Atomistic and *Ab Initio*
Prediction and Optimization
of Thermoelectric and Photovoltaic Properties

by

Maria Kai Yee Chan

Submitted to the Department of Physics
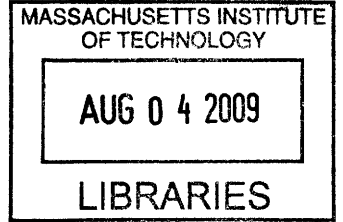in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

Author . . . . .
Department of Physics
May 1, 2009

Certified by . . . .
Professor Gerbrand Ceder
R. P. Simmons Professor of Materials Science and Engineering
Thesis Supervisor

Certified by . . . .
Professor John Joannopoulos
Francis Wright Davis Professor of Physics
Thesis Supervisor

Accepted by . . . .
Professor Thomas J. Greytak
Department Head for Education

Atomistic and *Ab Initio*

Prediction and Optimization

of Thermoelectric and Photovoltaic Properties

by

Maria Kai Yee Chan

Abstract

The accurate prediction of physical properties in the vast spaces of nanoscale structures and chemical compounds is made increasingly possible through the use of atomistic and *ab initio* computation. In this thesis we investigate lattice thermal conductivities κ_L and electronic band gaps E_g, which are relevant to thermoelectric and photovoltaic applications, respectively, and develop or modify computational tools for predicting and optimizing these properties.

For lattice thermal conductivity, we study SiGe nanostructures, which are technologically important for thermoelectric applications. From computing κ_L for various SiGe nanostructures, we establish that the Kubo-Green approach using classical molecular dynamics (MD) gives additional quantitative predictions not available from phenomenological models, such as the existences of a minimum value of κ_L as the nanostructure size is varied and of configurational dependence of κ_L. We carry out the minimizatin of κ_L in the space of atomic configurations in SiGe alloy nanowires and demonstrate the feasibility of using the cluster expansion technique to parameterize κ_L. We find that the use of coarse graining and a meta cluster expansion approach is effective, in conjunction with a genetic algorithm, to find configurations which drastically lower κ_L. The low values of κ_L obtained, close to the bulk amorphous limit, are due to the absence of long-range order, and such absence allows a local cluster expansion approach to optimize κ_L.

We examine *ab initio* bandgap prediction for semiconductor compounds, and address the large errors of Kohn-Sham band gaps in density functional theory (DFT). We apply corrections using the self-energy approach in the GW approximation, which includes non-local screened exchange and correlation, and find that the G_0W_0 approximation significantly reduces prediction errors compared to Kohn-Sham band gaps, though at much higher computational cost. We propose a new method involving total energies in DFT to predict the fundamental gap, by use of the properties of the screening or exchange-correlation hole in an electron gas. With this method, we are

able to efficiently predict band gaps that are in agreement with experimental values.

Thesis Supervisor: Professor Gerbrand Ceder
Title: R. P. Simmons Professor of Materials Science and Engineering

Thesis Supervisor: Professor John Joannopoulos
Title: Francis Wright Davis Professor of Physics

Acknowledgments

First of all I would like to express my sincere gratitude for the support, inspiration and guidance from my advisor, Professor Gerbrand Ceder. Without his leadership and wisdom, the research in this thesis would not have been possible. I would also like to thank my co-advisor, Professor John Joannopoulos, and members of my thesis committee Professor Mildred Dresselhaus and Professor Patrick Lee, for their guidance along the way. I am extremely fortunate to have been under the tutelage of these outstanding scientists.

I would like to acknowledge my collaborators: Professor Giulia Galli, Davide Donadio, John Reed, Tim Mueller, Shirley Meng, and Andrew Williamson in the project on thermal conductivity in SiGe nanostructures; Ruoshi Sun and Robert (Yabi) Wu in the band gap on semiconductors project. The work of Anubhav Jain, Chris Fischer, Geoffroy Hautier and Charles Moore in the construction of the Materials Genome database is also instrumental in the band gap project. In addition, I would like to thank Caetano Miranda, Fei Zhou, Rickard Armiento, Professor Nicola Marzari, and Professor Troy van Voorhis for helpful discussions. I would also like to thank DARPA, the Chesonis foundation, and ENI for funding support.

Other current and former members of the Ceder group have also provided an intellectually stimulating and vibrant research environment. I am particularly grateful to senior members Dane Morgan, Fei Zhou, Matteo Cococcioni, Anton Van Der Ven, Chris Marianetti, Kristin Persson, Chris Fischer, Tim Mueller and Caetano Miranda for their assistance and advice. I thank colleagues Kevin Tibbetts, Anubhav Jain, ShinYoung Kang, Denis Kramer, and Rahul Malik for motivational discussions. Together with Shyue Ping Ong, Robert Doe, Lei Wang, Byungchan Han, Byoungwoo Kang, Jae Chul Kim, Kisuk Kang, Xiaohua Ma, Lusann Yang, Sahak Patrosyan, Yoyo Hinuma and Thomas Maxsich, all of these colleagues make working in the Ceder group a rich educational experience. I also thank Kathryn Simons, administrative assistant to Professor Ceder, for her tireless efforts and amazing skills in ensuring that our group functions.

The MIT Graduate Women in Physics group has provided a supportive network which was very helpful during my graduate career. I thank all the current and former members of MIT Women in Physics for their comraderie and friendship, and our benefactor, Margaret Wong, for making it all possible. In particular, I thank Professor Michelle Povinelli for taking the time to assist in the preparation for my qualifying exam; as well as Kristin Burgess, Kendra Vant, Julia Steinberger and Professor Nergis Mavalvala for advice.

I owe many thanks to my colleague and dear friend Ashley Predith, to whom I can always turn for advice, for all the wisdom she imparted upon me, and all the inspirations she provided.

Last but not least, I thank my husband James Osborn for his loving support, tireless encouragement, and unfailing optimism.

Contents

List of Figures

9

List of Tables

Chapter 1

Introduction

1.1   A lot more is a lot different

Since the famous exclamation that "More is Different" by Anderson [9] some forty years ago, the study of condensed matter has expanded considerably both in scale and in complexity. Some of the most scientifically and technologically significant discoveries, e.g. Giant Magnetoresistance for which the 2007 Nobel prize in physics was awarded, have been made in the phenomenologically rich space between the bulk thermodynamic limit and individual particles – the nanoscale. Just as the breaking of continuous translational symmetry leads to the wealth of crystal structures and accompanying properties, the breaking of discrete translational symmetry in nanoscale systems gives rise to additional complexity and new phenomena. Another dimension of complexity comes from the vast combinatorial expanse of the hundred or so elements, from which thousands of new compounds are discovered each year in the inorganic subspace alone [10].

In a twist of fate, however, increasingly such complexity is dealt with by constructionist means through the use of computation. In atomistic and *ab initio* computing, phenomena and properties emerge from mere inputs of initial conditions and fundamental laws, be they classical or quantum in nature. The broad (though far from universal) applicability of such powerful computational tools as molecular dynamics (MD) and density functional theory (DFT) has allowed us on many occasions

to answer precisely **how** "more" is "different". The ability to make quantitative or semi-quantitative predictions lends enormous scientific and societal relevance to these computational methods as well as their underlying theories.

The overarching theme of this thesis is the intersection between the understanding and prediction of physical properties (in the complex spaces of nanosystems and chemical compounds), and the development of computational tools to facilitate this understanding. We focus on phenomena with practical applications of broad societal impact, namely thermoelectric and photovoltaic effects. For each, we identify key properties, the space of possibilities, and existing computational tools. For each, we develop new or improved computational tools for more efficient exploration of the complexities or more accurate predictions of the properties.

1.2 Thermoelectric and photovoltaic effects

Thermoelectrics and photovoltaics hold significant potentials as alternative energy resources. Of the 100 quadrillion BTU (30 trillion kWh) of energy consumed in the US in 2006, 93% of which from non-renewable sources, over 60% is lost as waste heat [11]. Recovering even a small fraction of this heat as electricity with thermoelectrics could significantly reduce the dependence on fossil fuels. The amount of solar energy incident on the Earth is so vast that the coverage of 0.2% of the land surface on Earth with 15% efficient solar cells could provide all human energy needs. In both thermoelectric and photovoltaic applications, there are key physical properties of the materials in use that critically affect their efficiencies. The scientific problem, then, is to understand and develop the ability to predict and optimize these properties. Economical considerations, which favor the use of compounds containing earth-abundant elements, serve as constraints in the property optimization problems.

Thermoelectric phenomena involve the transport of thermal and/or electric currents under the driving forces of electric fields and/or temperature gradients. The Seebeck effect occurs when a voltage (ΔV) is generated from a temperature difference (ΔT), and is the principle behind thermoelectric power generation. The Seebeck coef-

ficient S is defined as the ratio between voltage generated and temperature difference, i.e.

$$S = \frac{\Delta V}{\Delta T} \tag{1.1}$$

The Peltier effect stems from the differences in ratios of heat to electrical currents in different materials, and is the basis of thermoelectric cooling (refrigeration), the effectiveness of which also depends on the Seebeck coefficient. To obtain a high efficiency for power generation or a high coefficient of performance for refrigeration, it is desirable to have a high thermoelectric figure of merit ZT, which is given by

$$ZT = \frac{S^2 \sigma}{\kappa} T \tag{1.2}$$

where σ is the electrical conductivity, κ the thermal conductivity, and T the temperature. Currently the highest ZT achieved for bulk materials is roughly 1.5 [12], which gives an efficiency 30% that of the Carnot efficiency operating between reservoirs of 300K and 900K.

The photovoltaic effect describes the generation of electrical current from charge carriers created through the absorption of photons. The effect occurs in a material, typically a semiconductor or organic compound, with an energy gap between occupied and unoccupied electronic states. The fundamental limitation on efficiency is imposed by this energy gap, although electronic transport and defect properties are also important. Currently the highest efficiencies of $\sim$40% are obtained with multi-junction semiconductor photovoltaic cells [13].

1.3   Overview of this thesis

In chapter 2, background information involving computational tools used and key thermoelectric and photovoltaic properties considered are given. We motivate the specific study of lattice thermal conductivities κ_L and electronic band gaps E_g, using a combination of molecular dynamics (MD), density functional theory (DFT), and cluster expansion techniques.

Chapter 3 is divided into two parts, namely the study of lattice thermal conductivities in various SiGe nanostructures, and the prediction and optimization of κ_L in SiGe nanowires. In the former we seek to establish the validity of classical MD simulations as a predictive tool for κ_L, study the factors that affect κ_L in these nanostructures, and identify the configurational degrees of freedom for which optimization is possible. In the latter we use the predictive tool of MD and the cluster expansion technique to carry out this optimization to devise nanowire configurations with low lattice thermal conductivities.

Chapter 4 tackles the issue of *ab initio* band gap prediction. We discuss and demonstrate the large errrors involved with band gap predictions in DFT and other first principles methods. Corrections to these errors are attempted in two approaches, the first being the many body perturbation theory self-energy correction, in the *GW* approximation. An alternative approach to predict the band gap, making use of the fundamental properties of screening or exchange correlation in an electron gas, is also suggested and tested.

Chapter 2

Background

2.1 Tools for materials properties prediction

Properties of matter are largely governed by electrons, ions[1], and their collective excitations. Due to the large mass difference between electrons and ions, in most cases it is possible to separately treat the electronic and ionic degrees of freedom using the Born-Oppenheimer approximation. In this section we give a birds-eye view of some of the computational tools used for the respective study of ions and electrons, focusing especially on the physical principles behind these methods, and the ways in which we will use, extend, or augment them for the purpose of this thesis.

2.1.1 Classical molecular dynamics (MD)

With the exception of a few light elements or at very low temperatures, the thermal deBroglie wavelengths of ions are much smaller than their spatial separations in normal matter, meaning that they can be described adequately using classical dynamics. In classical molecular dynamics (MD) simulations, Newton's equations of motion are integrated numerically to produce classical trajectories of ions. The forces governing the motion are typically obtained from a system-specific potential model. The simulations can be performed adiabatically, in which the system energy E is conserved, or

---

[1]We will use the word "ions" to denote atoms, molecules and charged ions.

at constant temperature T through the use of a thermostat. Analogously, either the volume V can be kept constant or the pressure P maintained using a barostat. Typically the particle number N is conserved. The common types of MD simulations are called NVE, NVT, and NPT where the symbols denote the quantities kept constant.

Classical MD is arguably the most widely-used computational tool for the study of condensed matter. The advantages of classical MD include its versatility, simplicity, computationally expediency, and physical transparency. From ionic trajectories, quantities such as diffusion rates and lattice thermal conductivities can be derived from pair correlation functions. In addition, mechanical, chemical or physical phenomena such as cracking, docking, and melting can also be studied. One disadvantage of classical MD, however, is the fact that a potential model needs to be developed for each type of system. There are usually parameters in each model that have to be fitted either to experimental or accurately-calculated quantities. Sometimes the models are not transferrable to different environments of the same species (e.g. amorphous vs. crystalline). Therefore, MD is most suited for the repetitive study of different behaviors and/or configurations of the same chemical species in similar environments. We employ classical MD in the study of the lattice thermal conductivity of a variety of different SiGe nanostructures (Chapter 3), a chemical system with well-characterized potential models.

2.1.2   Cluster expansion

Configurational freedom refers to the possibility of different spatial arrangements of species of ions, the simplest case being two species of atoms on a crystal lattice (e.g. a binary alloy) which we use as an example. In this case the number of configurations is 2^N, N being the number of lattice sites, so that explicit accounting quickly becomes intractable. The cluster expansion technique [14] is used to overcome this difficulty. The set of all 2^N configurations is recast into $\sum_M {}^N C_M$ terms, each representing a connected cluster of M sites (hence the name). A quantity that depends on configurations, e.g. the total energy, can be expressed as a sum of contributions from each of these clusters. If the dependence on configuration falls off sufficiently

rapidly with distance, then a small number of clusters would be sufficient to describe the configurational dependence.

In a typical application of the cluster expansion technique, a configurational-dependent quantity Q is initially calculated for a sample set of configurations by other means, and the contribution of each cluster to Q is then determined by fitting to the set of calculated values of Q. The fitting coefficients, called effective cluster interactions (ECIs), characterize the configurational dependence of Q. The value of Q for any other configurations can be determined from the ECIs. We stress that in order to use the cluster expansion method, it is necessary to have an initial method for computing Q. In the case of Q being the total energy E, the initial method is typically density functional theory (DFT). The advantage of the cluster expansion is that Q can be evaluated much more rapidly as a sum of cluster contributions using the ECIs than with the initial method, so that a sampling of configurational space is possible. In the case of Q being the total energy, sampling can be used to identify ground state configurations. In other words, cluster expansion enables a practical optimization to be made over the enormous space of 2^N configurations.

We extend the cluster expansion technique (Chapter 3) to study the configurational dependence of thermal conductivities in SiGe nanowires, using MD as the initial method of computation. We introduce new coarse-graining techniques to the cluster expansion method in order to account for the stochastic nature of heat transfer. Using the sampling capability of the cluster expansion, we carry out optimization in order to find low thermal conductivity configurations, which would be desirable for thermoelectric applications.

2.1.3 Density functional theory (DFT)

Density functional theory (DFT), based on universal quantum mechanical principles, reduces the intractable Schrödinger equation governing 10^{23} electrons in a typical physical system to a single-particle effective theory amenable to computational solutions. DFT has allowed an unprecedented breadth and depth of understanding of a wide range of materials properties – structural, mechanical, electronic, spectroscopic,

and magnetic, among others. The *ab initio* nature and hence broad applicability of DFT, together with its accuracy and computational efficiency, all contribute towards its popularity as of this writing, there are over 47,000 published scientific articles containing the term "density functional theory" in the Web of Science.

The foundation of DFT lies in the Hohenberg-Kohn theorem [15], which states that the ground state charge density of a system of interacting electrons uniquely determines the total Hamiltonian, including the external potential due to ions $V(\mathbf{r})$, and hence the total ground state energy. As a consequence, the total kinetic and Coulomb energy of a system of electrons is a universal functional $E[n(\mathbf{r})]$ of the charge density $n(\mathbf{r})$. Although this universal functional is not known, in the Kohn-Sham scheme [16], DFT can be recast into a familiar Schrödinger-like Kohn-Sham equation. The effects of electrons being identical fermions (exchange) and their interactions being many-body in nature (correlation) are encapsulated in an exchange-correlation potential. While the exact exchange-correlation potential is also not known, approximations such as the Local Density Approximation (LDA) and Generalized Gradient Approximation (GGA) are highly successful in producing quantitatively-accurate ground state energies and related properties such as lattice parameters.

The solutions to the Kohn-Sham equation are single-particle wavefunctions that formally only serve to reproduce the ground state density. In practice, however, the corresponding eigenvalues for solids are often compared with the band structures measured by photospectroscopy because there are often qualitative agreements. A major challenge facing DFT today, however, is the fact that for non-metals the energy gaps between the filled and occupied bands predicted using the Kohn-Sham eigenvalues are significantly below that of experimental values. This has been referred to as the "band gap problem" of DFT, and an oft-used comment is that DFT is a ground state theory and thus cannot be used to treat excited state properties. We will investigate, and propose a solution for, the band gap problem in DFT in Chapter 4.

2.2 Key thermoelectric and photovoltaic properties

2.2.1 Lattice thermal conductivity κ_L

Definition

The conduction of thermal energy in a system is governed by the Fourier Law, which states that the heat flux, i.e. thermal energy transferred per unit time per unit area, is proportional to the temperature gradient. The direction of the heat flux is towards lower temperature, so we have

$$\mathbf{J} = -\kappa \nabla T \tag{2.1}$$

where the (positive) constant of proportionality is called the thermal conductivity κ.

Importance of κ_L for thermoelectric efficiency

Recall that the thermoelectric figure of merit ZT which governs the thermoelectric power generation efficiency is given by:

$$ZT = \frac{S^2 \sigma}{\kappa} T \tag{2.2}$$

To increase ZT, it is desirable to decrease the thermal conductivity. The thermal conductivity κ has contributions from ionic and electronic degrees of freedom. From the Wiedemann-Franz law, which is applicable in most normal matter, the electronic thermal conductivity κ_e is proportional to the electrical conductivity σ. Since σ and the Seebeck coefficient S both need to be maximized in order to increase the figure of merit ZT, the lattice contribution to the thermal conductivity, which we will refer to as κ_L, becomes the only property tunable independently of the electronic transport properties. The focus on reducing the thermal conductivity in order to increase ZT is exemplified in the development of "phonon-glass-electron-crystal" [17] materials, in which the scattering of phonons is increased and the thermal conductivity decreased due to special characteristics of the crystal structure such as rattling atoms in cage-

like voids.

Reduction of κ_L in nanostructures

Since its proposal by Hicks and Dresselhaus in the 1990's [18, 19], the improvement of ZT by the use of nanostructures has been a subject of intense research. This improvement is accomplished in large part through the reduction of κ_L. In phenomenological models of the lattice thermal conductivity [20], this reduction can be explained by the fact that increasing the presence of boundaries reduces phonon lifetimes and hence deceases κ_L. For structures that are at true nanoscale, however, there are fundamental changes to heat transport due to phonon confinement, such as a reduction in phonon density of states and group velocities. Additional complexity of configurational dependence may appear in systems with very low values of κ_L and reduced long-range order. In order to ultimately devise nanostructures which minimize κ_L, it is important to understand these effects of quantum confinement and to investigate any configurational dependence. Since in nanostructures it is often necessary to go beyond bulk models and parameters, the use and further development of atomistic computational tools is an integral part of this investigation.

2.2.2 Electronic band gaps

Definition

Electronic structures govern vital properties of a photovoltaic material: optical absorption, radiative and non-radiative recombination rates, charge transport, and ultimately, efficiency. Arguably the most important electronic property of a photovoltaic material is the energy gap E_g, which determines the maximum conversion efficiency. The optical energy gap is defined as the lowest energy of a photon that can be absorbed to produce an electron and hole excitation, from which current and hence energy can be extracted. In a bulk semiconductor, the optical gap is approximately equal to the energy gap between the occupied and unoccupied states in the band structure, i.e. the band gap, since typical electron-hole (exciton) binding energies are

small (of order 0.05 eV).

Importance of band gaps for photovoltaic efficiency

Intuitive arguments can be used to predict the existence of an optimal band gap E_g for photovoltaic conversion efficiency. First of all, solar energy can only be transferred to the electrons and used to do work when the photon energy exceeds E_g. Because electrons thermalize to the band edges within femto-seconds, the energy derived per absorbed photon is no more than E_g. Intuitively, if E_g is too small, then the percentage of incident photons absorbed is large, but the energy extracted from each is small; conversely, if E_g is large, then the energy from each absorbed photon is large, but the number absorbed is small. Therefore, the maximum efficiency occurs at an intermediate range for E_g relative to the solar spectrum. A formal version of this intuitive argument is given below.

From purely thermodynamic arguments, originally by Shockley and Queisser [21], it can be shown that given a particular irradiation spectrum, there exists an optimal band gap that gives maximum ideal efficiency of a single-junction photovoltaic device[2]. The limitation is fundamentally due to detailed balance, i.e. the equivalence between matrix elements in the processes of absorption and spontaneous emission of photons through radiative recombination. A manifestation of detailed balance is that at any energy E, the absorptivity $a(E)$ (probability of photon absorption) is equal to the emissivity $\epsilon(E)$ (probability of photon emission). We assume the following for an ideal photovoltaic device:

- every photon with energy above the band gap E_g is absorbed $a(E) = \epsilon(E) = \Theta(E - E_g)$;
- absorption of one photon produces one electron-hole pair;
- there is no non-radiative recombination;
- current is transmitted without loss (i.e. the resistance in the circuit is zero);
- the device is at thermal equilibrium at ambient temperature T_a; and

---

[2]The derivation here closely follows that in [22].

- the difference between quasi-Fermi levels of electrons and holes $\Delta\mu$, also known as the chemical potential of light, is uniform throughout the device and equal to qV, where V is the bias potential and q the electron charge.

Let T_S denote the temperature of the sun and $f_S = \sin^2\theta_S \approx 2.16\times10^{-5}$ denote the solid angle subtended by the sun, and A be the surface area of the device. Assuming blackbody radiation for the sun, the ambient atmosphere and the device, the total current through the device is given by

$$I = qA\frac{2\pi}{h^3c^2}\int_{E_g}^{\infty} E^2\left[\frac{f_S}{e^{E/k_bT_S}-1} + \frac{(1-f_S)}{e^{E/k_bT_a}-1} - \frac{1}{e^{(E-qV)/k_bT_a}-1}\right]dE \qquad (2.3)$$

where the three terms inside the square brackets represent absorption from the sun, absorption from the ambient atmosphere, and emission of the device, respectively. The output power density as a function of bias voltage V is given by

$$\begin{aligned}
P(V) &= \frac{V\cdot I}{A}\\
&= qV\frac{2\pi}{h^3c^2}\int_{E_g}^{\infty} E^2\left[\frac{f_S}{e^{E/k_bT_S}-1} + \frac{(1-f_S)}{e^{E/k_bT_a}-1} - \frac{1}{e^{(E-qV)/k_bT_a}-1}\right]dE
\end{aligned} \qquad (2.4)$$

For any given E_g and T_a, $P(V)$ has a maximum for some value of the bias voltage $0 < V_m(E_g, T_a) < E_g$ (where V_m is measured in volts and E_g in eV). The maximum efficiency is given by

$$\eta_m(E_g, T_a, L) = \frac{P(V_m(E_g, T_a))}{f_S L} \qquad (2.5)$$

where $f_S L$ is the total solar energy density received by the device. For a blackbody sun (AM0) L is given by

$$L = \frac{2\pi}{h^3c^2}\int_0^{\infty} \frac{E^3}{e^{E/k_bT_S}-1}dE \qquad (2.6)$$

although a more realistic spectra which accounts for typical atmospheric absorption (called AM1.5) is usually used. If the solar radiation is concentrated, the factor f_S is multiplied by a constant C throughout. For a given irradiation spectrum L and operation temperature T_a, the efficiency $\eta_m(E_g, T_a, L)$ has a maximum value at a
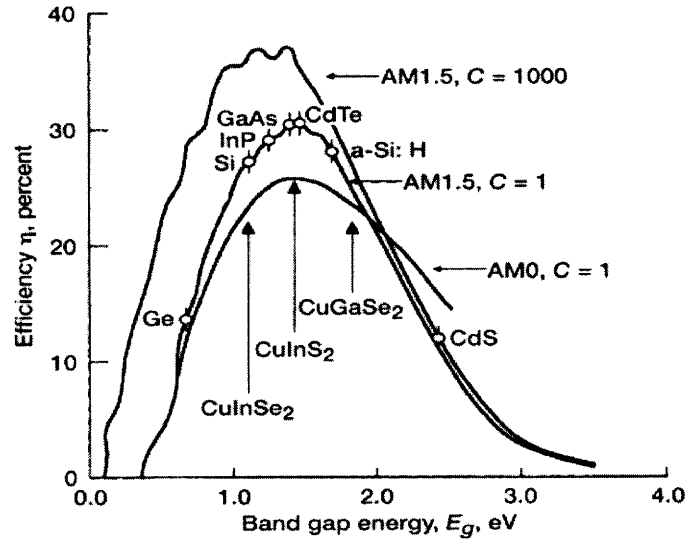
Figure 2-1: Theoretical maximum efficiency vs band gap for solar irradiation in space (AM0) and typical conditions on surface of the Earth (AM1.5). C represents concentration. The most commonly-used photovoltaic materials are also shown. Figure from [1].

certain E_g. The resultant maximum efficiencies η_m as a function of E_g for fixed L (AM0 and AM1.5) and T_a (300K) are shown in figure 2-1. The maximum efficiency of 31% without solar concentration is obtained for a band gap of about 1.3 eV. Similar arguments can be made for a multi-junction solar cell, in which light passes through layers of materials with successively lower band gaps such that the full energy from higher energy photons can be utilized in upper layers and yet lower energy photons can still be absorbed in lower layers. For two junctions, the maximum efficiency with full concentration is 55%, with optimal upper and lower gaps of 1.65 and 0.75 eV, respectively. Under the limit of an infinite number of layers, and full concentration, the thermodynamic limit of the efficiency is 86% [22].

Search for novel photovoltaic materials

There are at least tens of thousands of known stable inorganic compounds [10], and perhaps hundreds of thousands more that have yet to be discovered. It is quite possible that an excellent photovoltaic material lurks amongst them. If one has the computational ability to efficiently determine some of the pertinent properties, such

as band gaps, of these compounds, the search for a novel material may be carried out virtually. Thus the above-mentioned "band gap problem" in DFT is more than a mere academic curiosity. We find, again, a confluence of the needs for physical understanding and improvement of computational tools in our pursuit of a novel photovoltaic material.

Chapter 3

Thermal Transport in SiGe Nanostructures

3.1 Introduction

3.1.1 SiGe nanostructures for thermoelectric applications

Minimizing the thermal conductivity in silicon-germanium nanostructures is of special interest for thermoelectric applications. Bulk SiGe alloys have been used for thermoelectric power generation for several decades, most notably in NASA space missions since the 1970's. The mass disorder in Si-Ge causes enhanced phonon scattering, thereby reducing the thermal conductivity without significant modification of the electronic properties. Since proposed by Hicks and Dresselhaus in the 1990's [18, 19], lower dimensional systems such as quantum wells [23], thin films [24], superlattices [25], nanocomposites [26, 27] and nanowires [28] have been actively investigated for improved thermoelectric properties. Thermoelectric performance is measured by the figure of merit $ZT = S^2 \sigma T / (\kappa_e + \kappa_L)$, where S is the Seebeck coefficient, σ is the electrical conductivity, T is the temperature, and κ_e and κ_L are the electronic and lattice contributions to thermal conductivity, respectively. A key attribute of nanostructuring is the ability to achieve high ZT by reducing κ_L through the confinement and enhanced scattering of phonons, without a concomitant reduc-
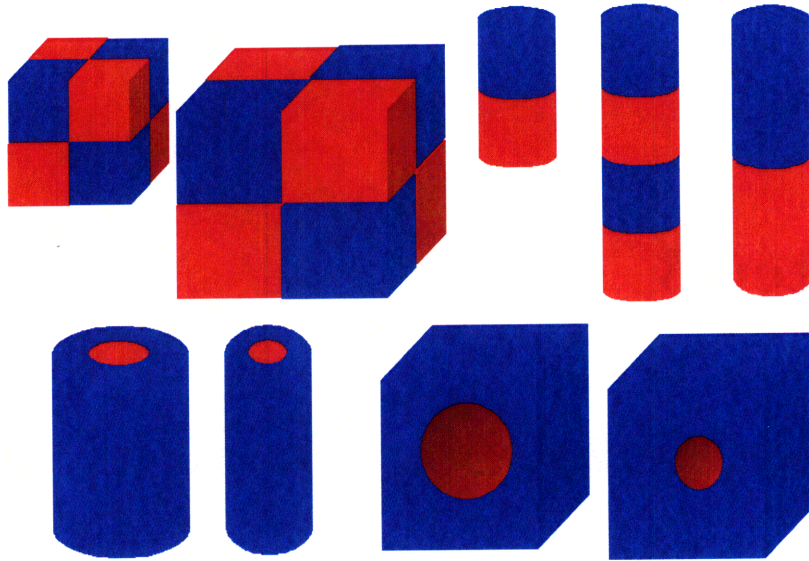
Figure 3-1: Types of nanostructures considered. Clockwise from top left: superlattices, superlattice nanowires, nanoparticle inclusions, and core-shell nanowires.

tion in σ or S. This is possible because of the disparate wavelengths and scattering lengths of electrons vs. phonons [29].

3.1.2 Measurements of κ_L in SiGe nanostructures

Experimental results show nanostructuring to be effective at reducing κ_L. For example, silicon nanowires with diameters from 20 to 100 nm show a thermal conductivity of 10 to 40 W/m-K at 300K [30], compared to 124 W/m-K for bulk Si [31]. Si/SiGe superlattice nanowires show a further 5-fold reduction compared to pure Si nanowires of similar diameters [32]. In SiGe nanocomposites consisting of 10 nm Si nanoparticles in a Ge host, κ_L has been measured to be less than 2 W/m-K [33]. Pure Si nanowires with roughened surfaces have been measured to have κ_L as low as 1.2 W/m-K [34]. The various experimental results show that κ_L depends sensitively on the type and size of nanostructures (see figure 3-1 for illustration), alloy compositions, and surface characteristics.

3.1.3 Computational prediction of κ_L in SiGe nanostructures

Computational studies using various approaches have been used to predict the dependence of the lattice thermal conductivity in SiGe nanostructures on types, size, shapes, nature of transport, and the characteristics of interfaces. Here we briefly describe representative studies and key findings from each approach. We shall focus our attention on diffusive transport, as it is the predominant mode of thermal transport within the ranges of interest of κ_L and of the temperature.

In the direct approach, the values of κ_L of Si nanowires in non-equilibrium molecular dynamics (MD) have been obtained [35] from the Fourier law, i.e. $\mathbf{J} = -\kappa_L \nabla_x T$ where T is the temperature, x is the direction of heat flow, and $\mathbf{J}$ is the heat flux. The resultant values of κ_L at 300K for 4-nm-diameter Si nanowires are 20-30 W/m-K for tetrahedrally-coordinated nanowires and 5-10 W/m-K for clathrate nanowires. In reference [35], an unexplained anomalous increase in κ_L as the diameter is decreased from 3 nm to 1 nm is found at 100K. Note that the temperature gradients in the simulations are of order 10^8 K/m.

In the continuum approach, the Boltzmann transport equation has been used together with bulk parameters in a heat-transfer model to describe the thermal transport in SiGe superlattices [36], nanocomposites [37], superlattice nanowires [38] and core-shell nanowires [37]. The values of κ_L were found, as expected, to decrease as the size of the nanostructure decreases, although the amounts depend on model parameters and assumptions. In particular, both the applicability of the model assumptions and predicted values become uncertain when the structures approach true nano-scale values. The specularity of the interfaces, for example, was found to influence κ_L by a factor of 2-4 for core-shell nanowires 10-500 nm in diameter and superlattices 10-200 nm in periodicity, but by two orders of magnitude under some model assumptions for superlattices with a period of 1-5 nm.

