COMPUTATIONAL STUDY OF ADVANCED SEMICONDUCTING MATERIALS FOR NEXT-GENERATION ELECTRONIC DEVICES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jingyang Wang

December 2019

COMPUTATIONAL STUDY OF ADVANCED SEMICONDUCTING MATERIALS FOR NEXT-GENERATION ELECTRONIC DEVICES

Jingyang Wang, Ph.D.

Cornell University 2019

The semiconductor industry enabling numerous electronic devices that empower the digital era has arrived at a turning point in recent years. Traditional silicon (Si)-based devices are about to reach the material limit as a result of extreme scaling of gate length below 10nm, leading to the forthcoming end of Moore's law. One of the proposed ways to advance beyond Moore's law is to use alternative, advanced semiconducting materials such as indium gallium arsenide (InGaAs) and gallium oxide (Ga_2O_3), mainly for their superior electronic properties compared to traditional semiconductors. However, a series of major challenges need to be overcome in order for these materials to live up to their promise, including insufficiently high dopant activation and insufficiently low contact resistivity for InGaAs, and strong degree of dopant segregation towards the surface for Ga_2O_3. In this thesis, we address these challenges with a computational modeling approach, based on *ab initio* density functional theory calculations. Our main findings include: (1) negatively-charged cation vacancies are the major contributor of charge compensation in heavily Si-doped InGaAs; (2) composition, surface termination, doping concentration, compositional grading and metal-semiconductor alloying can all affect the contact resistivity of InGaAs; (3) in random InGaAs, vibrational modes associated with shallow n-type dopants (Si, Se) and defects (cation vacancies) assume band-like distribution with many satellite peaks, in contrast with the single-peak signature of the

same impurities in binary compounds (GaAs, InAs); and (4) Sn has the strongest tendency to segregate towards the (010) surface of Ga_2O_3 among three common shallow n-type dopants (Si, Ge, Sn), and the presence of negatively-charged surface Ga vacancies drastically enhance the segregation effect due to Coulomb interaction. This work would serve as a guidance for further experimentation and engineering with advanced semiconducting materials in electronic device applications.

BIOGRAPHICAL SKETCH

Jingyang Wang is a native of China, born in a small town Jiande of Zhejiang Province and raised in the city of Beijing. After three years of high school, at the young age of eighteen, he embarked on a long adventure to attend Cornell University as a new freshman. During his first four years at Cornell, he studied Physics and Mathematics, had his first taste of research in three biophysics labs, and graduated as a proud member of the Class of 2013; little did he know until his last two months as a senior that he would continue his long academic journey at his dear Alma Mater. In his first year as a fresh Ph.D. student, he found a new interest in computational materials science; after an extended period of search, he realized that nowhere at Cornell allows him to pursue his newfound passion in this booming field better than in the group of Prof. Paulette Clancy, in the School of Chemical and Biomolecular Engineering. During his five years in the group, he has enjoyed an unprecedented degree of freedom in choosing his research topics and executing his projects, all with the generous and unfailing support of Prof. Clancy. This unique experience has allowed him to develop and hone a very diverse range of research skills, which would prove extremely useful in his two internships at IBM T. J. Watson Research Center in 2016/17 and one internship at Western Digital in 2018. After six fruitful years as a Ph.D. student, he has embarked on yet another grand journey; this time to the West, joining the groups of Prof. Yi Cui at Stanford University and Dr. Lin-Wang Wang at Lawrence Berkeley National Laboratory, to further expand his research portfolio in the realm of batteries, catalysis, and novel electronic devices. In his spare time, he enjoys music improvisation, reading, hiking, and writing poetry.

Being perfect is boring. It's the imperfections that make us perfect.

– Jessie J

Dedicated to my parents Mr. Puqu Wang and Ms. Xiangqin Fang

Dedicated to the memory of my grandparents Mr. Zhangwen Fang and Ms. Zhixian Yang

Dedicated to every day and night of my devotion to an ideal

ACKNOWLEDGEMENTS

I owe my first and foremost thanks to Prof. Paulette Clancy, who has been a great research advisor as well as a role model. She has provided great amount of freedom for my research, which is instrumental in cultivating my ability as an independent scholar to the fullest extent. She has encouraged me to conduct highly interdisciplinary projects, which has greatly expanded my gamut of research interest and perspectives. Her strong support for me to take on prolonged internships in industry has been truly helpful in shaping my problem-oriented mindset. Last but not least, her unwavering patience has provided me with the exact driving force I needed most to complete my Ph.D. Prof. Clancy, I will always remember how honored I am as your Ph.D. student.

