically motivated semiempirical potentials, as mentioned above and discussed in Sect. 3.2 for TB-based analytic BOPs. The parameters of the empirical force field have to be fitted to experimental data or first-principle calculations. First, the cohesive energy, lattice parameter and stability of the crystal structures have to be tested or fitted. The bulk modulus, elastic constants, and phonon spectra are very important properties for the fit. The following section describes some of the quantities that may be used for the fit or to be evaluated from the MD simulations if not fitted. Very important details concerning point defects and defect clusters – necessary to get the higher-order interaction terms – are given by the energy and structure of such defects. The data may be given by DFT or TB dynamics or geometry optimizations, as, e.g., [82–85].

2.5 Determination of Properties

Static properties of systems simulated by empirical MD can be directly calculated from the radial- or pair-distribution function (7). Dynamic properties follow from the trajectories using averages or correlations:

$$\langle A\rangle_t = \lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^{t} \mathrm{d}t A[\boldsymbol{r}(\boldsymbol{t})], \quad \mathcal{C}(t) = \langle A(\tau)B(t+\tau)\rangle_\tau, \tag{9}$$

which in principle correspond to time averages. However, they are sampled at discrete points, so that it is necessary to choose suitable sampling procedures to reduce the effects of the finite size of the system, stochastic deviations, and large MD runs.

Two basic relations are central for the analysis of the properties. The ergodic hypothesis states that the ensemble average $\langle A\rangle_e$ is equal to the time average $\langle A\rangle_t$, which relates the averages to the measurement of a single equilibrium system. The Green–Kubo formula $\lim_{t\to\infty}\frac{\langle[A(t)-A(t_0)]^2\rangle_t}{t-t_0} = \int_{t_0}^{\infty}\mathrm{d}\tau(\langle\dot{A}(\tau)\dot{A}(t_0)\rangle_e)$ relates mean square deviations with time correlations. The diffusion coefficient $D = \lim_{t\to\infty}\frac{1}{6Nt}\langle\sum[\boldsymbol{r}_j(t)-\boldsymbol{r}_j(0)]^2\rangle$ (Einstein relation) is equivalent to the velocity autocorrelation function, which is a special form of the Green–Kubo formula $D = \frac{1}{3N}\int_0^\infty\langle\sum\boldsymbol{v}_j(t)\cdot\boldsymbol{v}_j(0)\rangle$. Similarly, one uses the crosscorrelation of different stress components to obtain shear viscosities. Other transport coefficients may be derived analogously.

In thermodynamic equilibrium the kinetic energy $\mathcal{K}$ per degree of freedom is determined by the equipartition theorem $\mathcal{K} = \langle\sum_i m_i\dot{v}_i^2/2\rangle = 3Nk_\mathrm{B}T/2$ (k_B = Boltzmann constant) which yields a measure of the system temperature T. The strain tensor σ and the pressure P are obtained from the generalized virial theorem $\sigma_{kl} = 1/V[\sum_j\boldsymbol{v}_{jk}\cdot\boldsymbol{v}_{jl} + \sum_{ij}\boldsymbol{r}_{ijk}\cdot\boldsymbol{f}_{ijl}]$. Thus, the pressure is the canonical expectation value $P = 1/3V[2\mathcal{K}-\mathcal{M}]$ of the total virial $\mathcal{M}$. Alternatively, one can use the pair distribution in the form $P = \rho k_\mathrm{B}T - \rho^2\int_0^\infty g(r)\frac{\partial U}{\partial\boldsymbol{r}}4\pi r^3\mathrm{d}r$, which may be useful for correcting sampling errors.

Using the densities $\rho(\boldsymbol{r},\boldsymbol{p})$ as defined above and $\mathcal{Z} = \int\rho(\boldsymbol{r},\boldsymbol{p})\mathrm{d}\boldsymbol{r}$ as normalization, statistical mechanics deals with ensemble averages, which in general are written as

$$\langle Q\rangle_e = 1/\mathcal{Z}\int Q(\boldsymbol{r},\boldsymbol{p})\rho(\boldsymbol{r},\boldsymbol{p})\mathrm{d}\boldsymbol{r}\,. \tag{10}$$

All thermodynamic functions may be derived from the partition function $\mathcal{Z}$, for example the free energy $F(T,V,N) = -k_\mathrm{B}T\ln[Q(T,V,N)]$. In addition, one can obtain all the thermodynamic response coefficients. With the internal energy U, one computes the isochoric heat capacity $C_V = (\partial U/\partial T)_{N,V}$ and thus a measure for the quality of the temperature equilibration $(\langle T\rangle^2 - \langle T^2\rangle)/\langle T^2\rangle = 3/2N(1-3kN)/2c_V$. From the volume V follows the isothermal compression $\chi_T = -1/V(\partial V/\partial P)_{N,T}$, etc. One has to choose according to the different ensembles:

- Microcanonical: $\rho_{NVE} = \delta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) - \mathcal{E}]$ conserving entropy $S(E,N,V)$.
- Canonical: $\rho_{NVT} = e^{(\mathcal{H}/k_\mathrm{B}T)} \propto e^{V(\boldsymbol{r})/k_\mathrm{B}T}$ conserving free energy $F(T, V, N)$
- Isothermic-isobaric: $\rho_{NPT} = e^{(\mathcal{H}-TS)/k_\mathrm{B}T}$ conserving Gibbs free energy $G(T,P,N)$
- Grand canonical: $\rho_{T\mu V} = e^{(\mathcal{H}-\mu N)/k_\mathrm{B}T}$ conserving Massieu function $J(T,\mu,V)$ (Legendre transformation in entropy representation).

Finally, it should be mentioned that the Fourier transform of pair distributions is connected to the scattering functions in X-ray, neutron and electron diffraction. MD-relaxed structure models allow the simulation of the transmission electron microscope (TEM) or high-resolution electron microscope (HREM) image contrast and therefore make the contrast analysis more quantitative. For this purpose, snapshots of the atomic configurations are cut into thin slices, which are folded with atomic scattering amplitudes and each other to describe the electron scattering (multislice formulation of the dynamical scattering theory), cf. the applications in Sect. 4 using this technique to interpret HREM investigations of quantum dots and bonded interfaces.

3 Extensions of the Empirical Molecular Dynamics

Coupling of length and timescales in empirical MD means bridging the first-principles particle interactions and the box environment (1). It can be done either using embedding and handshaking or by a separate treatment and a transfer parameter between the subsystems. MD simulations of the crack propagation [86] and the analysis of submicrometer MEMS [87] are successfully applications of the FEM coupling between MD and an environmental continuum. In Sect. 3.1 enhanced boundary conditions for MD are discussed: where the coupling between MD and an elastic continuum is a handshaking method based on an extended Lagrangian [88, 89]. The main steps in the development of an analytic TB-based BOP [90] are sketched in Sect. 3.2 as an example of using enhanced potentials.

3.1 Modified Boundary Conditions: Elastic Embedding

Elastic continua may be coupled to MD when the potential energy of an infinite crystal with a defect as shown in Fig. 3a is approximated in the outer region II by generalized coordinates a_k [89]:

$$E(\{\boldsymbol{r}_i\},\{\boldsymbol{r}_j\}) = E(\{\boldsymbol{r}_i\},\{a_k\})\,. \tag{11}$$

In the defect region I, characterized by large strains, the positions of atoms $\boldsymbol{r}_i$ $(i = 1,\ldots,N)$ are treated by empirical MD. The atomic positions $\boldsymbol{r}_j$ $(j > N)$ in the outer regions II and III result from the linear theory of elasticity

$$\boldsymbol{r}_j = \boldsymbol{R}_j + \boldsymbol{u}^{(0)}(\boldsymbol{R}_j) + \boldsymbol{u}(\boldsymbol{R}_j,\{a_k\})\,, \tag{12}$$

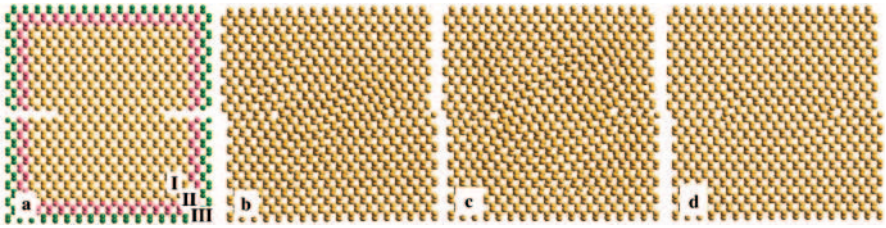

Fig. 3. Dislocation geometry ((**a**), I = MD region, II = elastic, III = overlap, cf. text) to apply elastic boundary conditions for a dipole of 60° dislocations, and snapshots during MD annealing: 500 K (**b**), 600 K (**c**), 0 K (**d**)

where the fields $\boldsymbol{u}^0(\boldsymbol{R})$ and $\boldsymbol{u}(\boldsymbol{R})$ describe the displacements of atoms from their positions $\boldsymbol{R}_j$ in the perfect crystal, and satisfy the equilibrium equations of a continuous elastic medium with defects; $\boldsymbol{u}^0(\boldsymbol{R})$ is the static displacement field of the defect and independent of the atomic behavior in I; $\boldsymbol{u}(\boldsymbol{R})$ is related to the atomic shifts in region I and can be represented as a multipole expansion:

$$\boldsymbol{u}(\boldsymbol{R},\{a_k\}) = \sum_{k=1}^{\infty} a_k \boldsymbol{U}^{(k)}(\boldsymbol{R}) \tag{13}$$

over homogeneous eigensolutions $\boldsymbol{U}^{(k)}(\boldsymbol{R})$ [88]. The $\boldsymbol{U}^{(k)}$ are rapidly decreasing with $\boldsymbol{R}$, thus the sum in (13) is truncated to a finite number K of terms.

The equilibrium positions of atoms in the entire crystal are obtained from the minimization of the potential energy given in (11) with respect to $\boldsymbol{r}_i$ and a_k. This is equivalent to a dynamic formulation based on the extended Lagrangian as discussed above (Sect. 2.2 and Sect. 2.3) with the extension (11):

$$\mathcal{L} = \sum_{i=1}^{N} \frac{m_i \dot{\boldsymbol{r}}_i^2}{2} + \sum_{k=1}^{K} \frac{\mu \dot{a_k}^2}{2} - E(\{\boldsymbol{r}_i\},\{a_k\})\,. \tag{14}$$

Here, m_i are the atomic masses and μ is a parameter playing the role of mass for the generalized coordinates a_k. If μ is properly chosen, the phonons are smooth from I to II and the outer regions oscillate slower than the MD subsystem as demonstrated in the snapshots Figs. 3b–d. The corresponding system defining the forces reads

$$\boldsymbol{F}_i = -\frac{\partial E(\{\boldsymbol{r}_i\},\{a_k\})}{\partial \boldsymbol{r}_i}\,, \tag{15}$$

$$F_k = -\frac{\partial E(\{\boldsymbol{r}_i\},\{a_k\})}{\partial a_k} = -\sum_{j>N} \frac{\partial E(\{\boldsymbol{r}_i\},\{\boldsymbol{r}_j\})}{\partial \boldsymbol{r}_j} \frac{\partial \boldsymbol{r}_j}{\partial a_k} = \sum_{j\in \mathrm{II}} \boldsymbol{F}_j \boldsymbol{U}^{(k)}(\boldsymbol{R}_j)\,. \tag{16}$$

These must vanish in the equilibrium $\boldsymbol{F}_i = F_k = 0$. The sum in (16) expands only over the atoms in region II surrounding region I, since in region III the linear elasticity alone is sufficient and the forces on these atoms are nearly vanishing. However, region II must be completely embedded in region III. The forces on atoms in region II must be derived from the interatomic potential, therefore the size of region II must be extended beyond the potential cutoff r_c. The Lagrangian (14) results in the equations of motion

$$m\ddot{\boldsymbol{r}}_i = \boldsymbol{F}_i\,, \quad (i = 1, \ldots, N)\,, \quad \mu\ddot{a}_k = F_k\,, \quad (k = 1, \ldots, K)\,, \tag{17}$$

thus extending Newton's equations by equivalent ones for the generalized coordinates.

3.2 Tight-Binding-Based Analytic Bond-Order Potentials

As discussed before, the use of TB methods allows much larger models than accessible to DFT. TB formulations exist in many forms (e.g., [2, 6–8, 91–93]). The application of an analytical BOP, developed mainly by *Pettifor* and coworkers [90, 94–98], may further enhance empirical MD by providing better justified force fields while allowing faster simulations than numerical TB-MD. The approximations to develop analytic TB potentials from DFT may be summarized by the following steps (cf. also [2, 99]): 1. Construct the TB matrix elements by Slater–Koster two-center integrals including s- and p-orbitals, 2. transform the matrix to the bond representation, 3. replace the diagonalization by Lanczos recursion, 4. get the momenta of the density of states from the continued fraction representation of the Green's function up to order n for an analytic BOP-n potential. The basic ideas sketched with a few more details are as follows.

The cohesive energy of a solid in the TB formulation can be written in terms of the pairwise repulsion U_{rep} of the atomic cores and the energy due to the formation of electronic bonds

$$\begin{aligned} U_{\mathrm{coh}} &= U_{\mathrm{rep}} + U_{\mathrm{band}} - U_{\mathrm{atoms}} \\ &= \frac{1}{2}{\sum_{i,j}}' \phi(r_{ij}) + 2 \sum_{n(\mathrm{occ})} \epsilon^{(n)} - \sum_{i\alpha} N_{i\alpha}^{\mathrm{atom}} \varepsilon_{i\alpha}\,, \end{aligned} \tag{18}$$

where the electronic energy ε of the free atoms has to be subtracted from the energy of the electrons on the molecular orbitals ϵ. Replacing $\epsilon^{(n)}$ with the eigenstates of the TB-Hamiltonian for orbital α at atom i,

$$\epsilon^{(n)} = \epsilon^{(n)}\langle n|n\rangle = \langle n|\epsilon^{(n)}|n\rangle = \langle n|\mathcal{H}|n\rangle = \sum_{i\alpha,j\beta} C_{i\alpha}^{(n)} H_{i\alpha,j\beta} C_{j\beta}^{(n)}\,, \tag{19}$$

the electronic contributions to the cohesive energy $U_{\mathrm{band}} - U_{\mathrm{atom}} = U_{\mathrm{bond}} + U_{\mathrm{prom}}$ can be rearranged and separated in the diagonal and offdiagonal parts:

$$U_{\mathrm{bond}} = 2{\sum_{i\alpha,j\beta}}' \rho_{j\beta,i\alpha} H_{i\alpha,j\beta}, \quad U_{\mathrm{prom}} = \sum_{i\alpha} \left[2\rho_{i\alpha,i\alpha} - N_{i\alpha}^{\mathrm{atom}}\right] \varepsilon_{i\alpha}\,. \tag{20}$$

The third contribution, the promotion energy U_{prom} compares the occupancy of the atomic orbitals of the free atoms with the occupation of the corresponding molecular orbitals [100]. The bond energy U_{bond} describes the energy connected with the exchange or hopping of electrons between arbitrary pairs of atomic neighbors $\{i(j), j(i)\}$, the factor 2 is due to the spin degeneracy. The transition matrix elements of the Hamiltonian are hopping energies, and their transition probability is given by the corresponding element of the density matrix. The contribution for one bond between the atoms i and j is thus characterized by the part of the density called the *bond order* $\Theta_{j\beta,i\alpha}$ that may be expressed as a trace:

$$U_{\mathrm{bond}}^{i,j} = \sum_{\alpha,\beta} H_{i\alpha,j\beta}\Theta_{j\beta,i\alpha} = Tr(\mathcal{H}\Theta)\,. \tag{21}$$

Besides the bond energy, the force exerted on any atom i must be given analytically and therefore one needs the gradient of the potential energy in (18) at the position of the atom i:

$$-\boldsymbol{F}_i = \frac{\partial U_{\mathrm{coh}}}{\partial \boldsymbol{r}_i} = \frac{\partial U_{\mathrm{bond}}}{\partial \boldsymbol{r}_i} + \frac{1}{2}\sum_{j}^{j\neq i} \frac{\partial \phi(r_{ij})}{\partial \boldsymbol{r}_i}\,. \tag{22}$$

This expression includes electronic bonds and ionic pairwise repulsions from all atoms of the system. The general form is still expensive to cope with for the simulation of mesoscopic systems. $\mathcal{O}(N)$ scaling behavior is provided if the cohesive energy of the system is approximated by the Tersoff potential in (8), where $b_{ij}e^{-\mu \boldsymbol{r}_{ij}}$ is replaced by $U_{\mathrm{bond}}^{i,j}$ and a suitable cutoff with resulting balanced interatomic forces is added.

To find an efficient way to obtain the bond energy in a manner that scales linearly with the system size, too, the derivative $\frac{\partial U_{\mathrm{bond}}}{\partial \boldsymbol{r}_i}$ of the bond energy in (22) is replaced assuming a stationary electron density ρ in the electronic ground state and applying the Hellmann–Feynman theorem. The forces can now be obtained without calculating the derivatives of the electronic states and leads to the Hellmann–Feynman force [101, 102]:

$$-\boldsymbol{F}_i^{\mathrm{HF}} = {\sum_{i\alpha,j\beta}}' \Theta_{j\beta,i\alpha}\frac{\partial H_{i\alpha,j\beta}}{\partial \boldsymbol{r}_i}\,. \tag{23}$$

Both the elements of the density matrix and the hopping elements of the Hamiltonian are functions of the relative orientation and separation of the bonding orbitals and have been calculated by *Slater* and *Koster* [103] assuming a linear combination of atomic orbitals (LCAO). To introduce the Slater–Koster matrix elements, usually denoted by $ss\sigma, sp\sigma, pp\pi$, etc., corresponding to the contributing orbitals and its interaction type, the TB Hamiltonian operator is transformed to a tridiagonal form. This can be done using the Lanzcos transformation [104]:

$$\langle U_m \mathcal{H}\rangle U_n = a_n\delta_{m,n} + b_n\delta_{m,n-1} + b_{n+1}\delta_{m,n+1}\,, \tag{24}$$

$$\mathcal{H}|U_n\rangle = a_n|U_n\rangle + b_n|U_{n-1}\rangle + b_{n+1}|U_{n+1}\rangle\,, \tag{25}$$

where the orthonormal basis at the higher level n is recursively developed from $|U_0\rangle$. The coefficients a_n, b_n are the elements of the continuous fraction of the Green's function below.