In the phenomenological approach [20], the thermal conductivity is decomposed into contributions by phonons with different group velocities (v_i) and lifetimes (τ_i), i.e. $\kappa_L = \sum_i c v_i^2 \tau_i$, where c is the heat capacity; τ is further decomposed into contributions

from different scattering processes via Matthiessen's rule:

$$\tau^{-1} = \tau^{-1}_{\text{impurity}} + \tau^{-1}_{\text{boundary}} + \tau^{-1}_{\text{umklapp}} \tag{3.1}$$

With group velocities $v_g = \partial \omega / \partial k$ obtained from bulk dispersion or continuum elastic wave theory, the predicted values of κ_L are in rough agreement with experimental measurements for Si nanowires larger than 40 nm in diameter [39]. Alternatively, one could derive v_g from dispersion curves calculated atomistically using potential models, and using bulk parameters for umklapp and impurity scattering. Doing so allows experimental values of κ_L and their temperature dependence to be reproduced for $37 - 115$ nm Si nanowires [40]. The phenomenological approach breaks down for nanowires with diameters less than 20 nm, when scattering rates become significantly altered from bulk behavior by inter-subband scattering [40].

In the linear response (Kubo-Green) approach, κ_L is obtained from equilibrium MD simulations via the fluctuation-dissipation theorem. Using the Stillinger-Weber potential, the values κ_L of 2-5 nm-wide Si nanowires were found to be in the 1-5 W/m-K range from 200 to 500 K, and to decrease with cross-sectional area as a power law [41]. For a Si/Ge superlattice with 2-3 nm periodicity, κ_L was found to range from 2 to 10 W/m-K at 200-300K.

3.1.4 Goals of this work

As evident from the previous section (3.1.3), many of the computational approaches used thus far are unable to tackle SiGe nanostructures approaching the 1-10 nm length scale (continuum and phenomenological approaches), require extreme conditions in the simulations (temperature gradient of 10^8 K/m in the direct approach) and/or suffer from diminished predictive abilities due to multiple fitting parameters and model assumptions (continuum approaches). The Kubo-Green approach using equilibrium molecular dynamics is a predictive method that does not require phenomenological

parameters such as phonon scattering rates, specularity factors, or group velocities [1]. Its atomistic nature allows the incorporation of nanoscale effects on phonon spectrum, dispersion, and interactions, as well as interfacial and alloying effects. The Kubo-Green approach has not been used previously to predict κ_L for many types of SiGe nanostructures such as core-shell nanowires and nanoparticle inclusions. Using the predictive power and parameter/model-free characteristics of the Kubo-Green approach, our first objective is to investigate size and interfacial effects on κ_L in the 1-10 nm scale for various nanostructure types, and to minimize κ_L accordingly. Corroborations with known experimental results or experimentally-verified computational predictions where applicable will be made to ascertain the accuracy of the approach. In addition, the configurational dependence of κ_L has not been previously studied, and with MD as a predictive tool we will be able to develop a new methodology to carry out configurational optimization of κ_L.

3.2 Prediction of κ_L in SiGe nanostructures – size and interfacial effects

3.2.1 κ_L of SiGe nanostructures from molecular dynamics

We compute the thermal conductivity for various SiGe nanostructure types in section 3-1, using the Kubo-Green approach with classical molecular dynamics (MD). The SiGe nanostructures used in the simulations are chosen to reflect varieties in size (at least two sizes are chosen for each type), dimensionality (e.g. superlattice and superlattice-wire), types of interfaces (e.g., coherent and incoherent nanoparticle inclusions), direction of the interface relative to that of the heat flow (e.g., core-shell nanowire, spherical inclusion and cubical superlattice), the interfacial-area-to-volume ratio (as suggested in [43] as a key quantity), and the proportion of Si to Ge. For each component of a nanostructure, e.g. a segment of a superlattice wire, pure Si or

---

[1]A more detailed review and evaluation of the Kubo-Green approach to κ_L prediction in bulk systems is given in [42].

Ge is used in order to avoid confusing these effects with those of alloying, which will be addressed separately in section 3.3.3.

The MD simulations are performed using the DLPOLY [44] code, modified to output the heat flux as described in [45]. Interactions are described by a bond order Tersoff potential [46] designed for C-Si-Ge systems that varies chemical bond strength according to the local coordination environment, given by

$$V_{ij} = f_c \left(r_{ij} \right) \left[f_r(r_{ij}) + b_{ij} f_a(r_{ij}) \right].$$ (3.2)

Here f_r and f_a are repulsive and attractive pair potential terms, and the bond order term b_{ij} is a function of the number of neighboring atoms as well as the bond angles and lengths. The thermal conductivity for each wire was calculated using the Kubo-Green formula

$$\kappa_{\alpha\beta} = \frac{V}{k_B T^2} \lim_{t \to \infty} \int_0^t \langle J_\alpha \left(t' \right) J_\beta \left(0 \right) \rangle dt'$$ (3.3)

where J is the heat current. The surfaces of nanowires are free, with surface reconstruction done either by hand (superlattice nanowires) or by annealing at 1000K (core-shell nanowires).

As a test of the suitability and accuracy of the Tersoff potential and of the Kubo-Green approach, we calculate the κ_L of bulk isotopically-pure silicon from a cubic simulation cell containing 1728 atoms. The simulation cell is 3.28 nm on each side, and periodic boundary conditions are applied in all directions. The MD simulations are carried out at 300K and 1000K, with time steps of 0.8 fs, for 16 ns. At 300K, the value of κ_L for naturally-occurring Si (92% ^{28}Si, 5% ^{29}Si, and 3% ^{30}Si) is reported to be 148 [31] to 156 [47] W/m-K. A range of values of κ_L has been measured for isotopically-enriched silicon at 300K: from 10 ± 2% above naturally-occurring Si (99.983% ^{28}Si) [48] to 237(8) W/m-K (99.8588% ^{28}Si) [49] and 250 W/m-K (99.7% ^{28}Si) [50]. From our MD simulations we obtain $\kappa_{L,300K} = 230 \pm 50$ W/m-K at 300K, which is within the range of measured values for isotopically-pure Si. Because of the uncertainty in the measurements at 300K, and also in order to avoid any possible issues with classical statistics , we use instead the values at 1000K for comparison.

38

At 1000K, umklapp scattering is the dominant contribution to κ_L and the difference between isotopically-pure and naturally-occurring silicon is expected to be of order 1% [47, 40]. We compute the value of κ_L to be 35 ± 4 W/m-K at 1000K, in reasonable agreement with the measured value of 31 W/m-K for natural silicon [31]. In addition, we note that the absolute and relative uncertainties are smaller for the calculated value of $\kappa_{L,1000K}$ than that for $\kappa_{L,300K}$. This is because a larger value of κ_L implies lower degrees of anharmonicity and phonon-phonon scattering rates, as well as longer relaxation times, so that the values of κ_L calculated from MD are more dependent on the initial conditions. Far longer simulation times than computationally feasible are required to significantly reduce the uncertainties in $\kappa_{L,300K}$.

Having established the validity of the Kubo-Green approach and Tersoff potential, we construct three types of SiGe nanostructures for the κ_L computation: cubic superlattices, nanoparticle inclusions, and nanowires. The cubic superlattices are formed with tiling cubes of Si and Ge alternately, with each cube consisting of 3^3, 6^3, or 12^3 atoms, giving simulation cells of 1728 to 13824 atoms. The nanoparticle inclusions are constructed out of pure-Si supercells with 1728 or 13824 atoms, with spherical inclusions of radius 5 to 20 Å as Si embedded in a Ge matrix or vice versa. The nanoparticles are either coherently embedded, i.e. maintaining the same crystal orientation as the matrix, or rotated with respect to the matrix, in which case the boundary is amorphized as described below. There are two main types of nanowires: core-shell nanowires are formed by cutting out concentric cylinders from a bulk 1728-atom Si cell and assigning atoms in the inner/outer cylinders as Ge/Si; superlattice nanowires are constructed using Si nanowire structures found by *ab initio* calculations [51] to be stable, and assigning alternative segments of 2-4 nm long as Si/Ge.

Prior to obtaining κ_L of the constructed nanostructures from MD, it is necessary to perform equilibration in order to relax the strain due to the difference in lattice constants of Si and Ge, and to reconstruct surfaces or interfaces. For cubic superlattices and coherent nanoparticle inclusions, equilibration is carried out by NPT (constant pressure and temperature) at 1000K for 400 ps in order to obtain the appropriate cell

size. For rotated nanoparticle inclusion, formed by rotating a spherical nanoinclusion of Si within a Si/Ge matrix, the system is annealed with NPT runs in steps between 1600K and 1000K to ensure strain relaxation as well as amorphization of the interface between the nanoinclusion and the matrix. For the smaller rotated nanoparticle inclusions (<15 Å radius), the atoms in the inclusion are held fixed during the annealing process to prevent realignment with the matrix. Finally, for core-shell and superlattice nanowires, the lengths of the simulation cells are changed manually to several values prior to NVT (constant volume and temperature) equilibration runs. The length of the cell that gives rise to the lowest average pressure is chosen to minimize the residual strain during the subsequent NVE simulation. This procedure is necessary because the barostat in DLPOLY has not been implemented to be applied along one direction only. The surface of core-shell nanowires are reconstructed using annealing at 1000K.

With equilibrated nanostructures we perform NVE (constant volume and energy) simulations to obtain the heatflux and hence κ_L according to equation (3.3). The NVE runs are carried out at 300K (nanowires) or 800-1000K (nanoparticle inclusions), in time steps of 0.8 fs, for a total of 8-16 ns. For each structure, two to five NVE simulations are done; within each run, the heat flux autocorrelation function is integrated for 40-120 ps. Different runs and, in some cases, different integration times give varying values of κ_L which are then averaged and the standard deviation reported as $\delta\kappa_L$. The results are shown in table 3.1.

3.2.2 Nanostructure size effects on κ_L

From a phenomenological point of view, nanostructuring decreases the thermal conductivity by increasing the interfacial area and hence boundary scattering rate of phonons. In the bulk, the boundary scattering relaxation time $\tau_{boundary} \sim \ell/v$, where ℓ is the system size and v is the phonon velocity. For nanostructures, we can identify ℓ with the boundary or interfacial separation. Following [33], we use the ratio of volume to interfacial area, which has a unit of length, as our boundary length scale ℓ. Since $\kappa_L = \sum cv^2\tau$, if boundary scattering is the dominant scattering mechanism one

Label	Type	Tempera-ture (K)	Cross-sectional area (nm²)	Simulation Cell Length (nm)	Volume (nm³)	Interfacial area (nm²)	Interfacial Area/Volume (1/nm)	% Ge	Kappa (W/m-K)	Uncertainty (W/m-K)
1a		1000	44.9	6.70	300.8	269.3	0.90	50	2.6	0.4
1b	Cubic superlattice	1000	11.2	3.35	37.6	67.3	1.79	50	1.4	0.1
1c		1000	11.2	3.35	37.6	134.7	3.58	50	1.8	0.1
2a		1000	10.8	3.28	35.3	2.1	0.06	1	26.4	1.5
2b		1000	10.8	3.28	35.3	3.3	0.09	2	23.1	2.2
2c		1000	10.8	3.28	35.4	4.8	0.14	3	15.2	1.4
2d	Nanoparticle inclusion (Ge in Si, coherent interface)	1000	10.8	3.29	35.6	8.2	0.23	6	10.2	0.5
2e		1000	10.9	3.30	35.9	13.1	0.36	11	7.0	0.3
2f		1000	10.9	3.31	36.2	19.4	0.54	21	4.5	0.3
2g		1000	11.2	3.34	37.3	29.9	0.80	40	2.7	0.2
3a		1000	11.7	3.42	40.0	3.3	0.08	98	12.2	0.5
3b		1000	11.7	3.42	40.0	4.8	0.12	97	10.6	0.9
3c	Nanoparticle inclusion (Si in Ge, coherent interface)	1000	11.6	3.41	39.7	8.5	0.21	94	7.8	0.9
3d		1000	11.6	3.41	39.7	13.3	0.34	89	6.0	0.7
3e		1000	11.5	3.39	39.1	18.6	0.48	79	4.3	0.3
3f		1000	11.3	3.36	37.9	29.0	0.76	60	3.1	0.2
4a	Rotated nanoparticle inclusion (Si in Ge, amorphous interface)	1000	11.5	3.40	39.1	18.6	0.48	79	1.4	0.2
4b		1000	11.6	3.40	39.4	18.6	0.47	79	1.9	0.2
5a	Rotated nanoparticle inclusion (Si in Si, amorphous interface)	300	43.0	6.56	282.3	50.3	0.18	0	18.2	0.7
5b		300	43.0	6.56	282.3	50.3	0.18	0	20.0	0.9
6a	Core-shell nanowire	300	12.57	6.56	82.4	12.4	0.15	25	13.0	1.0
6b		300	7.07	6.55	46.3	10.3	0.22	44	7.8	0.6
7a	Superlattice nanowire	300	2.01	3.90	7.8	6.0	0.76	50	1.1	0.2
7b		300	2.01	7.80	15.7	7.9	0.51	50	1.9	0.2

Table 3.1: Results of MD calculations on various SiGe nanostructures.
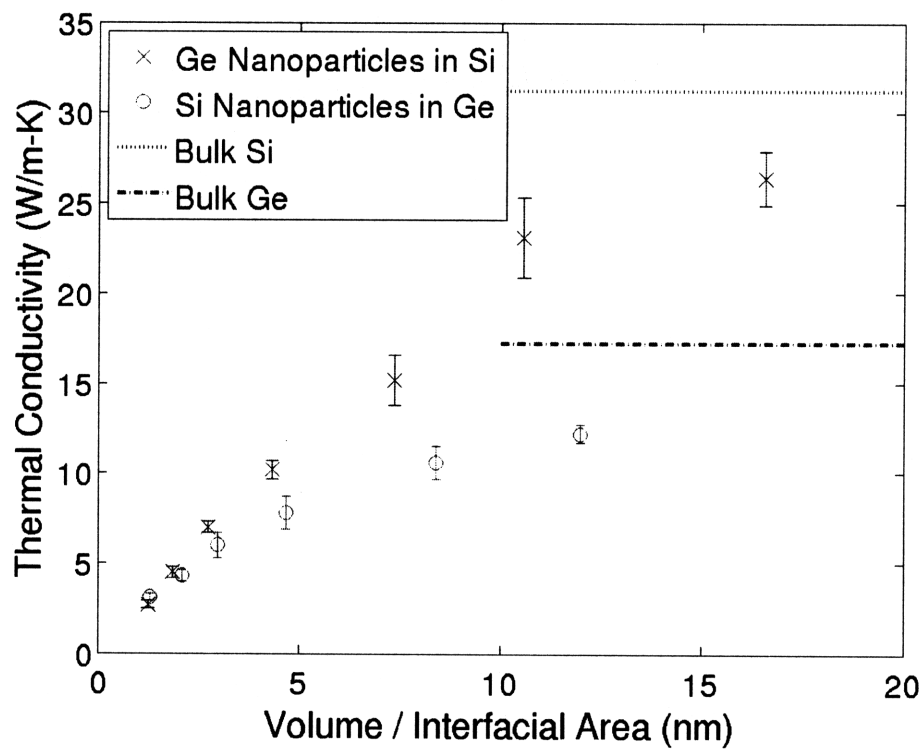
Figure 3-2: Dependence of κ_L on volume-interfacial area ratio for Ge/Si nanoparticle inclusions in a Si/Ge matrix.

would expect κ_L to scale linearly with ℓ. Figure 3-2 shows the computed values of κ_L for coherent nanoparticle inclusions as a function of volume-to-interfacial area ratio ℓ. The (blue) crosses denote Ge nanoparticles of 4 to 15 Å radii embedded in a Si matrix, and the (red) circles denote Si nanoparticles of 5 to 15 Å radii in a Ge matrix. We can see that at small values of ℓ, up to ~ 3 nm, this linear dependence of κ_L on ℓ is indeed observed. At larger values of ℓ, deviation from linear behavior indicates significant contributions from umklapp scattering. As expected, κ_L approaches bulk Si/Ge values in the limit of infinite ℓ.

The linear relation between the thermal conductivity and ℓ cannot be justified, however, when ℓ reaches the scale of interatomic distances. In fact, from the direct method using non-equilibrium MD [35] on perfect Si nanowires, and from the Kubo-Green method on Si nanowires with roughened surface [52], the value of κ_L has been found to increase with decreasing radius once the radius decreases below 2-3 nm. From the cubic superlattices (rows 1a-c of table 3.1), we see that κ_L increases from 1.4 ± 0.1 to 1.8 ± 0.1 W/m-K when the periodicity is decreased from 1.7 to 0.8 nm. This means that there is a size for each nanostructure for which κ_L is a minimum.

3.2.3 Interfacial and configurational effects on κ_L

In addition to the interfacial area to volume ratio, the characteristics of the interface also affects the thermal conductivity of the nanostructures. In particular, an amorphous, disordered interface is expected to be able to reduce κ_L more effectively than a coherent, epitaxial one. This difference is usually characterized by an *a priori* unknown specularity parameter in phenomenological models. From MD calculations, however, it is possible to predictively quantify the effect of such interfacial characteristics by performing the simulation including a realistically-amorphized interface. In table 3.1, rows 3e, 4a and 4b show the results at 1000K for 2.4-nm-diameter Si nanoparticle inclusions in a Ge matrix with both types of interfaces, with $\kappa_L = 4.3 \pm 0.3$, 1.4 ± 0.2, and 1.9 ± 0.2 W/m-K respectively. The sample 3e includes a coherent interface, whereas in 4a and 4b the inclusions are rotated by different angles with respect to the crystalline matrix, and the systems annealed in steps from 1600K

43

to 1000K to amorphize the interfaces. The Ge concentration is 79% in all three cases. We see that, all else being equal, an amorphous interface lowers κ_L by a factor of 2-3 compared to a coherent one. Calculations are also performed at 300K on Si-in-Si rotated inclusions with amorphous interfaces (rows 5a and 5b in table 3.1), where we found κ_L reduced by an order of magnitude compared to bulk values. Both results validate the importance of interfacial roughness as a mechanism to lower κ_L.

To investigate the effects of the shape of the interface, we compare the values of κ_L of cubic superlattices to those of spherical nanoparticle inclusions. Rows 1a, 2g and 3f in table 3.1 show cubic superlattice, Ge-in-Si and Si-in-Ge nanoparticles respectively, with similar values of ℓ (1.12, 1.25 and 1.31 nm), and similar Ge concentrations (50%, 40%, 60%). The values of κ_L are 2.6 ± 0.4, 2.7 ± 0.2, and 3.1 ± 0.2 W/m-K respectively. Using linear scaling with ℓ, there is no significant difference between cubic and spherical interfaces. This is consistent with previous reports [33] based on Monte Carlo simulations using bulk parameters which found no shape- and orientation-dependence for nanoparticle inclusions.

Having established the effects of volume-to-interfacial-area ratio and roughness of the interface on κ_L, we investiage the possibility of additional degrees of freedom with which to further lower κ_L. We find that with rotated nanoparticle inclusions (rows 4a,b and 5a,b in table 3.1), the values of κ_L are dependent on the angle of rotation, i.e. relative crystal orientation of the nanoparticle and the matrix. For 4-nm-diameter Si-in-Si nanoparticles, the relative difference between κ_L of two arbitrarily chosen orientations (18.2 ± 0.7 vs 20.0 ± 0.9 W/m-K) is roughly 10%; for 2.4-nm-diameter Si-in-Ge nanoparticles, the relative difference (1.4 ± 0.2 vs 1.9 ± 0.2 W/m-K) between the same two orientations is about 30%. In both cases the absolute differences are significant. This means that even for nanostructures with large roughened interfaces, for which κ_L is already greatly reduced from the bulk values, there are still configurational degrees of freedom involved, e.g., in the orientation of the nanoparticles, that can be optimized to further lower κ_L. In the next section we investigate an ultimate form of this configurational optimization by treating the Si and Ge atomic configurations in alloy nanowires.

3.3 Optimization of κ_L in SiGe nanowires – local ordering effects

In bulk materials it is well-known that the minimum thermal conductivity is achieved by complete disordering, i.e. in an amorphous environment [53]. Whether the same is true in nanostructures has not been established. In this section we investigate the effects of local ordering, or the lack thereof, on κ_L. We ask whether there are arrangements of Si and Ge atoms in otherwise-identical SiGe nanowires that give the lowest possible κ_L, and if so, we seek to find such arrangements, i.e. an ultimate form of nanostructuring. Since phenomenological parameters, such as umklapp scattering times or the specularity of nanowire surface, do not contain configurational dependence. Therefore, it is essential to model the dynamics at the atomistic level, via MD simulations.

3.3.1 Extending the cluster expansion technique to κ_L

Cluster expansion [14], whereby properties are expanded in terms of the distribution of atoms on a topology of sites, is a powerful technique for the optimizing or ensemble averaging of properties. Its most common use is in the parameterization of the total energy and derivatives thereof, although there have been extensions to other properties, such as band gaps [54] and, more recently, tensorial quantities [55]. The cluster expansion of the total energy has been used extensively, often with *ab initio* calculations, to build effective Hamiltonians for the prediction of thermodynamic [56] and kinetic properties [57]. Unlike atomic potential models, which are rapidly evaluated but require extensive chemistry-dependent parameterization and are not universally available, or *ab initio* calculations, which are (almost) universally available but are computationally-intensive, the cluster expansion approach is a widely-applicable parameterization method which allows rapid evaluations for a large number of different atomic configurations. In addition, the fitting parameters of the expansion, called effective cluster interactions (ECIs), often give important physical insights into the

system.

It is desirable to investigate the applicability of the cluster expansion approach to properties of crystalline materials not based on total energy, such as thermal conductivity. Since the evaluation of thermal conductivity by even classical molecular dynamics is computationally intensive, the ability to parameterize the results for subsequent rapid evaluations would enable the otherwise-prohibitive sampling of atomic configurations for optimization purposes. While the cluster expansion is formally exact with the inclusion of all possible clusters, for practicality the expansion is necessarily truncated to a finite number of clusters. Since the inclusion of each cluster requires the inclusion of all subclusters for completeness [58, 59], such truncation generally leads to an inclusion of only short-ranged effects. Unlike atomic potentials, which are largely local, it is not *a priori* clear whether the thermal conductivity can be parameterized by local configurational variables alone.

The purpose of this part of our work is hence two fold: to optimize atomic configurations in SiGe nanowires for low lattice thermal conductivity κ_L, and to evaluate the applicability of the cluster expansion technique to the treatment of κ_L. Our strategy is to determine the values of κ_L for a test set of atomic configurations using equilibrium molecular dynamics (MD) simulations, use the cluster expansion technique to allow rapid evaluation of κ_L for any configuration, and predict structures with the lowest κ_L through genetic algorithm optimization. The validity of the technique is finally checked with the direct evaluation of κ_L for the predicted optimal structures using MD.

3.3.2 κ_L of SiGe nanowires from molecular dynamics

Computational procedures

As in section 3.2, we obtain from equilibrium classical MD and the Kubo-Green formalism the thermal conductivity of $Si_{1-x}Ge_x$ nanowires. The nanowires in our study have circular cross sections, [111] orientation and a diameter of 1.6 nm, but varying Ge concentrations ($0.03 < x < 0.2$) and Si/Ge configurations. The simulation cells
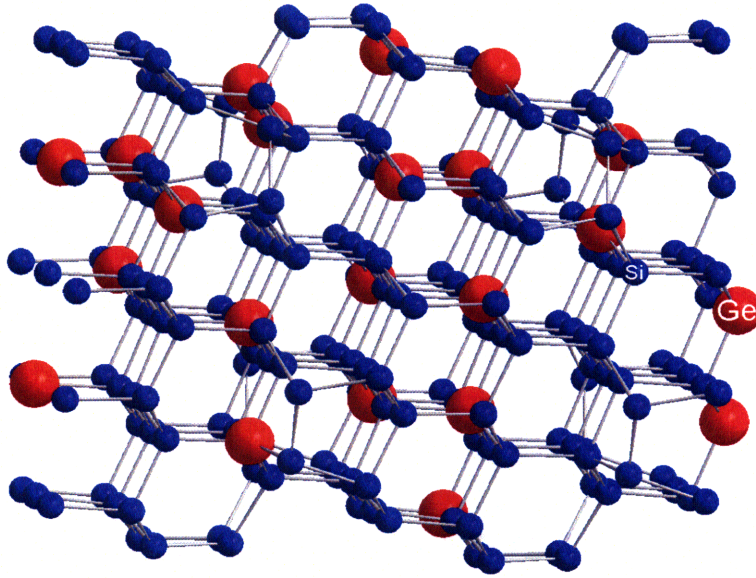
Figure 3-3: An example SiGe nanowire simulation cell (side view). Red, larger spheres denote Ge while blue, smaller spheres denote Si.

are 2 nm long along the axial direction with periodic boundary conditions applied. Simulations are carried out at 300K. An example of a simulation cell is shown in Figure 3-3.

For this part of our work, the XMD Molecular Dynamics Program developed by Jon Rifkin [60], modified to output the heat flux for systems incorporating C, Si, Ge and H, is used to perform the MD simulations. Upon initial NVT (constant-temperature) equilibration, NVE (constant-energy) simulations are performed with a time step of 0.8 fs, for a total simulation time of 1-10 ns. The thermal conductivity for each wire is calculated from the heatflux using the Kubo-Green formula (3.3). The surfaces of the nanowires are free. Previously reported [61] errors arising from unpassivated surfaces are corrected in the manner described in [62]. More details on the MD simulation procedures are available from Reference [62].

Summary of MD results

The calculated thermal conductivities of the 104 SiGe nanowires in the training set are shown in Figure 3-4. The training set configurations are selected to cover a range
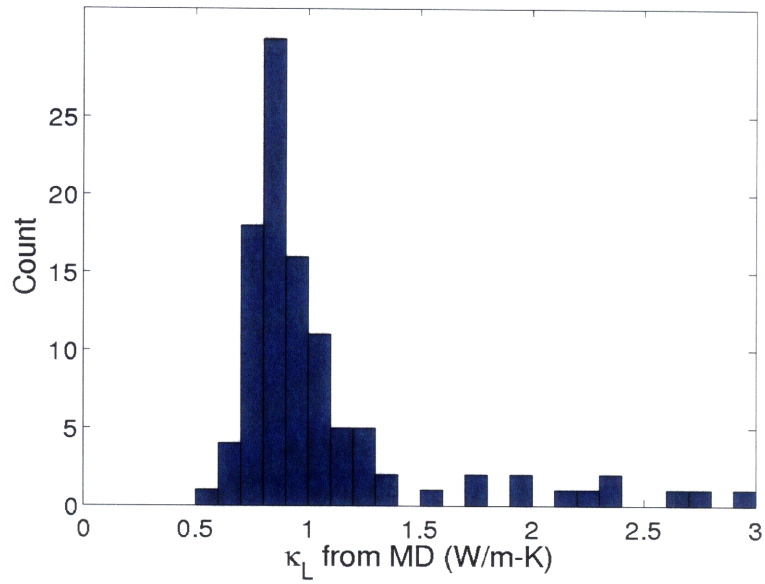
Figure 3-4: Histogram of lattice thermal conductivity calculated using MD for the training set of 104 nanowires with different Si/Ge configurations.
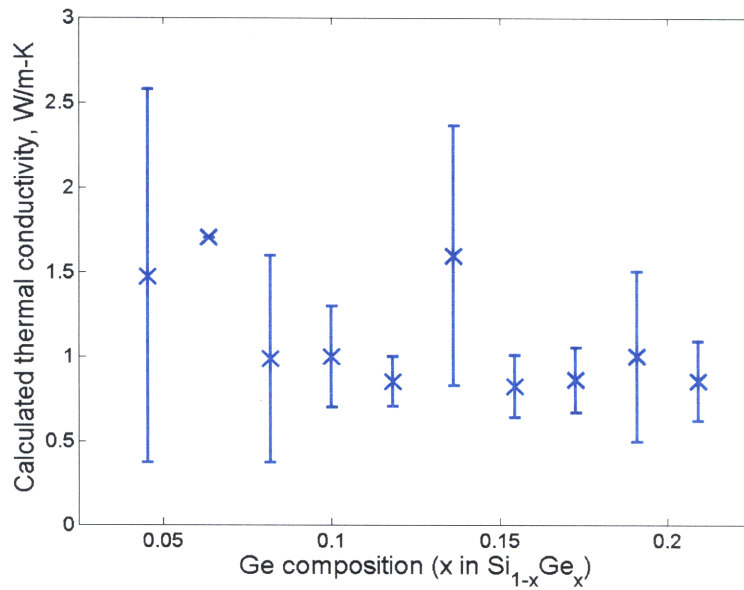
Figure 3-5: Thermal conductivity of nanowires with different Si/Ge configurations vs Ge composition.

of compositions up to 22% Ge and provide a large variety in the distribution of Si and Ge atoms. The calculated values of κ_L range from 0.55 to 3.0 W/m-K, with a mean and standard deviation of 1.1 and 0.5 W/m-K respectively. For comparison, the measured values of κ_L at 300K are 9 to 25 W/m-K for bulk $Si_{1-x}Ge_x$ alloys (0.05 < x < 0.22) [63], 15 to 40 W/m-K for 37-115 nm diameter Si nanowire with smooth surfaces [30], and 1-8 W/m-K for 50-150 nm diameter Si nanowires with roughened surfaces [34]. Figure 3-5 shows the mean and standard deviation of κ_L in each Ge compositional range, and shows that for a given Ge concentration, the arrangements of Si and Ge atoms (configurations) affect the values of κ_L significantly. The dependence of κ_L on the concentration of Ge is smaller than the configurational dependent.

3.3.3 Cluster expansion of κ_L

Traditional cluster expansion

A traditional cluster expansion [14, 55] for a binary alloy is an Ising-like model in which each site i in a lattice is assigned a value $\sigma_i = \pm 1$ depending on the occupying species. Polynomials of σ_i of all orders form a complete orthonormal basis set in which to expand any configurational-dependent physical quantity Q, i.e.

$$Q = \sum_\alpha V_\alpha \left(\prod_{i \in \alpha} \sigma_i \right) \tag{3.4}$$

where the sum is over all possible distinct clusters of sites α and the coefficients of the expansion V_α are fitting parameters known as effective cluster interactions (ECIs). In practice, the ECIs are obtained by fitting Equation 3.4 to the calculated (e.g. by MD simulations in the case of κ_L) values of Q for a number of sample configurations. The expansion can then be used to predict values of Q for any configuration. As mentioned, a cluster expansion is necessarily truncated to clusters of reasonably small order. Symmetry is used to reduce the number of ECIs.

Coarse-grained cluster expansion

For both physical and practical reasons, we perform a coarse-graining of the cluster expansion. Physically, it is expected that clusters similar in location, size and orientation would give a similar contribution to the thermal conductivity. Practically, we are limited by the number of relevant parameters that can be extracted. The low symmetry in a nanowire with surface reconstruction leads to a large number of symmetrically-inequivalent clusters and therefore many ECIs (V_α's). Together with the inherent noise in the values of κ_L due to the stochastic nature of MD simulations, such a large number of parameters leads to over-fitting and diminished predictive power of the cluster expansion. Recently, approaches have been developed to deal with such low symmetry situations by imposing non-uniform prior probability densities on the ECIs [64]. We will, however, use a simpler coarse-graining approach as described below.

Coarse-graining procedures are performed to group physically-similar, but symmetrically-inequivalent, clusters (points, pairs, and triplets) in the nanowire. All clusters in a group are considered equivalent in the coarse-grained cluster expansion and their ECIs have the same value. Equivalent clusters have (a) similar distance from the axis of the nanowire and, for pairs and triplets only, (b) similar extents along the length of the nanowire. For classification (a), clusters entirely within the inner 10% of the cross sectional area of the nanowire are considered *core*, those entirely outside the inner 90% are considered *surface*, and the rest are considered *intermediate*. For (b), clusters contained in planes parallel/nearly-parallel to nanowire cross-sections are considered *across*, while those in planes parallel/nearly-parallel to the nanowire axis are considered *along*, and all others are *oblique*. The coarse-graining procedure gives, for the simulation nanowires described in Section 3.3.2, 40 coarse-grained point-, pair- and triplet- clusters, with a maximum spatial extent of 4.5Å. Figure 3-6 shows examples of clusters that are considered equivalent in the coarse graining procedure.