My next thanks goes to Dr. Binit Lukose, now at Wiley Publishing. Before I joined this group, I was just a new Ph.D. student knowing nothing about atomistic simulation. Were it not for Binit's numerous detailed hands-on explanations, the barrier of learning computational materials science for me would have been so much higher. He has been a master teacher for whatever question I had regarding density functional theory and molecular dynamics simulations. Binit, I truly wish you the best luck in your new position and in life!

Also, I must express my deep gratitude to every mentor who helped me in one way or another during my research projects. My thanks goes to: Prof. Mike Thompson at Cornell, whose vast knowledge of semiconductor processing has given me many important insights from a practical point of view; Dr. Phil Oldiges and Dr. Pranita Kerber at IBM, whose deep experience in device simulation has opened another door for my research and made my internship so enjoyable; and Dr. Zhaoqiang Bai and Dr. Derek Stewart, whose deep knowledge in phase change materials adds a novel and important dimension to my

research portfolio. I would also like to thank my Ph.D. committee (Prof. Thompson, Prof. Clancy, Prof. Craig Fennie, and Prof. Francis DiSalvo) for their many useful comments in my defense and thesis preparations. There have been so many great mentors and colleagues I'm indebted to for various reasons, including (but certainly not limited to): Prof. David Muller, Prof. Debdeep Jena, Prof. Grace Xing (Cornell); Dr. Siyuranga Koswatta, Dr. Christian Lavoie, Alexander Hsu (IBM); Dr. Alan Kalitsov, and Dr. Gerardo Bertero (Western Digital).

Next, my big thanks goes to the members in the Clancy group. They have been top-notch scholars, wonderful companions, and best friends to laugh with both in work and in bar. Their diverse talents have always inspired me to go higher in my own research, and give it back to whomever needs my help. In particular, I would like to thank Dr. Henry Herbol for being a very sweet and outgoing friend and a great teacher in any and all things about computer and programming; Haili Jia for her great contribution in our paper and her strong faith in my project design; Dr. Jonathan Saathoff and Dr. James Stevenson for their vast and deep knowledge regarding molecular simulations, chemistry, and programming; Dr. Victoria (Tori) Sorg for her many helpful insights from an experimentalist's point of view; Dr. Mardochee Reveil for his strong talent and many thoughtful discussions on III-V; Nikita Sengar for her inexhaustible enthusiasm; Isaiah Chen for his effervescent humor; and every other member in the group from whom I have benefited a great deal in my research as well as in life.

I must also express my deep gratitude for several people who have been instrumental in helping me adjust my life at Cornell during my undergraduate and graduate years, including: Dr. Zhongwu Wang, Dr. Xiaoyun Lu, Mr. Frank and Ms. Susan Eggleston.

Last but not least, I am deeply thankful for the endless love of my parents and grandparents who have made it all possible with their fullest support. Without them, I could not arrived where I am right now.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The modern age is properly termed the "Information Age". [1] Human life has been so fundamentally changed since the invention of information-processing devices, such as computers, laptops and cell phones, that it is almost unimaginable today how our predecessors lived in a time without them. At the very heart of this tremendous technology advancement lies the innovation of semiconductor materials. The prototypical example is the usage of semiconductors such as silicon (Si) and germanium (Ge) in the invention of transistors, the basic building block of all computing devices today. [1] In fact, the *de facto* rule that the semiconductor industry has been following in the past half century, the so-called "Moore's law" (Fig. 1.1(a)), [2] was enabled in large part by the exceptional electronic properties as well as the low cost of silicon, and has given birth to the extraordinary growth of electronics and computer industry in the latter half of 20th century. Nothing emblems the role of silicon in this achievement better than the name "Silicon Valley".