The offdiagonal elements of the density matrix are related to the Green's function [105],

$$\rho_{i\alpha,j\beta} = -\frac{1}{\pi} \lim_{\eta \to 0} \Im \int^{E_f} \mathrm{d}E \, G_{i\alpha,j\beta}(Z) \,, \tag{26}$$

where the complex variable $Z = E + i\eta$ is the real energy E with a positive, imaginary infinitesimal to perform the integration (theorem of residues). This intersite Green's function can be connected to the site-diagonal Green's function [105] via

$$G_{i\alpha,j\beta}(Z) = \frac{\partial G^{\Lambda}_{00}(Z)}{\partial \Lambda_{i\alpha,j\beta}} + G^{\Lambda}_{00}(Z)\delta_{i,j}\delta_{\alpha,\beta} \,, \tag{27}$$

and the latter can be evaluated recursively [106] using the coefficients of the Lanczos recursion algorithm as mentioned above:

$$G^{\Lambda}_{00}(Z) = \cfrac{1}{Z - a^{\Lambda}_0 - \cfrac{\left(b^{\Lambda}_1\right)^2}{Z - a^{\Lambda}_1 - \cfrac{\left(b^{\Lambda}_2\right)^2}{Z - a^{\Lambda}_2 - \dots}}} \,. \tag{28}$$

The bond order can now be expressed in terms of the derivatives of the recursion coefficients a^{Λ}_n and b^{Λ}_n,

$$\Theta_{i\alpha,j\beta} = -2 \left[\sum_{n=0}^{\infty} \chi^{\Lambda}_{0n,n0} \frac{\partial a^{\Lambda}_n}{\partial \Lambda_{i\alpha,j\beta}} + 2 \sum_{n=1}^{\infty} \chi^{\Lambda}_{0(n-1),n0} \frac{\partial b^{\Lambda}_n}{\partial \Lambda_{i\alpha,j\beta}} \right] \,, \tag{29}$$

with the response function $\chi_{0m,n0}(E_f) = \frac{1}{\pi} \lim_{\eta \to 0} \Im \int^{E_f} \mathrm{d}E G^{\Lambda}_{0m}(Z) G^{\Lambda}_{n0}(Z)$ and the elements $G_{0n} = G_{n0}$ following from the system of equations $(Z - a_n)G_{nm}(Z) - b_n G_{n-1,m}(Z) - b_{n+1}G_{n+1,m}(Z) = \delta_{n,m}$. The more recursion coefficients included in (29), the more accurately the bond order will be approximated. The recursion coefficients are related to the moments of the local density of states (LDOS) [105] and the site-diagonal Green's function of (26) and (27) relates to the LDOS itself. Therefore, the recursive solution of (28) gives an approximation to LDOS in terms of its moments [107]

$$\mu^{(n)}_{i\alpha} = \int E^n n_{i\alpha}(E) \mathrm{d}E = \sum_{\mathrm{all} j_k \beta_k} H_{i\alpha,j_1\beta_1} H_{j_1\beta_1,j_2\beta_2} \dots H_{j_{n-1}\beta_{n-1},i\alpha} \tag{30}$$

and may be interpreted as self-returning (closed) loops of hops of length n for electrons over neighboring atoms. The local atomic environment defines the LDOS via the moments (30), which in turn is used to calculate the bond

order in (21) and the (local) atomic force (23). The remaining free parameters in the analytic form (8) with $U^{i,j}_{\mathrm{bond}}$ instead of b_{ij} may be fitted in the usual way. This is still a hard task, because the bond and the promotion energy involve different parameters as the repulsive part and cutoff parameters and screening functions for all terms have to be included. Besides an accurate fit, the BOP requires well-parameterized TB matrix elements or parameter optimizing, and the problem of transferability [99, 108–110] has to be considered separately. For BOP of order $n = 2$ the bond-order term in the TS-representation reads $b_{ij} = (-ss\sigma_{ij} + pp\sigma_{ij})\Theta_{i\sigma,j\sigma} - 2pp\pi_{ij}\Theta_{i\pi,j\pi}$ and the numerical behavior of BOP2 and TS are approximately equivalent. The details for higher-order BOP are given in the papers of Pettifor's group [2, 99]. The b_{ij} terms of the analytic BOP4 involve complex angular dependencies, partially beyond those neglected in Pettifor's formalism. For structures with defects as well as the wafer bonding of diamond surfaces where π-bonds cannot be neglected, BOP can be found in [111–114] and will soon be published elsewhere with detailed derivations.

4 Applications

It is impossible to review the rapidly growing number of successful applications of empirical MD in materials research. A few representative examples may give an impression of the wide range of problems considered. Isolated point defects are mainly simulated to check the quality of the fit of the potential parameters. Surface reconstructions, adatom and absorption phenomena, growth processes, and especially extended defect structures and interactions can be investigated in detail. The analysis of dislocation core-structures [115], the use of core-structure data for studying dislocation kink motion [116] or the dislocation motion during nanoindentation [117] are examples. Interface investigations have a long tradition, as illustrated in the standard book for grain-boundary structure [118] and growth [119]. Heterophase interfaces, e.g., using a Khor–de-Sama potential for Al and TS for SiC [120] to simulate Al/SiC interfaces, demand special attention to the correct description of the misfit [121] to get good interface energies. Simulations with the Tersoff potential and its modifications yield the correct diameter of the critical nuclei for the growth of Ge nanocrystals in an amorphous matrix [122], allow the study of growth, strains, and stability of Si- and C-nanotubes [123, 124], and SiC surface reconstructions to propose SiC/Si interface structures [125]. A review of atomistic simulations of diffusion and growth on and in semiconductors [126] demonstrates the applicability of SW and TS potentials in comparison with data from TB and DFT calculations.

In the following, two examples of our work are discussed. Empirical MD simulations of first, high-resolution electron microscopy (HREM) image contrasts in quantum dots (Sect. 4.1) and second, the physical processes at interfaces during wafer bonding (Sect. 4.2).

4.1 Quantum Dots: Relaxation, Reordering, and Stability

A quantum dot (QD) is a nanometer-scaled island or region of suitable material free-standing on or embedded in semiconductor or other matrices. The possibility to arrange QDs into complex arrays implies many opportunities for scientific investigations and technological applications. However, depending on the growth techniques applied (mainly MBE and MOCVD), the islands differ in size, shape, chemical composition and lattice strain, which strongly influences the confinement of electrons in nanometer-scaled QDs. The shape, size and strain field of single QDs, as well as quality, density, and homogeneity of equisized and equishaped dot arrangements determine the optical properties, the emission and absorption of light, the lasing efficiency, and other optoelectronic device properties [127, 128]. A critical minimum QD size is required to confine at least one electron/exciton in the dot. A critical maximum QD size is related to the separation of the energy levels for thermally induced decoupling. Uniformity of the QD size is necessary to ensure coupling of states between QDs. The localization of states and their stability depend further on composition and strain of the QDs. The strain relaxation at facet edges and between the islands is the driving force behind self-organization and lateral arrangement, vertical stacking on top or between buried dots, or preordering by surface structuring. In addition, an extension of the emission range towards longer wavelengths needs a better understanding and handling of the controlled growth via lattice mismatched heterostructures or self-assembling phenomena (see, e.g., [129]).

A wide variety of imaging methods are used to investigate growth, self-assembly, and physical properties of quantum dots. Among these the cross-sectional HREM and the plan-view TEM imaging techniques are suitable methods to characterize directly shape, size, and strain field [130]. But the HREM and TEM techniques images are difficult to interpret phenomenologically, especially when separating shape and strain effects. Modeling is essential to uniquely determine the features and provide contrast rules.

In Fig. 4, the atomic structure of an InGaAs-QD in a GaAs matrix (for other systems cf. [130, 131]) is prescribed by geometric models and relaxed by MD simulations. Very different dot shapes have been proposed and theoretically investigated: lens-shaped dots, conical islands, volcano-type QDs, and pyramids with different side facets of type {011}, {111}, {112}, {113}, {136}, and both $\{011\}+\{111\}$ mixed, etc. Some of these and a spherical cap are schematically presented in Fig. 4a; in simulations one or two monolayers (ML) thick wetting layers are included, too. The most important difference between the various structures is the varying step structure of the facets due to their different inclination. There are at least two reasons to investigate these configurations [130]. First, small embedded precipitates always have facets; a transition between dome-like structure and pyramids due to changes in spacer distance, change the number and arrangement of the facets, and thus strain and electronic properties. Second, for highly faceted struc-

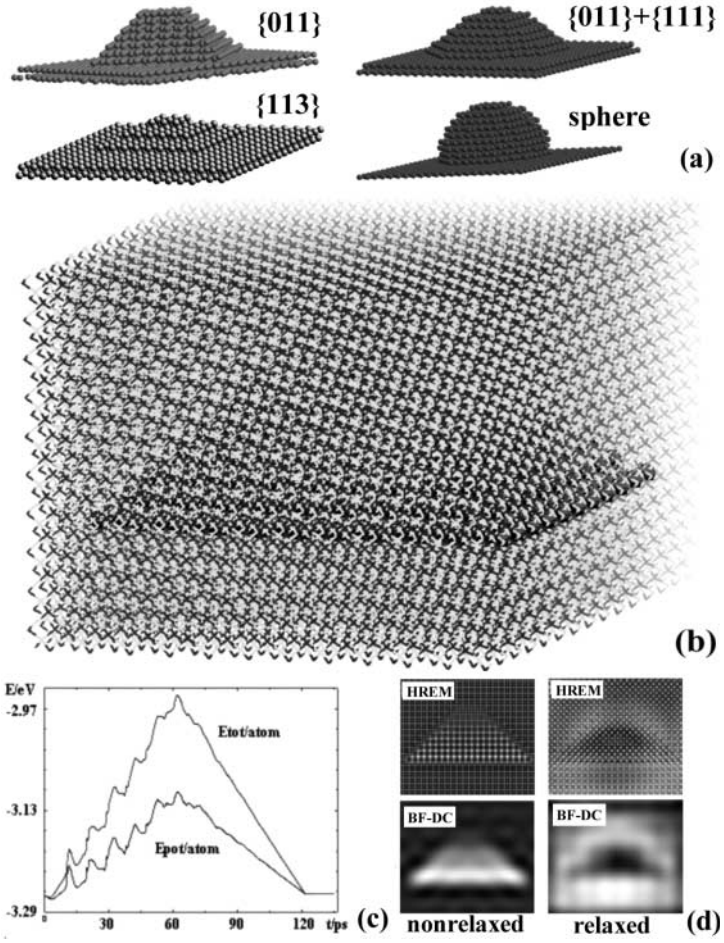

Fig. 4. Structure modeling and image simulation of different pyramidal-shaped quantum-dot configurations: (**a**) different faceting, truncation, and wetting of pyramidal start models (matrix removed, models related to (001)-base planes), (**b**) relaxed complete model of a {011} pyramid (base length about 6 nm, $10 \times 10 \times 10$-supercell length 10 nm), (**c**) energy relaxation of a {011} quantum dot (potential E_{pot} and total E_{tot} energy versus time steps), (**d**) cross-sectional HREM and bright-field diffraction contrast simulated for model (**b**) before and after relaxation assuming a standard 400 kV microscope at the Scherzer focus (i.e., $\Delta = -40\,\mathrm{nm}$, $C_{\mathrm{s}} = 1\,\mathrm{mm}$, $\alpha = 1.2\,\mathrm{nm}^{-1}$, $t = 9\,\mathrm{nm}$, $\delta = 8\,\mathrm{nm}^{-1}$, cf. [130])

tures the continuum elasticity is practically inapplicable and finite element calculations must be done in 3-D instead of 2-D. However, the technique of ab-initio MC provides a more accurate prediction of the QD shapes than empirical MD [132]. The embedding of one perfect {011} pyramid in a matrix is demonstrated in Fig. 4b after prerelaxation. Figure 4c shows typical annealing behavior during empirical MD calculation, characterized by the potential E_{pot} and the total energy E_{tot} per atom. The energy difference $E_{\mathrm{tot}} - E_{\mathrm{pot}}$ is equal to the mean kinetic energy and is thus directly related to the temperature of the system. After the prerelaxation of 5 ps at 0 K, an annealing cycle follows, 60 ps stepwise heat up to 600 K and cool down to 0 K, equilibrating the system at each heating step. The example here demonstrates a short cycle; most of the embedded QDs are relaxed at each T-step for at least 10 000 timesteps of 0.25 fs, followed by annealing up to about 900 K (the temperature is not well defined with empirical potentials but is below the melting temperature). Whereas the structure in Fig. 4b is less strained, highly strained configurations occur due to the self-interaction of the QD in small supercells that correspond to a stacked sequence with very small dot distances. The extension of the supercell chosen in the simulations depends on the extension of the QD to be investigated, as well as on the overlap of the strain fields at the borders, to avoid self-interactions. Whereas Fig. 4b shows a relatively small supercell, investigations are made for supercells of up to $89 \times 89 \times 89$ unit cells with a base length of the QD of 9 nm containing several million atoms. For the image simulations, subregions of the supercells are used, sliced into one atomic layer and applying the multislice image simulation technique. By comparing imaging for structures before and after relaxation, Fig. 4d demonstrates the enormous influence of the relaxations on the image contrast in cross-sectional HREM and TEM [130, 131]. In summarizing some of the results one can state that MD calculations with SW (CdZnSe) and TII (InGaAs, Ge) allow us to obtain well-relaxed structures, in contrast to the static relaxations performed with the *Keating* potential [133, 134]. With this insight into the atomic processes of rearranging and straining QDs at an atomic level the growth conditions for quantum dots may be enhanced as a first step to tailoring their properties.

4.2 Bonded Interfaces: Tailoring Electronic or Mechanical Properties?

Wafer bonding, i.e., the creation of interfaces by joining two wafer surfaces, has become an attractive method for many practical applications in microelectronics, micromechanics or optoelectronics [135]. The macroscopic properties of bonded materials are mainly determined by the atomic processes at the interfaces (clean and polished hydrophobic or hydrophilic surfaces, as schematically shown in Fig. 5) during the transition from adhesion to chemical bonding. For this, elevated temperatures or external forces are required, as could be revealed by MD simulations of hydrogen-passivated interfaces [136]

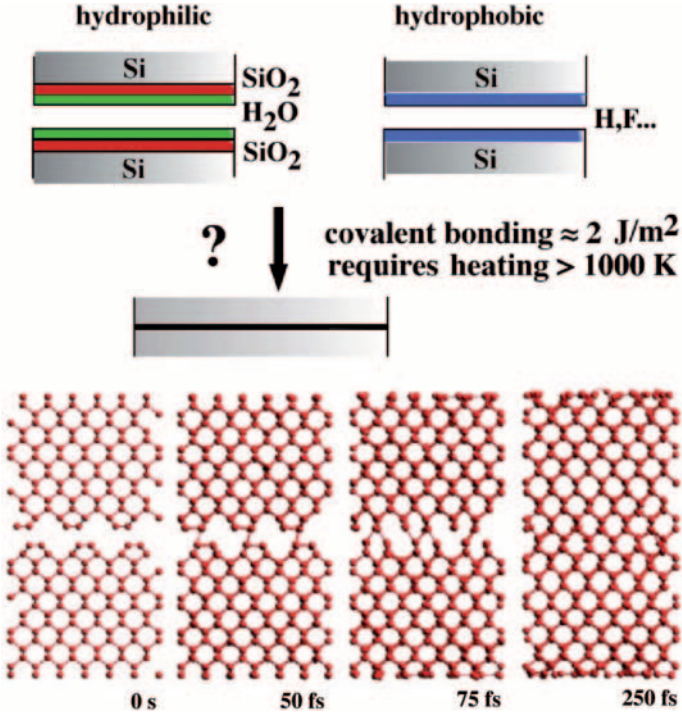

Fig. 5. Wafer bonding at Si[100] surfaces: simulations of surface rearrangement for perfect alignment and slow heating. The figure describes the evolution of the atomic relaxation leading to bonded wafers

or of silica bonding [65]. Thus, describing atomic processes is of increasing interest to support experimental investigations or to predict bonding behavior. Already, slightly rescaled SW potentials predict the bond behavior via bond breaking and dimer reconfiguration as shown in the snapshots explaining the possibility of covalent bonding at room temperature for very clean hydrophobic surfaces under UHV conditions [137].

Whereas the bonding of two perfectly aligned, identical wafers gives a single, perfectly bonded wafer without defects, miscut of the wafer results in steps on the wafer surfaces and edge dislocations at the bonded interfaces are created. In Fig. 6 the red color describes the potential energy above the ground state during wafer bonding over two-atomic steps. The upper terraces behave like perfect surfaces and the dimerized starting configuration of Fig. 6a is rearranged and new bonds are created. The energy gained is dissipated and for slow heat transfer the avalanche effect leads to bonding of the lower terraces, too. Two 60° dislocations remain and, depending on the rigid shift of the start configuration, an additional row of vacancies [137]. The dislocations may split as simulated and observed in experiments [138].

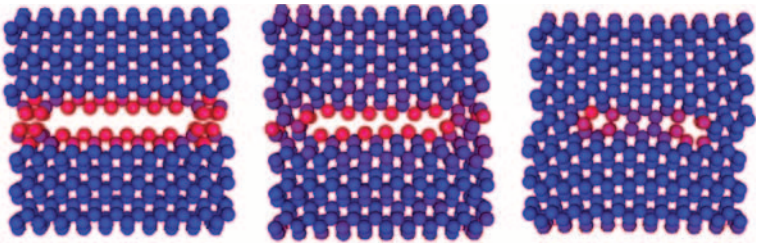

Fig. 6. MD simulation of wafer bonding over surface steps. Note the dislocation relaxation for the initial configuration (*left*), a snapshot during annealing up to 900 K, after 12.5 ps (*middle*) and the final relaxed state (*right*)

Bonding wafers with rotational twist leads to a network of screw dislocations at the interface. A special situation is the 90° twist, which always occurs between monoatomic steps. A Stillinger–Weber potential applied to a 90°-twist bonded wafer pair [16] yields a metastable fivefold-coordinated interface with a mirror symmetry normal to the interface characterized by a $Pmm(m)$-layer group (cf. Fig. 2a). Using the Tersoff or BOP-like potentials [113] and metastable or well-prepared starting configurations allows further structure relaxation and energy minimization. Figure 2b shows this relaxed configuration, which is (2×2) reconstructed and consists of structural units with a $\overline{4}2m$-(D2d) point group symmetry, called the $\overline{4}2m$-*dreidl*. It should be emphasized that the *dreidl* structure is found to be the minimum energy configuration also in DFT-LDA simulations [16]. However, the energies differ from those given in [139, 140]. Much more important is the modification of the band structure due to the different interface relaxation that may enable engineering of the electronic properties: whereas the metastable configuration (cf. Fig. 2c) yields semimetallic behavior, the *dreidl* structure (cf. Fig. 2d) yields a larger bandgap than in perfect lattices. The *dreidl* interface structure and its band-structure modification is very similar to the essential building blocks proposed by *Chadi* [141] for group-IV materials. They describe geometry and properties of the transformation of Si and Ge under pressure and the special *allo*-phases as a new class of crystalline structures.

A small misalignment of the wafers during wafer bonding yields bonded interfaces with twist rotation resulting in a checkerboard-like interface structure [142, 143]. Figure 7 shows some of the resulting minimum structures gained for higher annealing temperatures and different twist rotation angles. Before the bonding process takes place, the superposition of the two wafers looks like a Moiré pattern in the projection normal to the interface. After bonding and sufficient relaxation under slow heat-transfer conditions, almost all atoms have a bulk-like environment separated by misfit screw dislocations, which may have many kinks. The screw dislocation-network of the bonded wafer has a period half that of the Moiré pattern. One finds more imperfectly bonded regions around the screw dislocations for smaller twist

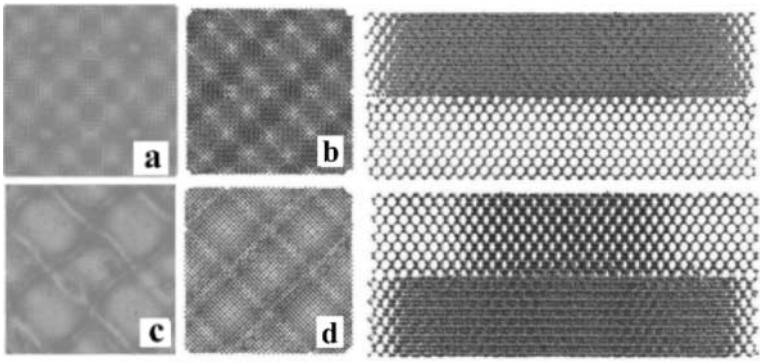

Fig. 7. MD relaxations of bonding rotationally twisted wafers ([001] and [110] views) with different angles: (**a**) 2.8°, 22 nm box, orthogonal dimers, [001] view; (**b**) 6.7°, 9.2 nm box, orthogonal dimers, [001] and [110] views; (**c**) as (**a**) and (**d**) as (**b**) with parallel dimers

angles, whereas bonding at higher angles results in more or less widely spread strained interface regions. In Fig. 7 bonding with orthogonal start configurations (a,b) is compared to those with parallel dimer start configurations (c,d). Thus the bonding of Figs. 7a and b may be considered as bonding with an additional 90° twist rotation. Clearly the periodicity of the defect region is twice those of Figs. 7b and c smoothing out the interface, but creating additional shear strains. Irrespective of the chosen twist angles and box dimensions all final structures yield bond energies of approximately 4.5 eV/atom at 0 K, however, varying slightly with the twist angle. A maximum occurs between 4° and 6° twist related to a change of the bonding behavior itself. The higher the annealing temperature the better the screw formation. In conclusion, simulation of the atomic processes at wafer-bonded interfaces offers not only the tailoring of electronic properties, it is also an important step in understanding how to control the creation of special interface structures for strain accommodation or prepatterned templates (compliant substrates) [135] by rotational alignment.