The expansion coefficients (ECIs) in the coarse-grained cluster expansion are obtained by the following procedures. We expand the values of κ_L for the N wires in
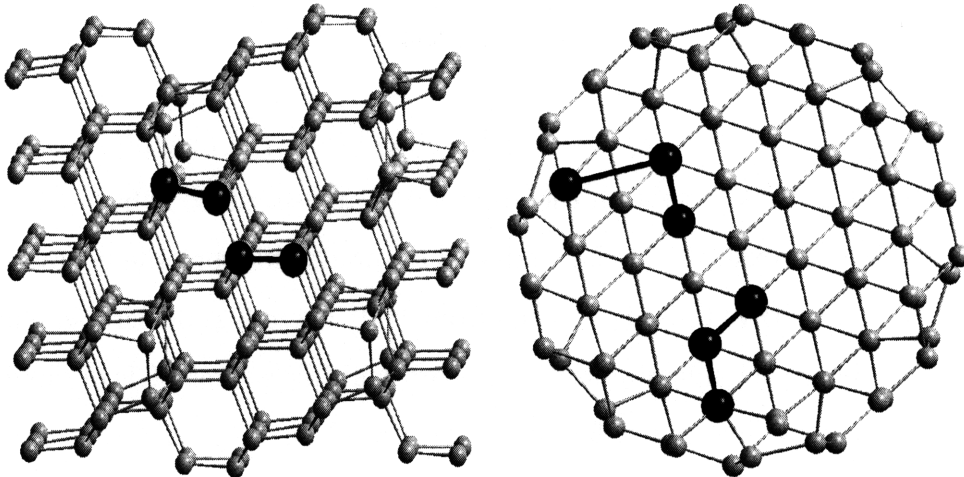
Figure 3-6: Examples of clusters considered equivalent in the coarse-grained cluster expansion: (left, side view) two pair clusters that are both *surface* and *along*, (right, end view) two triplet clusters that are both *intermediate* and *across*.

the training set as

$$\kappa_{Ln} = \sum_{\beta} V_\beta \langle \prod_{i \in \alpha} \sigma_{in} \rangle_\beta \equiv \sum_{\beta} V_\beta \Phi_{\beta n} \tag{3.5}$$

where n = 1,..., N, β labels a coarse-grained cluster, and $\langle . \rangle_\beta$ refers to average over all clusters α coarse-grained into β. The $\Phi_{\beta n}$'s are then orthonormalized, i.e. (3.5) is transformed into

$$\kappa_{Ln} = \sum_{\beta'} \tilde{V}_{\beta'} \tilde{\Phi}_{\beta' n}, \tag{3.6}$$

where $\sum_{\beta'} \tilde{\Phi}_{\beta' m} \tilde{\Phi}_{\beta' n} = \delta_{mn}$. Note that each β' now represents a linear combination of coarse-grained clusters. The values of the trasnformed ECIs $\tilde{V}_{\beta'}$ are obtained from least square fitting of (3.6). Unlike the ECIs in the cluster expansion of the total energy, which are expected to fall off as a function of distance, there is no *a priori* known behavior for the values of $\tilde{V}_{\beta'}$. In order to screen out irrelevant parameters, the least-square fit is repeated leaving one configuration out each time. For each β', the value of $\tilde{V}_{\beta'}$ varies with each leave-one-out fit, with a mean $\bar{V}_{\beta'}$ and a standard deviation $\sigma_{\beta'}$. A linear combination of clusters β' is considered irrelevant and removed if $|\bar{V}_{\beta'}| < \sigma_{\beta'}$. Roughly a quarter of the 40 ECIs (corresponding to the 40 coarse-grained clusters described above) are thus removed. The fit is then redone with the

remaining set of β's and the resultant $\tilde{V}_{\beta'}$s are transformed to obtain the original ECIs V_βs.

Because of the noise inherent in the MD values of κ_L, the magnitude of which is unknown, it is difficult to ascertain the accuracy of any set of ECIs from a single fit. Figure 3-7 shows an example of the values of κ_L predicted from the coarse-grained cluster expansion (3.6) vs those calculated from MD. There is considerable scatter as expected. In light of such uncertainties, we record multiple sets of ECIs obtained from different leave-one-out fits and different fitting procedures (e.g., different thresholds for choosing relevant linear combinations of clusters β', different subset of data by range of calculated κ_L values or Ge concentration). Instead of a single cluster expansion we obtained a group of expansions — a *meta* cluster expansion — with different sets of ECIs $\{V_\beta\}$, all of which are consistent with the given MD results. For any configuration, there is a range of predicted values of κ_L from the different sets of ECIs in the meta cluster expansion, which are used together in the optimization procedure described below.

Optimization by genetic algorithm

We use the thermal conductivity meta cluster expansion together with a genetic algorithm to evolve a trial population into configurations with optimal (lowest) κ_L. As in standard genetic algorithm implementations, pairs of nanowires with different Si/Ge configurations are mated by joining lateral or cross-sectional halves of each parent, and mutations are stochastically introduced to the resultant offspring. The fitness of each configuration, i.e. κ_L where a lower value is considered more fit, is evaluated by the meta cluster expansion obtained above. We require that the predicted values are robust, i.e. the different sets of ECIs in the meta cluster expansion predict values of κ_L within a certain range threshold, or else we consider the prediction unreliable and the configuration is discarded. At each generation, the fittest configurations are kept, and a random sample of the remainder added to ensure genetic diversity. Any exact duplicates are removed, and the genetic algorithm optimization is carried out until convergence of the mean value of κ_L among the wires in the population, which
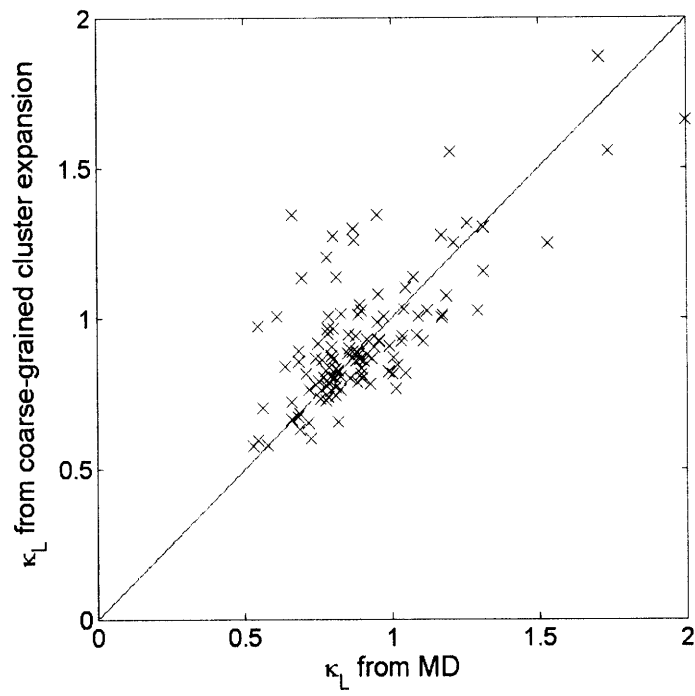
Figure 3-7: Example of the values of κ_L predicted from coarse-grained cluster expansion 3.6 vs those calculated from MD (both in W/m-K).
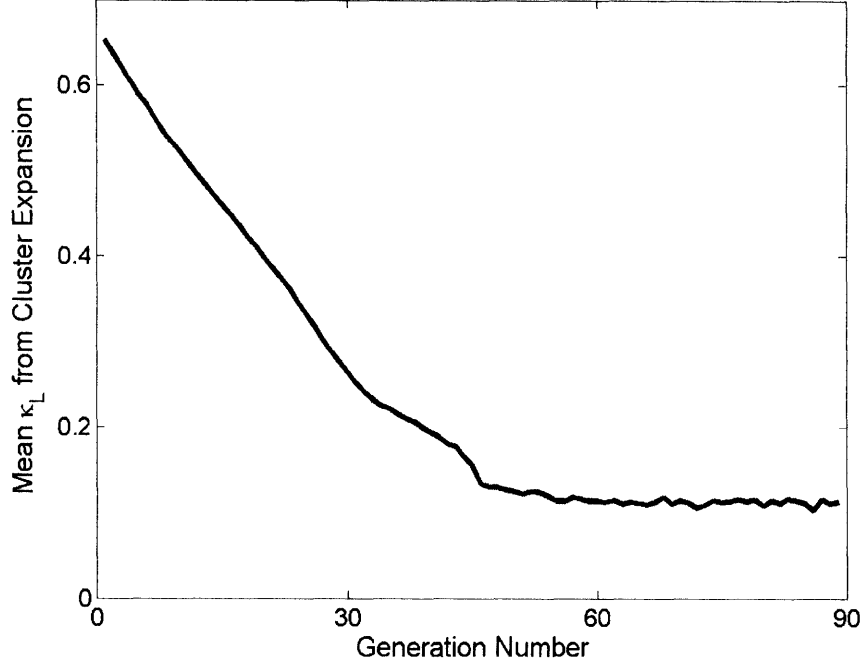
Figure 3-8: Convergence of κ_L (in W/m-K) with generation number in genetic algorithm optimization.

generally occurs within 100 generations. An example of the evolution of the mean value of κ_L with generation number is shown in Figure 3-8.

It is important to note that while the meta cluster expansion approach is able to predict low-κ_L structures, the predicted values of κ_L for these structures are often unphysical, i.e. very close to zero or negative. The error in the predicted values of κ_L can be traced back to errors in the ECIs, which in turn originate from the noise inherent in values of κ_L calculated from MD. Nonetheless, we use the meta cluster expansion and genetic algorithm optimization as a tool in finding the optimal configurations rather than to predict precise values of κ_L. To check the effectiveness of this approach, we use MD to obtain values of κ_L for a selected sample of predicted low-κ_L configurations.
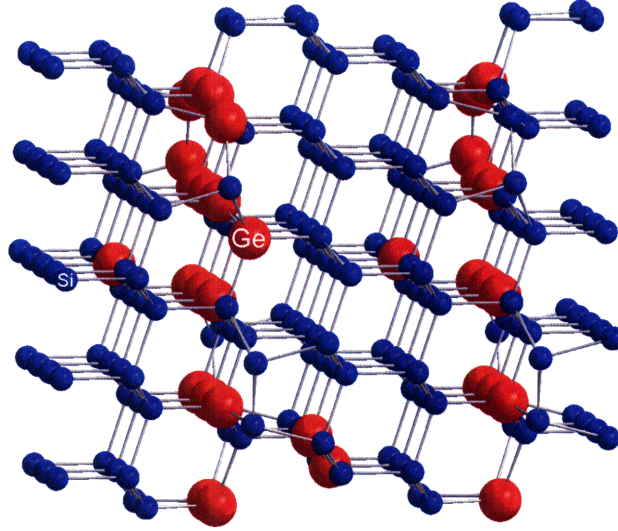
Figure 3-9: Example of configuration with low predicted κ_L.

Predicted low-κ_L configurations

Figure 3-9 shows an example of a configuration predicted to have low κ_L at the end of a genetic algorithm run. While one might expect disordered configurations to have lower κ_L, the meta cluster expansion and optimization algorithm predicts otherwise. Many predicted low-κ_L configurations consist predominantly of Ge clusters of the *across* type, i.e. have almost-complete planes of Ge perpendicular to the direction of the wire, instead of a randomized distribution. Such configurations are reminiscent of the Si/SiGe superlattice structures proposed in Reference [25], albeit one with single-atomic layers of Ge rather than segments.

To check the validity of the cluster expansion, the predicted low κ_L configurations, as well as those with perfect planes of Ge, are investigated using MD simulations. Figure 3-10 shows the κ_L values of these configurations obtained from MD as compared to those of the training set. Out of 28 configurations predicted to have low κ_L by the meta cluster expansion and genetic algorithm optimization, *one* has κ_L lower than all configurations in the training set, 16 (57%) have κ_L at or below the 10th percentile among the training set, and 24 (86%) are at or below the 50th percentile. Therefore, we can see that although the meta cluster expansion did not yield a particular configuration with drastically lower κ_L, it is effective in constructing a population of low-κ_L
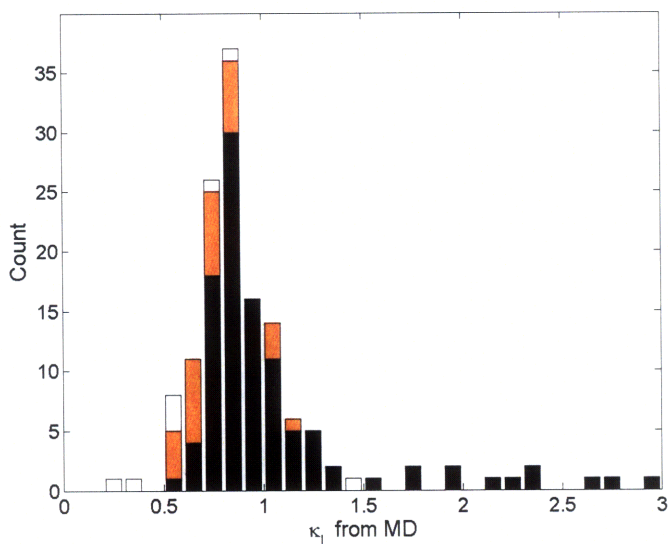
Figure 3-10: Histogram of κ_L (in W/m-K) calculated with MD for (a) predicted low-κ_L configurations from cluster expansion and genetic algorithm optimization (orange/grey), (b) configurations with perfect planes of Ge (white), and (c) the training data (black).

configurations, meaning that the expansion captures some physical factors governing configurational dependence of the thermal conductivity. More remarkably, one of the superlattice-like configurations with complete planes of Ge (see Figure 3-11), derived from idealizing the predicted configuration shown in Figure 3-9, is found to have κ_L = 0.23 ± 0.05 W/m-K, compared to a minimum of 0.55 W/m-K and a mean of 1.1 W/m-K for the training set. That this specific superlattice-like configuration with very low κ_L is not obtained from the genetic algorithm optimization may be due to noise in the ECIs or limitations of the optimization procedures; that it is so similar to a configuration predicted to have low κ_L confirms that cluster expansion is a viable approach to thermal conductivity optimization.

3.3.4 The role of short-range ordering in reducing κ_L

In this section we offer physical arguments for the mechanisms by which κ_L is reduced, and by which the coarse-grained cluster expansion is able to predict low-κ_L configurations. We note the similarity of the values of κ_L that we compute with
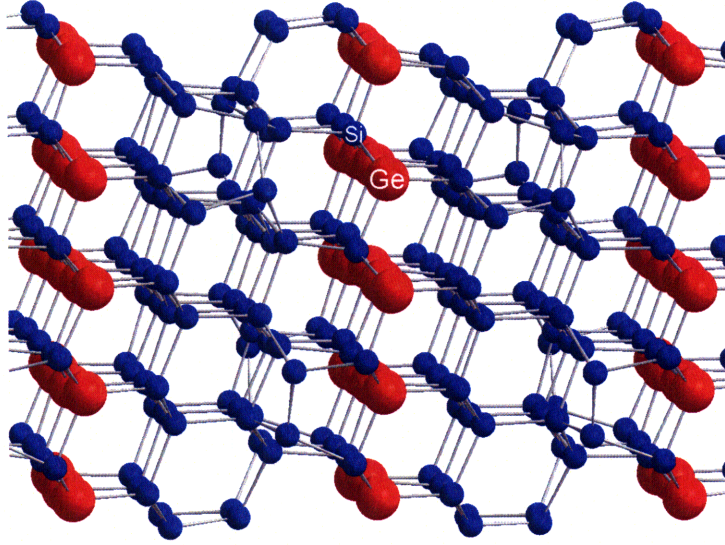
Figure 3-11: Superlattice-like configuration with the lowest value of κ_L as computed by MD (side view).

MD (mean = 1.1 W/m-K) with those measured in 52-nm-diameter Si nanowires with surface roughness on the scale of several nanometers (1.2 ±0.1 W/m-K) [34]. In both cases (our calculations and rough wire measurements), the κ_L values are near or below the bulk amorphous minimum thermal conductivity limits [65], despite unambiguous crystallinity. In both cases, the values of κ_L are an order of magnitude lower than in similar wires with longer-range order, i.e. nanowires with smooth surface and hence longer periodicity in measurements [34], or nanowires with longer simulation cells and hence longer periodicity in MD calculations [45, 52]. In the case of our MD simulations, using 2-nm-long simulation cells excludes thermal transport by longer wavelength phonons which have longer relaxation times; introducing these longer-wavelength phonons by increasing the simulation cell lengths to 12 nm increases κ_L by an order of magnitude, without any difference in surface features. Therefore we postulate that in [34], phonons with wavelengths longer than the length scale of surface roughness features do not exist due to the lack of long-range order (or are strongly scattered by the features) and their contribution to κ_L is greatly reduced, thus dramatically lowering the thermal conductivity. Given that extremely-low κ_L's near the amorphous limits are likely achieved with the exclusion of long-wavelength phonons,

it is reasonable that a coarse-grained cluster expansion approach which includes only local ordering would be able to treat the effects of further variations in κ_L arising from atomic configurations. We substantiate this argument by investigating the detailed properties of the phonon modes and heat flux.

To investigate the origin of the very low values of the thermal conductivity obtained, we compute explicitly the 0K phonon modes of the nanowires used in our MD calculations. The phonon densities of states (DOS) for the pure-Si, pure-Ge, SiGe with predicted-low-κ_L from the cluster expansion, and lowest-known-κ_L nanowires are shown in Figure 3-12. We can see that there is a suppression of the phonon DOS at very low-frequency, due to quantum confinement effects. In fact, the lowest non-zero phonon mode at Γ is at roughly 40 cm^{-1} (1.2 THz). Since low-frequency phonons are typically responsible for thermal transport owing to their low umklapp scattering rates, such suppression of the phonon DOS is partially responsible for the lowering of κ_L. The phonon dispersion curves for the pure-Si and lowest-κ_L nanowires are shown in Figure 3-13. The quadratic dispersion of the lowest two branches is a result of quantum confinement, i.e the acquiring of a mass of the acoustic phonons corresponding to the two finite dimensions perpendicular to the wire axis. The group velocities of the massless modes at Γ are found to be 2083 and 3960 m/s^{-1}, compared to 5100 and 9360 m/s^{-1} in bulk Si [66]. This is another factor that contributes to the lowering of κ_L compared to bulk SiGe alloys.

Within the population of SiGe nanowires, we can establish that there exist local ordering effects not yet accounted for by noting in Figures 3-12 and 3-13 that the phonon densities of states and dispersion curves are very similar for configurations that yield very different κ_L. The group velocities of the lowest massless branches differ by only 11% between pure-Si nanowire ($\kappa_L = 3.9 \pm 0.3$) and the lowest-κ_L ($\kappa_L = 0.23 \pm 0.05$) configuration. Thus the difference must lie within the relaxation times of individual phonon modes. Fourier analysis of the heat flux reveals persistent non-dispersive (zero-velocity) torsional and shear modes at $1 - 4$ THz. Local ordering affects the frequencies and characters of these modes, as well as their coupling to the heat-carrying dilatation modes. Differences in coupling constants affect the relaxation
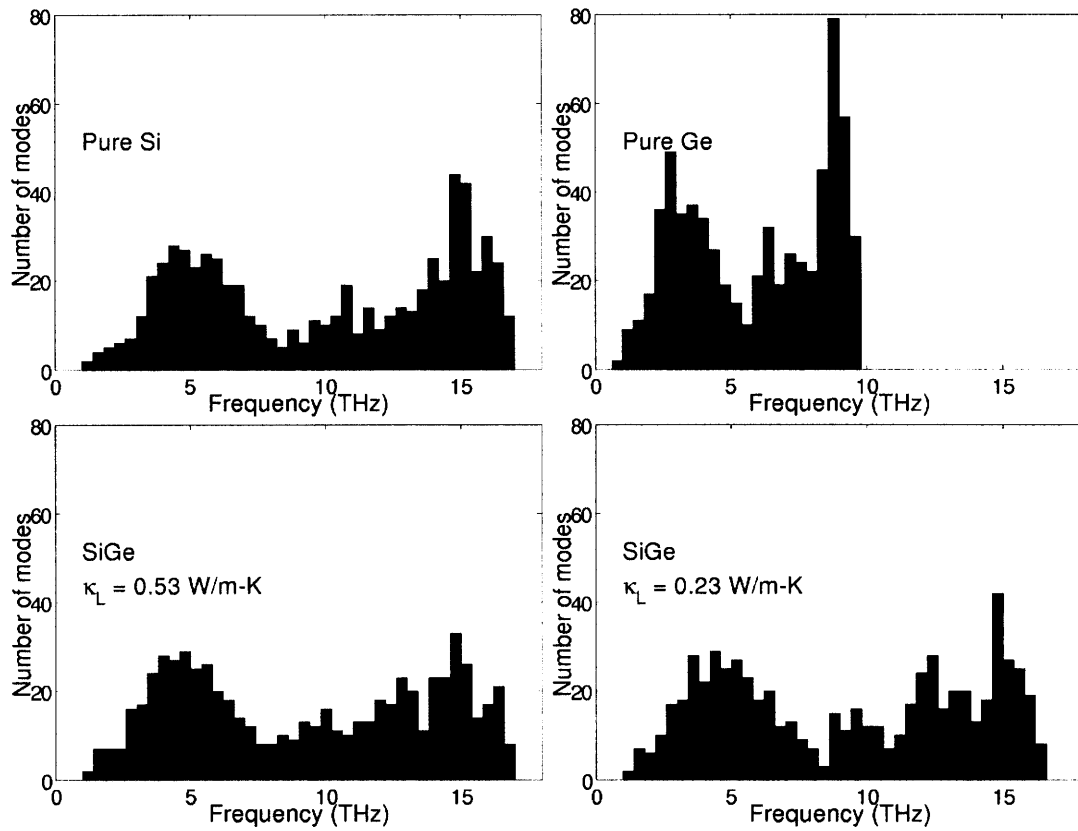
Figure 3-12: Phonon density of states for several nanowires: pure Si (upper left), pure Ge (upper right), SiGe with predicted-low-κ_L from cluster expansion (lower left), and SiGe with lowest-known-κ_L (lower right).
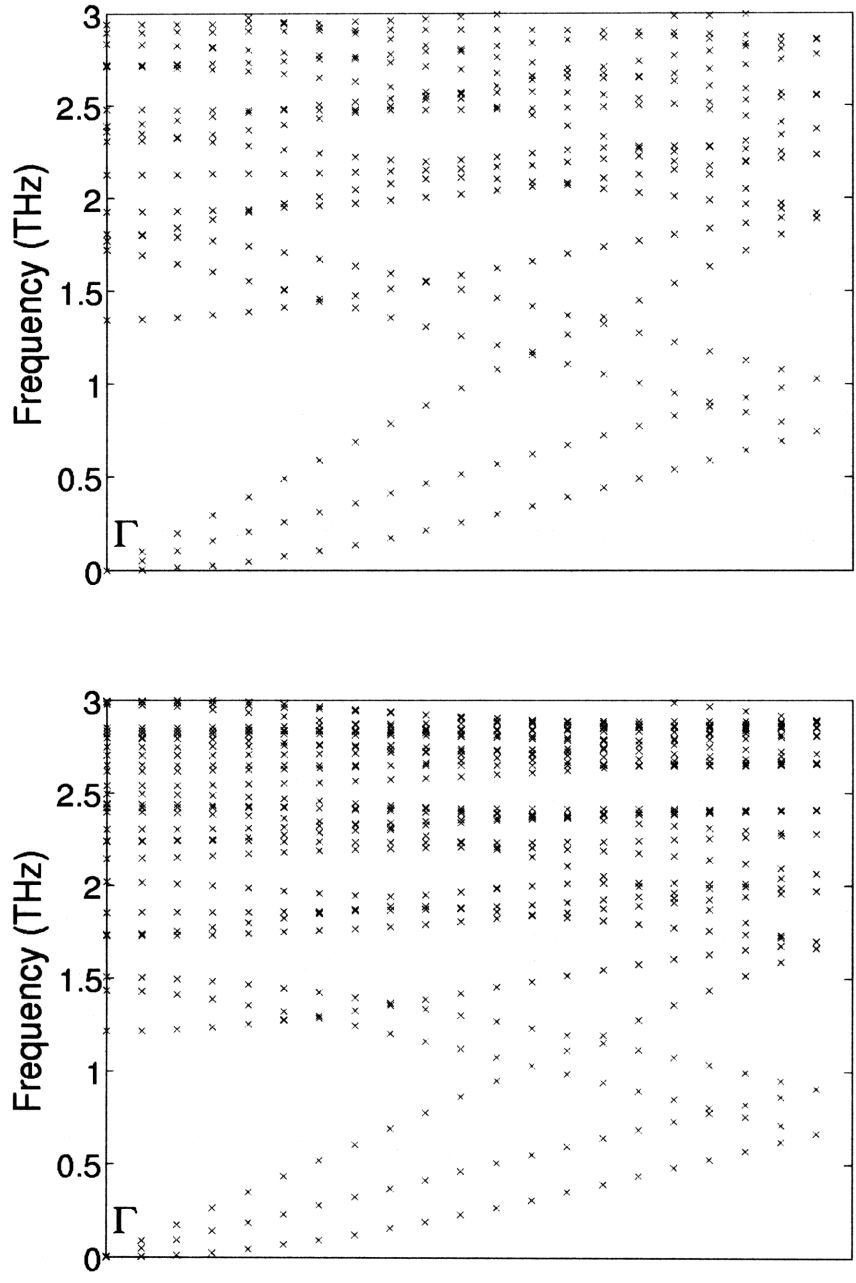
Figure 3-13: Phonon dispersion curve for pure-Si (top) and lowest-known-κ_L (bottom) nanowires.

times of the heat-carrying modes and ultimately κ_L. The precise configurational dependence of the coupling constants is a subject of further research.

3.4 Conclusion

From considering the thermal conductivity of a wide range of SiGe nanostructures, we established that the Kubo-Green approach using molecular dynamics reproduces the expected dependence of the lattice thermal conductivity on typical interfacial/boundary length scales ℓ and interface disorder, but also gives additional quantitative predictions not available from phenomenological models. We found that in the limit of high interfacial area per unit volume, κ_L is proportional to ℓ as expected for boundary scattering being the dominant relaxation mechanism. However, the trend reverses and κ_L increases with smaller ℓ when ℓ approaches a few times the interatomic distance. The presence of an amorphized, roughened interface decreases the thermal conductivity by a factor of 2-3 compared to a coherent, epitaxial interface. By considering the relative crystal orientation of nanoparticle inclusions and the matrix, we saw that additional configurational degrees of freedom can be used to further optimize κ_L.

We carry out the optimization of the thermal conductivity in the space of atomic configurations in SiGe alloy nanowires and demonstrate the feasibility of using the cluster expansion technique to parameterize κ_L. Compared to the total energy E, the computation of the lattice thermal conductivity κ_L is much more expensive, which means an efficient parameterization is highly desirable; yet the computation of κ_L is also fraught with more uncertainties than that of E, which makes such a parameterization difficult. We found that instead of using a traditional cluster expansion, in which each symmetrically inequivalent cluster enters, a coarse-grained approach is effective; instead of using a single set of cluster expansion parameters, a meta cluster expansion approach can be used to take into account variations due to noise in the data as well as the choice of fitting parameters. Using these approaches, we are able to discover populations with generally lower κ_L as well as configurations similar to those which drastically lower thermal conductivity. We find that configurations with

complete planes of Ge atoms have the lowest κ_L compared to other configurations. We postulate that the low κ_L values obtained, close to the bulk amorphous limit, are due to the absence of long-range ordering in the simulation, and that such an absence allows a local cluster expansion approach to be successful in optimizing κ_L.

Chapter 4

Electronic Band Gap Prediction

4.1 Introduction

4.1.1 Band gap prediction for photovoltaic materials design

Photovoltaics form a vital renewable energy source. For the past several decades, development and applications in bulk and thin film inorganic photovoltaics have focused on a few well-known materials such as Si, CdTe, GaAs, and $CuIn(Ga)Se_2$, with crystalline silicon being the predominant material to this date. As with thermoelectric materials, a myriad of properties — opto-electronic, transport, structural and defect-related, to name a few — need to be optimized for a photovoltaic material. Of those properties, the band gap is a key characteristic because it determines the maximum efficiency, as explained in Section 2.2.2. If we are able to make quantitative predictions of the band gap and other pertinent properties, it may be possible to discover or design an optimal photovoltaic material. Fortunately, nature provides us with a wide search space, as there are over 40,000 inorganic compounds in the Inorganic Crystal Structure Database (ICSD) [10] alone. With the multitude of chemistries and structures available, computational tools that accurately account for atomistic and electronic details are necessary; with the vast number of compounds, efficiency is paramount. The abundance and costs of elements also serve as important constraints determining feasibility of wide-spread application of any compound.

Since computational efforts often scale exponentially with the precision level, it is important to define the desired level of accuracy. We argue that in order to identify promising materials with band gaps in the range of interest for both single-junction and tandem photovoltaic cells, i.e. $1.0-2.5$ eV, an accuracy of $0.2-0.3$ eV is sufficient. The reasons are as follow:

- The experimental measurements of optical band gaps consist of extrapolation of intensities from photoluminescence or transmission measurements. The uncertainty in measured values of band gaps originating from such extrapolation is typically in the range $0.05 - 0.2$ eV.

- For typical semiconductors, the difference between band gaps measured at 300K and $\sim$0K is of order 0.1 eV [2].

- Variations in carrier concentrations can change the band gap by an amount of order 0.1 eV via the Moss-Burstein effect [67, 68].

- Exciton binding energies in typical semiconductors are of order 0.1 eV.

- The change in maximum photovoltaic efficiency near the optimal band gap value is about one percentage point per 0.1 eV [22].

The importance of computational efficiency cannot be overstated. Apart from the tens of thousands of candidate compounds for which properties need to be evaluated, additional degrees of freedom in the optimization of photovoltaic performance may come in the form of surface- and nano-structuring. For example, TiO_2 and other oxide nanoparticles have been used in dye-sensitized solar cells [69], and semiconductor nanowires have been investigated for photovoltaic applications [70, 71]. Typical computations involving surface and nanostructures are several orders of magnitude more intensive than bulk properties, and ground state *ab initio* calculations on nanostructures have become widely-feasible only within the past several years.

4.1.2 Band gap errors in Hartree-Fock and density functional theory

The band gap is also a basic property of a solid whose efficient and (even semi-quantitatively) accurate *ab initio* prediction has proven elusive. In this section we aim to develop a physical understanding of the reasons for such difficulties. The fundamental gap, the energy required to create a non-interacting electron-hole pair, is defined by

$$E_{\text{gap,fundamental}} = E(N+1) + E(N-1) - 2E(N) \tag{4.1}$$

where $E(m)$ represents the total energy of a system of m electrons. The physical interpretation of equation (4.1) is apparent: the fundamental gap is the energy required to excite an electron, $E(N+1) - E(N)$, plus that to excite a hole $E(N-1) - E(N)$, from the ground state with energy $E(N)$. The band gap measured from optical absorption is usually equal to the fundamental gap minus the electron-hole (exciton) binding energy. Exceptions occur when the optical transitions between states with the closest energies in the valence and conduction bands are forbidden.