Despite the remarkable advancement Moore's law has contributed to the Information Age, in recent years it has started to show signs of weariness. Moore's law, which states that the number of transistors on an integrated circuit doubles roughly every two years, [2] depends on aggressive and constant miniaturization of individual transistor size. After more than fifty years of exponential scaling, the characteristic size of a transistor ("gate length") has entered the sub-10 nanometers regime (Fig. 1.1(b)). At this extreme scale, silicon-based transistors is quickly approaching its physical limit, and are faced with unprecedented challenges on both materials level and device level. On materials level, the in-

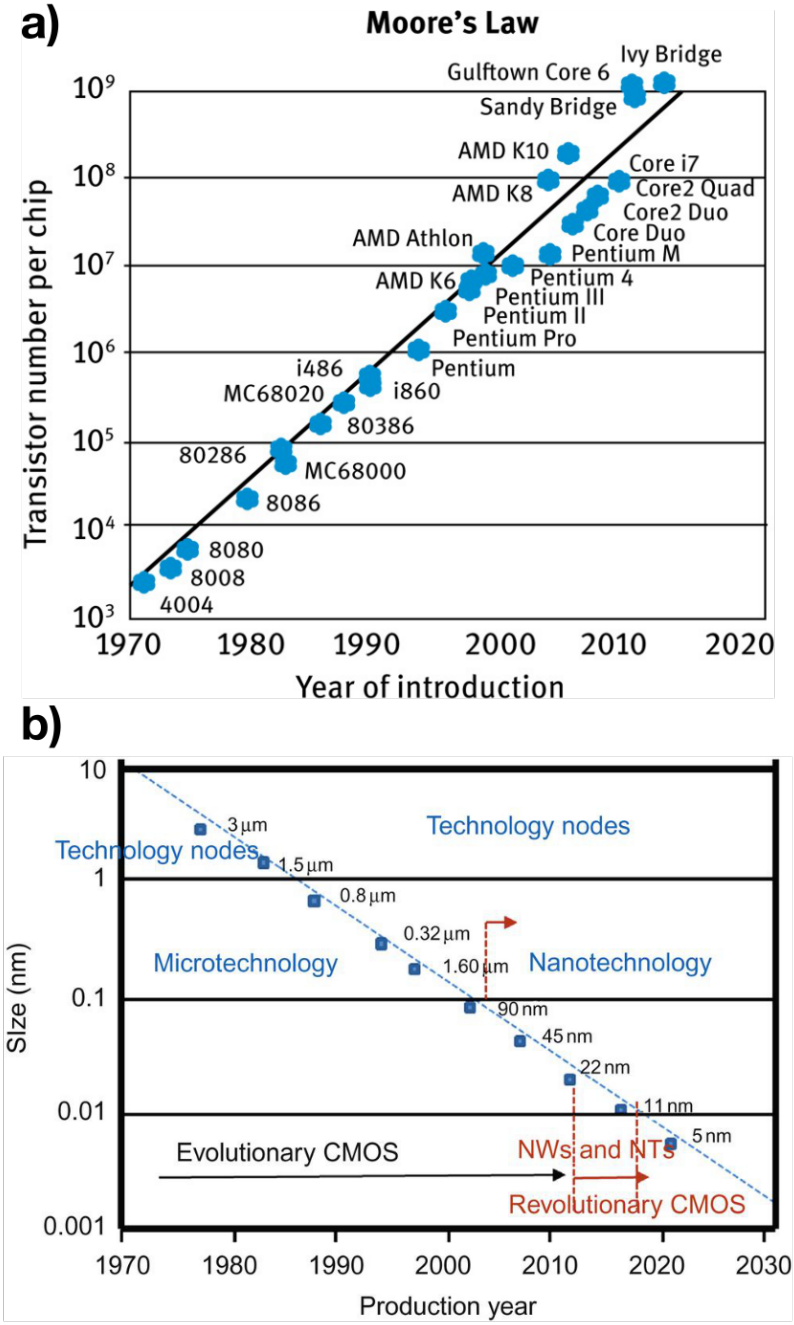

Figure 1.1: (a) Number of transistors per chip as a function of year of introduction (Moore's law; reproduced from Fig. 3, [3]); (b) transistor gate length in technology nodes and production year (reproduced from Fig. 4.1, [4]).

sufficient charge carrier mobility of silicon starts to limit the switching speed of scaled-down transistors. [5] On device level, the parasitic resistances and capacitances, especially at the material interfaces, limit respectively the maximum current and the maximum operating frequency of the device. [6] These challenges will only become more pronounced as the device size continues scaling down. In order to continue enabling future technological advances such as Internet of Things (IoT) and Artificial Intelligence (AI), such performance-limiting hurdles must be overcome with innovative strategies built upon fundamental understanding of the physics of semiconducting materials and devices. [7]