5 Conclusions and Outlook

Molecular dynamics simulations based on empirical potentials provide a suitable tool to study atomic processes that influence macroscopic materials properties. The applicability of the technique is demonstrated by calculating quantum-dot relaxations and interaction processes with defect creation at wafer-bonded interfaces. A brief overview describes the method itself and its advantages and limitations, i.e., the macroscopic relevance of the simulations versus the limited transferability of the potentials. The quality of the simulations and the reliability of the results depend on the coupling of the

atomistic simulations across length and timescales. Whereas the Lagrange formalism is well established for the embedding into suitable environments at the continuum level, the approximations of first-principles-based potentials need continuous future work.

References

[1] W. Kohn: Rev. Mod. Phys. **71**, 1253 (1999) 214

[2] M. W. Finnis: Interatomic forces in condensed matter, in *Oxford Series on Materials Modelling* (Oxford University Press, Oxford 2003) 214, 215, 221, 222, 227, 230

[3] K. Ohno, K. Esforjani, Y. Kawazoe: *Computational Materials Science: From ab initio to Monte Carlo Methods* (Springer, Berlin, Heidelberg 1999) 214

[4] M. C. Payne, M. P. Teter, D. C. Allen, et al: Rev. Mod. Phys. **64**, 1045 (1992) 214

[5] G. Galli, A. Pasquarello: First-principles molecular dynamics, in M. P. Allen, D. J. Tildesley (Eds.): *Computer Simulation in Chemical Physics* (Academic Publisher, Dordrecht, Netherlands 1993) pp. 261–313 214

[6] T. Frauenheim, G. Seifert, M. Elstner, et al: phys. stat. sol. B **217**, 41 (2000) 214, 227

[7] D. R. Bowler, T. Miyazaki, M. J. Gillan: J. Phys.: Condens. Matter **14**, 2781 (2002) 214, 227

[8] J. M. Soler, E. Artacho, J. D. Gale, et al: J. Phys.: Condens. Matter **14**, 2745 (2002) 214, 227

[9] M. D. Segall, P. J. D. Lindan, M. J. Probert, et al: J. Phys.: Condens. Matter **14**, 2717 (2002) 214

[10] R. Car, M. Parinello: Phys. Rev. Lett. **55**, 2471 (1985) 214

[11] G. C. Csany, T. Albaret, M. C. Payne, A. DeVita: Phys. Rev. Lett. **93**, 175503 (2004) 215

[12] P. Ruggerone, A. Kley, M. Scheffler: Bridging the length and time scales: from ab initio electronic structure calculations to macroscopic proportions, Comm. Condens. Matter Phys. **18**, 261 (1998) 215

[13] R. M. Nieminen: J. Phys.: Condens. Matter **14**, 2859 (2002) 215, 216

[14] R. E. Rudd, J. Q. Broughton: phys. stat. sol. B **217**, 251 (2000) 215, 216

[15] B. I. Lundquist, A. Bogicevic, S. Dudy, et al: Comput. Mater. Sci. **24**, 1 (2002) 215, 216

[16] K. Scheerschmidt, D. Conrad, A. Belov: Comput. Mater. Sci. **24**, 33 (2002) 216, 235

[17] J. M. Haile: *Molecular Dynamics Simulation: Elementary Methods* (Wiley, New York 1992) 216

[18] D. C. Rapaport: *The Art of Molecular Dynamics Simulation* (Cambridge Univ. Press, Cambridge 1998) 216

[19] R. Haberlandt: *Molekulardynamik: Grundlagen und Anwendungen* (Vieweg, Braunschweig, Wiesbaden 1995) 216

[20] W. G. Hoover: Molecular dynamics, in *Lecture Notes in Physics*, vol. 258 (Springer, Berlin, Heidelberg 1986) 216

[21] I. G. Kaplan: Theory of molecular interactions, in *Studies in Physical and Theoretical Chemistry*, vol. 42 (Elsevier Science Ltd., Amsterdam 1986) 216
[22] M. Sprik: Introduction to molecular dynamics methods, in K. Binder, G. Cicotti (Eds.): *Monte Carlo and Molecular Dynamics of Condensed Matter Systems*, vol. 49, Conf. Proc. Vol. (SIF, Bologna 1996) Chap. 2, pp. 47–87 216
[23] D. W. Heermann: *Computer Simulation Methods in Theoretical Physics* (Springer, Berlin, Heidelberg 1990) 216
[24] J. M. Ziman: *Principles of the Theory of Solids* (Cambridge Univ. Press, Cambridge 1972) 218
[25] S. W. deLeeuw, J. W. Perram, E. R. Smith: Proc. Roy. Soc. London A **373**, 27 (1980) 218
[26] D. Wolf, P. Keblinski, S. R. Phillpot, J. Eggebrecht: J. Chem. Phys. **110**, 8254 (1999) 218
[27] A. F. Voter: Phys. Rev. B **57**, 13985 (1998) 218
[28] A. F. Voter, F. Montalenti, T. C. Germann: Ann. Rev. Mater. Res. **32**, 321 (2002) 218
[29] W. Rullmann, H. Gunsteren, H. Berendsen: Mol. Phys. **44**, 69 (1981) 221
[30] S. Nosé: Mol. Phys. **52**, 255 (1984) 221
[31] S. Nosé: J. Chem. Phys. **81**, 511 (1984) 221
[32] W. G. Hoover: Phys. Rev. A **31**, 695 (1985) 221
[33] H. C. Andersen: J. Chem. Phys. **72**, 2384 (1980) 221
[34] M. Parinello, A. Rahman: J. Appl. Phys. **52**, 7182 (1981) 221
[35] L. Yang, D. J. Srolowitz, A. F. Yee: J. Chem. Phys. **107**, 4396 (1997) 221
[36] G. P. Morriss, D. J. Evans: Phys. Rev. A **35**, 792 (1987) 221
[37] D. J. Evans, G. P. Morriss: *Statistical Mechanics of Nonequilibrium Liquids* (Academic, London, New York 1990) 221
[38] J. D. Gale: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 3 (1996) 221
[39] A. Horsfield: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 85 (1996) 221
[40] S. C. Price: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 95 (1996) 221
[41] J. N. Murrell: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 163 (1996) 221
[42] D. L. Cooper, T. Thorsteinsson: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 175 (1996) 221
[43] M. C. Payne, I. J. Robertson, D. Thomssen, V. Heine: Special Issue: Interatomic Potentials, Philos. Mag. B **73**, 191 (1996) 221
[44] Guest Editor A. F. Voter, V. Vitek: MRS-Bull. **21**, 20 (1992) 221
[45] S. M. Foiles: MRS-Bull. **21**, 24 (1992) 221
[46] M. Stoneham, J. Hardy, T. Harker: MRS-Bull. **21**, 29 (1992) 221
[47] D. W. Brenner: MRS-Bull. **21**, 36 (1992) 221
[48] A. P. Sutton, P. D. Goodwin, A. P. Horsefield: MRS-Bull. **21**, 42 (1992) 221
[49] N. A. Marks, M. W. Finnis, J. H. Harding, N. C. Pyper: J. Chem. Phys. **114**, 4406 (2001) 222
[50] K. W. Jacobsen, J. K. Norskov, M. J. Puskas: Phys. Rev. B **35**, 7423 (1987) 222
[51] N. A. Marks: J. Phys. Condens. Matter **14**, 2901 (2002) 222
[52] M. S. Daw: Phys. Rev. B **39**, 7441 (1989) 222

[53] Y. C. Wang, K. Scheerschmidt, U. Gösele: Phys. Rev. B **61**, 12864 (2000) 222
[54] H. R. Trebin: Comp. Phys. Commun. **121–122**, 536 (1999) 222
[55] A. K. Rappe, C. J. Casewit, K. S. Colwell, et al: J. Am. Chem. Soc. **114**, 10024 (1992) 222
[56] W. F. van Gunsteren, H. J. C. Berendsen: Angew. Chem., Int. Ed. Engl. **29**, 992 (1990) 222
[57] F. H. Stillinger, T. A. Weber: Phys. Rev. B **31**, 5262 (1985) 222
[58] R. Biswas, D. R. Hamann: Phys. Rev. B **36**, 6434 (1987) 222
[59] T. Takai, T. Halicioglu, W. A. Tiller: Scr. Metal. **19**, 709 (1985) 222
[60] C. P. Herrero: J. Mater. Res. **16**, 2505 (2001) 222
[61] A. K. Nakano, B. J. Berne, P. Vashista, R. K. Kalia: Phys. Rev. B **39**, 12520 (1989) 223
[62] S. H. Garofalini: J. Non-Cryst. Solids **120**, 1 (1990) 223
[63] D. Timpel, K. Scheerschmidt, S. H. Garofalini: J. Non-Cryst. Solids **221**, 187 (1997) 223
[64] D. Timpel, K. Scheerschmidt: J. Non-Cryst. Solids **232–234**, 245 (1998) 223
[65] D. Timpel, M. Schaible, K. Scheerschmidt: J. Appl. Phys. **85**, 2627 (1999) 223, 234
[66] M. I. Baskes: Phys. Rev. B **46**, 2727 (1992) 223
[67] M. W. Finnis, J. E. Sinclair: Philos. Mag. A **50**, 45 (1984) 223
[68] J. Tersoff: Phys. Rev. Lett. **56**, 632 (1986) 223
[69] J. Tersoff: Phys. Rev. B **38**, 9902 (1989) 223
[70] J. Tersoff: Phys. Rev. B **39**, 5566 (1989) 223
[71] W. Dodson: Phys. Rev. B **35**, 2795 (1987) 223
[72] G. C. Abell: Phys. Rev. B **31**, 6184 (1985) 223
[73] D. W. Brenner: phys. stat. sol. B **217**, 23 (2000) 223
[74] M. V. R. Murty, H. A. Atwater: Phys. Rev. B **51**, 4889 (1995) 223
[75] D. W. Brenner: Phys. Rev. B **42**, 9458 (1990) 223
[76] A. J. Dyson, P. V. Smith: Surf. Sci. **355**, 140 (1996) 223
[77] K. Beardmore, R. Smith: Philos. Mag. A **74**, 1439 (1996) 223
[78] K. Nordlund, J. Nord, J. Frantz, J. Keinonen: Comput. Mater. Sci. **18**, 504 (2000) 223
[79] T. J. Lenosky, B. Sadigh, E. Alonso, et al: Model. Simul. Mater. Sci. Eng. **8**, 825 (2000) 223
[80] S. D. Chao, J. D. Kress, A. Redondo: J. Chem. Phys. **120**, 5558 (2004) 223
[81] H. Balamane, T. Halicioglu, W. A. Tiller: Phys. Rev. B **46**, 2250 (1992) 223
[82] M. Kohyama, S. Takeda: Phys. Rev. B. **60**, 8075 (1999) 223
[83] M. Gharaibeh, S. K. Estreicher, P. A. Fedders: Physica B **273–274**, 532 (1999) 223
[84] S. K. Estreicher: phys. stat. sol. B **217**, 513 (2000) 223
[85] W. Windl: phys. stat. sol. B **226**, 37 (2001) 223
[86] S. Kohlhoff, P. Gumbsch, H. F. Fischmeister: Philos. Mag. A **64**, 851 (1991) 225
[87] R. E. Rudd, J. Q. Broughton: Phys. Rev. B **58**, R5893 (1998) 225
[88] A. Y. Belov: Dislocations emerging at planar interfaces, in V. L. Indenbom, J. Lothe (Eds.): *Elastic Strain Fields and Dislocation Mobility* (Elsevier, Amsterdam 1992) pp. 391–446 225, 226
[89] J. E. Sinclair: J. Appl. Phys. **42**, 5321 (1971) 225

[90] D. G. Pettifor: Phys. Rev. Lett. **63**, 2480 (1989) 225, 227
[91] C. M. Goringe, D. R. Bowler, E. Hernandez: Rep. Prog. Phys. **60**, 1447 (1997) 227
[92] A. P. Horsfield, A. M. Bratkovsky: J. Phys. Condens. Matter **12**, R1 (2000) 227
[93] D. A. Papaconstantopoulos, M. J. Mehl: J. Phys. Condens. Matter **15**, R413 (2003) 227
[94] D. G. Pettifor, I. I. Oleinik, D. Nguyen-Manh, V. Vitek: Comput. Mater. Sci. **23**, 33 (2002) 227
[95] D. Pettifor: *From Exact to Approximate Theory: The Tight Binding Bond Model and Many-Body Potentials.*, vol. 48, Springer Series (Springer, Berlin, Heidelberg 1990) 227
[96] D. G. Pettifor, I. I. Oleinik: Phys. Rev. B **59**, 8487 (1999) 227
[97] D. G. Pettifor, I. I. Oleinik: Phys. Rev. Lett. **18**, 4124 (2000) 227
[98] D. G. Pettifor, M. W. Finnis, D. Nguyen-Manh, et al: Mater. Sci. Eng. A **365**, 2 (2004) 227
[99] A. P. Sutton: *Electronic Structure of Materials* (Clarendon Press, Oxford 1994) 227, 230
[100] A. P. Sutton, M. W. Finnis, D. G. Pettifor, Y. Ohta: J. Phys. C **35**, 35 (1988) 228
[101] H. Hellmann: *Einführung in die Quantenchemie* (Deuticke, Leipzig 1937) 228
[102] R. P. Feynman: Phys. Rev. **56**, 340 (1939) 228
[103] J. C. Slater, G. F. Koster: Phys. Rev. **94**, 1498 (1954) 228
[104] C. Lanczos: J. Res. Natl. Bur. Stand. **45**, 225 (1950) 228
[105] A. P. Horsfield, A. M. Bratovsky, M. Fear, et al: Phys. Rev. B **53**, 12694 (1996) 229
[106] R. Haydock: *Recursive Solution of the Schrödinger Equation*, vol. 35, Solid State Physics (Academic, London, New York 1980) 229
[107] F. Ducastelle: J. Phys. **31**, 1055 (1970) 229
[108] O. F. Sankey, D. J. Niklewski: Phys. Rev. B **40**, 3979 (1989) 230
[109] I. Kwon, R. Biswas, C. Z. Wang, et al: Phys. Rev. B **49**, 7242 (1994) 230
[110] T. J. Lenosky, J. D. Kress, I. Kwon, et al: Phys. Rev. B **55**, 1528 (1997) 230
[111] D. Conrad: *Molekulardynamische Untersuchungen für Oberflächen und Grenzflächen von Halbleitern*, Thesis, Martin-Luther-Universität, Halle (1996) 230
[112] V. Kuhlmann: *Entwicklung analytischer bond-order Potentiale für die empirische Molekulardynamik*, Thesis, Martin-Luther-Universität, Halle (2006) 230
[113] D. Conrad, K. Scheerschmidt: Phys. Rev. B **58**, 4538 (1998) 230, 235
[114] D. Conrad, K. Scheerschmidt, U. Gösele: Appl. Phys. Lett. **77**, 49 (2000) 230
[115] T. Harry, D. Bacon: Acta Mater. **50**, 195 (2002) 230
[116] V. V. Bulatov, J. F. Justo, W. Cai, et al: Philos. Mag. A **81**, 1257 (2001) 230
[117] E. T. Lilleodden, J. A. Zimmermann, S. M. Foiles, W. D. Nix: J. Mech. Phys. Solids **51**, 901 (2003) 230
[118] D. Wolf, S. Yip (Eds.): *Materials Interfaces: Atomic-Level Structure and Properties* (Chapman & Hall, London 1992) 230
[119] A. J. Haslam, D. Moldovan, S. R. Phillpot, et al: Comput. Mater. Sci. **23**, 15 (2002) 230
[120] X. Luo, G. Qian, E. G. Wang, C. Chen: Phys. Rev. B **59**, 10125 (1999) 230

[121] R. Benedek, D. N. Seidman, C. Woodward: J. Phys. Condens. Matter **14**, 2877 (2002) 230
[122] K. J. Bording, J. Tafto: Phys. Rev. B **62**, 8098 (2000) 230
[123] J. W. Kang, J. J. Seo, H. J. Hwang: J. Nanosci. Nanotechnol. **2**, 687 (2002) 230
[124] D. Srivasta, D. W. Brenner, J. D. Schall, et al: J. Phys. Chem. B **103**, 4330 (1999) 230
[125] C. Koitzsch, D. Conrad, K. Scheerschmidt, U. Gösele: J. Appl. Phys. **88**, 7104 (2000) 230
[126] E. Kaxiras: Comput. Mater. Sci. **6**, 158 (1996) 230
[127] D. Bimberg, M. Grundmann, N. N. Ledentsov: *Quantum Dot Heterostructures* (John Wiley & Sons., Chicester 1999) 231
[128] L. W. Wang, A. Zunger: Pseudopotential theory of nanometer silicon quantum dots, in P. V. Kamat, D. Meisel (Eds.): *Semiconductor Nanoclusters – Physical, Chemical, and Catalytic Aspects* (Elsevier, Amsterdam, New York 1997) 231
[129] H. Lee, J. A. Johnson, M. Y. He, et al: Appl. Phys. Lett. **78**, 105 (2001) 231
[130] K. Scheerschmidt, P. Werner: Characterization of structure and composition of quantum dots by transmission electron microscopy., in M. Grundmann (Ed.): *Nano-Optoelectronics: Concepts, Physics and Devices* (Springer, Berlin, Heidelberg, New York, Tokyo 2002) Chap. 3, pp. 67–98 231, 232, 233
[131] K. Scheerschmidt, D. Conrad, H. Kirmse, R. Schneider, W. Neumann: Ultramicroscopy **81**, 289 (2000) 231, 233
[132] E. Pehlke, N. Moll, A. Kley, M. Scheffler: Appl. Phys. A **65**, 525 (1997) 233
[133] P. N. Keating: Phys. Rev. **145**, 637 (1966) 233
[134] Y. Kikuchi, H. Sugii, K. Shintani: J. Appl. Phys. **89**, 1191 (2001) 233
[135] Q. Y. Tong, U. Gösele: *Semiconductor Wafer Bonding: Science and Technology* (Wiley, New York 1998) 233, 236
[136] D. Conrad, K. Scheerschmidt, U. Gösele: Appl. Phys. Lett. **71**, 2307 (1997) 233
[137] D. Conrad, K. Scheerschmidt, U. Gösele: Appl. Phys. A **62**, 7 (1996) 234
[138] A. Y. Belov, R. Scholz, K. Scheerschmidt: Philos. Mag. Lett. **79**, 531 (1999) 234
[139] A. Y. Belov, D. Conrad, K. Scheerschmidt, U. Gösele: Philos. Mag. **77**, 55 (1998) 235
[140] A. Y. Belov, K. Scheerschmidt, U. Gösele: Phys. Status Solidi A **159**, 171 (1999) 235
[141] D. J. Chadi: Phys. Rev. B **32**, 6485 (1985) 235
[142] K. Scheerschmidt: MRS Proc. **681E**, I2.3 (2004) 235
[143] K. Scheerschmidt, V. Kuhlmann: Interf. Sci. **12**, 157 (2004) 235

Index

Defects in Amorphous Semiconductors: Amorphous Silicon

D.A. Drabold and T.A. Abtew

Department of Physics and Astronomy, Ohio University, Athens, Ohio
drabold@ohio.edu

Abstract. Defects in disordered (amorphous) semiconductors are discussed, with an emphasis on hydrogenated amorphous silicon. The general differences between defect phenomena in crystalline and amorphous hosts are described, and the special importance of the electron–phonon coupling is stressed. Detailed calculations for amorphous Si are presented using accurate first principles (density-functional) techniques. The various approximations of ab-initio simulation affect aspects of the network structure and dynamics, and suitably accurate approximations are suggested. Defect dynamics and the motion of hydrogen in the network are reported.

1 Introduction

This book is primarily concerned with defects in crystalline materials. In this Chapter, we depart from this and discuss defects in *amorphous* materials. In the Chapter of Simdayankin and Elliott, a most interesting feature of amorphous materials, photoresponse, is discussed in detail. As in the case of crystals, defects determine key features of the electronic, vibrational and transport properties of amorphous materials [1–3]. These are often precisely the properties that are relevant to applications. In this Chapter we strongly emphasize amorphous silicon (a-Si), while not focusing on it exclusively. This is significantly due to the background of the authors and constraints on the length of the Chapter, but it is also true that a-Si offers at least a somewhat generic theoretical laboratory for the study of disorder, defects, and other aspects of amorphous materials in general.