In Hartree-Fock (HF) and density functional theory (DFT) in the Kohn-Sham implementation, the fundamental gap is associated with the difference in eigenvalues. Using Koopman's theorem, $E_{gap} = \epsilon_{N+1}(N) - \epsilon_N(N)$, where $\epsilon_i(M)$ is the ith lowest eigenvalue of eigenfunctions evaluated in a system of M electrons. In DFT, the difference in energy between the lowest unoccupied and the highest occupied Kohn-Sham eigenstates is called the Kohn-Sham gap. In molecular systems, the term HOMO-LUMO gap (Highest-Occupied and Lowest-Unoccupied Molecular Orbitals) is also used. The identification of Kohn-Sham and HOMO-LUMO gaps with fundamental gaps is justified only in the limit of non-interacting electrons. This is because Koopman's theorem is derived assuming that the orbitals of all other electrons do not change upon the introduction or removal of one electron. As will be seen in section 4.2.2, DFT Kohn-Sham gaps are typically far below experimental band gaps, with errors of 30%-100% for semiconductors and insulators alike. On the other hand, Hartree-Fock typically overestimates semiconductor band gaps by a factor of 2-5 [72].

One of the most egregious examples of the failure of both DFT and HF is germanium, with DFT and HF gaps of 0.0 and 4.2 eV, respectively, compared to an experimental gap of 0.7 eV.

Direct computation of the electron/hole excitation energies and hence energy gaps from differences in total energy, i.e. the evaluation of equation (4.1), is possible. This method is sometimes called the "ΔSelf Consistent Field" (ΔSCF) or "ΔHartree-Fock" method. The ΔSCF/HF method consists of doing multiple self-consistent DFT/HF calculations with different number of electrons, and has been applied to atoms and molecules to successfully reproduce experimentally-measured ionization energies [73]. In a macroscopic solid, however, it is widely-believed that there is no suitable way to evaluate the fundamental gap and equation (4.1) reduces to the Kohn-Sham gap in the limit $N \to \infty$. We will revisit this claim in section 4.5.

The band gap errors in DFT and HF have their origins in the treatment of electronic interactions. Both HF and DFT treat electrons in a one-particle picture using single-particle wavefunctions $\psi_i(\mathbf{r})$. In both cases the average Coulomb interaction is accounted for by the mean-field Hartree term:

$$E_H = \frac{1}{2} \sum_{i,j} \int d\mathbf{r} d\mathbf{r}' \frac{\rho_i(\mathbf{r})\rho_j(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{4.2}$$

where $\rho_i = e \sum_i |\psi_i|^2$ is the charge density, and spin indices have been suppressed. The remaining interaction among electrons is treated by taking into account the exchange exactly in HF, and by the exchange-correlation potential in DFT. The exact treatment of exchange in HF

$$E_X = \frac{1}{2} \sum_{i,j} \int d\mathbf{r} d\mathbf{r}' \frac{\psi_j^*(\mathbf{r}')\psi_i^*(\mathbf{r}')\psi_i(\mathbf{r})\psi_j(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|} \tag{4.3}$$

is nonlocal and in fact, long-ranged and infrared divergent. In DFT, commonly-used exchange-correlation functionals are the local and semi-local LDA and GGA (Local Density Approximation and the Generalized Gradient Approximation). This distinction has been used to explain the band gap errors in HF and DFT [74]. In

HF, the addition and removal of electrons carry a high energy penalty because the exchange repulsion of each pair of electrons is considered separately and long-ranged. In DFT, in contrast, the penalty is low because the exchange-correlation is semilocal.

The band gap errors in HF and DFT can alternatively be understood in terms of the dielectric response of a system of electrons. Consider again the addition of an extra electron to the system. Nearby interacting electrons rearrange their charges to screen the addition, thus lowering the energy penalty as compared to when charges are absent, rigid (non-polarizable), or non-interacting. The degree of screening is characterized by the dielectric function, which is in general space- and time-dependent. The HF treatment is analogous to atoms in free space, with a dielectric function of ~ 1. In DFT, which can be viewed as a perturbative treatment of the homogeneous electron gas, the screening is biased towards the metallic limit. The DFT dielectric function is hence overestimated. The energy penalty and hence the band gap is therefore overestimated in HF and underestimated in DFT. Within these extremes lie the realities of solids and molecules. The key to reducing the band gap errors, therefore, is in finding a more realistic description of the dielectric response than in the overly-rigid HF and overly-soft DFT.

4.1.3 Hybrid approaches

Given that DFT overscreens and underestimates gaps, and HF does the reverse, it is plausible that a combination of the two treatment gives a more accurate answer than either extremes. This is the rationale behind the so-called hybrid functionals, in which non-local HF exchange is added to the DFT exchange-correlation functional. Justified by using the adiabatic connection formalism, which makes formal the interpolation between the extremes of HF and DFT by an integration in the coupling constant e, the following PBE0 hybrid functional was proposed [75]

$$E_{xc}^{hybrid} = E_{xc}^{DFT} + \frac{1}{4}(E_x^{HF} - E_x^{DFT}) \qquad (4.4)$$

which replaces 1/4 of the semilocal exchange in DFT with HF exact exchange. Because hybrid functionals interpolate between HF and DFT, the band gaps obtained are in better agreement with experiment, particularly for molecules [3]. However, the band gaps in semiconductors are directly proportional to the amount of exact exchange added [76], and are overestimated by 80% in silicon and 33% in GaAs [5]. Moreover, because of the extended range, hybrid functionals are computationally much more expensive compared to LDA/GGA.[1] An alternative to the mixing of long-ranged exact exchange with semi-/local functionals is screened hybrid exchange [76], which reduces the computational cost by exponentially damping the exact exchange term, such that it is only significant within a distance of $\sim 6-20$ times the Bohr radius. The resulting band gaps for semiconductors are in better agreement with experiment, given an underestimate of 0.23 eV (15%) and 0.1 eV (10%) in GaAs and Si respectively [5].

The seemingly *ad hoc* 1/4 factor of optimal exact exchange in PBE0 may have some deeper underpinnings. Although the exact form of the exchange-correlation energy functional $E_{xc}[n]$ is not known, Lieb and Oxford [77] established an upper bound for the magnitude of $E_{xc}[n]$ as a multiple of the Local Density Approximation (LDA) exchange energy E_x^{LDA}:

$$0 \geqslant E_{xc}[n] \geqslant \lambda_{LO} E_x^{LDA}[n] = -C_{LO} \int d^3\mathbf{r} n^{4/3} \tag{4.5}$$

where $\lambda_{LO} = 2.27$ and $C_{LO} = 1.68$. This bound applies for all non-relativistic fermion systems governed by the Coulomb interaction in three dimensions, and has been incorporated into the construction of approximate exchange-correlation functionals such as GGA-PBE [78]. Recently, systematic studies of the Lieb-Oxford bound were performed [79] using exact results computed from configuration interaction and quantum Monte Carlo. It was found that for realistic systems including atoms, molecules and a wide range of solids ranging from simple metals and semiconductors to strongly correlated systems, the exact values of $\lambda \equiv E_{xc}[n]/E_x^{LDA}[n]$ range between $\sim 1.10-1.33$.

---

[1] An exception is DFT+U, a particularly expedient type of hybrid that operates only in the vicinity of individual atomic sites, which is described in section 4.2.2.

The authors of [79] argue that the small range of λ may explain the success of hybrid functionals which incorporate a fraction of the exact exchange.

4.1.4 Goals of this work

We have seen that Hartree-Fock and density functional theories grossly overestimate and underestimate, respectively, the band gaps of solids in the spectrum of their energy eigenvalues. The reason for these errors can be thought of in terms of the range of the interaction, or equivalently, an inadequate accounting of the dielectric response. Hybrid functionals can be used to improve on the band gap predictions by interpolating between HF and DFT, and may well be justified on a theoretical basis, but their use on solids have some limitations. We are interested, therefore, in methods that also makes use of physical insights involving dielectric screening to make better band gap predictions, keeping in mind the need to balance accuracy with efficiency. To this end, we investigate the use of the many-body self-energy approach, specifically in the GW approximation, where the dielectric response function and its contribution to the band gap is explicitly calculated. In addition, we revisit the issue of fundamental gap calculations using the ΔSCF method, taking into account in an approximate way the effective dielectric screening of the system. Before we embark on these investigations, we construct a test set of compounds with known experimental gaps and perform calculations of the DFT Kohn-Sham gaps to form a baseline for our studies. The physical basis and current status of different existing *ab initio* approaches to band gap predictions are summarized in table 4.1.

4.2 Band gaps from density functional theory

4.2.1 Compounds with known experimental band gaps

In order to evaluate various *ab initio* band gap prediction methods, we need a set of reliable experimentally-measured band gaps on compounds with known crystal structures. We obtained such information for a test set of 131 binary compounds and 3

Method	Principle	Accuracy (relative to experiment)
Hartree-Fock	Exact exchange, no correlation	1.5-7 times overestimation
DFT with Local Density or Generalized Gradient Approximations (LDA/GGA)	Local or semilocal exchange-correlation	30-100% underestimation
DFT with hybrid functionals	Combine a fraction of exact exchange with LDA/GGA exchange-correlation	10-80% overestimation for semiconductors; accurate to 20% underestimation for insulators (a)
DFT with screened hybrid functionals	As above, but with screened instead of bare exchange	~10-25% and ~20% underestimation for semiconductors and insulators respectively (a,b)
G_0W_0	Evaluate self-energy as a convolution of Green function constructed from DFT wavefunctions and eigenvalues (G) with frequency-dependent screened Coulomb interaction (W)	Accurate to about 10-25% in semiconductors (b) and about 0.5 eV in mid-gap compounds (c); fails to correct 0 LDA gap in InN (c); 20% underestimation (c) or overestimation (d) in insulators
Self-consistent GW	As above, but also with self-consistent cycle to update wavefunctions and/or eigenvalues.	In dispute

Table 4.1: An overview of different methods for *ab initio* band gap predictions. References: (a) [5] (b) [6] (c) [7] (d) [8]. The theoretical basis and implementation details of the *GW* approximation are given in section 4.3 below.

elements (C, Si and Ge) from *Semiconductors: Data Handbook* [2]. The test set contains a wide variety of chemistries, including typical semiconductors, such as GaAs and CdTe, as well as main group and transition metal oxides, other chalcogenides, and halides. The experimental band gaps range from 0.2 to 10.6 eV. Information on the symmetry (space group) is corroborated with that from the Inorganic Crystal Structure Database (ICSD) [10], via the Materials Genome database developed by Anubhav Jain in the Ceder group [80]. For ease of computational testing, we select only compounds with unit cells containing less than 200 valence electrons. Compounds where there exist significant (> 0.3 eV) discrepancies in different experimental measurements are omitted. We take low-temperature values of the experimental gap whenever possible, and make extrapolations using experimentally-measured temperature coefficients to 0K. When temperature coefficients are not available, the 300K band gaps are used. The uncertainty in the experimental values are typically in the $0.05 - 0.3$ eV range. Since compounds with band gaps in the range of $1 - 4$ eV are most relevant for photovoltaic and related applications, specific analysis will be made for this group (henceforth called "medium-gap compounds") as well as for all compounds in the test set.

4.2.2 Computed DFT Kohn-Sham band gaps

We calculate the Kohn-Sham band gaps for compounds in the test set described above using ground state density functional theory (DFT). The DFT calculations are done using the Vienna Abinitio Simulation Package (VASP) [81], version 4.6, with accompanying Projector Augmented Wave (PAW) [82] atom data. The Generalized Gradient Approximation, as parametrized by Perdew, Burke and Ernzerhof (GGA-PBE) [78], is used as the exchange-correlation functional. The two spin densities are treated separately. Kinetic energy cut-offs for the plane wave basis sets are in the range 200-400 eV. Ionic positions and cell parameters are optimized using conjugate gradient relaxation. Brillouin zone sampling is carried out using $6^3 - 24^3$ gamma-centered k-point grids with an average of 80 k-points in the irreducible Brillouin zone (IBZ). The tetrahedron method with Blöchl corrections [83] is used for Brillouin zone

integration.

The Kohn-Sham band gaps thus obtained are plotted against the corresponding experimental values in figure 4-1. As expected, the Kohn-Sham band gaps severely underestimate the experimental values. The mean absolute error is 1.1 eV, with a standard deviation of 0.8 eV. The mean relative error is 56%, with a standard deviation of 30%. Among the 94 compounds with band gaps within the range of 1-4 eV, the mean absolute error is 1.1 eV, and the standard deviation is 0.6 eV. The mean relative error is 51%, with a standard deviation of 26%. Although some trends are noticeable, e.g. that the largest underestimations occur for wide-gap oxides (BeO, MgO, CaO, SrO), the errors are not in general systematic. This is shown by the large standard deviation for both the absolute and relative errors, both for all compounds and for medium-gap compounds. Given only the Kohn-Sham band gap, an inverse map to the experimental band gap is ill-defined. Out of the 134 test elements/compounds, six (BiI_3, Bi_2Se_3, Bi_2Te_3, PbS, PbSe and PbTe) have Kohn-Sham gaps that exceed the experimental gaps. The LDA/GGA overestimation of band gaps has been previously reported for lead [84] and bismuth [85] chalcogenides, and was attributed to the neglect of spin-orbit coupling. This is a reasonable explanation due to the large mass of Pb and Bi.

The Hubbard U correction to DFT, DFT+U, was introduced [86] to account for strong static correlations in systems with partially-filled d- and f-shells, i.e. transition metals. In addition to the usual Kohn-Sham Hamiltonian, an energy penalty is applied on partial projected occupancies of atomic d- or f-orbitals. The energy penalty is determined by the parameter U, which is typically in the range 1-5 eV. DFT+U can be thought of as a semi-local hybrid scheme, in which the non-locality is only applied on atomic sites, for which the Coulomb integrals are already pre-evaluated. Since DFT+U has been found to accurately predict total energies and remove qualitative errors in the electronic structures (for a review, see [86]), at little additional computational cost, we apply the DFT+U method for all transition metal oxides and halides. For transition-metal oxides, the values of the U parameter are taken from *ab initio* linear response calculations, as described in [87]. For halides, a standard
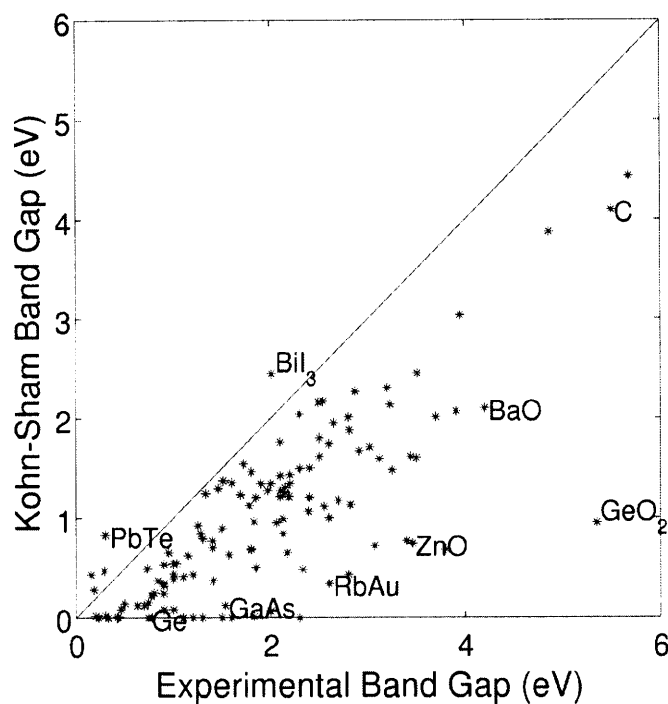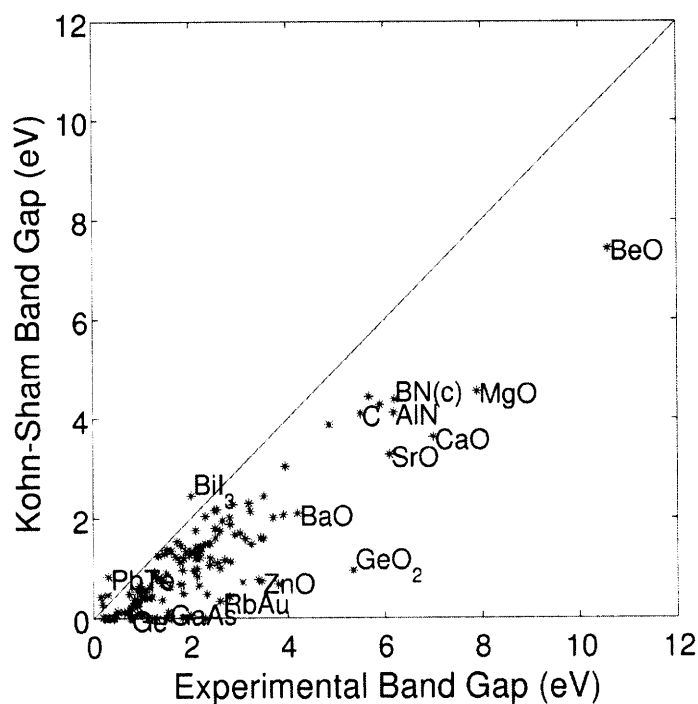
Figure 4-1: Calculated Kohn-Sham band gaps, defined as the difference between eigenvalues of lowest unoccupied state and highest occupied state, vs. experimentally-measured band gaps. The straight line represents one-to-one correspondence. Low-temperature values of the experimental band gaps are taken whenever possible, and extrapolations are made using provided temperature coefficients to 0K. The experimental band gaps and temperature coefficients are obtained from [2]. Apart from 6 compounds containing lead or bismuth, the Kohn-Sham gaps are smaller than experimental gaps. Top: all 131 binary compounds and 3 elements; Bottom: a close-up in the range 0-6 eV (122 compounds). Selected outlyers are labeled.

value of 3 eV is used. It should be noted, however, that the addition of U does not generally remove band gap errors. For AgF for example, which has an experimental gap of 2.8 eV, the Kohn-Sham gap with $1.5 \gtrsim U \geqslant 0$ eV is zero, and increases to 1.3 eV with $U = 7$ eV.

4.3 Self-energy approaches and the *GW* approximation

4.3.1 Theory

We begin by considering the Kohn-Sham equation of density functional theory (DFT), which is

$$H_{KS}\psi_{nk}(\mathbf{r}) \equiv \left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ion}} + V_{\text{Hartree}} + V_{\text{xc}}(\mathbf{r}) \right] \psi_{nk}(\mathbf{r}) = \epsilon_{nk}\psi_{nk}(\mathbf{r}) \qquad (4.6)$$

where n and k are the band and k-point indices. The exchange-correlation potential V_{xc} is local, real and time-independent, reflecting the non-interacting nature of the Kohn-Sham theory. A similar equation can be written for Hartree-Fock or hybrid theories, in which the only difference is that V_{xc} is non-local, but still real and time-independent. Eigenstates of H_{KS} or H_{HF} have real energies and are infinitely long-lived. [2] Hamiltonians of many systems of interacting electrons are, however, perturbatively connected to the non-interacting ones, such that one-particle wavefunctions are still approximately good degrees of freedom. In that case we call the interacting electrons quasiparticles, and rewrite equation (4.6) as

$$H_{KS}\phi_{nk}(\mathbf{r}) + \int \left[\Sigma(\mathbf{r}, \mathbf{r}', \epsilon'_{nk}/\hbar) - V_{xc}(\mathbf{r}') \right] \phi_{nk}(\mathbf{r}')d\mathbf{r}' = \epsilon'_{nk}\phi_{nk}(\mathbf{r}) \qquad (4.7)$$

where $\Sigma(\mathbf{r}, \mathbf{r}', \omega)$ is in general a non-local, complex, and time-dependent function called the self-energy. Since Σ is not Hermitian, the eigenvalues ϵ'_{nk} are in general

---

[2]Fortunately these Hamiltonians do not fully describe electrons in nature, or else chemistry or life would not be possible.

complex and the quasiparticles have finite lifetimes. The quasiparticle wavefunctions ϕ_{nk} are not the same as ψ_{nk}, but will be similar in the limit that $|\Sigma - V_{XC}|$ is small compared to H_{KS}.

The electron Green function $G(\mathbf{r}, \mathbf{r}', t, t')$, also called the propagator, gives the probability amplitude for propagation of an excitation from $(\mathbf{r}', t')$ to $(\mathbf{r}, t)$. The Dyson equation relates the full Green function G of an interacting system to that of a non-interacting system G_0 and the self-energy:

$$G(\mathbf{r}, \mathbf{r}', \omega) = G_0(\mathbf{r}, \mathbf{r}', \omega) + \iint G_0(\mathbf{r}, \mathbf{r}'', \omega)\Sigma(\mathbf{r}'', \mathbf{r}''', \omega)G(\mathbf{r}''', \mathbf{r}', \omega)d^3\mathbf{r}''d^3\mathbf{r}''' \quad (4.8)$$

where a Fourier transform in time has been made to express G and G_0 in the Lehmann representation, i.e. in terms of $(\mathbf{r}, \mathbf{r}', \omega)$ instead of $(\mathbf{r}, \mathbf{r}', t, t')$. If Σ is known, then so is G and any physically measureable quantites from the many-body system. In general, however, Σ cannot be solved explicitly, and perturbative iterations have to be used. The Dyson equation can be written in a perturbative series as, symbollically:

$$G = G_0 + G_0\Sigma G_0 + G_0\Sigma G_0\Sigma G_0 + G_0\Sigma G_0\Sigma G_0\Sigma G_0 + ... \quad (4.9)$$

The non-interacting Green function (bare propagator) G_0 is constructed from solutions of the non-interacting Hamiltonian. In the Lehmann representation, we have

$$G_0(\mathbf{r}, \mathbf{r}', \omega) = \sum_i \frac{\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')}{\hbar\omega - \epsilon_i \pm i\eta} \quad (4.10)$$

where the sum is over all occupied and unoccupied states i, and ψ_i and ϵ_i represent the eigenstates and eigenvalues of the non-interacting Hamiltonian.

In the GW approximation, first proposed by Hedin [88], the self-energy Σ is approximated by

$$\Sigma(\mathbf{r}, \mathbf{r}', \omega) = \frac{i\hbar}{2\pi} \int_{-\infty}^{\infty} d\omega' e^{-i\delta\omega'} G(\mathbf{r}, \mathbf{r}', \omega - \omega')W(\mathbf{r}, \mathbf{r}', \omega') \quad (4.11)$$

where W is the Coulomb interaction $\mathrm{v} = e^2/|\mathbf{r} - \mathbf{r}'|$ screened by the dielectric function
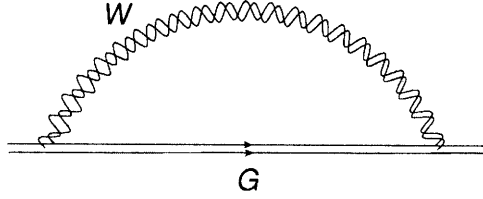
Figure 4-2: Diagrammatic representation of the approximation $\Sigma = iGW$. In the limit of bare electron and Coulomb propagators, this diagram reduces to Hartree-Fock exchange. Instead, W is a screened Coulomb interaction with a polarization evaluated in RPA.

$\varepsilon(\mathbf{r}, \mathbf{r}', \omega)$, i.e.

$$W(\mathbf{r}, \mathbf{r}', \omega) = \int d\mathbf{r}'' \varepsilon^{-1}(\mathbf{r}, \mathbf{r}'', \omega) v(\mathbf{r}'', \mathbf{r}', \omega) \tag{4.12}$$

The GW approximation is equivalent to factoring out the Coulomb and electron propagators, as shown in diagrammatic representation in figure 4-2, and specifically neglecting vertex corrections such as shown in figure 4-3. Consistent with the neglect of the vertex in Σ, the Coulomb propagator is treated at the random phase approximation (RPA) level, corresponding to the summation of an infinite series of 1-loop polarization bubbles, as shown in figure 4-4. This gives

$$\varepsilon(\mathbf{r}, \mathbf{r}', \omega) = \delta(\mathbf{r} - \mathbf{r}') - \int v(\mathbf{r}, \mathbf{r}'') P(\mathbf{r}'', \mathbf{r}', \omega) d^3\mathbf{r}'' \tag{4.13}$$

where P is the polarization given by

$$P(\mathbf{r}, \mathbf{r}', \omega) = \sum_{ij} (f_i - f_j) \frac{\psi_i(\mathbf{r})\psi_j^*(\mathbf{r})\psi_i^*(\mathbf{r}')\psi_j(\mathbf{r}')}{\omega - (\epsilon_j - \epsilon_i)/\hbar + i\eta} \tag{4.14}$$

with f_i being the Fermi occupation number of state i.

In self-consistent GW approximation, the self-energy obtained from G_0 is put into the Dyson equation (4.9) and iterated to obtain the full Green function G, as shown by double straight lines in figure 4-2. The validity of the self-consistent GW approach is under some dispute, particularly when the vertex corrections continue to be omitted. When $|\Sigma - V_{XC}|$ is small, however, the first-order G_0W_0 approximation can be used instead. In the G_0W_0 approximation, the bare propagator G_0, constructed from
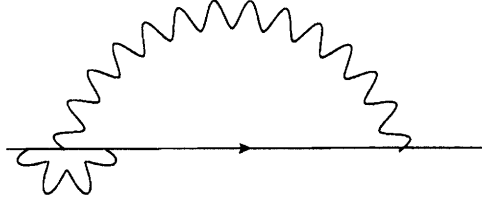
Figure 4-3: A contribution to the vertex correction Γ not included in the *GW* approximation.

Figure 4-4: The Coulomb propagator evaluated in the random phase approximation (RPA), corresponding to the summation of an infinite series of one-loop one-particle-irreducible (1PI) diagrams. Physically this means that individual particle-hole pairs are independent, or that electron-hole or exciton interactions are negligible.

Kohn-Sham eigenvalues and eigenfunctions by equation (4.10), is used to calculate $\Sigma_0 = iG_0W$. The self-energy Σ_0 (henceforth referred to simply as Σ) is used to evaluate corrections to non-interacting eigenvalues using

$$\epsilon'_{nk} - \epsilon_{nk} \approx \langle\psi_{nk}|\Sigma(\epsilon'_{nk}/\hbar) - V_{xc}|\psi_{nk}\rangle \tag{4.15}$$

where the evaluation of Σ is made at the corrected quasiparticle energy ϵ'. If $\epsilon' - \epsilon$ is small, we can again linearize to obtain

$$\Sigma(\mathbf{r},\mathbf{r}',\epsilon'_{nk}/\hbar) \approx \Sigma(\mathbf{r},\mathbf{r}',\epsilon_{nk}/\hbar) + \frac{\epsilon'_{nk} - \epsilon_{nk}}{\hbar}\frac{\partial\Sigma(\mathbf{r},\mathbf{r}',\epsilon_{nk}/\hbar)}{\partial\omega} \tag{4.16}$$

which gives

$$\epsilon'_{nk} - \epsilon_{nk} \approx Z_{nk}\langle\psi_{nk}|\Sigma(\epsilon_{nk}/\hbar) - V_{xc}|\psi_{nk}\rangle \tag{4.17}$$

in which the quasiparticle weight Z_{nk} is given by

$$Z_{nk} = \left(1 - \langle\psi_{nk}|\frac{\partial\Sigma(\epsilon_{nk}/\hbar)}{\hbar\partial\omega}|\psi_{nk}\rangle\right)^{-1}. \tag{4.18}$$

The quasiparticle weight describes the similarity of a quasiparticle to a free particle,

which has $Z = 1$. When Z differs signficantly from unity, it may no longer be justifiable to use the first order G_0W_0.

Finally, the band gap correction due to the GW approximation is given by

$$
\begin{aligned}
\Delta E_g &\equiv (\epsilon'_c - \epsilon'_v) - (\epsilon_c - \epsilon_v) \\
&= E_{g,KS} + Z_c\langle\psi_c|\Sigma(\epsilon_c/\hbar) - V_{xc}|\psi_c\rangle - Z_v\langle\psi_v|\Sigma(\epsilon_v/\hbar) - V_{xc}|\psi_v\rangle \quad (4.19)
\end{aligned}
$$

where c and v stand for states at the conduction band minimum and valence band maximum, respectively.

4.3.2   Computational details

We calculate the Kohn-Sham eigenvalues and eigenvectors and G_0W_0 corrections to the band edge eneriges using the plane wave pseudopotential code ABINIT [89], version 5.6. Due to the strenuous computational requirements, as explained below, a subset of the test compounds with small unit cells are selected for the G_0W_0 computation. We use GGA-PBE as the exchange-correlation functional, and Fritz-Haber-Institute (FHI) pseudopotentials generated in the Troullier-Martins scheme as obtained from the ABINIT website [90]. The optimized ground state ionic positions and cell parameters for each compound are used in subsequent G_0W_0 calculations. The kinetic energy cutoffs for the plane wave basis sets of the ground state calculations range from 15 to 45 Hartree, and are chosen for each compound such that the total ground state energy is converged to within 5 meV per atom. The DFT+U approximation has *not* been applied for these calculations.

The frequency integration in equation (4.11) can be performed numerically, but it is much more computationally efficient to use a plasmon pole model, in which the frequency dependence of ε is given by:

$$
\varepsilon^{-1}(\omega) = \delta + \frac{\Omega^2}{\omega^2 - \tilde{\omega}^2} \quad (4.20)
$$

The dielectric function is calculated by evaluating the RPA dielectric function (4.13)

from the polarization (4.14) at two frequencies, from which the matrix elements of Ω and $\tilde{\omega}$ are obtained. The plasmon pole model physically corresponds to the assumption that the loss spectrum (imaginary part of ε^{-1}) is dominated by excitations that have well-defined frequencies. The plasmon pole model is expected to describe typical semiconductors well but not strongly-correlated systems [91].

With the use of the plasmon pole model, the most computationally demanding step in the G_0W_0 calculation is the evaluation of the polarization (4.14) and the construction of the dielectric matrix. The computational time and memory requirements for this step scale roughly as the square of the number of k-points in the irreducible Brillouin zone, since the dielectric matrix is evaluated for each symmetrically-inequivalent $\mathbf{q} = \mathbf{k_i} - \mathbf{k_j}$. Since ε also carries the double indices $\mathbf{G}$ and $\mathbf{G}'$, the kinetic energy cutoff $\sim \hbar^2 G_{max}^2/2m$ of the plane wave basis for the dielectric matrix is typically set lower than that for the wavefunctions. The typical computational time for this step is a hundred to several thousand times that of a ground state calculation. The entire dielectric matrix written to (binary) file is of order tens to hundreds times the size of a wavefunction file for the same compound, which means that the memory and I/O requirements are strenuous. For example, for BeTe, a 2-atom cell with 8 valence electrons, the binary wavefunction file for a ground-state calculation with 16 k-points in the IBZ and kinetic energy cutoff of 20 Hartree is 1.7MB. With a kinetic energy cutoff of 15 Hartree, the dielectric matrix file is 370MB.