The work presented in this thesis is devoted to addressing the above-mentioned challenges by the means of computational modeling, from both materials and device perspectives. The thesis is organized as follows. Chapter 2 provides the necessary background for the subsequent chapters. Chapter 3 expounds the theoretical and computational methods employed in all the research works detailed in this thesis. Chapter 4 focuses on the calculation of formation energies of Si dopant and intrinsic defects of the ternary semiconducting compound $In_{0.5}Ga_{0.5}As$, and their consequences on the maximum carrier concentration and dopability of $In_{0.5}Ga_{0.5}As$. Chapter 5 focuses on the vibrational fingerprint of dopants and vacancies in random $In_{0.5}Ga_{0.5}As$. Chapter 6 focuses on the specific contact resistivity of various Ni-In(Ga)As interfacial systems. Chapter 7 focuses on the thermodynamics and atomistic mechanism of dopant and defect segregations towards the $Ga_2O_3(010)$ surface. Chapter 8 concludes the thesis with limitations of our approach and future works. The theoretical and computational approach outlined in the thesis is striven to be comprehensive and general in nature, and thus can find wide applications in aspects of materials research not limited to those covered in this thesis.

Bibliography

[1] Manuel Castells. The information age: Economy, society and culture (3 volumes). *Blackwell, Oxford*, 1997:1998, 1996.

[2] Gordon E Moore et al. Cramming more components onto integrated circuits.

[3] Karsten König and Andreas Ostendorf. *Optically induced nanostructures: biomedical and technical applications*. Walter de Gruyter GmbH & Co KG, 2015.

[4] Henry Radamson and Lars Thylen. *Monolithic Nanoscale Photonics-Electronics Integration in Silicon and Other Group IV Elements*. Academic Press, 2014.

[5] Jesús A Del Alamo. Nanometre-scale electronics with iii–v compound semiconductors. *Nature*, 479(7373):317, 2011.

[6] Thomas Skotnicki, James A Hutchby, Tsu-Jae King, H-SP Wong, and Frederic Boeuf. The end of cmos scaling: toward the introduction of new materials and structural changes to improve mosfet performance. *IEEE Circuits and Devices Magazine*, 21(1):16–26, 2005.

[7] International roadmap for devices and systems (irds™) 2018 edition. `https://irds.ieee.org/editions/2018`. Accessed: 2019-08-16.

CHAPTER 2

BACKGROUND

This chapter provides a self-contained exposition of the materials science background for all the research works presented in this thesis. The chapter is divided into two parts. Sec 2.1 gives a bird's eye view of the fundamentals of semiconductor materials, with a focus on two main factors – doping and defects – that influence the materials' electrical behavior. Sec 2.2 turns to two semiconductor material systems of wide interest – III-V compounds and gallium oxide (Ga_2O_3) – that comprise the basis of our research works in this thesis.

2.1 Fundamentals of semiconductor materials

Semiconductor is an important class of materials that possess unique electronic properties. For practical purposes, semiconductors are characterized by their ability to conduct electricity *in a controllable manner*. Such control is typically achieved by intentional incorporation of external chemical element(s) in the crystal lattice of the semiconductor material in question, a procedure known as **"doping"** (see Sec 2.2). Although the process of doping introduce impurity atoms in the crystal, it is an essential technique in semiconductor processing, as it can increase the charge carrier concentration in a semiconductor material and hence enhance the electrical performance of the device. [1] In contrast, unintentionally present impurities, called **"defects"**, often dictates the maximum charge carrier concentration in a semiconductor. [2] For a particular type of semiconductor material, the effectiveness of doping is often limited by the type and concentration of defects present. Hence, in order to develop strategies for future improvement of semiconductor device performance, it is critical to gain a

fundamental understanding of the interplay between doping and defect inside semiconductor materials. Such is a main theme of this thesis.

2.1.1 Structure of crystals

Many semiconductor materials of technological importance have the atomic structures of **crystals**, meaning that the atoms arrange themselves in an ordered, periodic fashion in space, forming a lattice pattern; each atom of the crystal occupies exactly one **lattice site** (Fig.2.1). Any repeating unit in a crystal is called a **unit cell**; the smallest of all unit cells is called the **primitive cell**. It is also useful to define the **conventional cell** as the smallest unit cell that contains the same point group symmetries (see the following paragraph) as the whole lattice; the conventional cell typically has a more regular geometry than the primitive cell. Fig. 2.2 shows several representative crystal structures in semiconductor materials.