The outline of this Chapter is as follows: First, we briefly survey amorphous semiconductors. Next, we define the notion of a defect, already a somewhat subtle question in an environment that is intrinsically variable in structure. In the fourth section, we discuss the generic electronic and vibrational attributes of defects in amorphous solids, and briefly comment on the current methods for studying these systems. In Sect. 5, we discuss calculations of the *dynamics* of defects in the a-Si:H network, of critical importance to the stability and practical application of the material.

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 245–268 (2007)

2 Amorphous Semiconductors

Amorphous materials and glasses are among the most important to human experience with applications ranging from primitive obsidian weapons to optoelectronics. Today, every office in the world contains storage media for computers (for example DVD-RW compact disks) that exploit the special reversible laser-driven amorphization/crystallization transition properties of a particular GeSbTe glass. The same computer uses crystalline Si chips that depend critically upon the dielectric properties of a-SiO_2. With increasing pressure on global energy markets, it is notable that some of the most promising photovoltaic devices are based upon amorphous materials such as a-Si:H because of the low cost of the material (compared to crystalline devices) and the ability to grow thin films of device-quality material over wide areas. The internet depends upon fiber-optic glass light pipes that enable transmission of information with bandwidth vastly exceeding that possible with wires. This list is the proverbial tip of the iceberg.

Experimental measurements for structure determination, such as X-ray or neutron diffraction, lead easily to a precise determination of the structure of crystals. Nowadays, protein structures with thousands of atoms per unit cell are readily solved. The reason for this impressive success is that the diffraction data (structure factor) consists of a palisade of sharply defined spikes (the Bragg peaks), arising from reflection from crystal planes. For amorphous materials, the same experiments are rather disappointing; smooth broad curves replace the Bragg peaks. The wavelength-dependent structure factor may be interpreted in real space through the radial distribution function (RDF), which is easily obtained with a Fourier transform of the structure factor. In crystals, the RDF consists again of spikes, and the radii at which the spikes occur are the neighbor distances. In the amorphous case, the RDF is smooth, and normally has broadened peaks near the locations of the first few peaks of the crystal. This similarity in the small-r peak positions reflects a tendency of amorphous materials to attempt to mimic the local order of the crystalline phase in the amorphous network, but usually with modest bond-length and bond-angle distortions. At distances beyond several nearest-neighbor spacings, similarities between crystalline and amorphous pair correlations wane, and the pair correlations decay to zero after tens of Ångstroms in the amorphous material (this number is system dependent).

From an information theoretic point of view (for example, consideration of the Shannon–Jaynes information entropy [4]), it is clear that there is less information inherent in the (smooth) data for the amorphous case relative to the crystal. So, while it is possible to invert the diffraction data from crystals to obtain structure determination (with some assumptions to fix the phase problem), such a process fails for the amorphous case and many studies have emphasized the multiplicity of distinct structures possible that can reproduce measured diffraction data. In this sense we are in a situation rather like high-energy or nuclear physics where there are sum rules that must be obeyed,

but the sum rules by themselves offer an extremely incomplete description of the physical processes. The first goal of theoretical work in amorphous materials is to obtain an experimentally realistic model consistent with all trusted experiments and also accurate total energy and force calculations.

Early models of amorphous materials were made by hand, but this approach was soon supplanted by simulations with the advent of digital computers. A variety of modeling schemes based upon simulated atomic dynamics (so-called molecular-dynamics), Monte Carlo methods, methods based upon attempts to obtain the structure from the diffraction measurements (most sensibly done with constraints to enforce rules on bonding) and even hybrids between the last two are in current use [5–8]. Impressive progress has been made in the last 25 years of modeling, and highly satisfactory models now exist for many amorphous materials, among these: Si, Si:H, Ge, C, silica, elemental and binary chalcogenides [9–12]. Empirically, systems with lower mean coordination (floppier in the language of *Thorpe* [13]) are easier to model than higher-coordination systems. Multinary materials tend to be more challenging than elemental systems as the issue of chemical ordering becomes important and also these systems are harder to model accurately with conventional methods. Empirical potentials, tight-binding schemes and ab-initio methods are used for modeling the interatomic interactions [5, 14–16]. In the area of defects in amorphous semiconductors, it is usually the case that first-principles interactions are required.

Where electronic properties of amorphous semiconductors are concerned, k-space methods are not useful as the crystal momentum is not a good quantum number (the translation operator does not commute with the Hamiltonian). Rather, one focuses on the density of electron states, and on individual eigenvectors of the Hamiltonian, especially those near the Fermi level. The electronic density of states is usually qualitatively like that of a structurally related crystal, but broadened by disorder. In Fig. 1, we illustrate these features with a "real" calculation on a 10 000-atom model of a-Si [11, 17]. The sharp band edges expected in crystals are broadened into smooth "band-tails". The tails are created by structural disorder; early Bethe lattice calculations demonstrated that strained bond angles lead to states pushed out of the band (into the forbidden optical gap), and loosely, the more severe the defect, the further into the gap. It is known from optical studies of amorphous semiconductors that all such systems have band-tails decaying exponentially into the gap. While the optical spectrum is a convolution involving both tails, photoemission studies have separately probed the valence tail and conduction tails particularly in a-Si:H [18], and found that both decay exponentially, albeit with different rates and a temperature dependence (quite different for the two tails) that suggests the importance of thermal disorder beside the structural order we have stressed so far [19].

The field of amorphous semiconductors is a vast area in which many books and thousands of papers have been written, and it is impossible to do justice even to the specialized topic of defects. For this reason, we emphasize a few

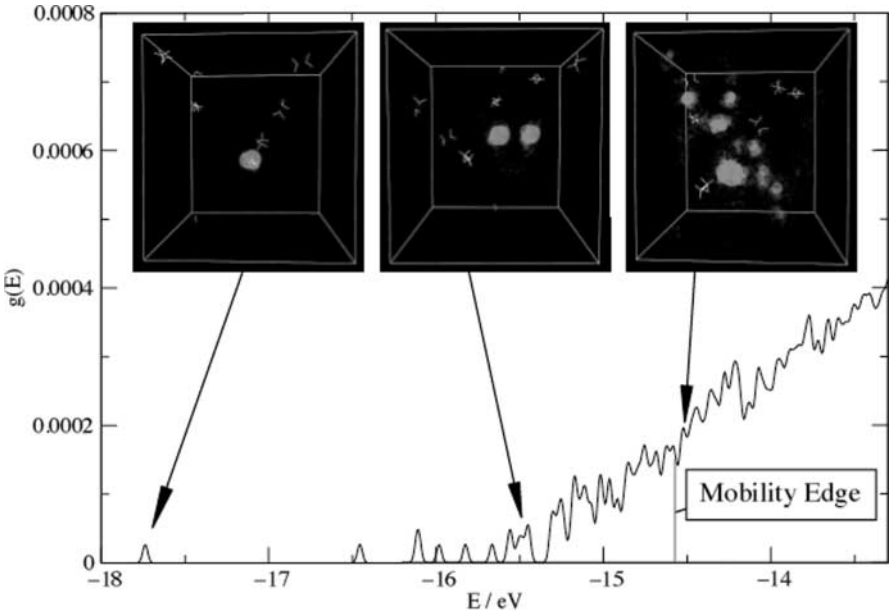

Fig. 1. Electronic density of states near a band edge in a 10 000-atom model of amorphous Si. *Insets* show $|\psi|^2$, indicating parts of the cell in which the defect wavefunctions are localized. The mobility edge separating localized and extended states is indicated. The "band-tail" extends from about −15.5 eV to the mobility edge [17]

questions of particular current interest, and recommend the comprehensive treatments of *Elliott* and *Zallen* [20, 21] for general discussion of amorphous materials including defects.

3 Defects in Amorphous Semiconductors

3.1 Definition of Defect

Static Network

A complete discussion of defects in amorphous materials may be found in Chap. 6 of the book by *Elliott* [20]. Evidently, the term *defect* implies a departure from the typical disorder of an amorphous network. Perhaps it is not surprising that it is not possible to produce a definition of defect without some arbitrary character. Thus, in defining a coordination defect, one introduces a distance that defines whether a pair of atoms is bonded or not. So long as the model structure contains no bonds very close to the critical length, the definition is unambiguous. In amorphous materials there is usually

a well-defined (deep) minimum in the pair-correlation function between the first- and second-neighbor peaks, which gives a reasonable justification for selecting the distance at which the minimum occurs as a cutoff length for definition of coordination. When one accounts for thermal motion of the atoms, the situation is murkier, as we describe below.

If the primary interest is electronic properties of the amorphous semiconductor, one can introduce an electronic criterion to identify defects. As for the geometrical criterion, such a definition is necessarily formulated as a sufficient departure from the mean. An electronic defect may be defined as a structural irregularity that produces a sufficiently localized electronic eigenstate. The ambiguous point here is the modifier "sufficiently": isolated states in the middle of the gap (such as a threefold "dangling-bond" state in a-Si) are easily recognized as defects with a spatially confined wavefunction. Other less-localized states stemming from other network defects may be better characterized as part of the band-tailing. In the same spirit, the chemical bond order (essentially offdiagonal elements of the single-particle density matrix) may be used to identify defects, but a cutoff will be required as always to define the line between bonded and not! A simple quantitative measure of localization is given below.

To contrast the geometrical and electronic definitions of defects, well-localized electronic eigenstates do correspond to network irregularities (either structural, chemical or both) and sufficiently large geometrical irregularities manifest themselves as localized states in a spectral gap in the electronic spectrum (note that this does not necessarily have to be the optical gap (the gap that contains the Fermi level) – other spectral gaps or even the extremal band edges can exhibit these states).

Dynamic Network

The schemes to deine defects outlined above are consistent for sufficient departure from the mean structure. To a surprising degree, however, room-temperature thermal disorder implies the existence of thermally induced geometrical defect fluctuations. Thus, in an ab-initio simulation of a-Si, the number of fivefold "floating" bonds varied between 0 and 10 in a 2 ps room-temperature run for a 216-atom unit cell with coordination defined by a 2.74 Å bond distance [19]. At face value this is astonishing, suggesting that 5 % of the atoms change their coordination in thermal equilibrium at room temperature! If the electronic and geometrical definitions were identical, it would imply that states should be jumping wildly into and out of the gap. Of course this is not the case, though it is true that instantaneous Kohn–Sham energy eigenvalues in the gap do thermally fluctuate [19]. This example illustrates that geometrical defects are not necessarily electronic defects. One can also easily see that electronic defects are not necessarily coordination defects: while an atom may nominally possess ideal (8-N rule) coordination,

substantial bond-angle distortions can also induce localized electronic states. We present new simulations of defect dynamics in Sect. 5.

3.2 Long-Time Dynamics and Defect Equilibria

It is apparent from a variety of experiments in a-Si:H that there are reversible, temperature-dependent changes that occur in defect type and concentration. At its simplest, the idea is that particular defects have a characteristic energy in an amorphous material. It may be that these energies are not substantially removed from the typical conformations of the host. Because the amorphous network does not possess long-range order, it is possible for a small number of atoms to move to convert one kind of defect into another, without a huge cost in energy. These ideas can predict the temperature dependence of some electronic properties, for example quantities like the density and character of band-tail states [22–25]. By analyzing a variety of experiments, *Street* and coworkers [26] have demonstrated that the idea of defect equilibrium is essential. At present the approach is phenomenological to the extent that the microscopic details of the defects involved are not certainly known or even needed, but in principle these ideas could be merged with formation energies and other information from ab-initio simulations to make a first-principles theory of defect equilibria. The potential union between the phenomenological modeling and simulation is also potentially a good example of "bridging timescales" though there is much work still required on the simulation side to realize this goal.

3.3 Electronic Aspects of Amorphous Semiconductors

Localization

A remarkable feature of electron states in crystals is that the wavefunctions have support (are nonzero) throughout the crystal, apart from surface effects. Such states are called extended. The utility of Bloch's theorem is that it reduces the formulation of the electronic structure problem for the infinite periodic system to a simpler form: calculation within a single unit cell on a grid of $\boldsymbol{k}$ crystal momentum points. Thus, for crystals, information about the electronic structure is provided by band-structure diagrams. Such diagrams are not useful in amorphous systems, and furthermore, in amorphous materials, some states may *not* be extended.

To model the electronic consequences of disorder, it is usual to adopt a tight-binding (either empirical or ab initio) description of the electron states, and a Hamiltonian schematically of the form:

$$H = \sum_i \varepsilon_i |i\rangle\langle i| + \sum_{ij, i\neq j} J_{ij} |i\rangle\langle j| \,. \tag{1}$$

To keep the number of indexes to a minimum, we suppose that there is one orbital per site (indexed by i). In this equation, the ε_i specify the atomic energy at each site (in an elemental system, this would be constant for all sites) and J_{ij} is a hopping integral that depends on the coordinates of the atoms located at sites i and j. The topological disorder of a particular realization of an amorphous network (e.g., a structural model) is manifested through the offdiagonal or hopping terms J_{ij}. An enormous effort has been devoted to studying the properties of the eigenvectors of H, and also the critical properties (e.g., of quantum phase transitions [17]).

In realistic studies of the electronic localized–delocalized transition [17] (here "realistic" means large-scale electronic structure calculations in which the disorder in the Hamiltonian matrix is obtained from experimentally plausible structural models, rather than by appealing to a random number generator), it is found that the eigenstates near midgap are spatially compact (localized) and for energies varied from midgap into either the valence or conduction-bands, the states take "island" form: a single eigenstate may consist of localized islands of charge separated by volumes of low charge density. The islands themselves exhibit exponential decay, and the decay lengths for the islands increase modestly from near midgap to the mobility edge. The states become extended with approach to the mobility edge by proliferation (increasing number) of the islands [17, 22]. The qualitative nature of this localized to extended transition is similar for electronic and vibrational disorder of diverse kinds and also conventional Anderson models. Thus, in important ways the localized to extended transition is *universal* [17].

Characterizing Localized States

As we have discussed, the concept of localized states is a central one for amorphous materials and indeed defective systems in general. The question immediately arises: given an electronic eigenvector expressed in some representation, how can we quantify the degree of localization? The most widely adopted gauge of localization is the so-called "inverse participation ratio" (IPR), which is defined as:

$$I(E) = \sum_{n=1}^{N} q(n, E)^2 \,, \tag{2}$$

where N is the number of atoms and $q(n, E)$ is a Mulliken (or other) charge on atom n, and the analysis is undertaken for a particular energy eigenstate with energy E. This measure ranges from N^{-1} (N number of sites) for an ideally extended state to 1 for a state perfectly localized to one site. One can use other measures, such as the information entropy, but at a practical level, we have found that the IPR and information entropy produce qualitatively (not quantitatively) similar results for the localization. IPR is a simple tool for categorizing the zoo of extended, localized and partly localized states of

amorphous semiconductors with defects [27]. As quantum chemists know, all such definitions are basis dependent, and thus care is needed in the interpretation of local charges, IPR and related quantities [28].

Locality of Interatomic Interactions

For atomistic simulation of any materials, a fundamental measure of the spatial nonlocality of interatomic interactions is the decay of the single-particle density matrix, or equivalently, the decay characteristics of the best localized Wannier functions that may be constructed in the material. The density matrix is defined as:

$$\rho(\boldsymbol{r}, \boldsymbol{r}') = \sum_{i \text{occupied}} \psi_i^*(\boldsymbol{r}')\psi_i(\boldsymbol{r})\,, \tag{3}$$

where ψ_i are eigenfunctions of the Hamiltonian in position representation. In a typical condensed-matter system, the ψ_i are oscillating functions, almost all of which will be delocalized through space. Consequently, one can expect "destructive interference" effects as in other wave phenomena when many wiggling functions are superposed as in (3), which can make $\rho(\boldsymbol{r}, \boldsymbol{r}')$ decay rapidly for large $|\boldsymbol{r} - \boldsymbol{r}'|$. If $\hat{H}$ is the Hamiltonian, then the electronic contribution to the total energy is $E = Tr(\hat{\rho}\hat{H})$. If the trace is carried out in the position representation, one can then see that the decay of the density matrix provides information about the locality of the interatomic potential. The details of the chemistry and structure of the material determine this decay length, and the full results even for the asymptotic decay in a greatly simplified two-band model are too complicated to reproduce here [29]. For a material with a finite optical gap as we assume in this Chapter, the decay is ultimately exponential and the decay is faster for larger optical gaps.

Recently, the density matrix has been explicitly computed for a structurally realistic 4096-atom model [10] of a-Si that was fourfold coordinated, but with some large bond-angle distortions. There are two key conclusions from this work: 1. the spatial nonlocality of interatomic interactions is very similar in a-Si and c-Si (because the density matrix decays in a similar fashion for both) and 2. defect centers in a-Si (in this model associated with bond-angle distortions) have Wannier functions with asymptotic decay similar to ordinary (tetrahedral) sites (of course the short-range behavior involving the first few neighbors may be quite different) [22].

3.4 Electron Correlation Energy: Electron–Electron Effects

For a localized single-particle defect state, one must consider the possibility of the state being occupied by zero, one or two (opposite spin) electrons. Consideration of the energetics of these various occupations led in the 1970s and 1980s to insights into the nature of defects, particularly in chalcogenide materials. Typically the electron–electron Coulomb repulsion costs a net energy

(usually labeled "U") to enable two electrons to occupy the same localized spatial state. This is called a "positive U" defect center (the name comes from the symbol used in a Hubbard Hamiltonian to model the electron–electron interaction). In a-Si:H, $U > 0$ for the dangling-bond. It is worth mentioning that the accurate calculation of U is a challenging task for the conventional density-functional methods, as these are essentially mean-field methods and appear to consistently overestimate the delocalization of defect states.

When the possibility of structural relaxation is considered as the second opposite-spin electron is added to the localized state, a negative *effective* U, implying a net energy lowering for double occupation, can emerge. This is observed in chalcogens, and as pointed out by early workers, explains the pinned Fermi level and the diamagnetism of the materials (that is, lack of unpaired spins and so the absence of an ESR signal). It follows that double occupancy implies that defects are charged, and the chemistry (p bonding) of the chalcogens like S or Se leads to the occurrence of "valence alternation pair" defects. Thus, careful simulations reveal that isolated dangling-bonds in a-Se are unstable, and are likely to convert instead into VAPs [30]. *Elliott* discusses these points thoroughly [20].

4 Modeling Amorphous Semiconductors

4.1 Forming Structural Models

For amorphous materials we immediately face a challenging problem: *What is the atomistic structure of the network?* Usually there is a strong tendency for particular local structure (chemical identity of neighbors, and geometrical bonding characteristics), but this preference is only approximately enforced: there is a characteristic distribution of bond lengths and bond angles that is dependent upon both the material and its preparation. Thus, in "better-quality" a-Si samples, almost all of the atoms are four-coordinated, and the bond angle between a reference Si atom and two neighbors is within about 10 degrees of the tetrahedral angle: a strong echo of the chemistry and structure of diamond. In binary glasses (for example $GeSe_2$), Ge is essentially always fourfold coordinated, and Ge only rarely bonds to Ge. Also, as one would expect from crystalline phases of $GeSe_2$ or simple considerations of chemical bonding, the glass is made up predominantly of $GeSe_4$ tetrahedral, again with Se–Ge–Se bond angles close to the tetrahedral angle [31–37].

As reviewed by *Thorpe* [38], the first attempt to understand glasses was based upon the idea that amorphous materials were microcrystalline with a very fine grain size. Eventually, it became clear that this model could not explain the structural experiments that were available. The idea that was ultimately accepted was advanced in 1932 by *Zachariasen* [39] for amorphous SiO_2, the "continuous random network" (CRN) model. Here, the local chemical requirements (four-coordinated Si bonded only to two-coordinated O, and

no homopolar bonds) were enforced, but there was local disorder in bond angles and to a smaller degree, bond lengths. Such networks do not possess long-range structural order. Zachariesen's ideas were taken seriously much later starting in the 1960s, when comparisons to experiment showed that the idea had real promise.

The next step was to use computers to help in making the models. An early computation was performed using a Monte Carlo approach. The idea was to put atoms inside a simulated box and move the atoms at random [40]. At each step the radial distribution function was computed and if a random move pushed the model closer to the experimental data, the move was kept, otherwise moves were retained with Metropolis probability. This is very similar in spirit to a current method, the so-called reverse Monte Carlo (RMC) method [8]. These approaches are information theoretic in spirit: use the available information to produce the model. While this is an eminently reasonable idea, the problem is that merely forcing agreement with diffraction data (pair correlations) grossly underconstrains the model: there are many configurations that agree beautifully with diffraction data but make no sense chemically or otherwise.