Convergence testing with respect to the kinetic energy cutoff used for the dielectric matrix is performed for MgO. The results are shown in figure 4-5. For the ground state calculations, an energy cutoff of 40 Hartree is required to obtain the total energy to within 5 meV per atom. Such a large energy cutoff is not realistic for the dielectric matrix calculations, as explained above. From the convegence testing, we note that if the cutoff for the dielectric function is set at 15 Hartree, the G_0W_0 band gap obtained is 0.26 eV below the final converged value of 6.69 eV at 30 Hartree. The experimental gap of MgO is 7.8 eV. The computational time required for the dielectric function calculation is 1.5 hours and 15.6 hours for energy cutoffs of 15 and 30 Hartree, respectively. Therefore, achieving final convergence to within 0.1 - 0.2
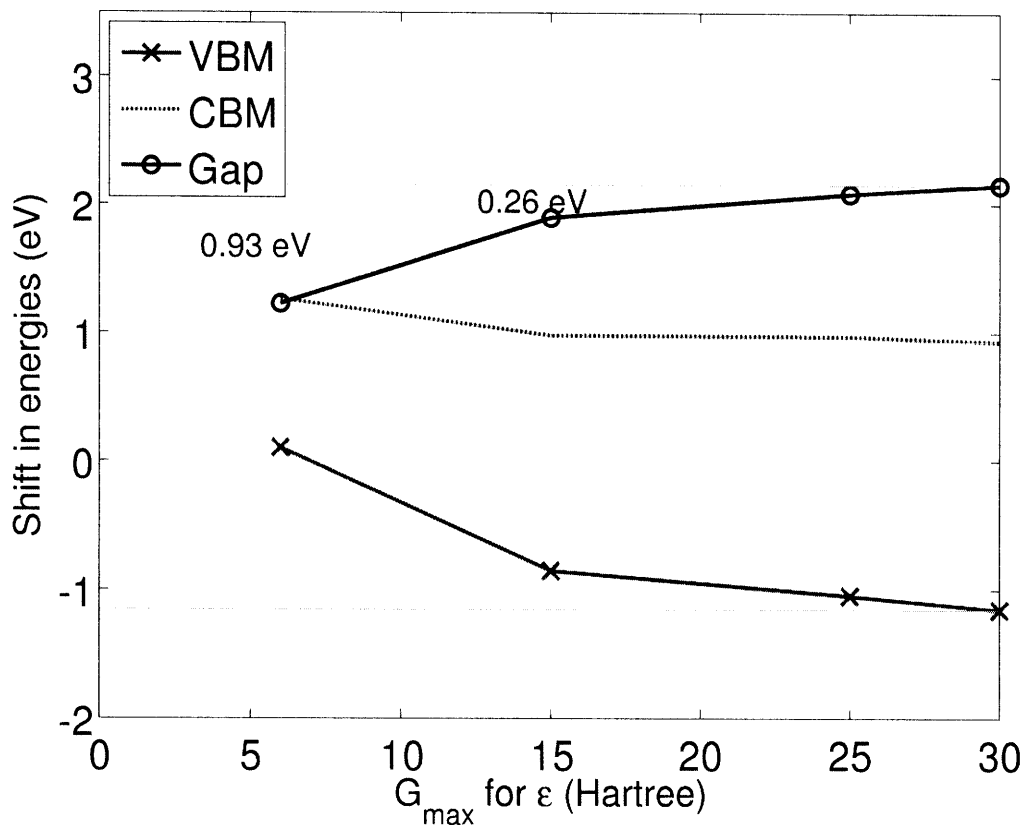
Figure 4-5: The convergence of G_0W_0 corrections to the valence band maximum energy, conduction band minimum energy, and the band gap, as a function of the kinetic energy cutoff for the plane wave basis set used to represent ε. Note that the main contribution to the error when using a smaller energy cutoff in this case comes from the valence band. The final value for the G_0W_0 band gap correction (2.2 eV) to the Kohn-Sham gap comes in equal parts from the downward adjustment of the valence band maximum and the upward adjustment of the conduction band minimum.

eV comes at rather large computational cost. Because it is not feasible to perform such convergence testing for each compound, we use 15 Hartree as the kinetic energy cutoff of the plane-wave basis set for all compounds where feasible, and reduce it to 10 or 6 Hartree if the larger calculation fails to finish (typically because of insufficient memory).

Apart from the convergence with respect to basis set size (kinetic energy cutoff), we should also note that there may also be convergence issues with respect to the number of empty bands included in the calculation. Both in the expressions for the

polarization matrix in equation (4.14) and the bare Green function in equation (4.10), there is a sum over all the occupied and unoccupied states. Because the output of the Kohn-Sham states into a file used for subsequent polarization and self-energy calculations has not been implemented in parallel, we find it difficult to include more than about 100-200 empty bands. This is found to be sufficient for compounds with smaller than 10 occupied bands, but the effects on compounds with more occupied bands is unknown.

4.3.3 Results of G_0W_0 computations

The band gaps of 36 compounds computed using the G_0W_0 approximation using the plasmon pole model are shown in figures 4-6 and 4-7. The mean absolute errors for all 36 compounds and the 21 medium-gap compounds within this set compared to the experimental values are 0.57 and 0.39 eV, respectively, with standard deviations of 0.61 and 0.38 eV, respectively. This is in comparison to the errors for the Kohn-Sham gap of 1.2 and 1.0 eV (with standard deviations of 0.96 and 0.55 eV) for the same compounds. The Kohn-Sham band gap errors are reduced by 50-60% on average by using the G_0W_0 approximation within the parameters of our calculations. For the medium-gap compounds, 50% have G_0W_0 gaps that are within 10% of the experimental value, and 70% are within 25% of experimental value. For an addditional 21 compounds for which the calculations are completed, the GW band gaps are at or below the Kohn-Sham gaps and not plotted. Results for subgroups of compounds transition metal compounds, medium-gap compounds, and insulators are discussed separately below.

We note that the values of the quasiparticle weight Z for all compounds calculated fall in the range 0.69-0.84, similar to the RPA values for the homogeneous electron gas [88] with the same range of r_s ($\sim$1-3). This is because of the use of the plasmon pole model for the frequency-dependence, as Z is derived from the frequency dependence of Σ and the plasmon pole model omits satellites due to strong correlations. Since we start from the non-interacting Hamiltonian and unit "quasiparticle" weights, the validity of the G_0W_0 approximation would be suspect if the values of Z obtained
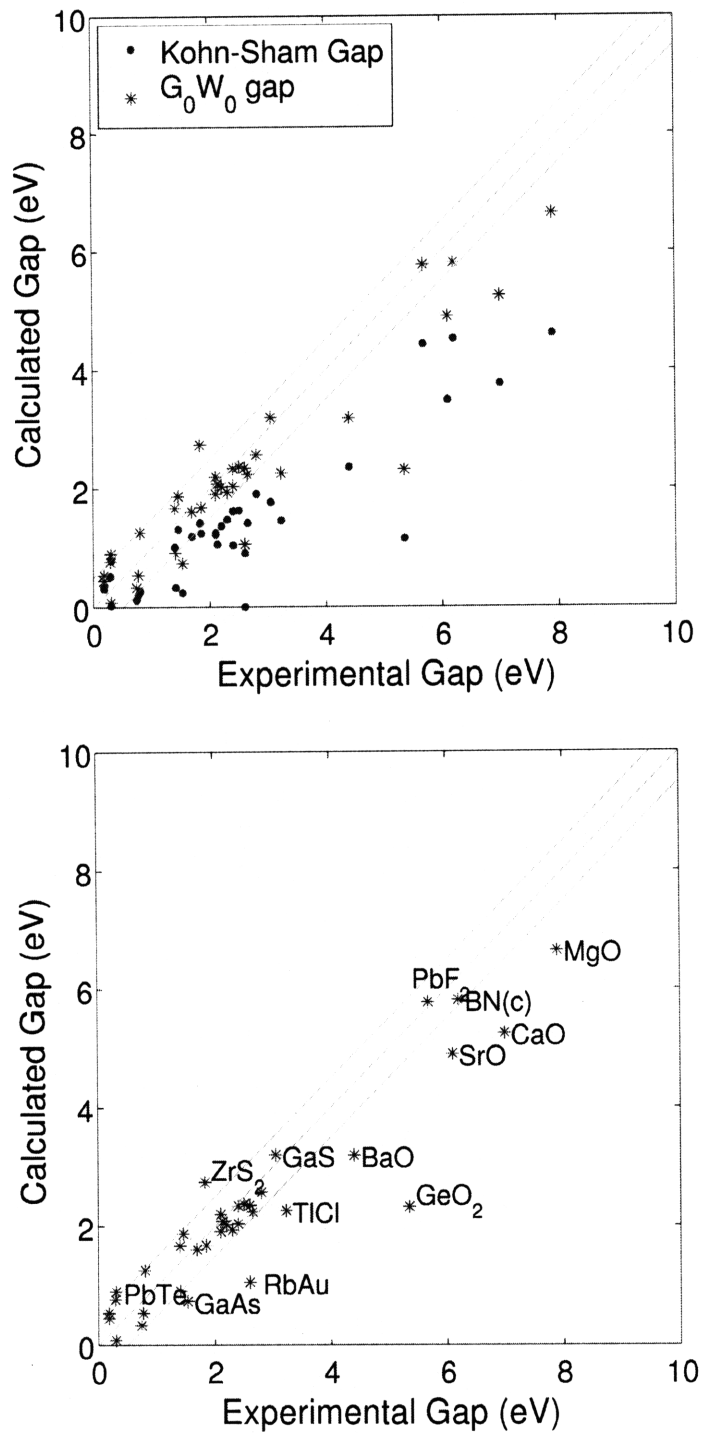
Figure 4-6: Band gaps of 36 compounds predicted by the G_0W_0 approximation, plotted against the experimental values. The straight lines denote one-to-one correspondence and $\pm\,0.5$ eV deviations. Top: Also plotted are the Kohn-Sham band gaps of the same compounds. Bottom: Selected compounds are labeled.
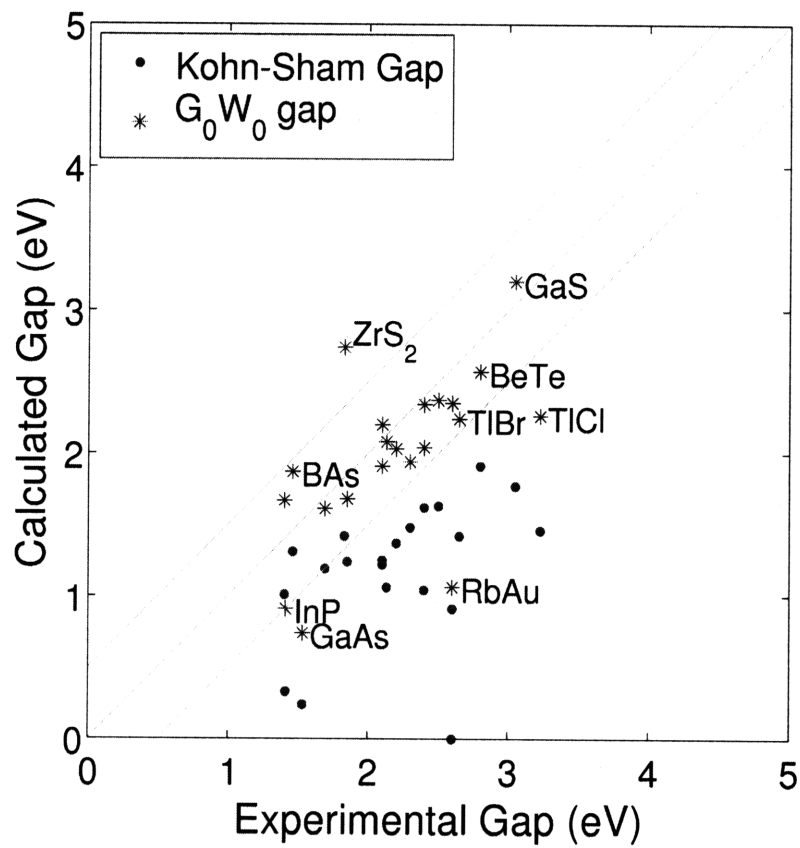
Figure 4-7: An expanded view of figure 4-6 for medium-gap compounds (1-4 eV). The largest absolute error comes from RbAu, which is discussed in more detail in section 4.3.3.

were very different from 1.

We would also like to remark upon the validity of the so-called "scissors operator", in which the valence and conduction bands are manually shifted apart by the same amount regardless of $\mathbf{k}$. Because we calculate the self-energy corrections to the eigenvalues at two k-points for indirect gap compounds, we can compare the shifts at the valence band maximum and conduction band minimum at those two k-points. The results for 23 compounds are shown in figure 4-8. We find that RbAu and CsAu are outliers, in having a difference of 1.2 and 0.90 eV, respectively, between the G_0W_0 correction of the gap at two k-points. For the remaining compounds, the differences are generally in the 0.05-0.2 eV range. The results imply that simply shifting the bands by the same amount at all k-points may be approximately accurate for most compounds, though may not be a robust procedure.

Transition metal compounds

For many transition-metal compounds, the G_0W_0 corrections to the gap are essentially zero (less than 0.05 eV) or negative. These include all the chalcogenides and halides of Cu (Cu_2O, Cu_2S, Cu_2Se, CuBr, CuCl, CuI) and Ag (AgF, AgCl, AgBr, AgI, Ag_2O). The origin of this negative correction is the shifting of the valence band maximum upwards by 2-3 eV (for Cu) and 1-2 eV (for Ag). For both Yb compounds considered (YbS and YbSe), the upward shift of the valence band maximum is about 5 eV. All these compounds have strong d- or f-characters in the valence bands maxima. For these localized states, contributions to the screened Coulomb interaction in the GW correction are mostly diagonal, and in applying equation (4.15) to compute the GW correction to the valence band energy, we are simply evaluating the on-site Coulomb interaction, or the U parameter that is used in DFT+U calculations. Indeed, the 2-3 eV for Cu, 1-2 eV for Ag and 5 eV for Yb are typical values for the U parameter used in DFT+U calculations. The values obtained are listed in Table 4.2. The failure of G_0W_0 in these instances to correct the band gap reflects the fact that the quasiparticles are not close to the Kohn-Sham eigenstates. However, if one updates the eigenvalues and performs the iteration again, i.e. in the next step towards self-consistency, these
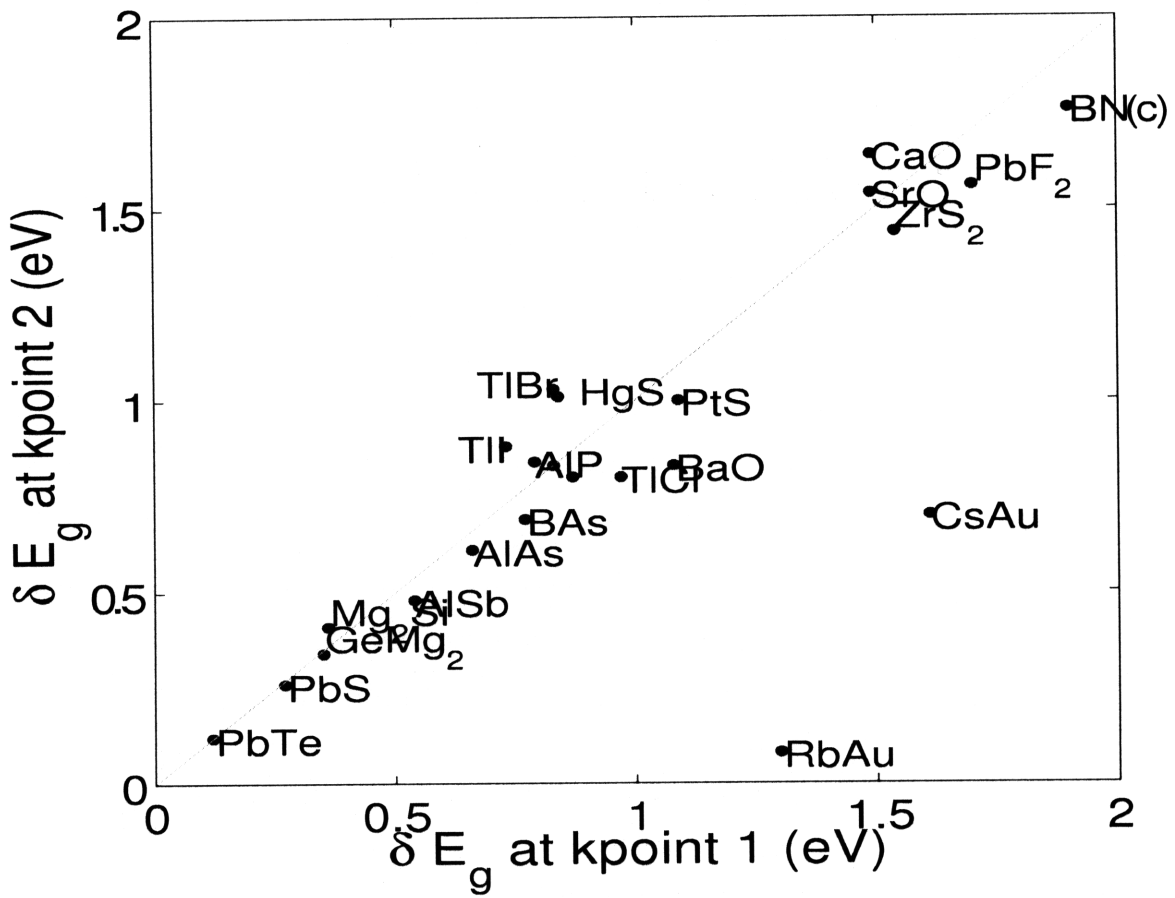
Figure 4-8: The G_0W_0 correction to valence band-conduction band separation at two k-points.

Table 4.2: The self-energy correction to the valence band maximum calculated from G_0W_0, for Cu, Ag and Yb compounds, which can be identified as the value of the U parameter in DFT+U calculations.

Compound	U from G_0W_0 (eV)
CuCl	3.3
CuBr	3.2
CuI	2.7
Cu_2O	3.8
Cu_2S	3.3
Cu_2Se	2.8
AgO	1.8
AgF	1.7
AgCl	1.0
AgBr	1.0
AgI	0.9
YbS	5.2
YbSe	4.8

d- and f-like states with high positive eigenvalue corrections $\epsilon' - \epsilon$ will no longer be at the band edge. This has the same effect as the DFT+U correction in pushing the Hubbard bands away from the Fermi level, except that the GW iteration does not require an *a priori* knowledge of the parameter U. Of course, there is enormous computational expense that comes with performing multiple iterations of the GW calculation.

Medium-gap compounds ($1 \leq E_g \leq 4$ eV)

The results for medium-gap compounds are shown in Table 4.3. The G_0W_0-corrected gaps for III-V and III-VI compounds are in general in good agreement with experimental values, with a mean absolute error of 0.26 eV. The largest errors in these groups came from the arsenides, without which the mean errors are only 0.16 eV.

The largest absolute error in the G_0W_0 values of the band gap for medium-gap compounds is found in RbAu (1.1 vs 2.6 eV), which is metallic in LDA/GGA. The series of alkali-metal gold compounds undergo a metal-semiconductor transition between RbAu and KAu [92, 93]. The band gap of CsAu, which has been more ex-

Table 4.3: Experimental and computed G_0W_0 band gaps of compounds with experimental band gaps between 1 and 4 eV. The second decimal digit is generally not significant.

Compound	Experimental Gap (eV)	Kohn-Sham Gap (eV)	G_0W_0 Gap (eV)
AlAs	2.30	1.48	1.94
AlP	2.50	1.63	2.37
AlSb	1.69	1.19	1.61
BAs	1.46	1.31	1.87
BP	2.20	1.37	2.03
GaAs	1.53	0.24	0.74
GaP	2.40	1.62	2.34
InP	1.41	0.33	0.91
GaS	3.05	1.77	3.20
GaSe	2.13	1.06	2.08
GeS	1.85	1.24	1.68
BeTe	2.80	1.91	2.57
K3Sb	1.40	1.01	1.67
CsAu	2.60	0.91	2.35
RbAu	2.60	0.00	1.06
HgI	2.40	1.04	2.04
HgS	2.10	1.25	2.20
ZrS2	1.83	1.42	2.74
TlBr	2.65	1.42	2.24
TlCl	3.23	1.46	2.26
TlI	2.10	1.22	1.91

tensively studied, was measured to be 2.6 eV in thin film form [94], and in *Semiconductors: Data Handbook* [2], RbAu was also listed as having a gap of 2.6 eV. However, we were unable to find experimental reports of this value for the measurement. Moreover, because the metal-semiconductor transition is accompanied by large volume changes which affects the band gap [95], it is quite possible that the experimental value for RbAu is indeed different from 2.6 eV. We find in general that the experimental values of the band gap for some compounds are either not well-known or in dispute, for reasons ranging from difficulties in sample preparation to effects of surface adsorbants.

Table 4.4: Experimental and computed G_0W_0 band gaps of compounds with experimental band gaps greater than 4 eV.

Compound	Experimental Gap (eV)	G_0W_0 Gap (eV)
GeO_2	5.4	2.3
PbF_2	5.7	5.8
BN (cubic)	6.2	5.8
BaO	4.4	3.2
SrO	6.1	4.9
CaO	7.0	5.3
MgO	7.8	6.7

Insulators

The absolute Kohn-Sham band gap errors are largest for insulators. Since the electronic screening is weak, the screened interaction W is expected to be similar to the bare interaction v, in which case the GW correction should be large and the linear treatment G_0W_0 may not be an appropriate treatment. From Table 4.4, we see that for all the alkali-earth oxides, G_0W_0 underestimates the gap by 1.2-1.7 eV. The underestimation is larger than previously-reported values of about 0.5-1 eV [7]. The trend, however, is consistent, as the whole series is shifted by the same amount and the previously-reported underestimation is also largest for CaO. For cubic BN and PbF_2, however, the G_0W_0 band gap values are in good agreement with experiment.

The severe underestimation in GeO_2 (2.3 eV vs 5.4 eV) is possibly due to a lack of convergence with respect to the size of the plane wave basis set for ε, as in MgO, but computation with a kinetic energy cutoff larger than 15 Hartree fails to complete due to memory problems. It should be noted that the kinetic energy cutoff for ε required to produce accurate results is affected by the "hardness", or the minimum required energy cutoff, of the pseudopotential used. This is because the polarization P and Green function G_0 are both constructed from the wavefunctions, so that if wavefunctions require a large number of planewaves to construct accurately, so do P and G_0. Therefore, it is desirable to optimize the pseudopotentials for use in GW calculations such that the kinetic energy cutoff requirements are less stringent.

4.4 Interlogue

We saw that with the explicit inclusion of the non-local and frequency-dependent interactions, and explicit construction of the dielectric response function of a system from the polarization, we are able to correct a large portion of the DFT band gap error using the GW approximation. G_0W_0 also gives the Hubbard U parameter that can serve as a useful input for DFT+U calculations. Our errors are larger than in typical work involving GW computations, which usually state an accuracy of 0.1-0.2 eV, for two reasons. The first is that in the interest of general applicability, it is desirable to avoid the need for the construction of pseudopotentials and fine-tuning of parameters for each calculation. The issue of pseudopotentials may be avoided all-together by performing all-electron calculations, though at an added computational cost. Secondly, again in the interest of being able to perform calculations for a large number of systems, we attempted to keep the computational requirements to a minimum. In principle, the full frequency-dependent dielectric matrix and Green function contains all the information available, but in practice the results are quite sensitive to the resolution at which we carry out our computations. This is because, as in DFT, the numerical result comes from the addition of several terms of similar size: $\langle\psi|\Sigma_x|\psi\rangle$, $\langle\psi|\Sigma_c|\psi\rangle$ and $-\langle\psi|V_{xc}|\psi\rangle$.

At this juncture, we would like to ask whether simplifications can be made that capture the physical insights in the GW approach that are relevant to the prediction of the band gap, and yet render the problem less computationally exacting. A large body of work exists on the building of model dielectric functions and other simplifications to the GW approach, and reviews can be found in [?]. For us, the most important realization is that the band gap is essentially a static property. This is because the dielectric response, which is plasmon-pole-like, is relatively featureless below the plasma frequency $\omega_p = \sqrt{\frac{4\pi Ne^2}{mV}}$, where N, e and m are the number, charge and mass of electrons, respectively, and V is the volume. Typical plasma frequencies correspond to an energy of 15 eV, which is far above the energy scale of the separation between the conduction and valence bands in our systems of interest. Therefore, it is only
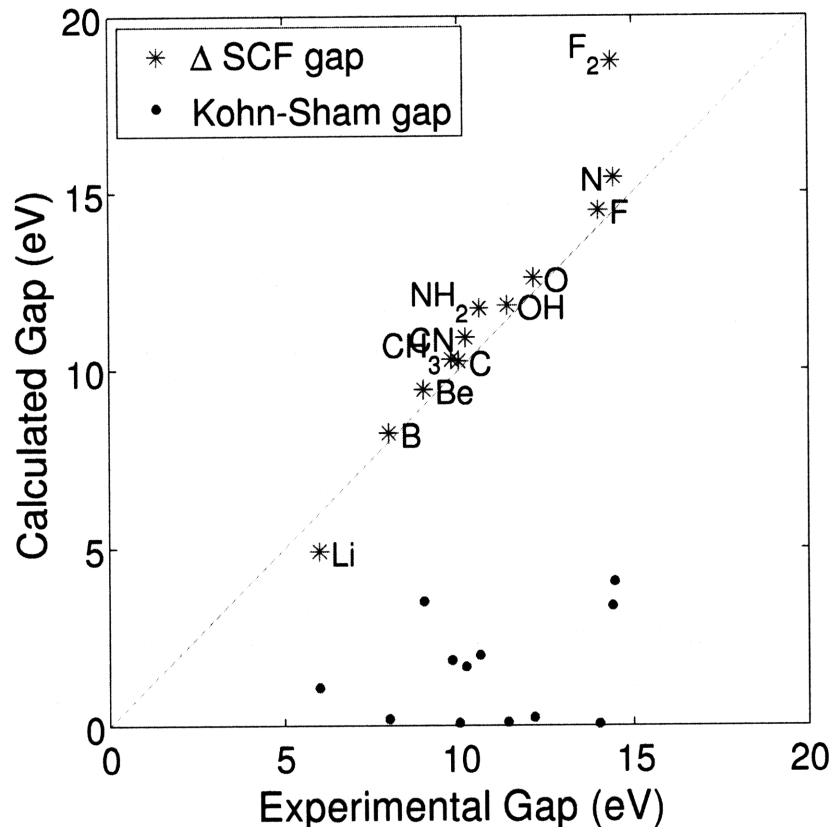
Figure 4-9: A plot of gaps calculated from DFT (LDA) vs. experimental values for atoms and molecules, using data from [3] and reference 14 therein. The calculations done with the ΔSCF method, i.e. direct evaluation of $E_{\text{gap,fundamental}} = E(N+1) + E(N-1) - 2E(N)$, show reasonable agreement with experiment, in stark contrast to the extreme underestimation of the Kohn-Sham gaps, which are differences in eigenvalues of the highest-occupied and lowest-unoccupied states.

necessary to consider the static dielectric response.

4.5 Fundamental gaps from effective screening

4.5.1 ΔSCF revisited

As described in the introduction to this chapter, the direct evaluation of the fundamental gap from differences in energy $E_{\text{gap,fundamental}} = E(N+1) + E(N-1) - 2E(N)$, i.e. the ΔSCF method, has been applied to atoms and molecules with good results.

For example, figure 4-9 shows the energy gaps for a series of atoms and molecules computed from ΔSCF using DFT-LDA (plotted with data from [3] and reference 14 therein). Unlike the Kohn-Sham gaps which are gross underestimations, the fundamental gaps are in reasonable agreement with experiments. The problem that arises, in doing the same for a solid, is in determining the meaning of N. In a standard electronic structure textbook [74], Martin claims that "it is not obvious how to carry out such a calculation... [and] there is no effect if the state is delocalized in an infinite system". In a recent paper on the issue of band gap corrections in the context of defect calculations [96], Lany and Zunger carried out the fundamental gap calculation in ZnO as a function of the number of electrons N per added/removed electron and concluded that the results so obtained "converge [to the Kohn-Sham band gap] in the limit of a dilute gas of free electrons and holes". A claim that the bulk band gap of silicon can be deduced from ΔSCF calculations of clusters of increasing size [97] was met with intense debate [54, 98].

In all these discussions, it is assumed that the added and removed electrons occupy and vacate Bloch states with definite crystal momentum **k** and infinite spatial extent, and hence the limit $N \to \infty$ must be made in order to recover the bulk value of the fundamental gap. However, we can equivalently transform to the Wannier basis and view the added and removed electrons as having a finite extent. It is typically possible to localize Wannier functions to atomic scale. Since the Bloch and Wannier states are related by a unitary transformation, physical observables are unaffected. Then we can recast the problem of finding the fundamental gap from DFT as that of the energy related to the addition/removal of a localized charge distribution to/from the existing charge density in DFT. We recall that DFT can be thought of as perturbations of a homogeneous electron gas, and that the inclusion of screening is crucial to the accurate description of the band gap. Our strategy, then, is to investigate the screening behavior of a homogeneous electron gas and to attempt to find transferrable quantities that would help in the determination of the fundamental gap. Although one might argue that semiconductors do not have metallic screening, one must remember that what is important is the screening behavior in the DFT calculations.
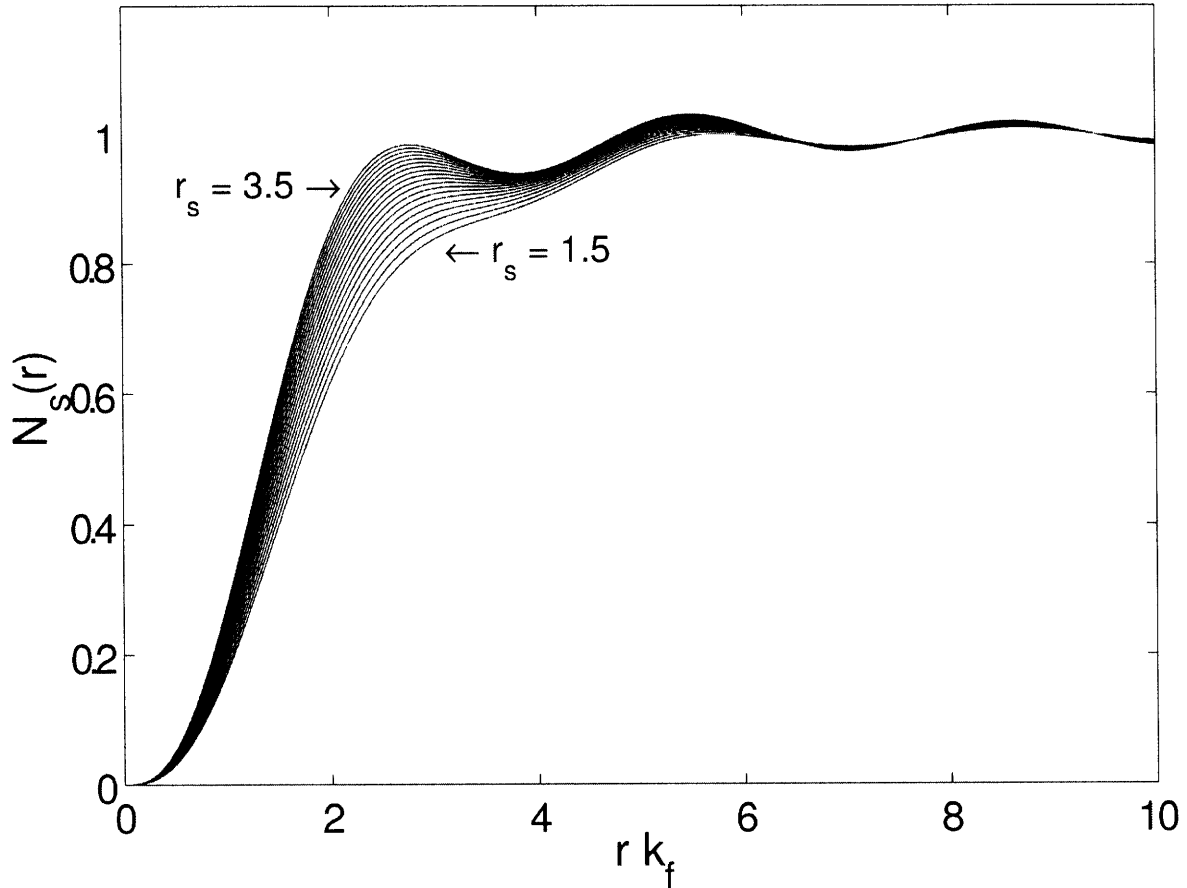
Figure 4-10: The integrated screening charge $N_s(r)$ within a radius r of a point perturbation in a homogeneous electron gas, plotted against the dimensionless distance rk_f, where $k_f = (3\pi^2 V/N)^{1/3}$ is the Fermi wavevector. The functional form is given by (4.22), as derived from the static RPA dielectric function in [4]. Different curves correspond to different values of the density parameter r_s, which ranges from 1.5 to 3.5. This range of r_s is representative of that of the compounds in our test set. Note that a smaller r_s implies a larger distance, in units of k_f^{-1}, before $N_s(r)$ first approaches unity, and that the differences between different values of r_s become negligible beyond a distance of 5-8 k_f^{-1}.

In general, once charges are added/removed from the computational cell, the system is metallic.