All unit cells of a crystal that fill up three-dimensional space must have the shape of a parallelepiped, which is uniquely defined by the three side lengths a, b, c (called **"lattice constants"**) and three internal angles α, β, γ; it can also be represented by the lattice vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ (or denoted as $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$), as shown in Fig. 2.3. [3] Based on the relations among these structural quantities, all crystals can be grouped into seven distinct **crystal systems**. These seven crystal systems, together with the non-equivalent point lattice sites inside the unit cell, comprise 14 **Bravais lattices** in three-dimensional space (see Table 2.1). These lattices result in 230 non-equivalent ways of repeating atoms throughput the space that satisfy the required translational and rotational symmetry; these arrangements

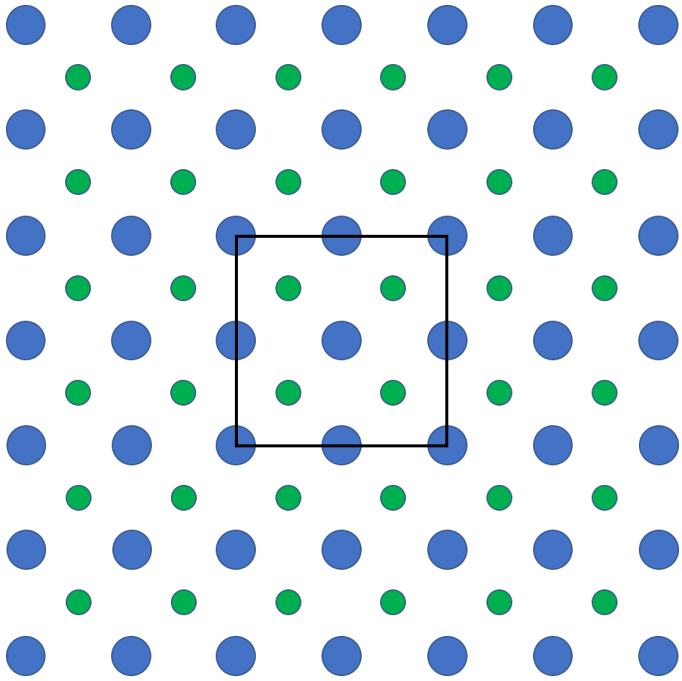

Figure 2.1: Schematics of part of a crystal lattice. The atoms (blue and green spheres) are arranged in an ordered and periodic fashion in all directions. The black square represents one possible unit cell of the crystal.

are known as **space groups**, in addition to 32 non-equivalent symmetries of a *local* atomic environment, known as the **(crystallographic) point groups**. [2] Hence, each space group uniquely represents the spatial symmetry of the crystal structure.

Although a real crystal sample contains extremely large number of atoms, it is nevertheless not infinite. In many systems of interest, the crystal sample is grown on a substrate oriented towards a particular direction. The crystal thus is terminated by a surface of atoms in perpendicular to this direction. In any crystal, there exist certain directions along which the atoms are aligned more regularly forming atomic planes, and therefore are favored by growth and processing. To denote such directions in a consistent manner, a convenient notation