In the 1980s, *Wooten*, *Weaire* and *Winer* (WWW) [9], introduced a Monte Carlo scheme for tetrahedral amorphous materials with special bond-switching moves, and energetics described by Keating springs and by requiring the network to be fourfold coordinated. They applied their method to a-Si and a-Ge with remarkable success. Here, the disorder of the real material was somehow captured by what appeared then to be a completely ad-hoc procedure. This method and improved versions of it are still the "gold standard" for creating models of amorphous Si. Years later, *Barkema* and *Mousseau* [11] showed that the likely reason for the success of the WWW scheme is that on long time-scales, the WWW moves occur quite naturally!

Nowadays, the great majority of computer simulations are done using the molecular-dynamics method. The idea is to mimic the actual process of glass formation. To start with, one needs an interatomic potential that describes the interactions in the material. Typically, a well-equilibrated liquid is formed not far above the melting point, then the kinetic energy of the system is gradually reduced (by using some form of dissipative dynamics, such as velocity rescaling at each time step). Eventually there is structural arrest (the computer version of the glass transition), and a structural model results that may be useful for further study. Such calculations are often useful, but it needs to be clearly understood that there is little real similarity to the actual quenching process for glasses, which proceeds far more slowly in nature. As a final oddity, the "melt-quench" approach is often used even for materials that do not form glasses (for example, a-Si). Results are especially poor for this case.

4.2 Interatomic Potentials

To perform a molecular-dynamics simulation, or even a simulation of a mere relaxation of a network near a defect, it is necessary to compute the forces on the atoms in the material. This is never easy, and offers a particular challenge in amorphous materials, as there are usually a variety of bonding environments in the network, and empirical potentials tend to be most accurate/reliable near conformations that were used in a fitting process used to obtain them. The presence of defects exacerbates this further, as coordination or chemical order may be radically different for the defect relative to the rest of the material (and thus harder to describe with a simple interatomic interaction). Also, the complexity of interatomic potentials grows rapidly with the number of distinct atomic species, so that even for binary systems, there are very few reliable empirical potentials available.

The reason why it is difficult to compute accurate interatomic potentials is that the interatomic forces are obtained from the electronic structure of the material. Thus, the details of bonding, electron hybridization, all depend in minute detail upon the local environment (coordination, bond lengths, bond angles, on the environment of the neighboring atoms and so on). The way out is to adopt an approach in which the electronic structure is directly computed in some approximate form. This has been done with some success using empirical tight-binding Hamiltonians [41, 42].

A tremendous advance for defects in semiconductors, but a particular benefit for study of amorphous materials was the advent of practical density-functional codes. The union of density-functional methods and molecular-dynamics is now mature, and one can obtain excellent canned codes that can be used to undertake simulations of complex systems. Of course there is important background knowledge (of amorphous materials, electronic structure and simulation) needed.

4.3 Lore of Approximations in Density-Functional Calculations

Defect calculations typically must be carried out using robust approximations for the Kohn–Sham states. The density-functional basis set upon which the Kohn–Sham orbitals are represented, and spin polarization are the main quantities that need to be considered. Where the basis set is concerned, the prime issue is completeness: the adequacy of the basis functions to approximate the "true" (complete basis limit) Kohn–Sham orbitals. Spin polarization is of particular importance in simulations if there are unpaired spins in the model (as for example a singly occupied dangling-bond state at the Fermi level). The choice of density-functionals is most important for an accurate estimate of energetics: the use of gradient approximations tends to ameliorate the tendency of LDA to overbind.

One of the most important quantities for calculations involving defects is the positioning of defect energy levels, and also accurate estimates of the defect wavefunctions. DFT is in principle the wrong choice for either of these,

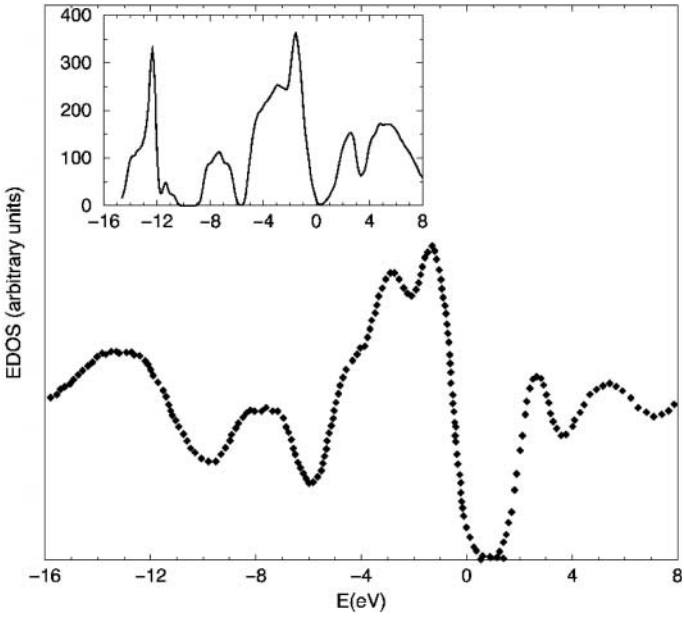

Fig. 2. $GeSe_2$ density of electron states: comparison of experiment and theory (Gaussian-broadened Kohn–Sham eigenvalues) [43]. The Fermi-level is at zero for both curves

since it is formally only a ground-state theory, and also because only the charge density (that is, sum of the squares of the occupied Kohn–Sham orbitals) is formally meaningful within the derivation of the Kohn–Sham equations. However, it is clear that this viewpoint is unduly restrictive, and the density of Kohn–Sham eigenvalues is useful for comparing to the single-particle density of states as measured for example in photoemission (see Fig. 2 taken from [43]). The Kohn–Sham DOS systematically underestimates the optical gap (often by a factor of 2). Curiously, local basis function calculations with a limited (single-zeta) basis tend to give approximately the correct gap as the incompleteness of the basis tends to exaggerate the gap thus partly (and fortuitously) fixing the underestimate intrinsic to the Kohn–Sham calculation. A proper job of describing these states requires methods beyond DFT. In this book, there are two relevant Chapters, that of *Scheffler* on the so-called "GW" methods (these provide self-energy corrections to DFT) and also the Chapter of *Needs* on quantum Monte Carlo. GW calculations of Blase and coworkers have shown that the Kohn–Sham orbitals tend to exaggerate the extent of localized states in crystalline systems; one should expect a similar effect for amorphous systems. Unfortunately, the understanding of these points is only empirical at present.

4.4 The Electron–Lattice Interaction

We have recently shown in some generality that localized electron states associated with defects exhibit a large coupling to the lattice [27]. Thus, for localized electronic eigenstates, deformations of the lattice in the vicinity of the localized state lead to significant changes in the associated electronic eigenvalue. This effect is easily tracked with first-principles MD (in such an approach, Kohn–Sham eigenvectors are computed for each instantaneous ionic configuration). Beside this empirical observation, it is possible to do a simple analysis of the vibration-induced changes in electron energies using the Hellmann–Feynman theorem [44] and exploiting the locality of the states to show that the electron–lattice coupling is roughly proportional to the localization (as gauged by the inverse participation ratio).

Previous thermal simulations with Bohn–Oppenheimer dynamics have indicated that there exists a large electron–phonon coupling for the localized states in the band-tails and in the optical gap [45]. Earlier works on chalcogenide glasses by *Cobb* and *Drabold* [46] have emphasized a strong correlation between the thermal fluctuations as gauged by root mean square (RMS) variation in the LDA eigenvalues and wavefunction localization of a gap or tail state (measured by the inverse participation ratio (2). *Drabold* and *Fedders* [24] have shown that localized eigenvectors may fluctuate dramatically even at room temperature. Recently, *Li* and *Drabold* relaxed the adiabatic (Born–Oppenheimer) approximation to track the time development of electron packets scattered by lattice vibrations [45]. We have recently shown for localized electron states, that the electron–phonon coupling $\Xi_n(\omega)$ (coupling electron n and phonon ω) approximately satisfies:

$$\Xi_n^2(\omega) \sim \mathcal{I}_n \times f(\omega)\,, \tag{4}$$

where $f(\omega)$ does not depend upon n, $\mathcal{I}_n$ is the IPR of electron state n. It is also the case that the thermally induced variance of electronic eigenvalue n, $\langle \delta\lambda_n^2 \rangle \propto \mathcal{I}_n$. It is remarkable that a simple correlation exists between a static property of the network (the IPR) and a dynamical feature of the system, the thermally induced fluctuation in the Kohn–Sham energy eigenvalues. These predictions are easily verified from thermal simulation as reported elsewhere. The main assumption in obtaining this connection is that the electron states under study are localized. Beside the thermally induced changes in electronic energies, there are also significant variations in the structure of the Kohn–Sham eigenstates, another consequence of the large electron–phonon coupling for localized states.

5 Defects in Amorphous Silicon

Geometry/Topology

Probably the most heavily studied amorphous semiconductor is Si. In its unhydrogenated form, the material is not electronically useful, as there are too many defect states in the optical gap. On the other hand, hydrogenated a-Si (a-Si:H) can be grown in a variety of ways, and can be prepared with defect concentrations of order 10^{-16}, which is small enough to enable many electronic applications [47]. Of critical importance for these applications, it is possible to dope a-Si:H $n(p)$ type with P(B) donors (acceptors). The doping efficiency of the material is very low (meaning that large amounts – of order 1 % impurity is needed to move the Fermi level significantly).

In a-Si, diffraction measurements show that these materials are very tetrahedral (typically more than 99.9 % of atoms are four-coordinated) and have bond angles distributed around the tetrahedral angle (typically with a FWHM of order 10 degrees). The key defect is the threefold-coordinated Si atom, the sp^3 dangling-bond, which is known on the basis of experiment and simulation to produce a midgap state. Such states are directly observable in electron spin resonance (ESR) measurements [48, 49].

Several papers have speculated on the importance of fivefold floating bonds [50], but their significance is still uncertain, and it appears that such defects would not produce midgap states, but rather states near the conduction-band edge. Depending upon the details of the local bonding environment, the levels associated with these states could move slightly from their ideal locations. Such configurations are popular in MD simulations of a-Si (quenches from the liquid). It is unclear, however, whether this implies the existence of floating bonds in a-Si:H or if there is an overemphasis on higher-coordinated sites because the liquid is roughly 6-fold coordinated.

Defects play many roles in a-Si:H, and one of the most interesting defect-dependent properties is associated with light-induced metastability, the "Staebler–Wronski" effect [51], which is usually interpreted as light-induced creation of defect centers (probably dangling-bonds). A remarkable collection of experiments and models have been undertaken to understand this effect that is obviously important for photovoltaic applications, but for thin-film device applications (like thin-film transistors) as well [52–65]. This effect is addressed in the Chapter of *Simdayankin* and *Elliott*.

Level of Approximations: A Cautionary Tale

We have undertaken a systematic study of the level of calculation needed to faithfully represent the electron states, total energies and forces in a-Si within the density-functional LDA approximations. This work was performed with the powerful local basis code SIESTA [66]. This calculation was carried out for a-Si:H, but we expect that many of the conclusions should at least be

considered when applying density-functional techniques to other amorphous or glassy materials. In fact studies analogous to ours for silicon should be implemented for more complex amorphous materials. For a unique "one-stop reference" to methodological issues for density-functional methods, see the book of *Martin* [67].

In a nutshell, a proper calculation of defect states and geometry, especially in amorphous materials, is difficult. Every approximation needs to be checked and optimized. Depending upon the type of question being asked more or less sophisticated approximations may be required. We divide the discussion into several categories:

(1) Pseudopotentials: One of the most important and fortunately reliable approximations used in electronic-structure calculations is the *pseudopotential*. This is a means to separate the atomic core and valence regions, and enables the use of only valence electrons in the calculation of the Kohn–Sham orbitals. Even for a relatively light atom like Si, this allows a calculation involving 4 electrons per atom rather than 14. When one reflects that mere diagonalization of the Hamiltonian scales as the cube of the number of electrons, the payoff is clear. After many years of work, the lore of pseudopotentials is fairly mature, though one must test a new potential carefully before using it widely.

(2) Basis Set: The most obvious approximation in any large-scale DF calculation is the use of a finite basis set to represent a set of continuous functions (the Kohn–Sham orbitals and the charge density). In a plane-wave calculation, it is easy to check for completeness, as the only "knob" is the plane-wave cutoff (or number of reciprocal lattice vectors). Care is needed with plane-wave calculations as defect states can be quite spatially compact and therefore difficult to approximate without a large collection of reciprocal lattice vectors.

For local basis codes, ab initio or empirical, completeness is a delicate question. Most current codes use basis orbitals much in the spirit of the linear combination of atomic orbitals (LCAO) method of chemistry, with s, p, d, f states. The minimal basis is defined to be a set of atom-centered functions that is just adequate to represent the occupied atomic orbitals on that atom. The minimal basis has very limited variational freedom. The first improvement on the minimal basis is introducing two functions with the symmetry of the original single-zeta (SZ) functions. Quantum chemists call this a "double-zeta" (DZ) basis. A suitably selected double-zeta basis can reproduce expansion and contraction in local bonding. The zeta proliferation can continue, though it is uncommon to proceed beyond triple-zeta in practical calculations. After adequately filling out the basis orbitals with the symmetries of the states for the ground-state atom, one proceeds to the next shell of states that are unoccupied in the atomic ground-state. These are called polarization functions (so named because the loss of symmetry caused by application of a (polarizing)

electric field distorts the ground-state functions into the next angular momentum shell).

In a-Si:H we find that the choice of basis affects virtually everything. The localization of the Kohn–Sham states depends dramatically on the basis. As one might guess, the less complete basis sets tend to overestimate the localization of the Kohn–Sham eigenvectors (as there are fewer channels for these states to admix into). What was surprising is that the degree of localization, measured by IPR for a well-isolated dangling-bond defect, varies by a factor of *two* between a single-zeta (four orbitals per site) and a double-zeta polarized (DZP) (thirteen orbitals per site) basis. Similar effects are seen for defects in crystals. Since the single-particle density matrix, the total energy and interatomic forces depend upon the Kohn–Sham eigenvectors, it is to be expected that defect geometry, vibrational frequencies and dynamical properties are all influenced by the choice of basis.

(3) Spin Polarization: *Fedders* and coworkers [68] have shown that, in order to correlate the degree of localization from dangling-bond states with ESR experiments, it is not enough to look at the wavefunctions, but to the net spin polarization near the danglin-bond. The reason is that the spin density includes contributions from electronic states other than the localized defect wavefunction, which contribute to make the spin polarization more localized than the specific localized state wavefunction. In order to confirm this result (obtained by Fedders et al. on cells of a-Si:H) in our structural models, we performed calculations allowing for spin polarization in our frozen lattice models, using the DZP basis set. We were not able to find a spin-polarized solution for any of the amorphous cells. The reason is the existence of two interacting dangling-bonds, which favors the formation of a spin singlet with two electrons paired. In order to force the appearance of a spin moment in our models, we introduce an unpaired spin by removing a single electron from the system. We find a contribution of almost 50 % to net spin by the central dangling-bond and its neighbors (the central atom alone contributing 38 %). However, the Mulliken charge contribution to the wavefunction of the corresponding localized state from the defect site is only $0.29e$. The hydrogen-terminated dangling-bond sites also contribute about 10 % of the net spin. The remainder is somewhat distributed uniformly at the other sites. For well-isolated dangling-bonds in a-Si, about 54 % of the net spin-localization sits on the dangling-bond and its nearest neighbors, in reasonable agreement with the experiment [49, 69].

The conclusion is that, for a dangling-bond defect state, there is a large difference between spin localization and wavefunction localization. The degree of spin localization is greater than that of the wavefunction localization at the dangling-bond site. To our knowledge, no experimental methods exist for measuring the extent of wavefunction localization on the dangling-bond orbital as opposed to spin.

(4) Gradient-Corrected Density-Functional: This has been inadequately studied, although there is no reason to expect large changes in structure with the use of gradient corrections. Typically, bond lengths may change modestly and cohesive energies improve when compared to experiment. Generally one expects gradient corrections to at least partly repair the tendency of LDA to overbind.

(5) Brillouin-Zone Sampling: Many current calculations of defects are carried out with a cell that is then periodically repeated to eliminate surface effects. In particular, this construction clearly yields a crystal with a unit cell with typically several hundred or more atoms (such a number is necessary to meaningfully sample the disorder characteristic of the material). Thus, the use of periodic (Born–Von-Karman) boundary conditions really amounts to consideration of a crystal with a large unit cell. Thus, there is a new (and completely artificial) band-structure ($\boldsymbol{k}$ dispersion) associated with the construction, a Brillouin-zone, etc. It is of course true that as the cell gets larger, the bands become flatter, thus reducing the significance of the periodicity. For computational convenience, total energies and forces are inevitably computed at the center ($\boldsymbol{k} = 0$) of the Brillouin-zone, though in principle these quantities involve quadrature over the first Brillouin-zone. However, if results for total energies and especially forces depend upon $\boldsymbol{k}$ in any significant way, then it is doubtful that the cell was selected to be large enough in the first place. For delicate energetics (an all too familiar state of affairs for defects), it is particularly important to test that the cell is big enough. In our experience a few hundred atoms in a cubic cell is adequate.

Defect Identification

From simulation studies, one finds the expected point defects of coordination type (threefold dangling-bonds and fivefold atoms floating bonds), and also strain defects (nominally four-coordinated structures with large deviations in the bond angles). The dangling-bond defect produces electronic states near the middle of the gap, and floating bonds near the conduction edge. Strain defects are associated with the valence- and conduction-band tails.

Defect Dynamics and Diffusion

An important, but underappreciated aspect of defects in amorphous semiconductors, is their *dynamics*. In hydrogenated amorphous silicon, the motion of defects and the motion of hydrogen (which are evidently related) are correlated with some of the most important physical properties of the amorphous matrix, such as the light-induced degradation of the material (Staebler–Wronski effect) [51, 70]. The expectation is that H motion consists of small oscillations in a particular potential well associated with a given local environment with rare escape events until the diffusing particle falls into

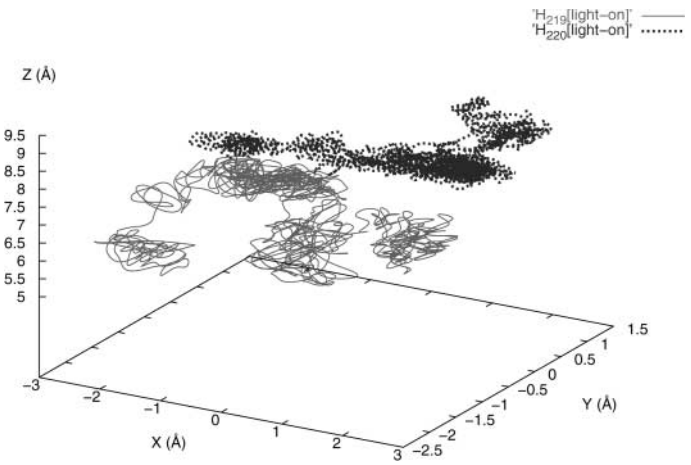

Fig. 3. Trajectory for two different hydrogen atoms, showing diffusion and trapping between silicon bond centers in a 223 atom model of hydrogenated amorphous silicon. The simulation time is 10.0 ps and the temperature is 300 K. 'Light-off' implies that the dynamics are in the electronic ground state

another trap. The time between such rare events depends upon the height of the barrier separating the two metastable configurations. Sophisticated methods exist for determining diffusion pathways and events, particularly the activation-relaxation technique (ART) of *Barkema* and *Mousseau* [71]. In a remarkable paper, these authors applied ART with a simple potential to directly compute the atomic "moves" in a model of a-Si. For large barriers and very rare events, there is no substitute for a study of "ART" type.