4.5.2 Effective screening radius

We consider the distribution of screening charges which gives a physical picture of the dielectric response. In the presence of a point charge Q, electrons rearrange to

shield, or screen, the charge. Beyond a certain distance, by the rearrangement of a total charge -Q, the screening is complete and the effects of Q are not felt. Beyond that distance, the integrated energy due to the distribution of screening charge does not increase appreciably. For a given wave-vector-dependent dielectric function $\varepsilon(\mathbf{q})$, the screening charge density per unit volume at a displacement $\mathbf{r}$ from Q is given by

$$\rho_s(\mathbf{r}) = Q \int_0^\infty \frac{d^3\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}\cdot\mathbf{r}} \left[1 - \frac{1}{\varepsilon(\mathbf{q})} \right] \qquad (4.21)$$

Here, we have taken the static (zero frequency) limit. Instead of calculating the dielectric function explicitly, which would bring us back to a static GW-like approach, we consider the RPA static dielectric function evaluated by Langer and Vosko [4]. Recall that the RPA corresponds to the neglect of vertex corrections and is used in the GW approximation. Using the RPA dielectric function in equation (4.21) and using spherical symmetry, the integrated screening charge as a function of distance from the perturbation r, which we will call $N_s(r)$, is given by

$$N_s(r) = \frac{2}{\pi} \int_0^r R\,dR \int_0^\infty q\,sin(qR) \left[\frac{1 + \frac{1}{q}(1 - \frac{q^2}{4})\ln\left|\frac{q+2}{q-2}\right|}{q^2/\eta + 1 + \frac{1}{q}(1 - \frac{q^2}{4})\ln\left|\frac{q+2}{q-2}\right|} \right] dk \qquad (4.22)$$

where $\eta = 0.3317r_s$. The radius parameter is defined as $r_s = \frac{3}{4\pi}\frac{V}{Na_o^3}$ with N, V and a_0 the number of electrons, the volume, and the Bohr radius, respectively. Plots of $N_s(r)$ for $r_s = 1.5 - 3.5$ obtained by numerical integration of (4.22) are shown in figure 4-10. It is clear that beyond a distance of about 5-8 k_f^{-1}, where $k_f = (3\pi^2 N/V)^{1/3}$ is the Fermi wavevector, the integrated screening charge is very close to unity. There is some ambiguity as to precisely what "close" means, of course, and different dielectric functions give slightly different details of the Friedel oscillations, first unity crossing, etc. There is also some dependence of the details on r_s, as shown in figure 4-11. Suppose we determine that the screening is complete at a distance of $r^* k_f^{-1}$, where r^* may depend on r_s, then the corresponding volume is given by

$$V^*(r_s) = \alpha r^*(r_s)^3 \frac{V}{3\pi^2 N} \qquad (4.23)$$
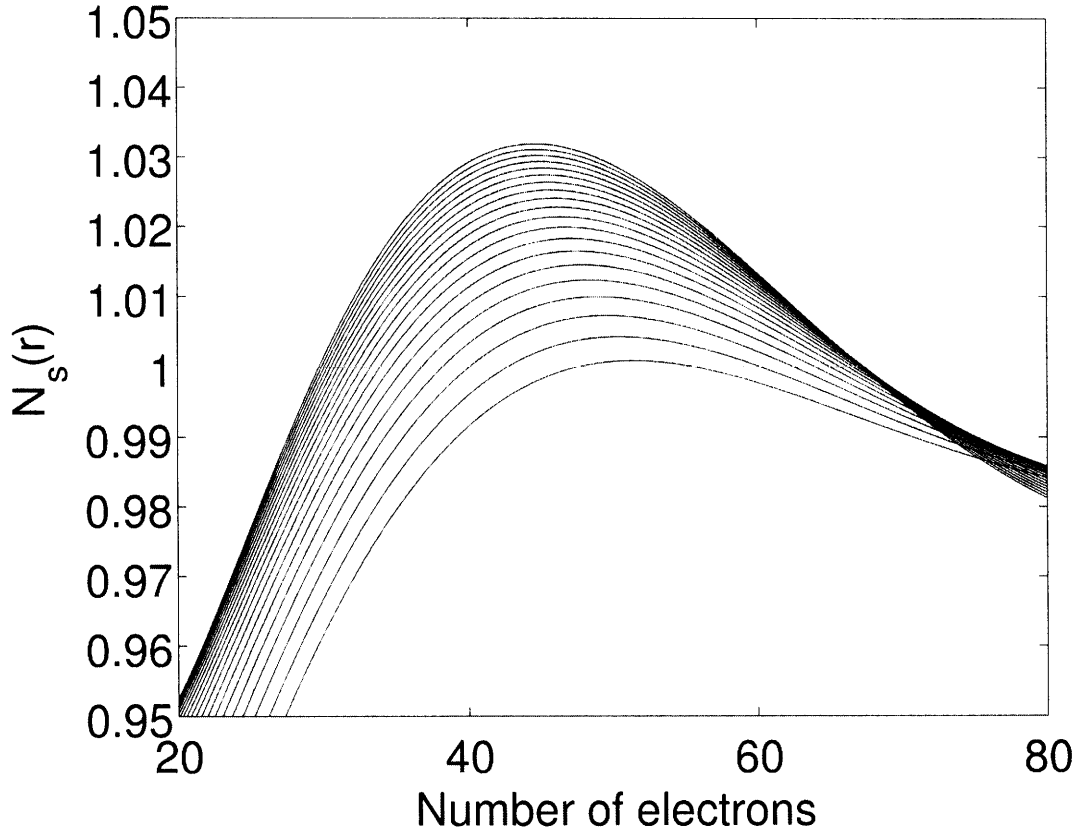
Figure 4-11: A close up of figure 4-10 in the region near complete screening, i.e. where $N_s(r) \approx 1$. Note the fine vertical scale. The screening charge is plotted against the dimensionless volume $(2rk_f)^3$, which corresponds to the number of electrons within the screening volume.

where α is a geometrical factor (equal to $4\pi/3$ for a sphere). Within V^*, the total number of electrons is

$$N^*(r_s) = \frac{\alpha}{3\pi^2} r^*(r_s)^3. \tag{4.24}$$

This means that for any given r_s, there is a fixed number of electrons $N^*(r_s)$ that completely screen the charge Q.

The preceding argument is derived in terms of a fixed point charge screened by electrons, but the screening charge distribution of electrons due to each other, i.e. the exchange-correlation hole, has a similar description [99]. Given an electron at the origin, the pair distribution function $g(r)$ describes the probability of finding another electron at a distance r, and is found from the wave-vector integral of the inverse

94

dielectric function in a way similar to equation (4.21). At small r, $g(r) \ll 1$ due to the Pauli exclusion principle (exchange) and at large r, $g \to 1$. This is referred to as the "exchange hole" because the depression of likelihood of finding another electron around one is as if an electron carries around with it a hole. Correlation refers to the remaining electronic interaction after subtracting the mean-field Hartree and exchange effects, and reduces the depth of the exchange hole without affecting its range. The exact pair distribution functions in a homogeneous electron gas calculated by quantum Monte Carlo (QMC) [100] have much the same form as figure 4-10 (except that the oscillations are very much damped) and saturate at about 5-7 k_f^{-1} regardless of r_s.

The precise form of the sreening charge density or the pair distribution function (exchange-correlation hole) is not expected to agree between DFT and exact results. However, as has often been remarked upon, the Local Density Approximation (LDA) owes its success largely to the fact that the spherical average, the extent, and the integral of the exchange-correlation hole reproduces reality well [99]. To obtain the fundamental gap, we are similarly interested in an integral quantity, namely the total energy due to the added/removed electron and its screening charge distribution. We do, however, know that DFT has a tendency to delocalize charges. Therefore we suspect that as long as we confine the added charge to a volume that is commensurate with the range of the screening effects, the integrated energies thus obtained will be reasonably correct. In other words, we propose a solution to finding the appropriate N for which to evaluate $E_{\text{gap,fundamental}} = E(N+1) + E(N-1) - 2E(N)$ in a solid, namely, N^* in equation (4.24), or, equivalently, the number of electrons contained in the exchange-correlation hole.

There are two quantities to determine, the geometric factor α and the radius where the charge is confined r^*. Real materials do not have spherical symmetry, of course, so there are some ambiguities in the choice of α, two reasonable choices being spherical ($\alpha = 4\pi/3$) and cubic ($\alpha = 8$). For r^*, in comparing the pair correlation functions from QMC with the integrated screening charge in figure 4-10, we find that the range 5 - 7 is reasonable. Alternatively, since r^* appears to depend weakly and

95

inversely on r_s, we can choose r^* to scale, e.g. by the distance at the first unity crossing of N_s. In figure 4-11 we have plotted N_s against the number of electrons using cubic geometry. Taken together, we find that N^* is roughly between 20 and 90. Because we do not know the precise value of r^*, we proceed with three possibilities: to use a fixed value for all compounds (which we will call the "$1N^*$" model), to use a r_s-dependent range based on first unit crossing of N_s (the "$N^*(r_s)$" model), and to use two values depending on a cutoff in r_s (the "$2N^*$" model). In reality, of course N^* will be somewhat different for each compound, possibly depending on the geometry, details of screening, and, especially, how far the charge density deviates from the homogeneous electron gas. The test of our hypothesis, then, is in whether the experimental band gaps are adequately predicted for a large number of compounds with a wide variety of chemistries from the use of one or two parameters which are present in the N^* models.

The above argument can be re-stated in terms of the derivative discontinuity of the exchange-correlation functional in DFT. The difference between the Kohn-Sham gap and the fundamental gap has been shown to be equal to the derivative discontinuity of the exchange correlation energy at integer particle number [101], i.e.

$$
\begin{aligned}
E_{\text{gap,fundamental}} &= E_{\text{gap,Kohn-Sham}} + \lim_{\delta N \to 0} \left[\left. \frac{\partial E_{xc}}{\partial N} \right|_{N_o + \delta N} - \left. \frac{\partial E_{xc}}{\partial N} \right|_{N_o - \delta N} \right] \\
&\equiv E_{\text{gap,Kohn-Sham}} + \Delta_{xc}
\end{aligned}
\tag{4.25}
$$

where N_o is the number of electrons in the original system. The fact that gaps are severely underestimated in DFT with LDA/GGA is attributable to two possibilities: that the $E_{\text{gap,Kohn-Sham}}$ from semilocal functionals are not the true Kohn-Sham gap, or that Δ_{xc} is a significant portion of the fundamental gap. Recent work in "exact Kohn-Sham" calculations favors the latter explanation [102]. For a local or semilocal functional like LDA/GGA, Δ_{xc} is zero, thus if Δ_{xc} of the exact exchange-correlation functional is large compared to the fundamental gap, then so is the error in LDA/GGA Kohn-Sham gap. We have seen in figure 4-9, however, that the total energies of atoms

and molecules are accurate at integer electron numbers, implying that

$$\Delta_{xc} = \int_{N_o}^{N_o+1} \frac{\partial E_{xc,LDA}}{\partial N} - \int_{N_o}^{N_o-1} \frac{\partial E_{xc,LDA}}{\partial N}. \qquad (4.26)$$

While the exact exchange-correlation functional has constant slope between, and derivative discontinuity at, integer particle numbers, LDA has a continuously-varying slope whose integral matches that of the exact exchange-correlation functional at integer numbers. Once again, we have an integral quantity that is physically accurate despite errors in the details of the integrand. Because the value of Δ_{xc} is obtained by the integral of the exchange-correlation potential, in a homogeneous electron gas the only plausible integration limits are placed by the extent of the exchange-correlation hole, which we have seen is 5-7 k_f^{-1}, and contains roughly 20-90 electrons. We arrive again at the previous conclusion, that in order to calculate the fundamental gap in a solid using the ΔSCF method, we should consider the addition and removal of one in N^* electron, where N^* is in the range stated above.

4.5.3 Computational details

From the above discussion, the relevant quantities to consider in order to calculate the fundamental gap in solids are the energy differences due to the addition and removal of one per N^* electron, where N^* is largely material-independent but may depend on the exchange-correlation functional. We do not, however, have the precise value of N^*, or know its dependence on r_s, if any. Our strategy is therefore to find this information based on the measured gaps of compounds in our test set. The fundamental gap is evaluated using

$$E_{\text{f}}(\delta) \equiv \frac{1}{\delta}[E(N_o + \delta) + E(N_o - \delta) - 2E(N_o)] \qquad (4.27)$$

where N_o is the number of valence electrons in a unit cell, in which δ electrons are added or removed. In the range of δ considered, this is essentially equal to

$$E_\mathrm{f}(\beta) \approx E(\beta N_o + 1) + E(\beta N_o - 1) - 2E(\beta N_o) \tag{4.28}$$

where $\beta = 1/\delta$. In the end we would like to have $\beta N_o = N^*$, the number of electrons within a screening volume as described above. In an effort to see if one- or two-parameter N^*-models are adequate, we calculate, for all 134 compounds in our test set, $E_\mathrm{f}(\delta)$ for a set of 8 values of δ ranging from $1/27$ to 1. Interpolations are made to determine the energies for intermediate values of δ. The energies are computed using ground state DFT in the plane wave pseudopotential code VASP as described in section 4.2.2. The contribution of each ion to the numbers of valence electrons N_o for main group elements are assigned according to the usual octet rule. For transition metals, all outermost d- and f-shell electrons are counted as valence electrons. There will be some scatter of errors due to the choice, but it is important to be consistent in order to maintain predictive power. The number of valence electrons used in the DFT calculations may not agree with this assignment, due to the necessity of including semicore electrons for some elements, e.g. the alkali and alkaline-earth metals.

The $1N^*$ model

We begin with the simplest model $1N^*$, in which a universal N^* is used to describe all compounds. This is a one-parameter model, with the *ab initio* energies and experimental band gaps as inputs. For each possible value of βN_o, we look up the fundamental gap values calculated using equation (4.27) from DFT and compare to the experimental values, and the error is minimized to find the optimal N^*. The mean absolute error is used to determine the goodness-of-fit, in lieu of the r.m.s. error. This is because we have a one-parameter model, so the risk of overfitting is minimal; further, we do not want to allow the value of N^* to be determined by a few outliers. The fitting is performed for all 134 compounds, 121 compounds with $E_g > 0.5$ eV, and 94 medium-gap ($1 \leq E_g \leq 4$) compounds. The results are shown in figure 4-12. In the
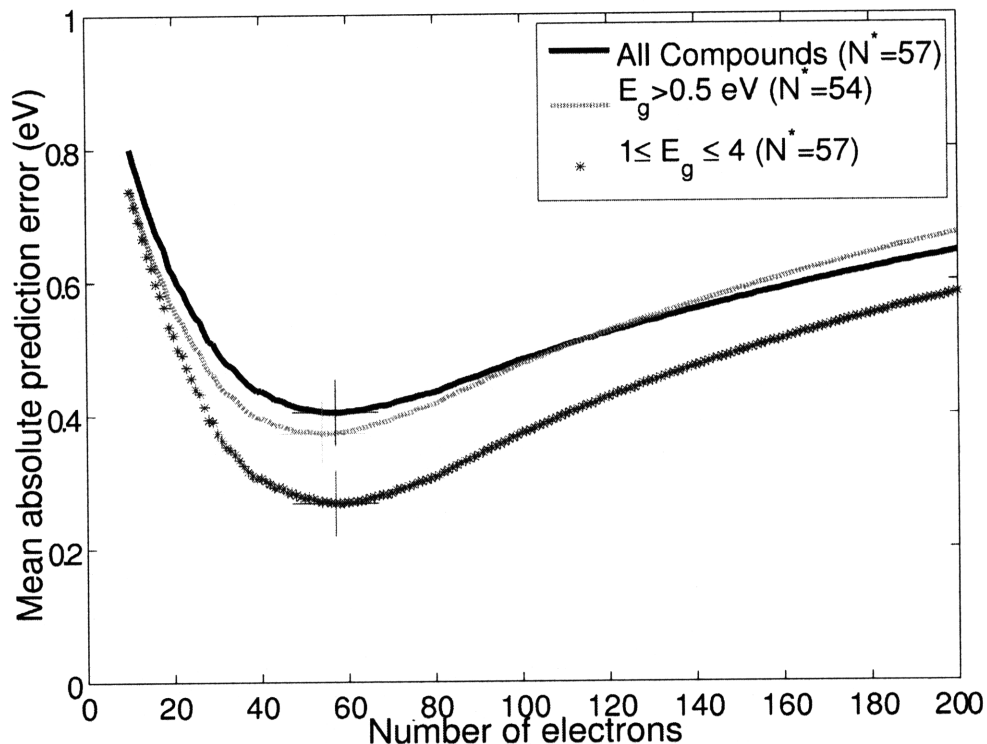
Figure 4-12: The mean absolute errors used to determine the value of N^* in the $1N^*$ model. The three curves do not differ appreciably in the location of the minimum.

three subsets we find the best N^* to be 57, 54 and 57 respectively, corresponding to $r^* \sim 6k_f^{-1}$ for a cubic configuration, which is reasonable. The mean absolute errors also do not change appreciably within a ± 10 range of N^*. In subsequent analysis we drop the middle group ($E_g > 0.5$ eV).

Cross validation

We perform leave-one-out and leave-N-out cross validation procedures in order to identify outliers and to verify that the parameter N^* is stable. In performing cross-validation, we leave one or N of the compounds out of the fit for N^*. If the values of N^* obtained are not strongly dependent on which compounds are left out, then the model parameter is stable. Figure 4-13 shows the results for the leave-one-out validation. A magnified view of the leave-one-out errors for the medium-gap group (bottom panel) shows that there is one compound whose omission significantly reduces the error (compared to the rest in the group). We find that it is again RbAu, the compound

that was the most severe outlier in the G_0W_0 calculations, that is near a metal-semiconductor transition, and for which experimental data were scarce. Therefore, we omit RbAu from subsequent cross-validation analyses, but include it in the fitting for N^*, the fundamental gap prediction and accuracy statistics. The spread in errors is larger for the all-compound group compared to the medium-gap group simply because the former contains the outliers in E_g. Indeed, the outliers similarly identified in the all-compound group were CaO, SrO, BaO and BeO (not shown).

We perform a leave-30%-out cross validation procedure on all compounds and medium gap compounds, and the results are shown in figure 4-14. In 100 iterations, the change in N^* upon the removal of a full 30% of the compounds was always less than 9, and the standard deviation was 2.8 and 2.6 for all compounds and medium-gap compounds, respectively. Again, the stability is due to the fact that there is only one parameter. There are no apparent clusters with different N^* values that we can discern from such random sampling.

The $N^*(r_s)$ model

We remarked upon the fact that the form of the integrated screening charge depends somewhat on the density parameter r_s. In a somewhat *ad hoc* way, we attempt to model the variation with the radius at which the screening density first reaches within a threshold of unity which we identify as r^*. There is a choice of the threshold, and also the geometric factor α. It was found, however, that such a variation of r^* with r_s does not improve on the one-parameter $1N^*$ model in terms of mean absolute errors, mostly because of a large increase in the errors of the outliers (see figure 4-22 for more details). Therefore, we do not find that the strongly r_s-dependent features in the RPA screening charge to be significant for our purpose.

The $2N^*$ model

Since we are unable to detect any disjoint subgroups in the dataset through a random leave-N-out cross validation, we attempt to divide the compounds into various categories by unit cell volume V_0, number of electrons in the unit cell N_0, r_s in different
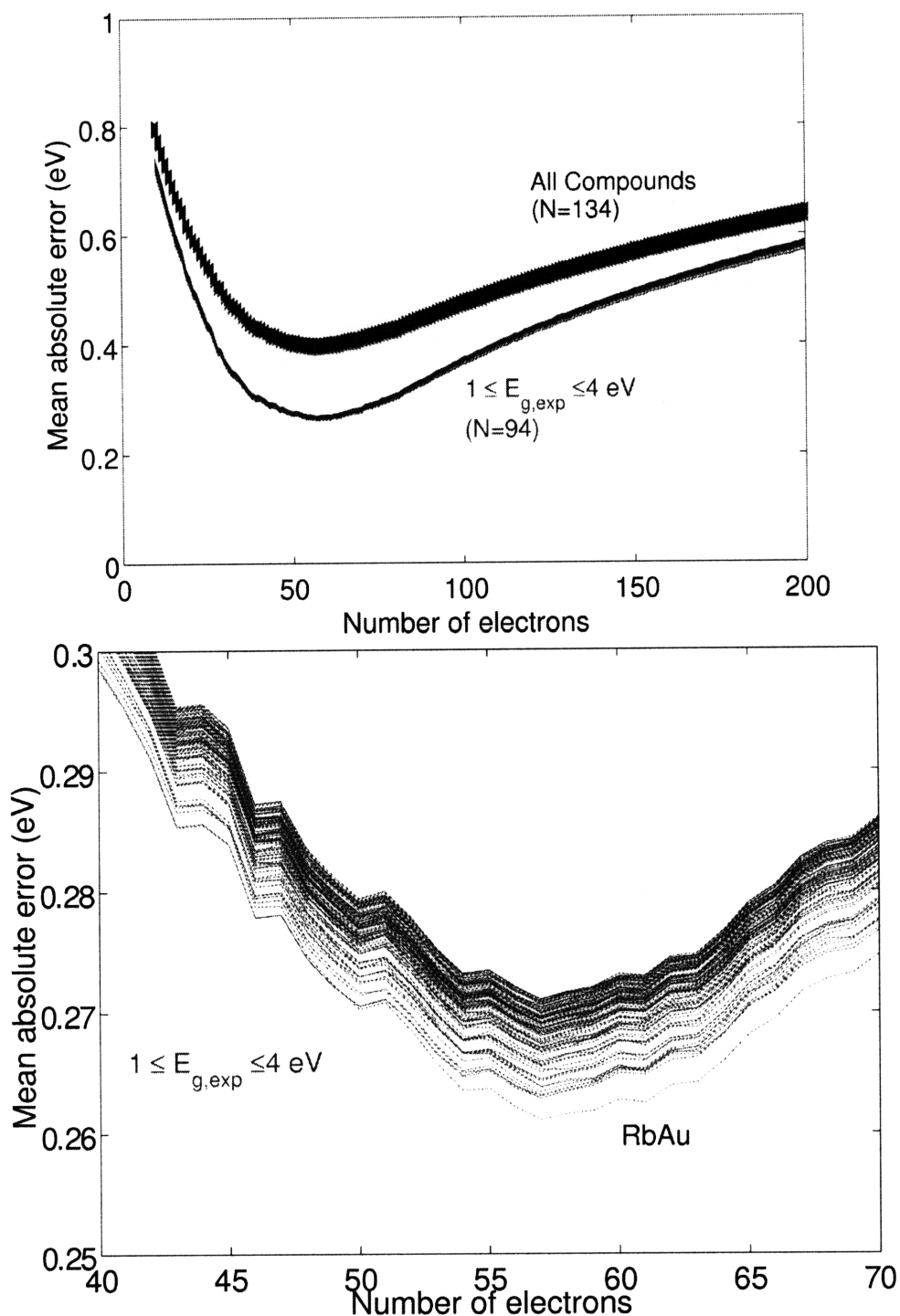
Figure 4-13: The leave-one-out cross validation curves, showing the change in the mean absolute error of the predicted gap as a function of the number of electrons per added/removed electron βN_o. The variation in the minimum error and best N^* is small when individual compounds are taken out of the fit. This is expected since there is only one parameter. The bottom panel shows a much expanded portion for the medium-gap compounds, showing that RbAu is a clear outlier, since when it is left out of the fit the mean errors are clearly much lower for most values of βN_o. However, leaving RbAu out does not seem to change the value of N^* appreciably.

101

Figure 4-14: The leave-30%-out validation curves. The best value of N^* is stable.

ranges of values, and whether the compound contains a transition metal. The parameters V_0 and N_0 do not seem to be significant in separating compounds into di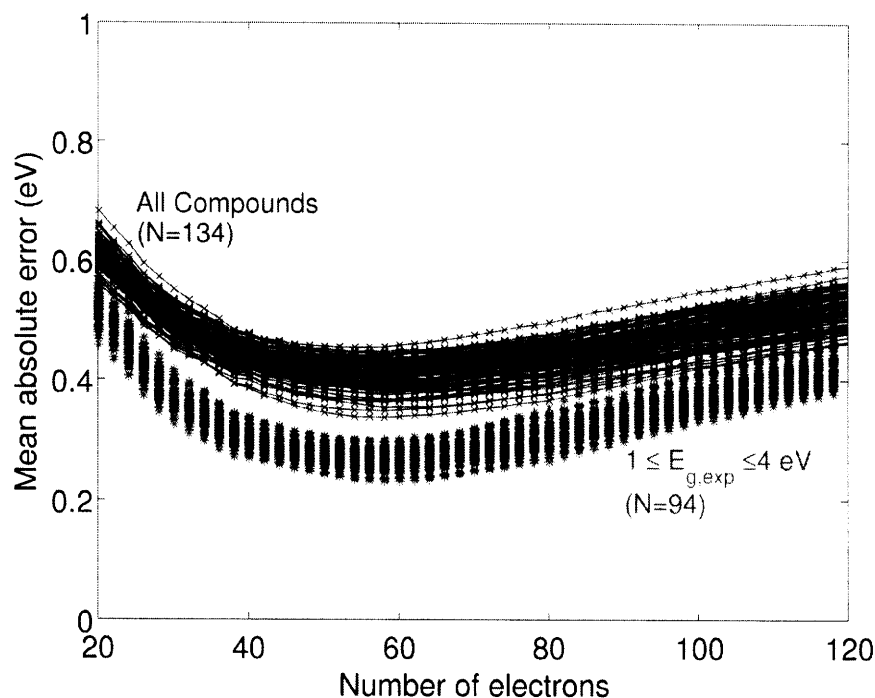sjoint groups. As shown in the top panel of figure 4-15, however, the set of compounds with a transition metal element is biased towards larger values of N^*. This is understandable since we have made the choice of including all the outermost d- and f-electrons instead of deciding on a case-by-case basis the effective number of valence electrons, and d- and f-electrons are clearly less homogeneous-electron-gas like. However, the alternative of including none of the d- or f-electrons would leave Cu, for example, with only one electron, which is clearly too few. We found, however, that the separation of transition metal vs non-transition metals is not as significant as that between $r_s \geq 2$ and $r_s < 2$, as can be seen from the bottom panel of figure 4-15. This leads us to choose the two ranges of values for r_s and we find the best value of N^* to be 38 for $r_s \geq 2$ and 60 for $r_s < 2$. The result gives a slight improvement over the $1N^*$ model in mean error, but reduces the maximum error significantly. We must caution, however, that this may be only true due to some bias in the test set of compounds.
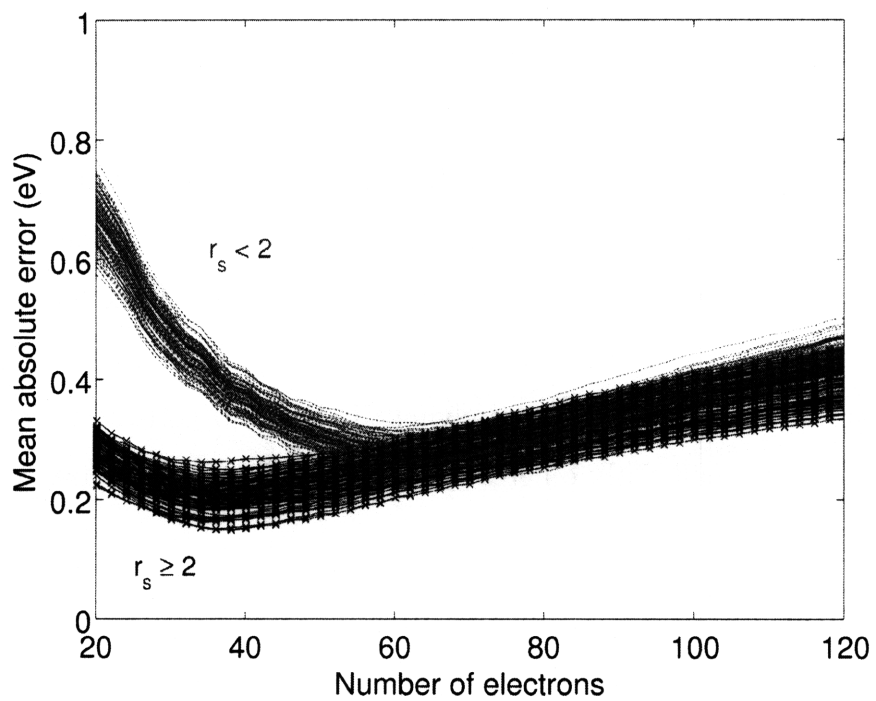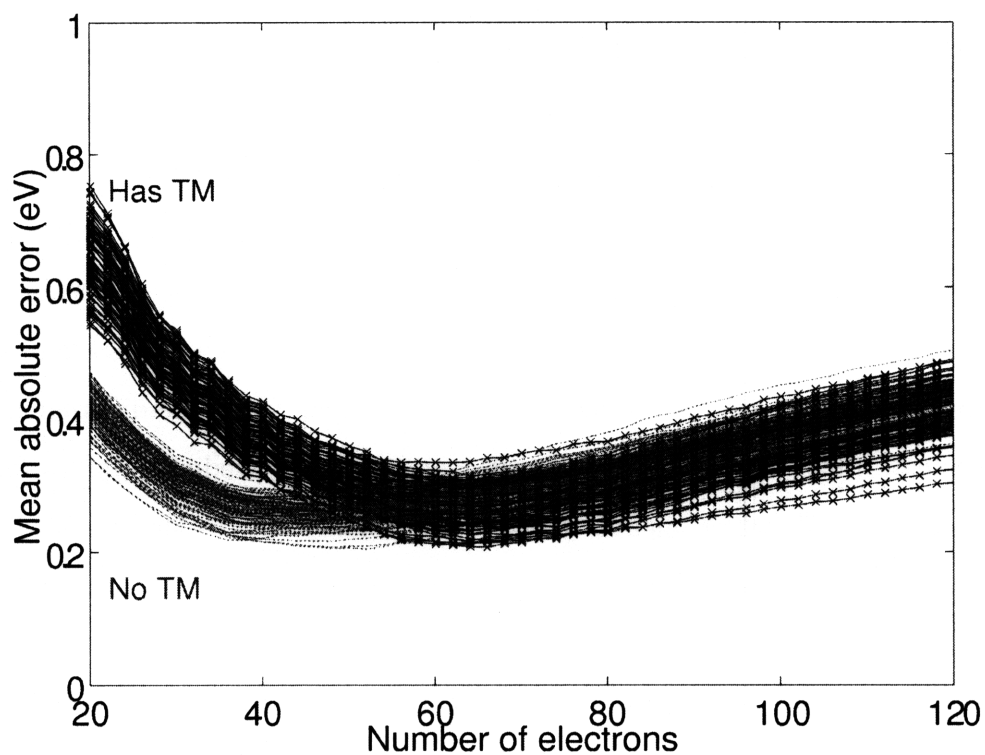
102

Figure 4-15: Leave-30%-out cross validation for compounds with and without transition metals, and for compounds with r_s greater than or smaller than 2.

4.5.4 Computed fundamental gaps

The fundamental gaps obtained from the above procedures are shown in figures 4-16 to 4-24. The main results are that any of the three different N^* models reduces the band gap prediction errors by 60-75% on average compared to the Kohn-Sham gaps, and also reduces the spread of the prediction errors by 40-65%. The improvement is especially marked in medium-gap compounds, where now 50% of the band gap values are predicted to within 10% of the experimental ones, and 90% are predicted to within 30%. There are no marked differences between the three different models, and the mean difference between the largest and smallest predicted values for each compound is 0.1 eV, which means that the uncertainty due to the use of different models is about the size of the symbols in the scatter plots. The predicted values of the band gaps in the 1 - 4 eV range are consistent with experimental values if one assumes standard deviations in experimental values of 0.2 - 0.3 eV, which is a reasonable assumption.

Effects of U

As mentioned in section 4.2.2 the ground state DFT calculations are performed with the DFT+U correction for transition metal oxides and halides. The fundamental gap values obtained vary linearly with the value of U, as we can see for the example of AgF from Table 4.5, which has an experimental gap of 2.8 eV. It is important to note, however, that even without applying the DFT+U correction, for AgF the fundamental gap is significantly different from zero (1.8 eV) unlike the Kohn-Sham gap.

Insulators

The band gaps for alkali-earth metal oxides (BeO, CaO, MgO, SrO and BaO), shown in Table 4.6, are again underestimated by 1.6 - 2.2 eV, with the underestimation being largest for CaO as in the G_0W_0 results discussed above. These compounds are far removed from the homogeneous electron gas limit and closer to the atomic limit,
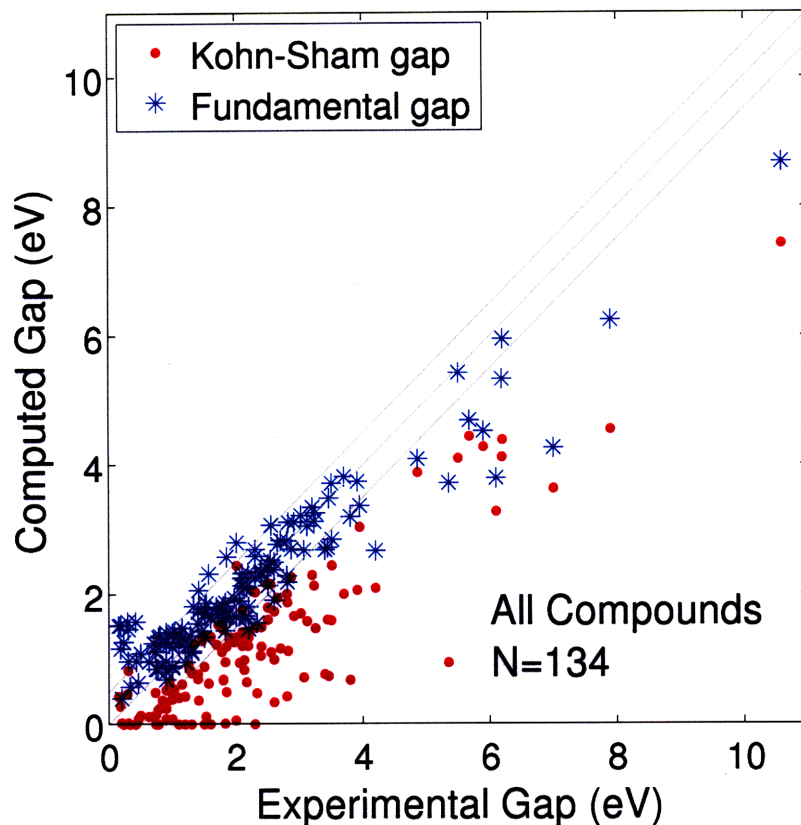
Figure 4-16: Calculated fundamental gaps for all compounds, plotted against the experimental gaps. Also plotted are the corresponding Kohn-Sham gaps. Straight lines are 1-to-1, and ± 0.5 eV deviations. The results shown are from the $2N^*$ model.

Table 4.5: The dependence of the Kohn-Sham and fundamental gaps on the U parameter in DFT+U calculations for AgF, which has an experimental gap of 2.8 eV. As is apparent, the value of the fundamental gap obtained, as with the Kohn-Sham gap, depends linearly on the U parameter. The fundamental gap does not suffer from the qualitative failure of the Kohn-Sham gap.

U (eV)	Kohn-Sham gap (eV)	Fundamental Gap (eV)
0.0	0.0	1.8
1.0	0.0	2.0
2.0	0.2	2.2
3.0	0.4	2.4
5.0	0.9	2.9
7.0	1.3	3.3
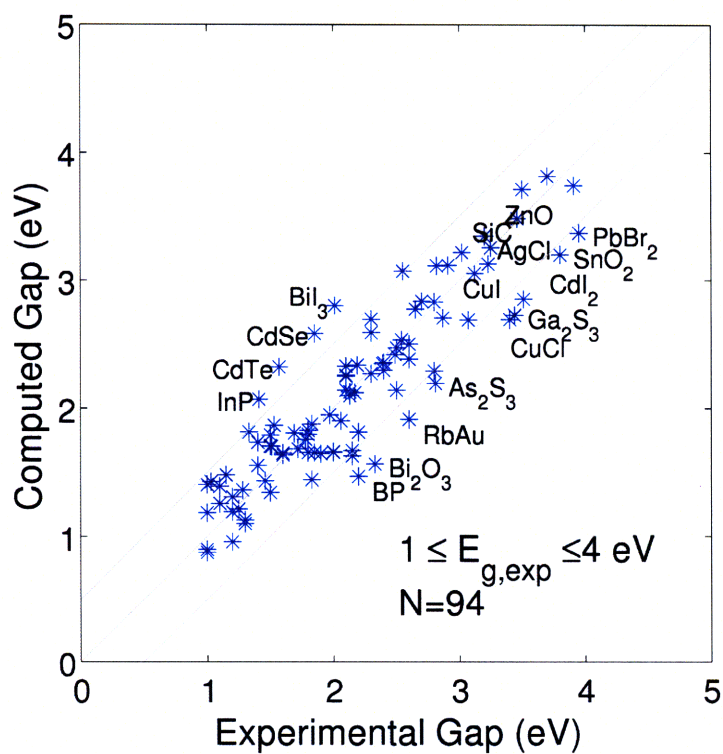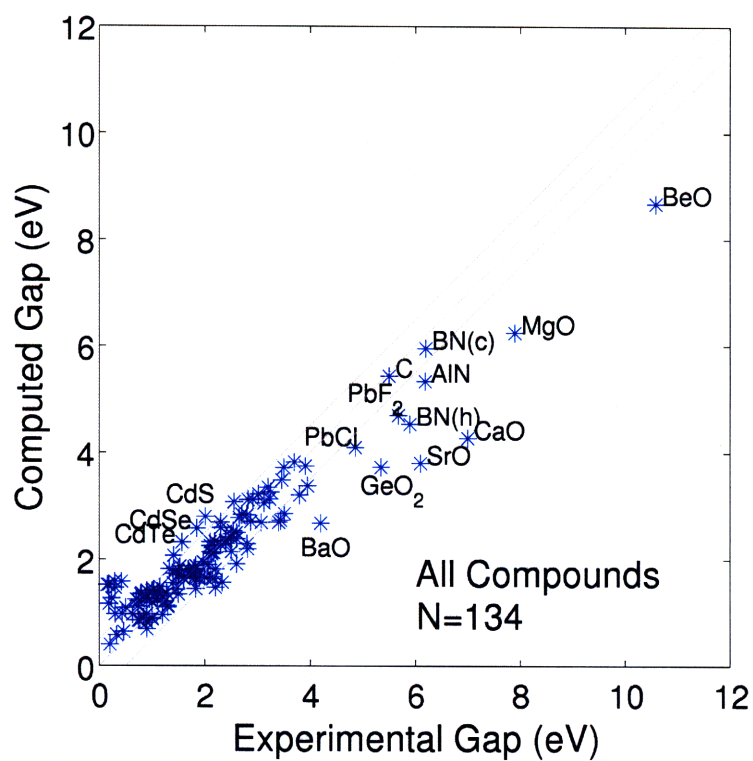
Figure 4-17: Previous figure annotated with compound names, for all compounds (top) and medium-gap compounds (bottom).
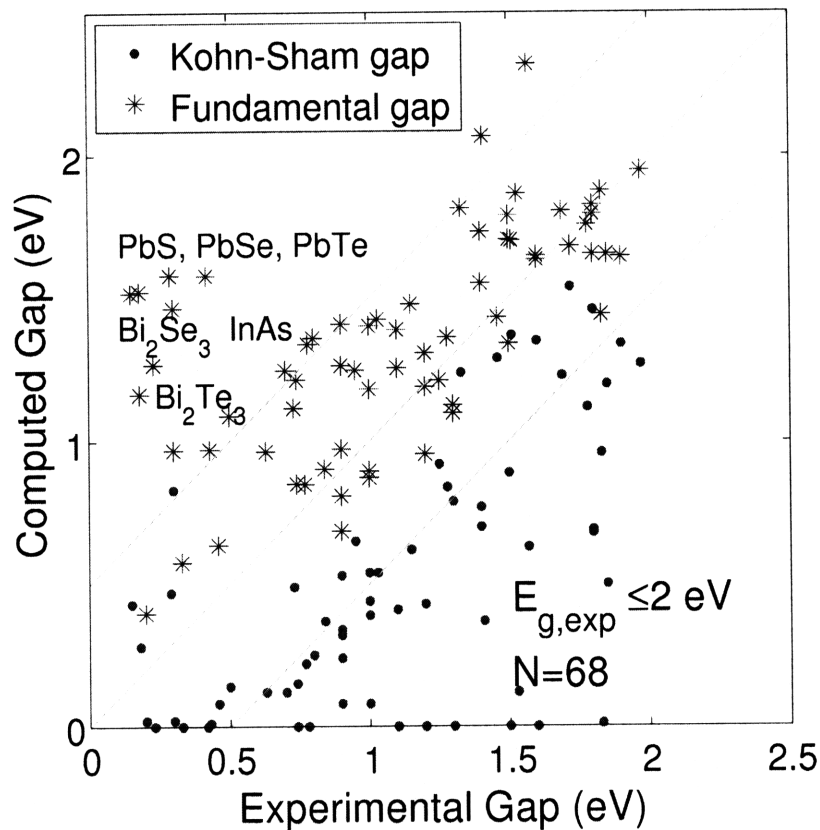
Figure 4-18: A close up of the small gap (0 - 2 eV) compounds. There is large overestimation for Bi and Pb compounds, for which the Kohn-Sham gaps are overestimated owing to the neglect of spin-orbit coupling, as explained in section 4.2.2. The mean absolute error of compounds with an experimental gap of 0.5-1 eV is similar to the medium gap compounds (0.29 eV), although the percentage error is of course higher for these small gap compounds.
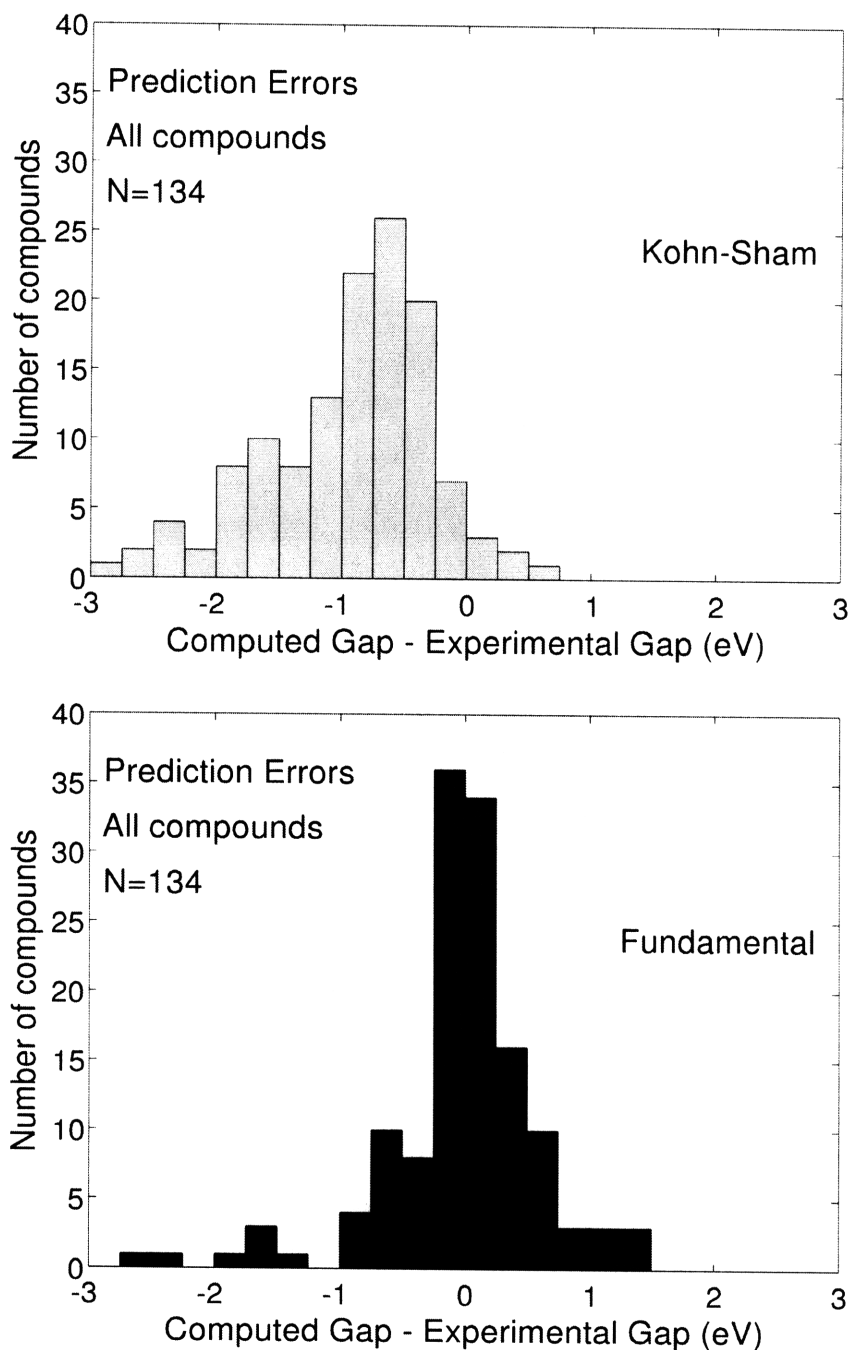
Figure 4-19: Histograms, for all 134 compounds, for the prediction error, i.e. calculated minus experimental values, for Kohn-Sham gaps (top) and fundamental gaps (bottom). The compounds that remain underestimated are wide-gap oxides. The horizontal and vertical scales are the same for both panels.
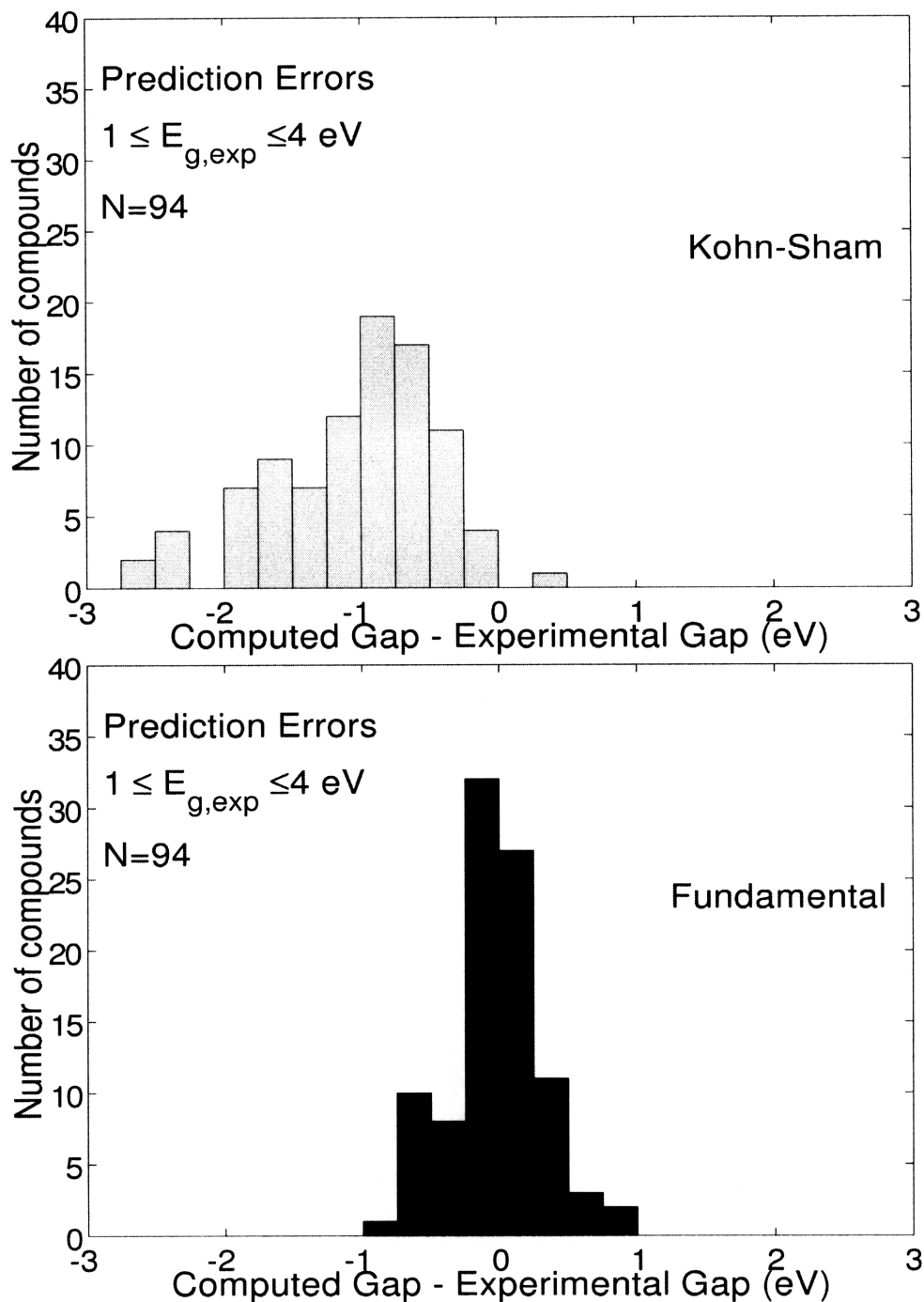
Figure 4-20: As in previous figure, for the 94 medium gap compounds. The center around 0 for fundamental gaps is a result of the fitting of N^*, but the much reduced width of the distribution (σ is reduced from 0.62 eV to 0.22 eV; see figure 4-23 for percentiles) indicates that significant information has been captured in our model.
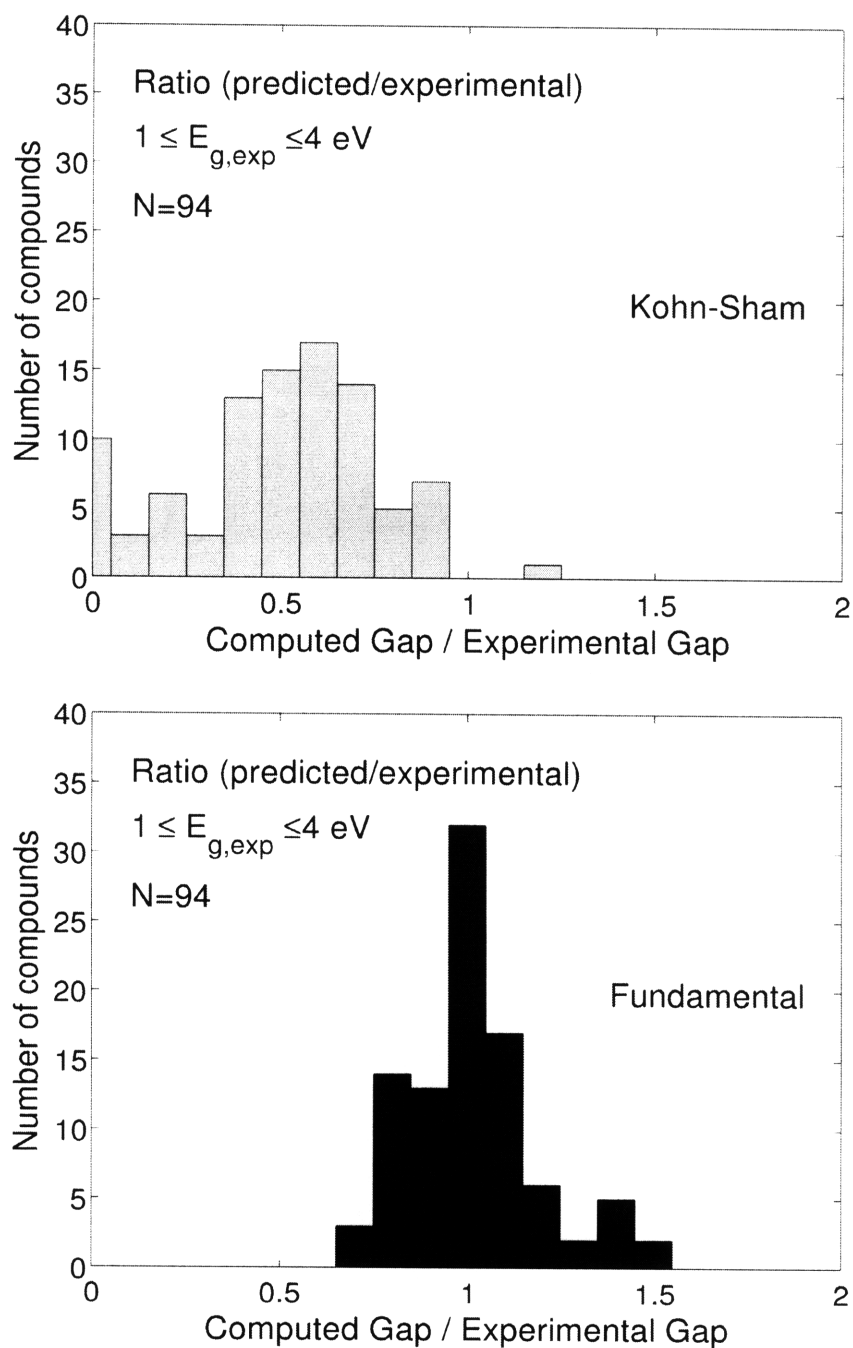
Figure 4-21: Histograms of the ratios of predicted to experimental values, for medium-gap compounds only. The mean relative prediction error is 50% for Kohn-Sham gaps and 13% for fundamental gaps.
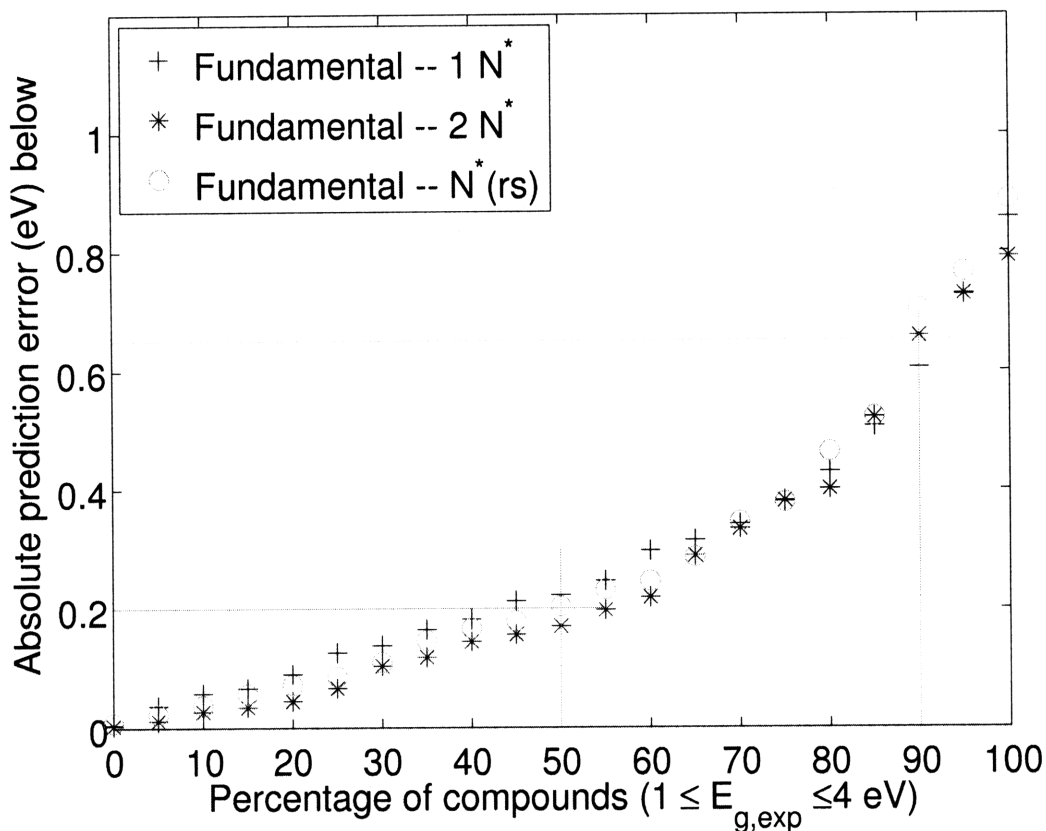
Figure 4-22: The prediction errors by percentiles using various effective screening models. A total of 94 compounds with experimental band gaps between 1 and 4 eV are included. The straight lines are guides to show that the prediction error is less than ~0.2 eV for 50% of compounds and less than ~0.65 eV for 90% of the compounds. The three models differ in the determination of optimal screening radius and hence number of electrons within, denoted by N^*. The notations $1N^*$, $2N^*$, and $N^*(r_s)$ correspond to using a constant number for all compounds, using 2 numbers for $r_s \geq 2$ and $r_s < 2$, and using the volume corresponding to the first unity crossing of the integrated screening charge in figure 4-11, respectively. The three models have one to two adjustable parameters, which are determined by fitting. The different models have very similar prediction errors.
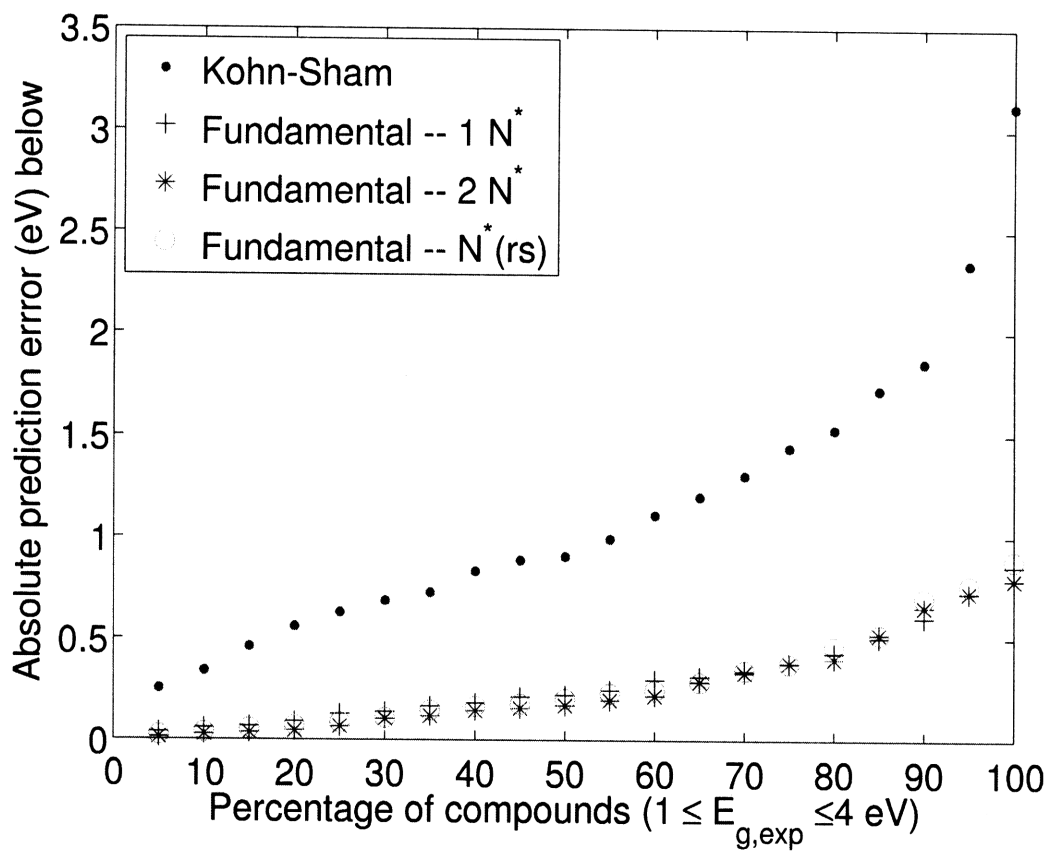
Figure 4-23: As in figure 4-22, but with the error by percentiles for the Kohn-Sham gaps included. Note the change in scale – the maximum error now exceeds 3 eV. Compared to the errors in the Kohn-Sham gap the different models give virtually identical results.
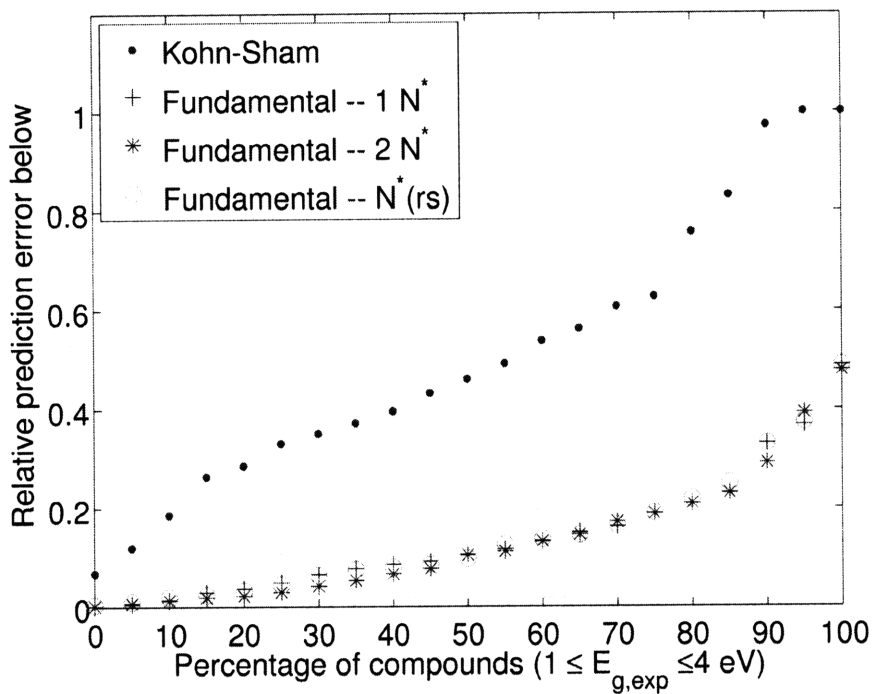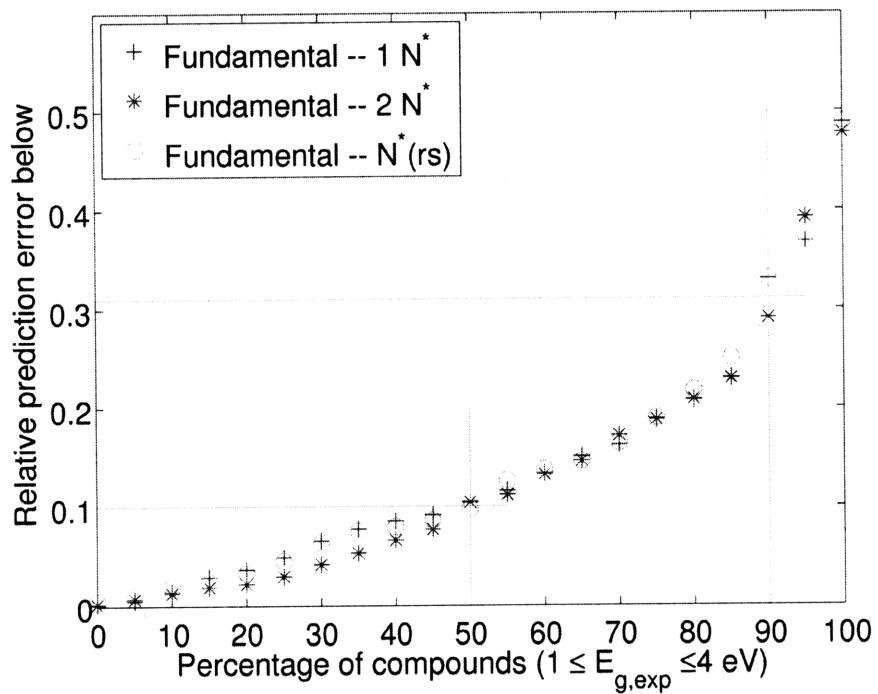
Figure 4-24: As in figures 4-22 and 4-23, but with relative instead of absolute errors. Note the two different vertical scales. Regardless of the model used, 50% of compounds have gaps predicted to better than 10% of their experimental values using the effective screening radius. The 90th percentile is a 30% error.