For smaller barriers, we have seen that it is possible to extract interesting short-time diffusive dynamics directly from MD simulation. In simulations of Ag dynamics in chalcogenide glass hosts, we found that it was not difficult to track the motion of the Ag atoms from simulations of order 50 ps. The existence of trapping centers, and even some information about trap geometry, and temperature-dependent residence times was obtained [72]. In a-Si:H, we have found that an analogous computation produces new insight into the motion of both Si and H atoms. Using a small cell (61 Si and 10 H atoms) with two dangling-bonds and no other defects, we employed SIESTA with high-level approximations (a double-zeta polarized basis) to monitor atomic motion. On the time-scale of 1 ps we have shown the trajectory of one of the H atoms in Fig. 3 [73]. Representative of the majority of H atoms in the cell, this particular trajectory shows the diffusion of the hydrogen atom, including trapping. While the H atoms are diffusing in the cell it is followed by breaking old bonds and forming new bonds. We have plotted the time evolution of undercoordinated and floating bonds formed as a consequence of hydrogen diffusion in Fig. 4 [73].

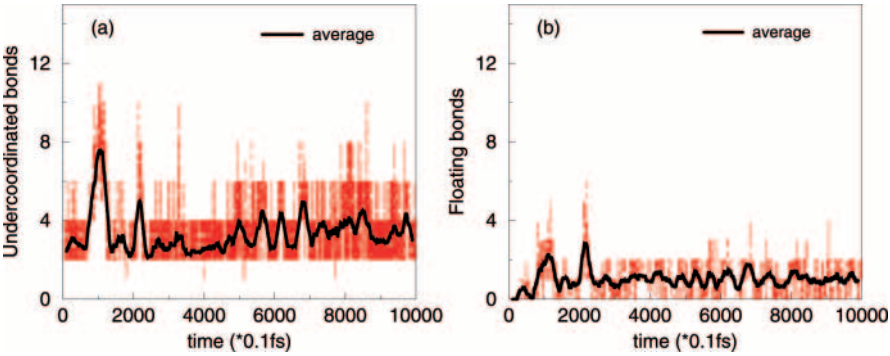

Fig. 4. (**a**) Total number of undercoordinated atoms and (**b**) Total number of overcoordinated atoms in our simulation at $T = 300\,\mathrm{K}$. The total time for the graph is 1.0 ps

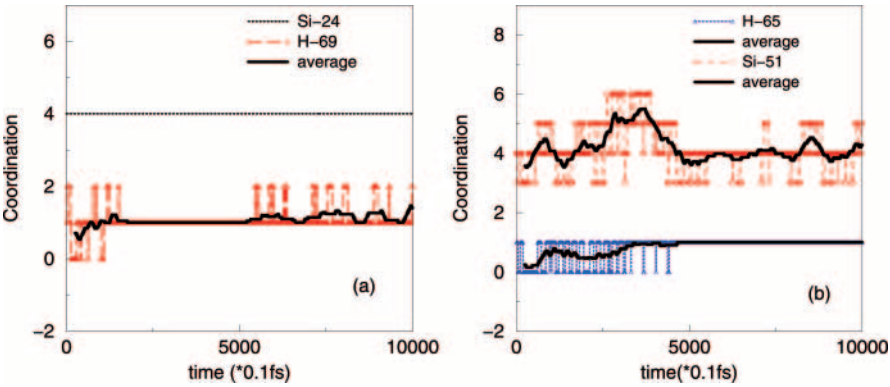

Fig. 5. Time evolution of coordination for a few selected atoms (H-69, H-65, Si-24, and Si-51) at room temperature $T = 300\,\mathrm{K}$. (**a**) time evolution of atoms, Si-24 which is far from the diffusing H with coordination 4 and H-69. (**b**) time evolution of selected atoms: Si-51, which is close to the diffusing H, and H-65

In our simulation we have observed rearrangements of the atoms while the hydrogen atoms are diffusing in the cell. The diffusion of H causes formation and breaking of bonds. We have also observed the formation of metastable states that trap the mobile hydrogen atoms. The mechanisms for the formation of these structures follow breaking of H atoms from the Si–H bonds and followed by diffusion in the cell. These mobile H atoms then collide with another Si+DB structure and form a bond. These processes continue until two hydrogens form a bond to a single Si atom or to two nearby Si atoms to form a metastable conformation. In Fig. 5, [73] we have shown the time evolution of the coordination for selected atoms. Figure 5a shows the stable coordination of a Si atom (Si-24) that is fully coordinated since there is no

diffusing atom in its vicinity and of H-69 that has an average coordination of 1. However, the Si atom (Si-51) and H-65, which are shown in Fig. 5b, shows change in the coordination as a function of time. This change in the coordination becomes more stable (1 for H and 4 for Si) after the formation of a metastable configuration. Our results suggest that atoms that are reasonably far from the diffusing hydrogen have an average coordination of 4 for Si and 1 for H. However, the coordination of the atoms, for instance Si-51, in the direction of the diffusing H atoms changes with time. This suggests that while the H atoms are diffusing in the cell there is breaking and formation of bonds.

Other researchers have emphasized the significance of defect and H motion. In particular, *Street* [74] has pioneered a phenomenological approach to the long-time dynamics with the defect-pool model. Our work on extremely short timescales should ultimately be smoothly connected to the inferred long-time defect dynamics from the defect-equilibria models. While we are very far from this goal at present, it is possible to imagine impacts of both approaches on each other within a few years. Ideally, the short-time simulations could provide ab-initio input into the energetics of the dynamics, and even information about residence times for various defects.

Acknowledgements

It is a pleasure to thank a number of collaborators for their contributions to the work reported here: R. Atta-Fynn, P. Biswas, J. Dong, S. R. Elliott, P. A. Fedders, H. Jain, J. Li, J. Ludlam, N. Mousseau, P. Ordejón, S. N. Taraskin and X. Zhang. We acknowledge the support of the National Science Foundation under NSF-DMR 0600073, 0605890 and the Army Research Office under MURI W911NF-06-2-0026.

References

[1] C. R. A. Catlow (Ed.): *Defects and Disorder in Crystalline and Amorphous Solids*, vol. 418 (Kluwer, Netherlands 1994) 245
[2] J. D. Joannopoulos, G. Lucovsky (Eds.): *The Physics of Hydrogenated Amorphous Silicon II*, vol. 56 (Springer, Berlin, Heidelberg 1984) 245
[3] M. Stutzmann, D. K. Biegelsen, R. A. Street: Phy. Rev. B **35**, 5666 (1987) 245
[4] E. T. Jaynes: *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge 2003) 246
[5] R. Car, M. Parrinello: Phy. Rev. Lett. **60**, 204 (1988) 247
[6] O. Gereben, L. Pusztai: Phys. Rev. B **50**, 14136 (1994) 247
[7] J. K. Walters, R. J. Newport: Phys. Rev. B **53**, 2405 (1996) 247
[8] R. L. McGreevy: J. Phys.: Condens. Matter **13**, R877 (2001) 247, 254
[9] F. Wooten, K. Winer, D. Weaire: Phy. Rev. Lett. **54**, 1392 (1985) 247, 254

[10] B. R. Djordjevic, M. F. Thorpe, F. Wooten: Phys. Rev. B **52**, 5685 (1995) 247, 252
[11] G. T. Barkema, N. Mousseau: Phy. Rev. B **62**, 4985 (2000) 247, 254
[12] R. L. C. Vink, G. T. Barkema: Phys. Rev. B **67**, 245201 (2003) 247
[13] M. F. Thorpe: J. Non-Cryst. Solids **57**, 355 (1983) 247
[14] A. E. Carlsson: *Solid State Physics, Advances in Research and Applications*, vol. 43 (Academic, New York 1990) p. 1 247
[15] F. Ercolessi, J. B. Adams: Europhys. Lett. **26**, 583 (1994) 247
[16] A. Nakano, L. Bi, R. K. Kalia, P. Vashishta: Phys. Rev. B **49**, 9441 (1994) 247
[17] J. J. Ludlam, S. N. Taraskin, S. R. Elliott, D. A. Drabold: J. Phys.: Condens. Matter **17**, L321 (2005) 247, 248, 251
[18] S. Aljishi, L. Ley, J. D. Cohen: Phys. Rev. Lett. **64**, 2811 (1990) 247
[19] D. A. Drabold, P. A. Fedders, S. Klemm, O. F. Sankey: Phys. Rev. Lett. **67**, 2179 (1991) 247, 249
[20] S. R. Elliott: *Physics of Amorphous Materials*, 2nd ed. (Longman Scientific, London; New York 1990) 248, 253
[21] R. Zallen: *The Physics of Amorphous Solids* (Wiley-Interscience, New York 1998) 248
[22] J. Dong, D. A. Drabold: Phys. Rev. Lett. **80**, 1928 (1998) 250, 251, 252
[23] P. A. Fedders, D. A. Drabold, S. Nakhmanson: Phys. Rev. B **58**, 15624 (1998) 250
[24] D. A. Drabold, P. A. Fedders: Phys. Rev. B **60**, R721 (1999) 250, 257
[25] D. A. Drabold: J. Non-Cryst. Solids **269**, 211 (2000) 250
[26] R. A. Street, J. Kakalios, T. M. Hayes: Phys. Rev. B **34**, 3030 (1986) 250
[27] R. Atta-Fynn, P. Biswas, P. Ordejón, D. A. Drabold: Phys. Rev. B **69**, 085207 (2004) 252, 257
[28] N. Szabo, N. S. Ostlund: *Modern Quantum Chemistry* (Dover, New York 1996) 252
[29] S. N. Taraskin, D. A. Drabold, S. R. Elliott: Phys. Rev. Lett. **88**, 196405 (2002) 252
[30] X. Zhang, D. A. Drabold: Phys. Rev. Lett. **83**, 5042 (1999) 253
[31] M. Cobb, D. A. Drabold, R. L. Cappelletti: Phys. Rev. B **54**, 12162 (1996) 253
[32] C. Massobrio, A. Pasquarello, R. Car: Phys. Rev. Lett. **80**, 2342 (1998) 253
[33] I. Petri, P. S. Salmon, H. E. Fischer: Phys. Rev. Lett. **84**, 2413 (2000) 253
[34] X. Zhang, D. A. Drabold: Phys. Rev. B **62**, 15 695 (2000) 253
[35] P. S. Salmon, I. Petri, J. Phys.: Condens. Matter **15**, S1509 (2003) 253
[36] D. A. Drabold, J. Li, D. Tafen: J. Phys.: Condens. Matter **15**, S1529 (2003) 253
[37] D. N. Tafen, D. A. Drabold: Phys. Rev. B **68**, 165208 (2003) 253
[38] M. F. Thorpe, D. J. Jacobs, B. R. Djordjevic: The structure and rigidity of network glasses, in P. Boolchand (Ed.): *Insulating and Semiconducting Glasses* (World Scientific, Singapore 2000) p. 95 253
[39] W. H. Zachariasen: J. Am. Chem. Soc. **54**, 3841 (1932) 253
[40] R. Kaplow, T. A. Rowe, B. L. Averbach: Phys. Rev. **168**, 1068 (1968) 254
[41] A. P. Sutton, M. W. Finnis, D. G. Pettifor, Y. Ohta: J. Phys. C: Solid State Phys. **21**, 35 (1988) 255
[42] N. Bernstein, E. Kaxiras: Phys. Rev. B **56**, 10488 (1997) 255

[43] P. Biswas, D. N. Tafen, D. A. Drabold: Phys. Rev. B **71**, 054204 (2005) 256
[44] R. P. Feynman: Phys. Rev. **56**, 340 (1939) 257
[45] J. Li, D. A. Drabold: Phys. Rev. B **68**, 033103 (2003) 257
[46] M. Cobb, D. A. Drabold: Phys. Rev. B **56**, 3054 (1997) 257
[47] J. D. Joannopoulos, G. Lucovsky (Eds.): *The Physics of Hydrogenated Amorphous Silicon I*, vol. 55 (Springer, Berlin, Heidelberg 1984) 258
[48] R. A. Street, D. K. Biegelsen, J. C. Knights: Phys. Rev. B **24**, 969 (1981) 258
[49] D. K. Biegelsen, M. Stutzmann: Phys. Rev. B **33**, 3006 (1986) 258, 260
[50] S. T. Pantelides: Phys. Rev. Lett. **57**, 2979 (1986) 258
[51] D. L. Staebler, C. R. Wronski: Appl. Phys. Lett. **31**, 292 (1977) 258, 261
[52] H. Fritzsche: Annu. Rev. Mater. Res. **31**, 47 (2001) 258
[53] T. Su, P. C. Taylor, G. Ganguly, D. E. Carlson: Phys. Rev. Lett. **89**, 015502 (2002) 258
[54] S. B. Zhang, W. B. Jackson, D. J. Chadi: Phys. Rev. Lett. **65**, 2575 (1990) 258
[55] D. J. Chadi: Appl. Phys. Lett. **83**, 3710 (2003) 258
[56] M. Stutzmann, W. B. Jackson, C. C. Tsai: Phys. Rev. B **32**, 23 (1985) 258
[57] R. Biswas, I. Kwon, C. M. Soukoulis: Phys. Rev. B **44**, 3403 (1991) 258
[58] S. Zafar, E. A. Schiff: Phys. Rev. B **40**, 5235 (1989) 258
[59] S. Zafar, E. A. Schiff: Phys. Rev. Lett. **66**, 1493 (1991) 258
[60] S. Zafar, E. A. Schiff: J. Non-Cryst. Solids **138**, 323 (1991) 258
[61] H. M. Branz: Phys. Rev. B **59**, 5498 (1999) 258
[62] R. Biswas, Y.-P. Li: Phys. Rev. Lett. **82**, 2512 (1999) 258
[63] N. Kopidakis, E. A. Schiff: J. Non-Cryst. Solids **269**, 415 (2000) 258
[64] S. B. Zhang, H. M. Branz: Phys. Rev. Lett. **87**, 105503 (2001) 258
[65] T. A. Abtew, D. A. Drabold, P. C. Taylor: Appl. Phys. Lett. **86**, 241916 (2005) 258
[66] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejón, D. Sanchéz-Portal: J. Phys.: Condens. Matter **14**, 2745 (2002) 258
[67] R. M. Martin: *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge 2004) 259
[68] P. A. Fedders, D. A. Drabold, P. Ordejón, G. Fabricius, E. Artacho, D. Sanchez-Portal, J. Soler: Phys. Rev. B **60**, 10594 (1999) 260
[69] T. Umeda, S. Yamasaki, J. Isoya, et al.: Phys. Rev. B **59**, 4849 (1999) 260
[70] T. A. Abtew, D. A. Drabold: J. Phys.: Condens. Matter **18**, L1 (2006) 261
[71] N. Mousseau, G. T. Barkema: Phys. Rev. E **57**, 2419 (1998) 262
[72] D. N. Tafen, D. A. Drabold, M. Mitkova: phys. stat. sol. (b) **242**, R55 (2005) 262
[73] T. A. Abtew, D. A. Drabold: Phys. Rev. B **74**, 085201 (2006) 262, 263
[74] R. A. Street: *Hydrogenated Amorphous Silicon* (Cambridge University Press, Cambridge 1991) 264

Index

Light Induced Effects in Amorphous and Glassy Solids

S. I. Simdyankin and S. R. Elliott

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
{sis24,sre1}@cam.ac.uk

Abstract. In this Chapter, we discuss how exposure to light can affect the properties of disordered materials and review our recent computational studies of these phenomena. Familiarity with the preceding contribution by *Drabold* and *Abtew* is beneficial for understanding this Chapter.

1 Photoinduced Metastability in Amorphous Solids: An Experimental Survey

1.1 Introduction

Electromagnetic (EM) radiation can interact with amorphous solids in a number of ways, depending on the wavelength of the radiation. EM radiation can be either absorbed or scattered by a solid. In the resonant process of absorption, the photon energy must match a transition energy between two quantum states in the material, e.g., for atomic vibrations corresponding to infrared (IR) wavelengths, or electronic transitions from valence to conduction band across a bandgap in a semiconductor, corresponding to visible/near ultraviolet (UV) wavelengths (depending on the magnitude of the bandgap energy, E_g). Scattering of EM radiation may be elastic, or inelastic when energy is exchanged between a photon and quantum states of a solid. Examples include X-ray scattering (diffraction) from the electrons in atoms, Raman (inelastic light) scattering from atomic vibrations (e.g. molecular-like modes) and Brillouin (inelastic light) scattering from acoustic phonon-like vibrational excitations. The photon–solid interaction can be interrogated in terms of either the ultimate state of the photon (as in the above examples), or in terms of changes in the internal state of the solid. In the latter case, it is often changes in the electronic degrees of freedom that are probed: for the case of most amorphous semiconductors with bandgaps in the range 1 eV $< E_g <$ 3 eV, this involves near-IR or visible-light excitation (although for more ionic oxides, e.g., vitreous (v-)SiO_2 with $E_g \approx 10$ eV, UV-light excitation is required). Electrons optically excited into the conduction band (CB), or holes excited into the valence band (VB), can be probed, for example, by their enhanced electrical transport (photoconductivity) or by their radiative recombination

D. A. Drabold, S. K. Estreicher (Eds.): Theory of Defects in Semiconductors,
Topics Appl. Physics **104**, 269–286 (2007)

(photoluminescence or fluorescence). At sufficiently high light intensities, the light–solid interaction can become nonlinear, this optical nonlinearity permitting many new phenomena, such as second-harmonic generation, three- and four-wave mixing, to occur. Electronic excitation of structural coordination defects, e.g., dangling bonds, can be probed by electron spin (paramagnetic) resonance (ES(P)R).

The optically induced phenomena outlined above can be exhibited by all kinds of semiconducting/insulating solid, whether crystalline or amorphous. However, certain features unique to the amorphous state in general, and to certain types of amorphous materials in particular, mean that some photoinduced phenomena are special to amorphous semiconducting materials, particularly those that are *metastable*, i.e., which remain after cessation of irradiation.

One general characteristic of amorphous semiconductors of relevance in this connection is the occurrence of disorder-induced spatial localization of electronic states in the band-tail states extending into the gap between the VB and CB [1]. The presence of continuous bands of localized states in these tail states, in the energy interval between the "mobility" edges in the VB and CB, marking the transition point between localized and delocalized (extended) electron states [2], as well as localized states deep in the bandgap arising from coordination defects [1], can have a profound influence on the nature of the photon–solid interaction. These localized states have an enhanced electron–lattice interaction (see the preceding Chapter by *Drabold* and *Abtew*), meaning that optical excitation of such electronic states can have a disproportionate effect on the surrounding atomic structure. Secondly, the radiative-recombination lifetime for an optically created electron–hole pair trapped in localized tail states can be very considerably longer than when the photoexcited carriers are in extended states (as is always the case for crystals), thereby allowing possible nonradiative channels to become significant (e.g., involving atomic-structural reconfiguration).

Other relevant aspects are more materials specific. Optically induced metastability associated with structural reorganization is likely to be more prevalent in those materials in which (some) atoms have a low degree of nearest-neighbor connectivity (e.g., being twofold coordinated, rather than fourfold coordinated), thereby imparting a considerable degree of local structural flexibility. In addition, structural reorganization following optically induced electronic excitation is more probable if the (VB) electronic states involved correspond to easily broken weak bonds.

One class of materials satisfying the above constraints consists of chalcogenide glasses, namely alloys of the Gp-VI chalcogen elements (S, Se, Te) with other (metalloid) elements, e.g., B, Ga, P, As, Sb, Si, Ge, etc. These materials are "lone-pair" semiconductors in which (if the chalcogen content is sufficiently high) the top of the VB comprises chalcogen p-π lone-pair (LP) states [3]. Interatomic, "nonbonding" (Van der Waals-like) interactions involving such LP states are appreciably weaker than for normal covalent

bonds (those states lying deeper in the VB). Since the ground-state electronic configuration of chalcogen atoms is s^2p^4, the occurrence of nonbonding LP electron states means that each chalcogen atom is ideally twofold coordinated by covalent bonds to its nearest neighbors. Thus, chalcogenide glasses, having a combination of localized electronic tail states, low atomic coordination and weak bonds associated with optically accessible states at the top of the VB, are ideal candidates for exhibiting photoinduced effects.

1.2 Photoinduced Effects in Chalcogenide Glasses

Amorphous chalcogenide materials exhibit a plethora of photoinduced phenomena. Such changes can be variously *dynamic* (i.e., present only whilst a material is illuminated) or *metastable* (i.e., the effects remain after cessation of illumination). Furthermore, the changes may be either scalar or vectoral in nature (respectively, independent or dependent on either the polarization state, or the propagation direction, of the inducing light). Finally, metastable changes may be irreversible, or reversible with respect to thermal or optical annealing.

Examples of dynamic effects are the afore-mentioned phenomena of photoluminescence and photoconductivity [1] and optical nonlinearity [4], which are common to all materials. (Chalcogenide glasses generally exhibit extremely large optical nonlinearities because of the highly electronically polarizable nature of chalcogen atoms present [4].) However, a (scalar) dynamic effect, which is characteristic of chalcogenide glasses, is photoinduced fluidity, wherein the viscosity of a glass (e.g. As_2S_3) decreases on illumination with subbandgap light or, in other words, light can cause viscous relaxation in a stressed glass [5, 6]. An example of a vectoral dynamic photoeffect in chalcogenide glasses is the optomechanical effect [7, 8], wherein linearly polarized light incident on a chalcogenide-coated clamped microcantilever causes it to displace upwards or downwards, depending on whether the polarization axis is, respectively, parallel or perpendicular to the cantilever axis, as a result of anisotropic photoinduced strains introduced into the chalcogenide–cantilever bimorph.

Metastable photoinduced changes are perhaps the most interesting (and applicable) of the effects observed in chalcogenide glasses. One such is photodarkening (or bleaching), wherein the optical absorption edge of the material shifts to lower energies (hence the material gets darker at a given wavelength), or to higher energies (bleaching), on illumination with (sub-) bandgap light. For reviews, see [9, 10]. Arsenic-based materials with higher sulfur contents and germanium sulfide materials seem to favor photobleaching, for reasons that are not clear at present. Irreversible shifts of the optical absorption edge occur in virgin (as-evaporated) thin films of amorphous chalcogenides containing structurally unstable molecular fragments from the vapor phase [11] that are particularly vulnerable to photoinduced (or thermal) change. Reversible photodarkening is observed in bulk glasses and well-annealed thin

films: optical illumination causes a redshift of the absorption edge, whilst subsequent thermal annealing to the glass-transition temperature, T_g, erases the effect [9, 10].

Photodarkening (bleaching) appears to be associated with photoinduced structural changes, such microscopic changes being manifested as macroscopic volume changes. Photocontraction is particularly prevalent in virgin obliquely deposited thin films of amorphous chalcogenides having a columnar-like microstructure [12], but photoexpansion is commonly associated with photodarkening in chalcogenide bulk glasses or thin films [9, 10]. Giant photoexpansion (up to 5 % expansion) occurs for subbandgap illumination [13]. Another metastable photoinduced structural change exhibited by chalcogenide materials is photoinduced crystallization and amorphization, as used in rewriteable "phase-change" CDs and DVDs [14, 15]. The amorphization process there is believed to result from a photoinduced melting and subsequent very rapid quenching of the material to the glassy state. However, illumination of certain crystalline chalcogenide materials (e.g., $As_{50}Se_{50}$) can also cause *athermal* photoamorphization [16]. Light can also cause "chemical" changes in chalcogenide glasses. For example, overlayers of certain metals (notably silver) diffuse into the undoped bulk glass on illumination [17, 18]. On the other hand, Ag-containing chalcogenide glasses, rich in Ag, exhibit the opposite effect, namely photoinduced surface deposition, wherein the metal exsolves from the glassy matrix on illumination [19, 20].

A particularly interesting metastable *vectoral* photoinduced effect exhibited by chalcogenide glasses is photoinduced optical anisotropy (POA), first observed in [21], and manifested in absorption, reflection (i.e., refractive index) and scattering (for a review, see [10, 22]). Illumination of an initially optically isotropic glass by linearly polarized light causes the material to become dichroic and birefringent. In the case of well-annealed thin films and bulk glasses, the effect is completely reversible optically (i.e., the induced optical axis is fully rotated when the light-polarization vector is rotated), and the POA can be annealed out thermally (but at a lower temperature than is the case for scalar photodarkening) or optically (using unpolarized or circularly polarized light).

2 Theoretical Studies of Photoinduced Excitations in Amorphous Materials

From the above very brief review, it is apparent that amorphous chalcogenide materials exhibit a wide range of interesting photoinduced phenomena. Although the experimental phenomenology is, for the most part, well developed, a proper theoretical understanding is still lacking. Until very recently, theoretical models have been confined to "hand-waving" models [9,10], involving simple notions of chemical bonding and the effects of light (e.g., bond breaking and defect formation). However, the predictive, and even descriptive, power

of such approaches is very limited, and for a microscopic (atomic-level) understanding of photoinduced phenomena in amorphous chalcogenides, proper quantum-mechanical calculations need to be performed. Obviously, this is an extremely challenging task, and two essentially opposing methods for making progress in this regard, using different approximations, can be employed. The first employs all-electron quantum-chemical calculations on small clusters (a few tens) of atoms, in which the optimized electronic and atomic configurations and energies are found for the ground and electronically excited states (see, e.g. [23–26]). This favors accuracy over dynamics. The other method, (discussed in the Chapter by *Drabold* and *Abtew*), namely ab-initio molecular-dynamics (MD) simulations, takes a converse approach: atomic dynamics are followed at the expense of accuracy. More detail will be given in the following, but the advantages and disadvantages of these two approaches can be summarized as follows. Quantum-chemical methods can only deal with very small clusters (particularly for excited-state calculations) and only initial (ground-state) and final (excited-state) configurations can be studied, not the intermediate dynamics, but the energetics are the most accurate. Ab-initio MD, on the other hand, provides full information about atomic dynamics, but only over very short timescales (typically a few picoseconds) and for relatively few atoms (typically less than a hundred). In order to make the calculations involved tractable, numerous more-or-less severe approximations need to be invoked (e.g., the local-density approximation (LDA) in density-functional theory (DFT)), which make certain results imprecise (e.g., underestimation of the bandgap). Moreover, the Kohn–Sham (KS) orbitals resulting from DFT are ground-state quantities and hence formally inadmissible for a consideration of excited-state behavior. Nevertheless, use of these theoretical methods has provided very useful atomistic information of help in understanding photoinduced phenomena in chalcogenide glasses. Some justification for using both occupied and virtual KS orbitals for qualitative, and sometimes partially quantitative, analysis of electronic structure is provided by comparing the shape and symmetry properties, as well as the energy order, of these orbitals with those obtained by wavefunction-based (e.g., at the Hartree–Fock level of theory [27]) and GW [28] calculations. There is a very strong similarity between GW and LDA states, and often identical DFT and GW quasiparticle wavefunctions are assumed in practical calculations [29].

2.1 Application of the Density-Functional-Based Tight-Binding Method to the Case of Amorphous As_2S_3

One approximate ab-initio-based MD scheme that has proved very useful in understanding the electronic behavior of chalcogenide glasses is the density-functional-based tight-binding (DFTB) method developed by Frauenheim and coworkers. Although this is reported in [30–32], since this approach is not described elsewhere in this book, we give here a brief description of this

method and review our recent results on the canonical chalcogenide glass a-As_2S_3 obtained by using the DFTB method [33–35].

Although the DFTB method is semiempirical, it allows one to improve upon the standard tight-binding description of interatomic interactions by including a DFT-based self-consistent second order in charge fluctuation (SCC) correction to the total energy [31]. The flexibility in choosing the desired accuracy while computing the interatomic forces brings about the possibility to perform much faster calculations when high precision is not required, and refine the result if needed.

As described in [36], the SCC-DFTB model is derived from density-functional theory (DFT) by a second-order expansion of the DFT total energy functional with respect to the charge-density fluctuations $\delta n' = \delta n(\mathbf{r}')$ around a given reference density $n_0' = n_0(\boldsymbol{r}')$:

$$
\begin{aligned}
E = & \sum_{i}^{\mathrm{occ}} \langle \psi_i | \hat{H}^0 | \psi_i \rangle \\
& + \frac{1}{2} \iint' \left(\frac{1}{|\boldsymbol{r} - \boldsymbol{r}'|} + \left. \frac{\delta^2 E_{\mathrm{xc}}}{\delta n \, \delta n'} \right|_{n_0} \right) \delta n \, \delta n' \\
& - \frac{1}{2} \iint' \frac{n_0' n_0}{|\boldsymbol{r} - \boldsymbol{r}'|} + E_{\mathrm{xc}}[n_0] - \int V_{\mathrm{xc}}[n_0] n_0 + E_{ii} \, ,
\end{aligned}
\tag{1}
$$

where $\int \mathrm{d}\boldsymbol{r}$ and $\int \mathrm{d}\boldsymbol{r}'$ are expressed by $\int$ and $\int'$, respectively. Here, $\hat{H}^0 = \hat{H}[n_0]$ is the effective Kohn–Sham Hamiltonian evaluated at the reference density and the ψ_i are Kohn–Sham orbitals. E_{xc} and V_{xc} are the exchange-correlation energy and potential, respectively, and E_{ii} is the core–core repulsion energy.

To derive the total energy of the SCC-DFTB method, the energy contributions in (1) are further subjected to the following approximations: 1. The Hamiltonian matrix elements $\langle \psi_i | \hat{H}^0 | \psi_i \rangle$ are represented in a basis of confined, pseudoatomic orbitals ϕ_μ,

$$
\psi_i = \sum_{\mu} c_\mu^i \phi_\mu . \tag{2}
$$

To determine the basis functions ϕ_μ, the atomic DFT problem is solved by adding an additional harmonic potential $(\frac{r}{r_0})^2$ to confine the basis functions [30]. The Hamiltonian matrix elements in this LCAO basis, $H^0_{\mu\nu}$, are then calculated as follows. The diagonal elements $H^0_{\mu\mu}$ are taken to be the atomic eigenvalues and the nondiagonal elements $H^0_{\mu\nu}$ are calculated in a two-center approximation:

$$
H^0_{\mu\nu} = \langle \phi_\mu | \hat{T} + v_{\mathrm{eff}}[n^0_\alpha + n^0_\beta] | \phi_\nu \rangle \quad \mu\epsilon\alpha, \nu\epsilon\beta, \tag{3}
$$

which are tabulated, together with the overlap matrix elements $S_{\mu\nu}$ with respect to the interatomic distance $R_{\alpha\beta}$. v_{eff} is the effective Kohn–Sham potential and n^0_α are the densities of the neutral atoms α.

2. The charge-density fluctuations δn are written as a superposition of atomic contributions δn_α,

$$\delta n = \sum_\alpha \delta n_\alpha, \tag{4}$$

which are approximated by the charge fluctuations at the atoms α, $\Delta q_\alpha = q_\alpha - q_\alpha^0$. q_α^0 is the number of electrons of the neutral atom α and the q_α are determined from a Mulliken-charge analysis. The second derivative of the total energy in (1) is approximated by a function $\gamma_{\alpha\beta}$, whose functional form for $\alpha \neq \beta$ is determined analytically from the Coulomb interaction of two spherical charge distributions, located at R_α and R_β. For $\alpha = \beta$, it represents the electron–electron self-interaction on atom α.
3. The remaining terms in (1), E_{ii} and the energy contributions, which depend on n_0 only, are collected in a single energy contribution E_{rep}. E_{rep} is then approximated as a sum of short-range repulsive potentials,

$$E_{\text{rep}} = \sum_{\alpha\neq\beta} U[R_{\alpha\beta}], \tag{5}$$

which depend on the interatomic distances $R_{\alpha\beta}$.

With these definitions and approximations, the SCC-DFTB total energy finally reads:

$$E_{\text{tot}} = \sum_{i\mu\nu} c_\mu^i c_\nu^i H_{\mu\nu}^0 + \frac{1}{2}\sum_{\alpha\beta}\gamma_{\alpha\beta}\Delta q_\alpha \Delta q_\beta + E_{\text{rep}}. \tag{6}$$

Applying the variational principle to the energy functional (6), one obtains the corresponding Kohn–Sham equations:

$$\sum_\nu c_{\nu i}(H_{\mu\nu} - \epsilon_i S_{\mu\nu}) = 0, \quad \forall\ \mu,\ i \tag{7}$$

$$H_{\mu\nu} = \langle\phi_\mu|H_0|\phi_\nu\rangle + \frac{1}{2}S_{\mu\nu}\sum_\zeta(\gamma_{\alpha\zeta} + \gamma_{\beta\zeta})\Delta q_\zeta, \tag{8}$$

which have to be solved iteratively for the wavefunction expansion coefficients c_μ^i, since the Hamiltonian matrix elements depend on the c_μ^i due to the Mulliken charges. Analytic first derivatives for the calculation of interatomic forces are readily obtained, and second derivatives of the energy with respect to atomic positions are calculated numerically.

The repulsive pair potentials $U[R_{\alpha\beta}]$ are constructed by subtracting the DFT total energy from the SCC-DFTB electronic energy (first two terms on the right-hand side of (6)) with respect to the bond distance $R_{\alpha\beta}$ for a small set of suitable reference systems.

To summarize, in order to determine the appropriate parameters for a new element, the following steps have to be taken. First, DFT calculations have to be performed for the neutral atom to determine the LCAO basis functions ϕ_μ

and the reference densities n_α^0. Here the confinement radius can, in principle, be chosen to be different for the density (r_0^n) and each type of atomic orbital ($r_0^{s,p,d}$). The value of r_0 is usually taken to be the same for s- and p-functions. In a minimal basis, this yields a total number of two adjustable parameters for elements in the first and second rows, while there are three if d-functions are included. After this, the different matrix elements can be calculated and the pair potentials $U[R_{\alpha\beta}]$ are obtained as stated above for every combination of the new element with the ones already parameterized.

The single-particle KS occupied, ψ_i, and unoccupied, ψ_j, orbitals and the corresponding KS excitation energies $\omega_{ij} = \epsilon_j - \epsilon_i$ are sometimes used for qualitative analysis of electronic excitations. Although the KS excitation energies can be considered as an approximation to true excitation energies ω_I [37] (see also end of the introduction to Sect. 2), the correction terms to this approximation can be very significant in many cases. Modeling electronic excitations by varying the occupation of the KS orbitals results in mixed quantum states with undefined contributions from singlet and triplet states. We have performed modeling of electronic excitations at the level of KS orbitals and energies and have attempted to validate the results by using more exact methods. One such method is based on time-dependent density-functional response theory (TD-DFRT) [38, 39], where the true excitation energies are found by solving the following eigenvalue problem:

$$\sum_{ij\sigma}[\omega_{ij}^2\delta_{ik}\delta_{jl}\delta_{\sigma\tau} + 2\sqrt{\omega_{ij}}K_{ij\sigma,kl\tau}\sqrt{\omega_{kl}}]F_{ij\sigma}^I = \omega_I^2 F_{kl\tau}^I. \tag{9}$$

Here σ and τ are spin indices. The indices i, k correspond to occupied, and j, l to unoccupied, KS orbitals, respectively. The coupling matrix $\boldsymbol{K}$ is defined as (see [40] for a more detailed description of the method):

$$K_{ij\sigma,kl\tau} = \iint{}' \psi_i(\boldsymbol{r})\psi_j(\boldsymbol{r})\left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r'}|} + \frac{\delta^2 E_{\rm xc}}{\delta n_\sigma\,\delta n'_\tau}\right)\psi_k(\boldsymbol{r'})\psi_l(\boldsymbol{r'}), \tag{10}$$

where we use a notation consistent with that in (2).

Our structural models of a-As_2S_3 were obtained by a "melt-and-quench" procedure [35] similar to the one described in the preceding Chapter by *Drabold* and *Abtew*. The structural quality of models can be assessed by examining the radial-distribution function (RDF) and the structure factor (see Fig. 1). Apart from general good agreement with experimental data, an interesting feature revealed by Fig. 1a is the shape of the first peak of the RDF. The shoulders on both sides of this peak indicate the presence of homopolar (As–As or S–S) bonds in the material. Such bonds (chemical defects), as well as coordination (or topological) defects, are easily detectable in computer models. An example of a configuration featuring an As–As bond is shown in Fig. 2. Special significance can be attributed to the presence of five-membered rings in models with an appreciable concentration of homopolar bonds (models with all-heteropolar bonds contain only an even number of

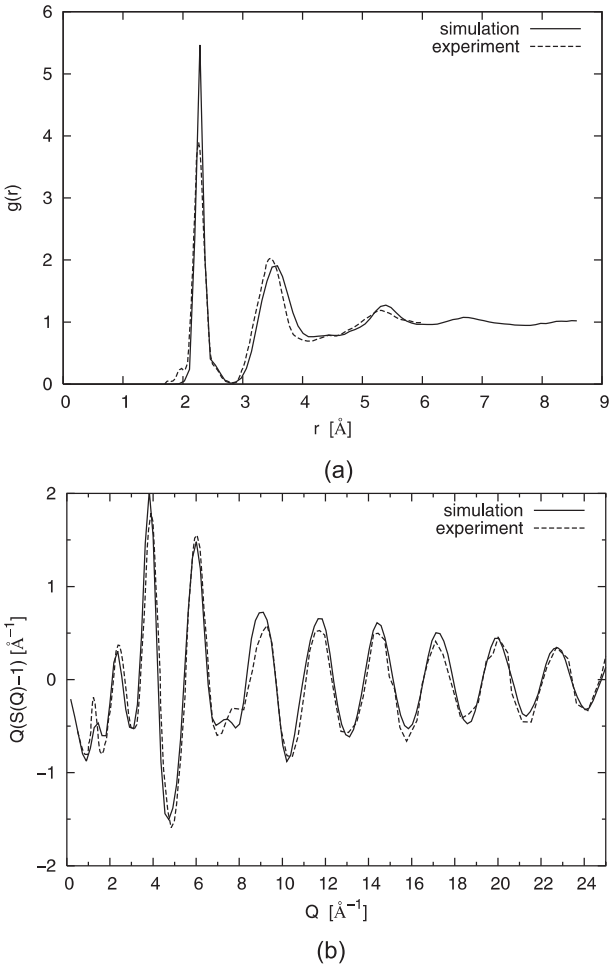

Fig. 1. (**a**) Pair-correlation functions for a 200-atom model of a-As_2S_3 and the neutron-diffraction experiment [41]. (**b**) Reduced structure factors (interference functions) for the same model and experiment

atoms in all rings). When such rings share some of the bonds, the resulting local structure is close to that of cage-like molecules (e.g., As_4S_4), as found in the vapor phase and in some chalcogenide molecular crystals. Figure 2a shows two such bond-sharing rings. Upon breaking the two bonds connecting the rings to the rest of the network, the distance between the two freed arsenic atoms (connected by a dashed line in Fig. 2a) could be reduced, thus producing another As–As homopolar bond and this group of atoms would then form an As_4S_4 molecule (shown in Fig. 2b). Evidence of the presence of such molecules in bulk $As_{\boldsymbol{x}}S_{1-\boldsymbol{x}}$ glasses from Raman-scattering experiments

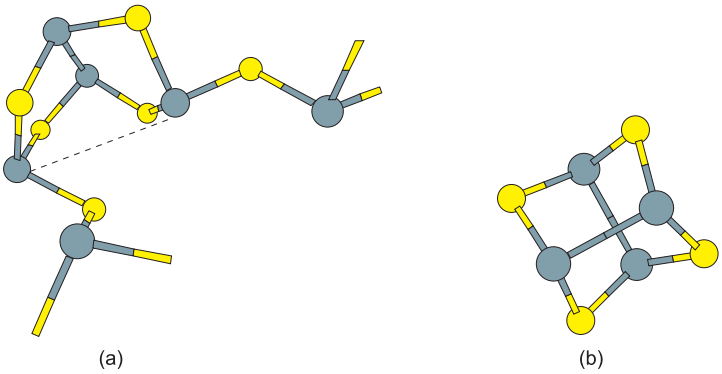

Fig. 2. (**a**) Fragment of a 200-atom model of a-As_2S_3: two bond-sharing five-membered rings and the two AsS_3 groups connected to this structure. The dangling bonds show where the displayed configuration connects to the rest of the amorphous network. The *dashed line* connects two As atoms, which, if brought nearer together, would close up to form an As_4S_4 molecule shown in panel (**b**). The shading of the As atoms (all with three neighbors) is darker than that of the S atoms (all with two neighbors)

has recently been reported in [42]. Our result shows that the As_4S_4 fragments may not only form discrete cage-like molecules but also may be embedded into the amorphous network. We verified that the vibrational signatures of the As_4S_4 fragment from models 1 and 2 are similar to those from an isolated As_4S_4 molecule, apart from a few very symmetric modes of the latter. The observed tendency for formation of quasimolecular structural groups suggests that amorphous chalcogenides can be viewed as nanostructured materials.