Table 4.6: Various gap values for alkali-earth oxides. All numbers are in eV. Note that the fundamental gaps are about 1 eV below the G_0W_0 gaps, which are themselves about 1 eV below the experimental gaps. In both cases the CaO gap is the most underestimated.

Compound	Experimental	G_0W_0	Fundamental	Kohn-Sham
BeO	10.6	NA	8.7	7.4
MgO	7.8	6.7	6.3	4.6
CaO	7.0	5.3	4.3	3.6
SrO	6.1	4.9	3.8	3.3
BaO	4.2	3.2	2.5	2.1

and therefore our reasoning in determining N^* is not well-justified.

For many other wide gap compounds, the errors are rather systematic. The lead halides PbF_2, $PbCl_2$, and PbF_2 are underestimated by 1.0, 0.8 and 0.6 eV, respectively, which is commensurate with the degree of ionicity and hence the expected dielectric contribution of the lattice. Note that for graphite-like hexagonal BN, the gap is underestimated by 1.4 eV, while for cubic BN the prediction is accurate. This is possibly due to the fact that screening is two-dimensional in hexagonal BN and is ill-described by the three-dimensional homogeneous electron gas.

Comparison of fundamental and G_0W_0 gaps

Figure 4-25 shows the comparison between band gaps calculated from G_0W_0 and the fundamental gaps calculated (with the $2N^*$ model, but the results are indistinguishable), both plotted against the experimental value. The performance of the two methods are comparable in the medium gap range in terms of reproducing experimental results, but the G_0W_0 gaps have a larger range of errors. It is certainly the case that the G_0W_0 results are more improvable, as it is a physically more accurate model compared to our procedure for obtaining the fundamental gap. However, as mentioned above, the computational cost of doing so is large.
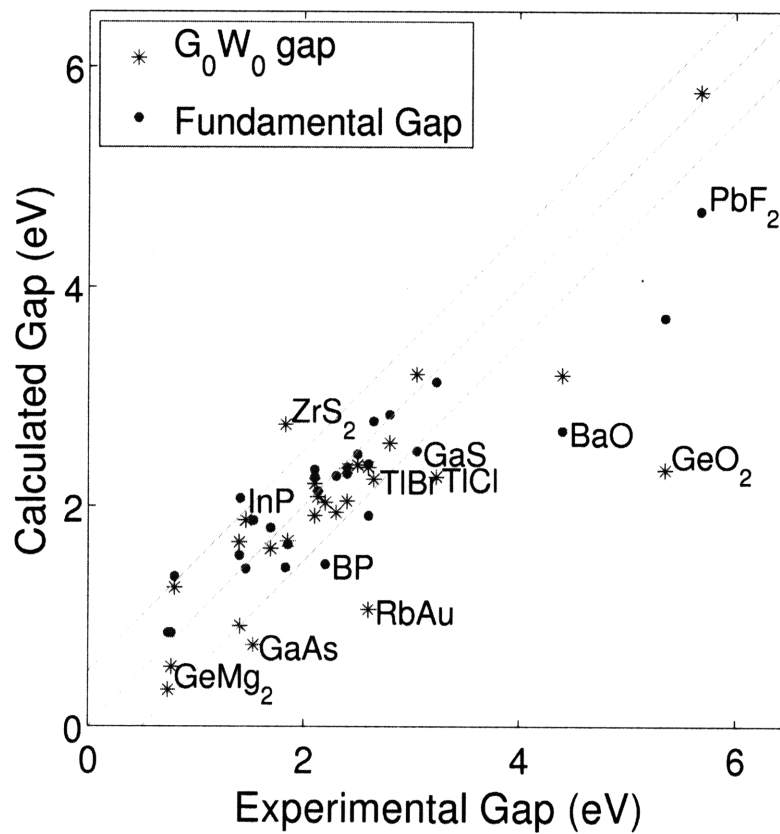
Figure 4-25: A comparison of the G_0W_0 and fundamental gaps, for compounds with experimental gaps in the range 0.5 to 6 eV. The labels correspond to calculated gaps that are further away from experimental values between the two methods.

4.5.5 Physical interpretations

The fact that we are able to make semi-quantitative predictions of the band gaps of a wide variety of compounds using such a simple argument as the screening radius or the extent of the exchange-correlation hole, with one or two parameters, is perhaps surprising and we would like to have a more physical understanding of the mechanism by which this method is able to do so. We return to the Kohn-Sham and quasiparticle equations:

$$H_{KS}\psi_{nk}(\mathbf{r}) \equiv \left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ion}} + V_{\text{Hartree}} + V_{\text{xc}}(\mathbf{r}) \right] \psi_{nk}(\mathbf{r}) = \epsilon_{nk}\psi_{nk}(\mathbf{r}) \qquad (4.29)$$

and

$$H_{KS}\phi_{nk}(\mathbf{r}) + \int \left[\Sigma(\mathbf{r}, \mathbf{r}', \epsilon'_{nk}/\hbar) - V_{KS}(\mathbf{r}') \right] \phi_{nk}(\mathbf{r}')d\mathbf{r}' = \epsilon'_{nk}\phi_{nk}(\mathbf{r}) \qquad (4.30)$$

In general the quasiparticles corresponding to an electron/hole excitation, ϕ_c and ϕ_v, will not have the form of the Kohn-Sham eigenstate of the conduction band minimum (CBM) ψ_c or valence band maximum (VBM) ψ_v. Because the Kohn-Sham eigenstates are complete, however, we can expand the quasiparticle wavefunctions in terms of the Kohn-Sham wavefunctions, for example,

$$\phi_c = \sum_i c_i \psi_i \qquad (4.31)$$

Wavefunction updates are sometimes performed in self-consistent GW approaches, but it is not involved in the G_0W_0 approximation. Assume that the self-energy operator does not mix states with very different energies, which is a valid assumption if we consider states at the band edges and if there is no significant transfer of spectral weight, i.e. no strong correlation effects. Then the sum in equation (4.31) is over a small set of Kohn-Sham states near the CBM and VBM. In other words, we expect the quantities $|c_i|$ to decrease as $|\epsilon_i - \epsilon_{c,v}|$ increases. When we add or remove a finite number δ of electrons from the system, the Kohn-Sham states are not changed

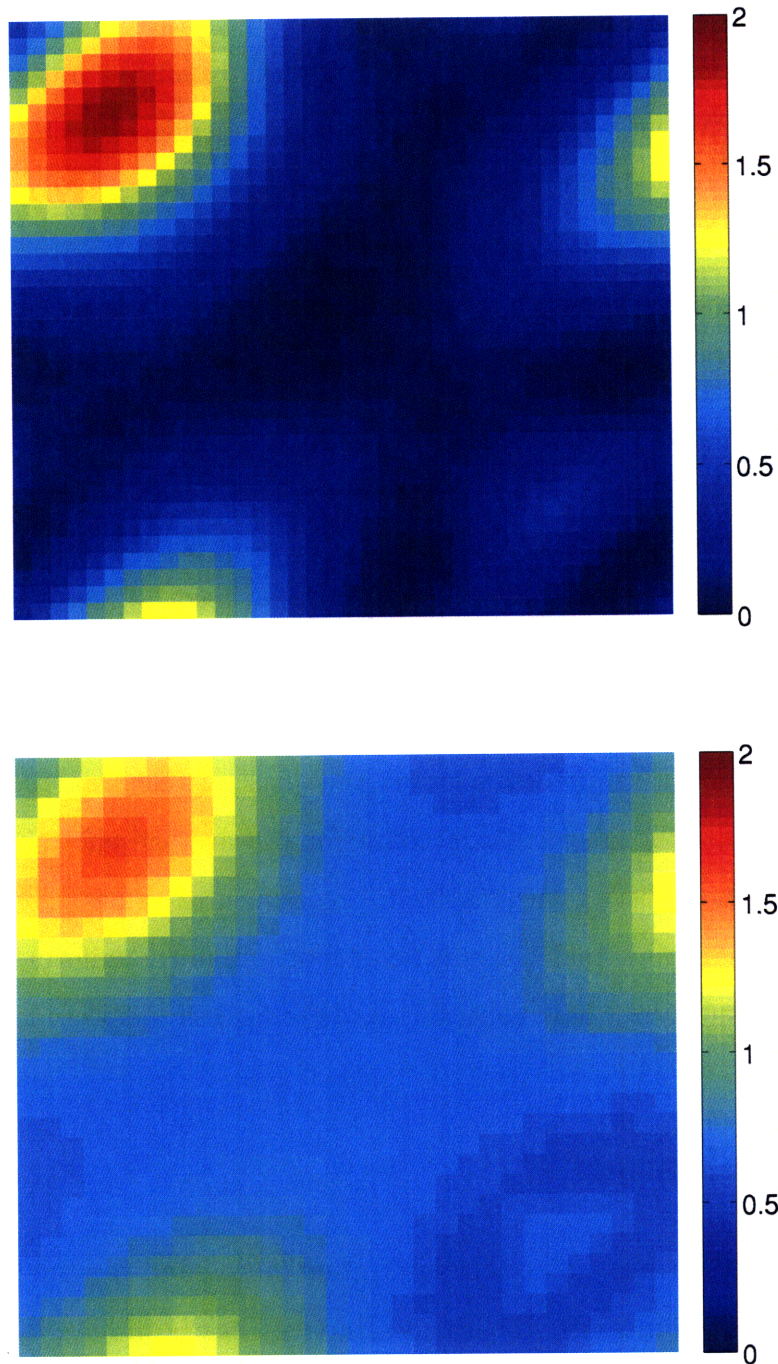Figure 4-26: The charge density from a unit occupation of the Kohn-Sham state at the valence band maximum at Γ for Mg_2Ge (top), compared with the charge density difference between the neutral cell and the cell with one per N^* fewer electrons (bottom). These are taken in the same spatial plane. Both charge densities are normalized to unity in the cell, though not necessarily in the plane shown.

drastically, but the occupations in a range of energies near the band edges do. The change in occupation from the original calculation also decreases with increasing $|\epsilon_i - \epsilon_{c,v}|$. What this means is that in occupying the Kohn-Sham eigenstates not just at the CBM/VBM but also in a range of nearby energies, which is what happens when we do not take the limit $N \to \infty$, we are making a crude approximation for the *density* that would have been produced by the quasiparticle wavefunctions $\phi_{c,v}$. Figure 4-26 shows the densities due to the Kohn-Sham state at the VBM and the density attributed to the hole, at the value of δ that corresponds to $N^* = 60$. We see that the charge is more "smeared out" for the added hole than the VBM state, which is reasonable if we consider that the quasiparticle is dressed by a screening cloud.

We do not claim, of course, that the procedure of occupying a number of states near the band edges reproduces the quasiparticle wavefunctions or even their exact densities. What we are missing are the effects of quantum mechanical interference, and exactly how the states are mixed due to exchange and correlation, i.e. the values of c_i and v_i. Nor can we predict *a priori* the quasiparticle weights renormalization away from unity, although we can account for the average effect by the scaling of N^*, which we do in fitting the results to experimental values. The fact that the quasiparticle weight renormalization is not strong and fairly uniform, as we have seen from our GW results, may have contributed to the success of our crude approximation.

4.6   Conclusion

In this chapter we have examined the issue of *ab initio* band gap prediction. It is well known that density functional theory and Hartree-Fock both have large errors in the prediction of band gaps, if one identifies differences in eigenvalues with the measured gap. The origin of these errors lies in the inadequate treatment of screening and non-local interactions. We investigate the GW approximation in the self-energy approach, which accounts for non-local screened exchange and correlations, and find that the G_0W_0 approximation significantly improves upon predicting experimental band gaps compared to the Kohn-Sham band gaps of DFT, though at a significant computation

cost.

We proposed a new method to use total energies in DFT to predict the fundamental gap. We calculate the total energy upon the addition and removal of a finite instead of infinitesimal amount of charge, with the amount determined by the screening charge or exchange-correlation hole extent, which are fundamental properties of the electron gas. We found that with this method, we are able to predict band gaps to within 10 and 30% for 50 and 90% of the medium gap ($1 \leq E_g \leq 4$) compounds in a test set that consists of a wide range of chemistries. The predicted values are consistent with experimental values if we assume a reasonable 0.2 - 0.3 eV standard deviation in the experimental measurements. The advantage of this method is that it is extremely efficient. Finally, we make connections between our new method of fundamental gap determination and the concept of quasiparticle excitations and found that there may be some physical correspondence.

Finally, we claim that density functional theory thus contains the essential information required for the calculation of band gaps, if one looks towards total energies instead of differences in the eigenvalues.

Chapter 5

Concluding Remarks

There are several general themes that run through the two seemingly-disparate subjects of thermal conductivities and electronic band gaps. We have already remarked upon the similarity in levels of complexity brought about by the enormous number of combinations of configurations or chemical species, and the fact that molecular dynamics and density functional theory are effectively constructionist approaches. In both cases general physical principles, not precise equations, guide our approach. In the case of lattice thermal conductivity, the general principle is that systems with low thermal conductivity brought about by the removal of long-range order should be amenable to a local cluster expansion model. In the case of fundamental gap prediction, the principle is that the integrated energy associated with a screened Coulomb interaction should converge within the screening radius, or equivalently, that effects due to exchange and correlation are confined to the extent of the exchange-correlation hole. In both cases, the general principles do not give us numerical factors, which have to be determined by fitting. The approaches we take differs from traditional empirical approaches in that we strive to maintain general applicability. If one is interested in the configurational dependence of thermal conductivity in a different system, for example, the effective cluster interactions would be different, but the principle of constructing the cluster expansion remains. Similarly, if a different functional is used for band gap prediction, the optimal value of N^*, the number of electrons within a screening radius, may be slightly different, but would be independent of the

compound.

There are, of course, many other properties which have to be considered before a nanostructure or a compound can be established as an effective thermoelectric or photovoltaic material. Beyond that, there are other properties that are of scientific and technological importance for which our fundamental theories are not yet adequate for a similar program of prediction and optimization to be carried out, high temperature superconductivity being an obvious example. Our work can only be considered a miniscule step in the immense task of understanding and predicting the properties of all condensed matter.

Bibliography

[1] http://www.grc.nasa.gov/WWW/RT/RT2001/5000/5410hepp2.html.

[2] Otfried Madelung. *Semiconductors: Data Handbook.* Springer-Verlag, New York, third edition, 2004.

[3] Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Fractional charge perspective on the band gap in density-functional theory. *Physical Review B (Condensed Matter and Materials Physics)*, 77(11):115123, 2008.

[4] J.S. Langer and S.H. Vosko. The shielding of a fixed charge in a high-density electron gas. *Journal of Physics and Chemistry of Solids*, 12(2):196 205, 1960.

[5] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Ángyán. Screened hybrid density functionals applied to solids. *The Journal of Chemical Physics*, 124(15):154709, 2006.

[6] Jochen Heyd, Juan E. Peralta, Gustavo E. Scuseria, and Richard L. Martin. Energy band gaps and lattice parameters evaluated with the heyd-scuseria-ernzerhof screened hybrid functional. *The Journal of Chemical Physics*, 123(17):174101, 2005.

[7] M. van Schilfgaarde, Takao Kotani, and S. Faleev. Quasiparticle self-consistent gw theory. *Physical Review Letters*, 96(22):226402, 2006.

[8] Giancarlo Cappellini, Sophie Bouette-Russo, Bernard Amadon, Claudine Noguera, and Fabio Finocchi. Structural properties and quasiparticle energies of cubic sro, mgo and srtio3. *Journal of Physics: Condensed Matter*, 12(15):3671 3688, 2000.

[9] P. W. Anderson. More is different. *Science*, 177(4047):393 396, 1972.

[10] http://icsdweb.fiz-karlsruhe.de/.

[11] http://www.eia.doe.gov.

[12] Terry M. Tritt and M.A. Subramanian. Introductory article:harvesting energy through thermoelectrics: Power generation and cooling. *MRS Bulletin*, 31(3), 2006.

[13] http://www.nrel.gov/pv/thin_film/docs/kaz_best_research_cells.ppt.

[14] J. M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A*, 128:334 350, 1984.

[15] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864 B871, Nov 1964.

[16] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133 A1138, Nov 1965.

[17] G.A. Slack. *Thermoelectric Handbook.* CRC, Boca Baton, 1995.

[18] L. D. Hicks and M. S. Dresselhaus. Effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B*, 47(19):12727 12731, May 1993.

[19] L. D. Hicks and M. S. Dresselhaus. Thermoelectric figure of merit of a one-dimensional conductor. *Phys. Rev. B*, 47(24):16631 16634, Jun 1993.

[20] Joseph Callaway. Model for lattice thermal conductivity at low temperatures. *Phys. Rev.*, 113(4):1046 1051, Feb 1959.

[21] William Shockley and Hans J. Queisser. Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics*, 32(3):510 519, 1961.

[22] Jenny Nelson. *The Physics of Solar Cells.* Imperial College Press, London, 2003.

[23] L. D. Hicks, T. C. Harman, X. Sun, and M. S. Dresselhaus. Experimental study of the effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B*, 53(16):R10493 R10496, Apr 1996.

[24] R Venkatasubramanian, E Siivola, T Colpitts, and B O'Quinn. Thin-film thermoelectric devices with high room-temperature figures of merit. *Nature*, 413(6856):597 602, Oct 2001.

[25] Yiying Wu, Rong Fan, and Peidong Yang. Block-by-block growth of single-crystalline si/sige superlattice nanowires. *Nano Lett.*, 2(2):83 86, 2002.

[26] Joseph P. Heremans, Christopher M. Thrush, Donald T. Morelli, and Ming-Cheng Wu. Thermoelectric power of bismuth nanocomposites. *Phys. Rev. Lett.*, 88(21):216801, May 2002.

[27] Bed Poudel, Qing Hao, Yi Ma, Yucheng Lan, Austin Minnich, Bo Yu, Xiao Yan, Dezhi Wang, Andrew Muto, Daryoosh Vashaee, Xiaoyuan Chen, Junming Liu, Mildred S. Dresselhaus, Gang Chen, and Zhifeng Ren. High-Thermoelectric Performance of Nanostructured Bismuth Antimony Telluride Bulk Alloys. *Science*, 320(5876):634 638, 2008.

[28] A.R. Abramson, Woo Chul Kim, S.T. Huxtable, Haoquan Yan, Yiying Wu, A. Majumdar, Chang-Lin Tien, and Peidong Yang. Fabrication and characterization of a nanowire/polymer-based nanocomposite for a prototype thermoelectric device. *J. Microelectromech. Syst.*, 13(3):505–513, June 2004.

[29] DG Cahill, K Goodson, and A Majumdar. Thermometry and thermal transport in micro/nanoscale solid-state devices and structures. *J. Heat Transfer*, 124(2):223 241, APR 2002.

[30] Deyu Li, Yiying Wu, Philip Kim, Li Shi, Peidong Yang, and Arun Majumdar. Thermal conductivity of individual silicon nanowires. *Appl. Phys. Lett.*, 83(14):2934–2936, 2003.

[31] David R. Lide, editor. *CRC Handbook of Chemistry and Physics, 88th Edition.* CRC Press/Taylor and Francis, Boca Raton, FL, June 2007.

[32] Deyu Li, Yiying Wu, Rong Fan, Peidong Yang, and Arun Majumdar. Thermal conductivity of si/sige superlattice nanowires. *Applied Physics Letters*, 83(15):3186–3188, 2003.

[33] M. S. Dresselhaus, G. Chen, M. Y. Tang, R. G. Yang, H. Lee, D. Z. Wang, Z. F. Ren, J. P. Fleurial, and P. Gogna. New directions for nanoscale thermoelectric materials research. In *Materials and Technologies for Direct Thermal-to-Electric Energy Conversion*, Warrendale, PA, 2005. Materials Research Society, Materials Research Society.

[34] Allon I. Hochbaum, Renkun Chen, Raul Diaz Delgado, Wenjie Liang, Erik C. Garnett, Mark Najarian, Arun Majumdar, and Peidong Yang. Enhanced thermoelectric performance of rough silicon nanowires. *Nature*, 451:163 167, 2008.

[35] Inna Ponomareva, Deepak Srivastava, and Madhu Menon. Thermal conductivity in thin silicon nanowires – phonon confinement effect. *Nano Letters*, 7(5):1155 1159, 2007.

[36] G. Chen. Thermal conductivity and ballistic-phonon transport in the crossplane direction of superlattices. *Phys. Rev. B*, 57(23):14958 14973, Jun 1998.

[37] Ronggui Yang, Gang Chen, and Mildred S. Dresselhaus. Thermal conductivity modeling of core-shell and tubular nanowires. *Nano Letters*, 5(6):1111 1115, 2005.

[38] C. Dames and G. Chen. Theoretical phonon thermal conductivity of si/ge superlattice nanowires. *Journal of Applied Physics*, 95(2):682–693, 2004.

[39] Yunfei Chen, Deyu Li, Jennifer R. Lukes, and Arun Majumdar. Monte carlo simulation of silicon nanowire thermal conductivity. *Journal of Heat Transfer*, 127(10):1129 1137, 2005.

[40] N. Mingo. Calculation of si nanowire thermal conductivity using complete phonon dispersion relations. *Phys. Rev. B*, 68(11):113308, Sep 2003.

[41] Sebastian G. Volz and Gang Chen. Molecular dynamics simulation of thermal conductivity of silicon nanowires. *Applied Physics Letters*, 75(14):2056 2058, 1999.

[42] Patrick K. Schelling, Simon R. Phillpot, and Pawel Keblinski. Comparison of atomic-level simulation methods for computing thermal conductivity. *Phys. Rev. B*, 65(14):144306, Apr 2002.

[43] M. S. Dresselhaus, G. Chen, M. Y. Tang, R. G. Yang, H. Lee, D. Z. Wang, Z. F. Ren, J.-P. Fleurial, and P. Gogna. New directions for low-dimensional thermoelectric materials. *Adv. Mater.*, 19:1043 1053, 2006.

[44] http://www.cse.scitech.ac.uk/ccg/software/DL_POLY.

[45] Davide Donadio and Giulia Galli. Thermal conductivity of isolated and interacting carbon nanotubes: Comparing results from molecular dynamics and the boltzmann transport equation. *Physical Review Letters*, 99(25):255502, 2007.

[46] J. Tersoff. New empirical approach for the structure and energy of covalent systems. *Phys. Rev. B*, 37(12):6991 7000, Apr 1988.

[47] C. J. Glassbrenner and Glen A. Slack. Thermal conductivity of silicon and germanium from 3 °k to the melting point. *Phys. Rev.*, 134(4A):A1058 A1069, May 1964.

[48] A A. V. Inyushkin, A. N. Taldenkov, A. M. Gibin, and H.-J. Pohl A. V. Gusev. On the isotope effect in thermal conductivity of silicon. *J Physica Status Solidi (c)*, 1:2995 2998, 2004.

[49] T. Ruf, R. W. Henn, M. Asen-Palmer, E. Gmelin, M. Cardona, H. J. Pohl, G. G. Devyatych, and P. G. Sennikov. Thermal conductivity of isotopically enriched silicon. *Solid State Communications*, 115(5):243 247, 2000.

[50] W. S. Capinski, H. J. Maris, E. Bauser, I. Silier, M. Asen-Palmer, T. Ruf, M. Cardona, and E. Gmelin. Thermal conductivity of isotopically enriched si. *Applied Physics Letters*, 71(15):2109 2111, 1997.

[51] T. Vo, A. J. Williamson, and G. Galli. First principles simulations of the structural and electronic properties of silicon nanowires. *Physical Review B (Condensed Matter and Materials Physics)*, 74(4):045116, 2006.

[52] D. Donadio and G. Galli. submitted.

[53] D. Cahill, H. Fischer, T. Klitsner, E. Swartz, and R. Pohl. *J. Vac. Sci. Technol. A*, 7:1259, 1989.

[54] Alberto Franceschetti and Alex Zunger. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature*, 402:60–63, 1999.

[55] Axel van de Walle. A complete representation of structure-property relationships in crystals. *Nat. Mater.*, 7:455–458, 2008.

[56] Fei Zhou, Thomas Maxisch, and Gerbrand Ceder. Configurational electronic entropy and the phase diagram of mixed-valence oxides: The case of li[sub x]fepo[sub 4]. *Phys. Rev. Lett.*, 97(15):155704, 2006.

[57] A. Van der Ven, G. Ceder, M. Asta, and P. D. Tepesch. First-principles theory of ionic diffusion with nondilute carriers. *Phys. Rev. B*, 64(18):184307, Oct 2001.

[58] Marcel H. F. Sluiter and Y. Kawazoe. Invariance of truncated cluster expansions for first-principles alloy thermodynamics. *Phys. Rev. B*, 71(21):212201, Jun 2005.

[59] Nikolai A. Zarkevich and D. D. Johnson. Reliable first-principles alloy thermodynamics via truncated cluster expansions. *Phys. Rev. Lett.*, 92(25):255702, Jun 2004.

[60] http://xmd.sourceforge.net.

[61] Lisa J. Porter, Ju Li, and Sidney Yip. Atomistic modeling of finite-temperature properties of [beta]-sic. i. lattice vibrations, heat capacity, and thermal expansion. *J. Nucl. Mater.*, 246(1):53 – 59, 1997.

[62] J. Reed, A. Williamson, E. Schwegler, and G. Galli. in preparation.

[63] B. Abeles. Lattice thermal conductivity of disordered semiconductor alloys at high temperatures. *Phys. Rev.*, 131(5):1906–1911, Sep 1963. and references therein.

[64] Tim Mueller and Gerbrand Ceder. *Phys. Rev. B*. submitted.

[65] G. A. Slack. *Solid State Physics*, volume 34. Academic Press, New York, 1979.

[66] http://www.ioffe.ru/SVA/NSM/Semicond/Si/mechanic.html.

[67] Elias Burstein. Anomalous optical absorption limit in insb. *Phys. Rev.*, 93(3):632–633, Feb 1954.

[68] T S Moss. The interpretation of the properties of indium antimonide. *Proceedings of the Physical Society. Section B*, 67(10):775–782, 1954.

[69] Michael Grätzel. Conversion of sunlight to electric power by nanocrystalline dye-sensitized solar cells. *Journal of Photochemistry and Photobiology A: Chemistry*, 164(1-3):3 14, 2004. Proceedings of the Dye Solar Cell Osaka ICP-21 Pre-symposium. Dedicated to Professor Shozo Yanagida on the occasion of his retirement.

[70] Wendy U. Huynh, Janke J. Dittmer, and A. Paul Alivisatos. Hybrid Nanorod-Polymer Solar Cells. *Science*, 295(5564):2425 2427, 2002.

[71] Bozhi Tian, Xiaolin Zheng, Thomas J. Kempa, Ying Fang, Nanfang Yu, Guihua Yu, Jinlin Huang, and Charles M. Lieber. Coaxial silicon nanowires as solar cells and nanoelectronic power sources. *Nature*, 449:885 889, 2007.

[72] A. Svane. Hartree-fock band-structure calculations with the linear muffin-tin-orbital method: Application to c, si, ge, and α-sn. *Phys. Rev. B*, 35(11):5496 5502, Apr 1987.

[73] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, 61(3):689 746, Jul 1989.

[74] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods.* Cambridge University Press, Cambridge, UK, 2004.

[75] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics*, 105(22):9982 9985, 1996.

[76] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics*, 118(18):8207 8215, 2003.

[77] Stephen Oxford Elliott H. Lieb. Improved lower bound on the indirect coulomb energy. 19(3):427 439, 1981.

[78] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865 3868, Oct 1996.

[79] Mariana M. Odashima and K. Capelle. How tight is the lieb-oxford bound? *The Journal of Chemical Physics*, 127(5):054106, 2007.

[80] Created and maintained by Anubhav Jain, with contributions from Christopher Fischer, Geoffroy Hautier, Charles Moore. For more detailed description, please see "Materials Design and an Electronic Structure Database using High-Throughput Density Functional Theory" by Anubhav Jain, MIT Department of Materials Science and Engineering thesis proposal, February 9, 2009.

[81] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169 11186, Oct 1996.

[82] P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50(24):17953–17979, Dec 1994.

[83] Peter E. Blöchl, O. Jepsen, and O. K. Andersen. Improved tetrahedron method for brillouin-zone integrations. *Phys. Rev. B*, 49(23):16223–16233, Jun 1994.

[84] Mohammed Lach-hab, Dimitrios A. Papaconstantopoulos, and Michael J. Mehl. Electronic structure calculations of lead chalcogenides pbs, pbse, pbte. *Journal of Physics and Chemistry of Solids*, 63(5):833–841, 2002.

[85] P M Lee and L Pincherle. The electronic band structure of bismuth telluride. *Proceedings of the Physical Society*, 81(3):461–469, 1963.

[86] V. I. Anisimov and O. Gunnarsson. Density-functional calculation of effective coulomb interactions in metals. *Phys. Rev. B*, 43(10):7570–7574, Apr 1991.

[87] F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan, and G. Ceder. First-principles prediction of redox potentials in transition-metal compounds with $lda + u$. *Phys. Rev. B*, 70(23):235121, Dec 2004.

[88] Lars Hedin. New method for calculating the one-particle green's function with application to the electron-gas problem. *Phys. Rev.*, 139(3A):A796–A823, Aug 1965.

[89] X. Gonze, G.-M. Rignanese, M. Verstraete, J.-M. Beuken, Y. Pouillon, R. Caracas, F. Jollet, M. Torrent, G. Zerah, M. Mikami, Ph. Ghosez, M. Veithen, J.-Y. Raty, V. Olevano, F. Bruneval, L. Reining, R. Godby, G. Onida, D.R. Hamann, and D.C. Allan. A brief introduction to the abinit software package. *Zeitschrift fr Kristallographie*, 220:558–562, May–June 2005.

[90] http://www.abinit.org/Psps.

[91] *Strong Coulomb Correlations in Electronic Structure Calculations: Beyond the Local Density Approximation*. Gordon and Breach Science Publishers.

[92] John Robertson. Electronic structure and the metal-nonmetal transition in liquid csau and $cs3sb$. *Phys. Rev. B*, 27(10):6322–6330, May 1983.

[93] A Hasegawa and M Watabe. Electronic structure of csau. *Journal of Physics F: Metal Physics*, 7(1):75–86, 1977.

[94] M. Skottke-Klein, A. Böttcher, R. Imbeck, S. Kennou, A. Morgante, and G. Ertl. Preparation and characterization of thin csau films. *Thin Solid Films*, 203(1):131–145, 1991.

[95] C. Koenig, N. E. Christensen, and J. Kollar. Electronic properties of alkali-metal gold compounds. *Phys. Rev. B*, 29(12):6481–6488, Jun 1984.

[96] Stephan Lany and Alex Zunger. Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for zno and gaas. *Physical Review B (Condensed Matter and Materials Physics)*, 78(23):235104, 2008.

[97] Serdar Öğüt, James R. Chelikowsky, and Steven G. Louie. Quantum confinement and optical gaps in si nanocrystals. *Phys. Rev. Lett.*, 79(9):1770 1773, Sep 1997.

[98] R. W. Godby and I. D. White. Density-relaxation part of the self-energy. *Phys. Rev. Lett.*, 80(14):3161, Apr 1998.

[99] Peter Fulde. *Electron Correlations in Molecules and Solids*. Springer, New York, 3 edition, 1995.

[100] G. Ortiz and P. Ballone. Correlation energy, structure factor, radial distribution function, and momentum distribution of the spin-polarized uniform electron gas. *Phys. Rev. B*, 50(3):1391 1405, Jul 1994.

[101] L. J. Sham and M. Schlüter. Density-functional theory of the energy gap. *Phys. Rev. Lett.*, 51(20):1888 1891, Nov 1983.

[102] Myrta Grüning, Andrea Marini, and Angel Rubio. Density functionals from many-body perturbation theory: The band gap for semiconductors and insulators. *The Journal of Chemical Physics*, 124(15):154108, 2006.