Localization of the electronic states near the optical bandgap edges is of great interest for studies of photoinduced phenomena. The inverse participation ratios (IPR, defined in the Chapter by *Drabold* and *Abtew*) for a model of a-As_2S_3 are shown in Fig. 3. The general picture is that, at the top of the valence band, the eigenstates are predominantly localized at what can be called sulfur-rich regions, where several sulfur atoms are closer than about 3.45 Å, i.e., their interatomic distances are on the low-r side of the second peak in $g(r)$ shown in Fig. 1a or some of these atoms form homopolar S–S bonds. For instance, most of the HOMO (highest occupied molecular orbital) level in this model is localized at two sulfur atoms separated by 3.42 Å and which are part of the molecule-like fragment depicted in Fig. 2a. By inspecting the projected (local) IPRs in Fig. 3b at the optical gap edges, it is seen that the IPRs are greatest for the S atoms. It appears that the localization at the top of the valence band is facilitated by the proximity of the lone-pair p orbitals in the sulfur-rich regions. This observation is consistent with a result for a-GeS_2 [44].

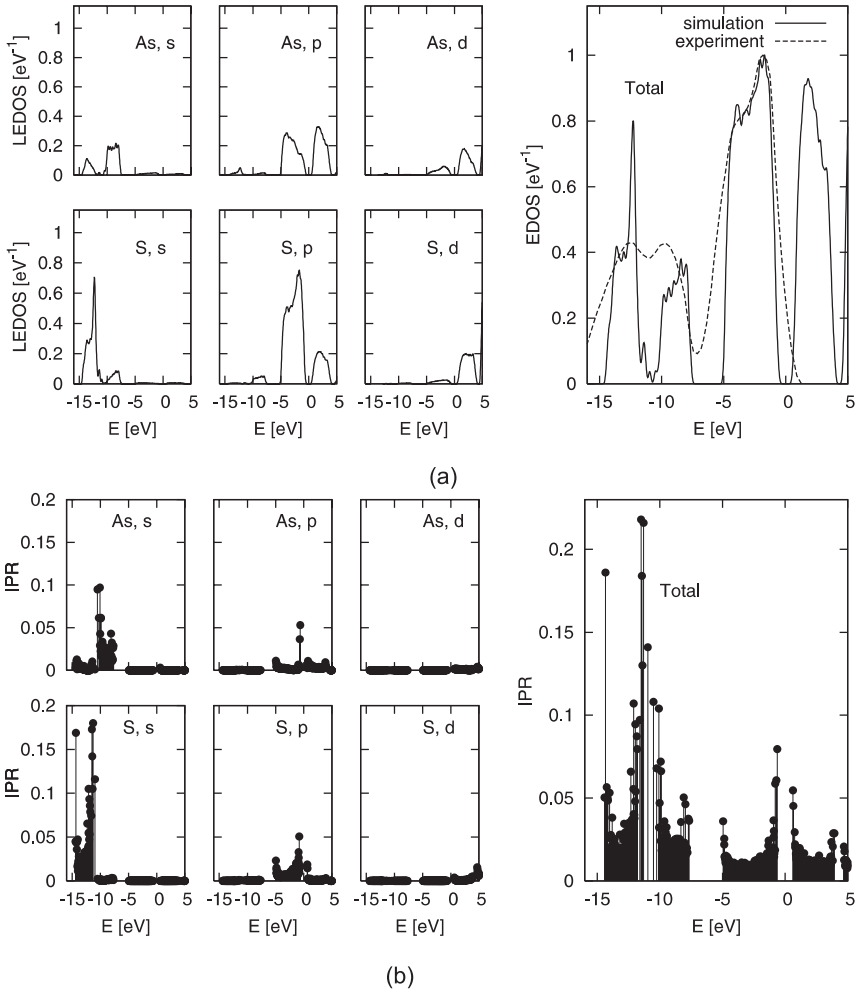

Fig. 3. Local and total electronic density of states (**a**) and inverse participation ratios (**b**) for a 200-atom model of a-As_2S_3. The Fermi energy is at the energy origin. The experimental data in (**a**) are obtained from [43]

At the bottom of the conduction band, the states tend to localize at various anomalous local configurations, such as four-membered rings, S–S homopolar bonds (some of these bonds are in five-membered rings) and valence-alternation pairs (coordination defects). For example, the LUMO (lowest unoccupied molecular orbital) state for the structure depicted in Figs. 5a,b is localized on an overcoordinated S atom (with three bonded neighbors).

We found that a basis set of s, p and d Slater-type orbitals for all atoms is an essential prerequisite for the observation of overcoordinated defects [36,45].

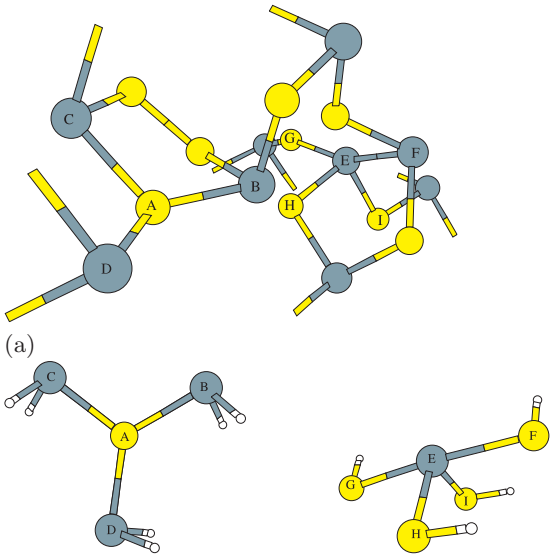

Fig. 4. Planar trigonal, $[S_3]^+$ (marked by the letter "A"), and "seesaw", $[As_4]^-$ (marked by letter "E"), configurations in (**a**) a fragment of a 60-atom model of a-As_2S_3 (the dangling bonds show where the displayed configuration connects to the rest (not shown) of the network) and, (**b**) and (**c**), charged isolated clusters (the dangling bonds are terminated with hydrogen atoms). The shading of the atoms is the same as in Fig. 2

It is possible to analyze individual contributions of orbitals of each type to the total EDOS and IPR, and these contributions are shown in Fig. 3.

As mentioned above, in the context of photoinduced metastability, a great deal of significance is attributed to the presence of topological and/or chemical defects [9]. It is therefore imperative to create models both with and without such defects in a theoretical investigation that attempts to be conclusive. Defect-free models can be produced by "surgical" manipulations. For example, atoms with undesired coordination can be removed from the model, and, in order to eliminate chemical defects, one can iteratively apply the following algorithm. First, a sulfur atom is inserted in the middle of each As–As homopolar bond. Second, each S–S bond is replaced by a single sulfur atom located at its midpoint so that each local As–S–S–As configuration is turned into As–S–As. Third, the distance between each newly introduced S atom and its two nearest arsenic atoms in the newly created As–S–As units is reduced in order to increase the bonding character of the As–S bonds stretched by the above manipulation. Fourth, the modified configuration is relaxed in an MD run. We have found that only a few iterations can be sufficient in order to obtain models with all-heteropolar (As–S) bonds.

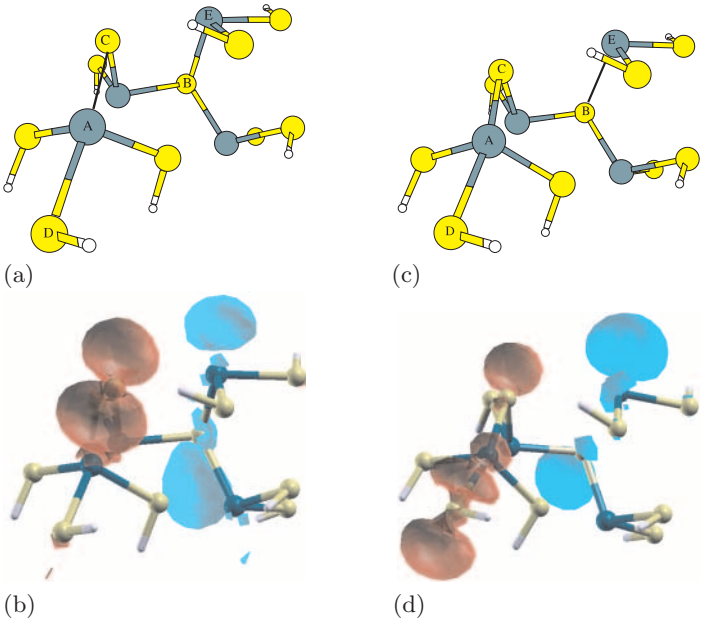

Fig. 5. An $As_4S_{10}H_8$ cluster containing both defect centers $[As_4]^-$ (marked by letter "A") and $[S_3]^+$ ("B"). The shading of the atoms is the same as in Fig. 2. The *black solid lines* signify elongated bonds. (**a**) Optimized ground-state geometry. Bond lengths are (Å): AC = 3.00, AD = 2.32, and BE = 2.4. (**b**) Isosurfaces corresponding to the value of 0.025 of electron density in the HOMO (darker red surface) and LUMO (lighter cyan surface) states for the structure shown in (**a**). (**c**) Optimized geometry in the first singlet excited state. Bond lengths are (Å): AC = 2.43, AD = 2.44, and BE = 2.82. (**d**) Same as (**b**), but for the structure shown in (**c**)

Elimination of the defects in our models removes some electronic states in the optical bandgap. As a result, the bandgap broadens, which can be viewed as an artificial bleaching of the material. The states at the band edges, however, are still localized due to disorder. It is possible that exposure to light may lead to bond breaking and defect creation even in such "all-heteropolar" models. Indeed, As–S bond elongation/breaking has been observed in all-heteropolar clusters [23,24] and cage-like molecules (unpublished). Such bond breaking under irradiation can lead to the creation of self-trapped excitons (STEs), i.e., oppositely charged defect pairs, as shown in [29] for silicon dioxide where, following Si–O bond breaking, the hole is localized at the oxygen defect center and the electron at the silicon defect center. In analogy with the case of SiO_2, it is possible that positively charged chalcogen defects and negatively charged nonchalcogen defects are introduced in amorphous chalcogenides by a similar mechanism. In the As–S system, such STEs can possibly lead to creation of the $[As_4]^-$–$[S_3]^+$ defect pairs described below.

These defects, as well as those that are always present in real materials and "as-prepared" models, may mediate local structural rearrangements upon photon absorption. Creation of additional states at optical bandgap edges (photodarkening) of a-As_2S_3 exposed to light has long been attributed to the generation by illumination of defects in excess of the thermal equilibrium concentration [9], but the exact mechanism of defect creation and the nature of the defects are still enigmatic. Possible candidate defect types, notably valence-alternation pairs, have been proposed over the years [9]. Normally, such defect pairs contain singly coordinated chalcogen atoms having distinct spectroscopic signatures [46]. Experimentally, the concentration of these defects is estimated to be rather small [47], i.e., $10^{17}\,\mathrm{cm}^{-3}$, compared with the atomic density of about $2 \times 10^{25}\,\mathrm{cm}^{-3}$, in order quantitatively to account for the observed magnitude of the photoinduced effects.

In our simulations, in addition to the charged coordination defects previously proposed to exist in chalcogenide glasses, a novel defect pair, $[As_4]^-$–$[S_3]^+$ (see Fig. 4), consisting of a fourfold coordinated arsenic site in a "see-saw" configuration and a threefold coordinated sulfur site in a near-planar trigonal configuration, was found in several models [33]. Such defect pairs are unusual in two ways. First, there is an excess of *negative* charge in the vicinity of the normally electropositive pnictogen (As) atoms and, second, there are no undercoordinated atoms with dangling bonds in these local configurations. The latter peculiarity may be why such defect pairs have not yet been identified experimentally. These defect pairs, however, are consistent with the STEs described above.

Although electronic excitations, where one electron is promoted from HOMO to LUMO Kohn–Sham states [48], are not especially realistic, we simulate such excitations in order qualitatively to assess defect stability with respect to (optically induced) electronic excitations. In some models, $[S_3]^+$–S_1^- defect pairs are converted into $[As_4]^-$–$[S_3]^+$ pairs as a result of the electronic excitation. The presence of these defect pairs introduces additional localized states at the optical bandgap edges with energies quantitatively consistent with the phenomenon of photodarkening [34].

A possible mechanism of conversion of $[S_3]^+$–S_1^- defect pairs into $[As_4]^-$–$[S_3]^+$ pairs is illustrated in Fig. 5, which shows an $As_4S_{10}H_8$ cluster containing an $[S_3]^+$–S_1^- pair in the ground state. Geometry optimization in the first singlet excited state within the linear-response approximation to time-dependent (TD) density-functional theory (which gives a much better description of excited states compared with HOMO-to-LUMO electron excitations [40]) leads to a redistribution of electron density, so that the singly coordinated S atoms become attached to a normally coordinated As atom, thus forming an $[As_4]^-$ defect. At the same time, the $[S_3]^+$ defect breaks up (we observed that bond breaking/elongation in all our models generally occurs at the groups of atoms where the LUMO is localized, indicating the expected antibonding character of LUMO states.). Perhaps in bulk materials similar rearangements can lead to the creation of an $[S_3]^+$ center at a differ-

ent location, which is suggested by the observation that, in our simulations of photoexcitations in supercell models, the $[S_3]^+$ centers after excitation are not necessarily located at the same S atom as before the excitation.

References

[1] S. R. Elliott: *Physics of Amorphous Materials*, 2nd ed. (Longman Scientific and Technical, London 1990) 270, 271
[2] J. J. Ludlam, S. N. Taraskin, S. R. Elliott: J. Phys. Condens. Matter **17**, L321 (2005) 270
[3] S. R. Elliott: Chalcogenide glasses, in J. Zarzycki (Ed.): *Materials Science and Technology*, vol. 9 (VCH, Weinheim 1991) p. 375 270
[4] A. Zakery, S. R. Elliott: J. Non-Cryst. Solids **330**, 1 (2003) 271
[5] H. Hisakuni, K. Tanaka: Science **270**, 974 (1995) 271
[6] S. N. Yannopoulos: Photo-plastic effects in chalcogenide glasses: Raman scattering studies., in A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) p. 119 271
[7] P. Krecmer, A. M. Moulin, T. Rayment, R. J. Stephenson, T. Rayment, M. E. Welland, S. R. Elliott: Science **277**, 1799 (1997) 271
[8] M. Stuchlik, S. R. Elliott: Proc. IEE A: Science, Measurement and Technology **151**, 131 (2004) 271
[9] A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) 271, 272, 280, 282
[10] K. Shimakawa, A. Kolobov, S. R. Elliott: Adv. Phys. **44**, 475 (1995) 271, 272
[11] S. A. Solin, G. V. Papatheodorou: Phys. Rev. B **15**, 2084 (1977) 271
[12] S. Rajagopalan, K. S. Harshavardhan, L. K. Malhotra, K. L. Chopra: J. Non-Cryst. Solids **50**, 29 (1982) 272
[13] H. Hisakuni, K. Tanaka: Appl. Phys. Lett. **65**, 2925 (1994) 272
[14] S. R. Ovshinsky: Phys. Rev. Lett. **21**, 1450 (1968) 272
[15] T. Ohta, S. R. Ovshinsky: Phase-change optical storage media., in A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) p. 310 272
[16] S. R. Elliott, A. V. Kolobov: J. Non-Cryst. Solids **128**, 216 (1991) 272
[17] A. Kolobov, S. R. Elliott: Adv. Phys. **40**, 625 (1991) 272
[18] T. Wagner, M. Frumar: Optically-induced diffusion and dissolution in metals in amorphous chalcogenides., in A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) p. 160 272
[19] S. Maruno, T. Kawaguchi: J. Appl. Phys. **46**, 5312 (1975) 272
[20] T. Kawaguchi: Photo-induced deposition of silver particles on amorphous semiconductors., in A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) p. 182 272
[21] V. G. Zhdanov, V. K. Malinovskii: Sov. Tech. Phys. Lett. **3**, 943 (1977) 272
[22] V. M. Lyubin, M. L. Klebanov: Photo-induced anisotropy in chalcogenide glassy semiconductors., in A. V. Kolobov (Ed.): *Photo-Induced Metastability in Amorphous Semiconductors* (Wiley-VCH, Weinheim 2003) p. 91 272
[23] T. Uchino, D. C. Clary, S. R. Elliott: Phys. Rev. Lett. **85**, 3305 (2000) 273, 281

[24] T. Uchino, D. C. Clary, S. R. Elliott: Phys. Rev. B **65**, 174204 (2002) 273, 281
[25] T. Uchino, S. R. Elliott: Phys. Rev. B **67**, 174201 (2003) 273
[26] T. Mowrer, G. Lucovsky, L. S. Sremaniak, J. L. Whitten: J. Non-Cryst. Solids **338**, 543 (2004) 273
[27] R. Stowasser, R. Hoffman: J. Am. Chem. Soc. **121**, 3414 (1999) 273
[28] M. S. Hybertsen, S. G. Louie: Phys. Rev. B **34**, 5390 (1986) 273
[29] S. Ismail-Beigi, S. G. Louie: Phys. Rev. Lett. **95**, 156401 (2005) 273, 281
[30] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, R. Kaschner: Phys. Rev. B **51**, 12947 (1995) 273, 274
[31] M. Elstner, D. Porezag, G. Jungnickel, J. Elstner, M. Haugk, T. Frauenheim, S. Suhai, G. Seifert: Phys. Rev. B **58**, 7260 (1998) 273, 274
[32] T. Frauenheim, G. Seifert, M. Elstner, T. Niehaus, C. Köhler, M. Amkreutz, M. Sternberg, Z. Hajnal, A. Di Carlo, S. Suhai: J. Phys. Condens. Matter **14**, 3015 (2002) 273
[33] S. I. Simdyankin, T. A. Niehaus, G. Natarajan, T. Frauenheim, S. R. Elliott: Phys. Rev. Lett. **94**, 086401 (2005) 274, 282
[34] S. I. Simdyankin, M. Elstner, T. A. Niehaus, T. Frauenheim, S. R. Elliott: Phys. Rev. B **72**, 020202 (2005) 274, 282
[35] S. I. Simdyankin, S. R. Elliott, Z. Hajnal, T. A. Niehaus, T. Frauenheim: Phys. Rev. B **69**, 144202 (2004) 274, 276
[36] T. A. Niehaus, M. Elstner, T. Frauenheim, S. Suhai: J. Mol. Struct. THEOCHEM **541**, 185 (2001) 274, 279
[37] A. Görling: Phys. Rev. A **54**, 3912 (1996) 276
[38] M. E. Casida: *Recent Advances in Density Functional Methods*, Part I (World Scientific, Singapore 1995) p. 155 276
[39] M. E. Casida: Developments and applications of modern density functional theory, in J. M. Seminario (Ed.): *Theoretical and Computational Chemistry*, vol. 4 (Elsevier Science, Amsterdam 1996) p. 391 276
[40] T. A. Niehaus, S. Suhai, F. Della Sala, P. Lugli, M. Elstner, G. Seifert, T. Frauenheim: Phys. Rev. B **63**, 085108 (2001) 276, 282
[41] J. H. Lee, A. C. Hannon, S. R. Elliott: eprint: cond-mat/0402587 277
[42] D. G. Georgiev, P. Boolchand, K. A. Jackson: Philos. Mag. **83**, 2941 (2003) 278
[43] S. G. Bishop, N. J. Shevchik: Phys. Rev. B **12**, 1567 (1975) 279
[44] S. Blaineau, P. Jund: Phys. Rev. B **70**, 184210 (2004) 278
[45] S. I. Simdyankin, S. R. Elliott, T. A. Niehaus, T. Frauenheim: Effect of defects in amorphous chalcogenides on the atomic structure and localisation of electronic eigenstates, in P. Vincenzini, A. Lami, F. Zerbetto (Eds.): *Computational Modeling and Simulation of Materials III*, vol. A (Techna Group s.r.l., Faenza 2004) p. 149 279
[46] M. Kastner, D. Adler, H. Fritzsche: Phys. Rev. Lett. **37**, 1504 (1976) 282
[47] A. Feltz: *Amorphous Inorganic Materials and Glasses* (VCH, Weinheim 1993) p. 203 282
[48] X. Zhang, D. A. Drabold: Phys. Rev. Lett. **83**, 5042 (1999) 282

Index

Index