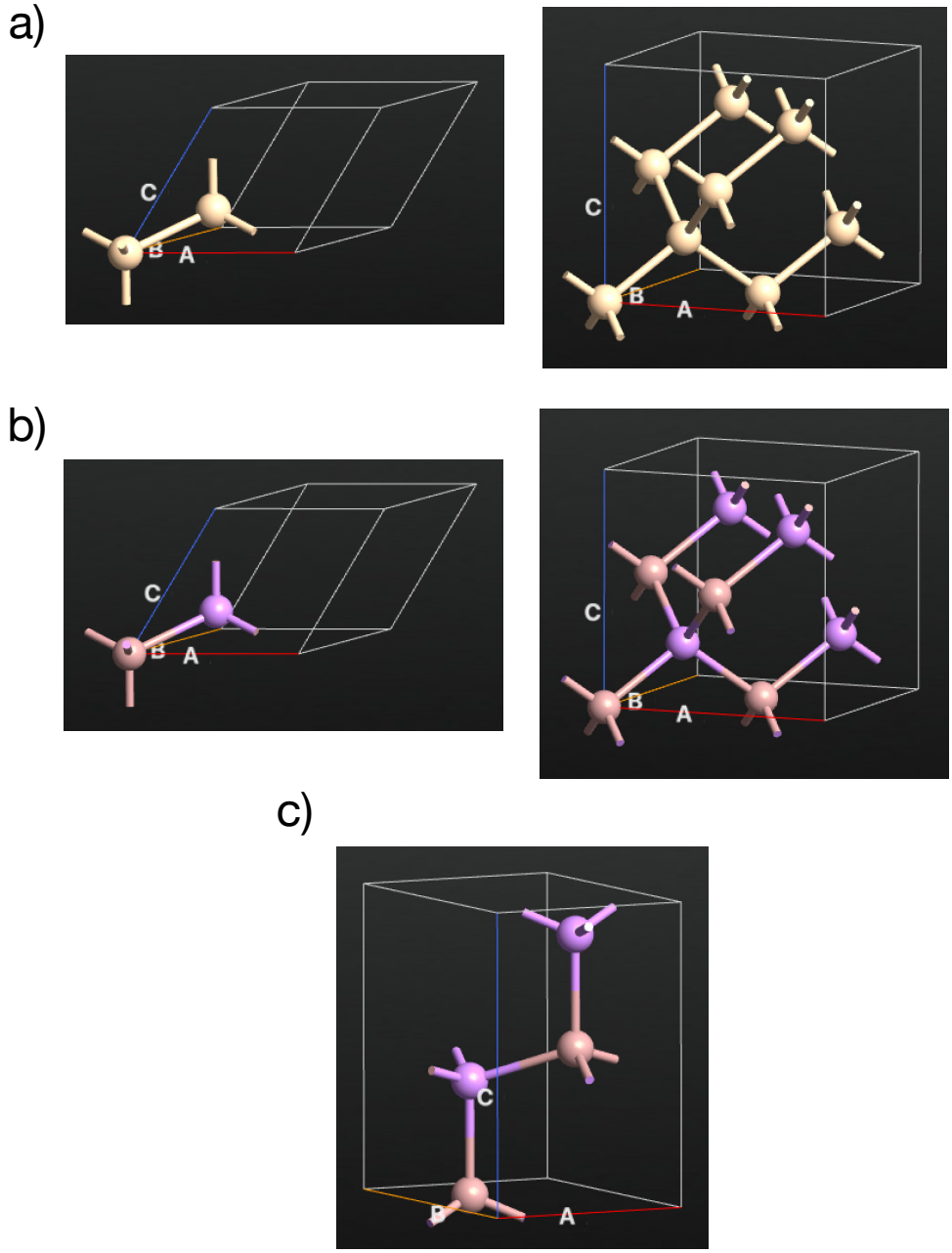

Figure 2.2: (a) Diamond crystal structure (left: primitive cell; right: cubic conventional cell); (b) zinc blende crystal structure (left: primitive cell; right: cubic conventional cell); (c) wurzsite crystal structure (hexagonal unit cell). Spheres with different colors represent different chemical species.

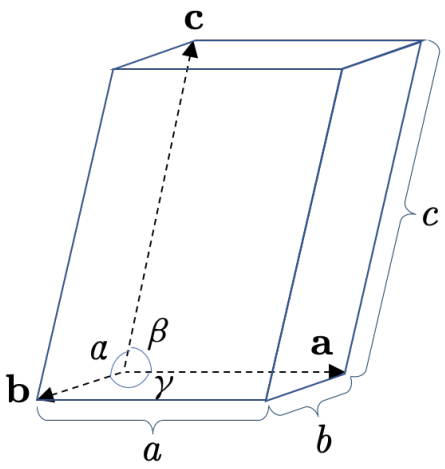

Figure 2.3: The parallelepiped shape of a crystal unit cell, defined by the side lengths a, b, c and internal angles α, β, γ, as well as by a set of lattice vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$.

called **"Miller indices"** are adopted. The Miller indices are a set of three integers (hkl). These integers, by convention, are designated to be the smallest set of integers such that the plane in question intercepts the three lattice vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ respectively at $a/h, b/k, c/l$. (If the plane does not intersect a given axis, then the corresponding Miller index equals 0.) Similarly, we denote the direction along $h\mathbf{a} + k\mathbf{b} + l\mathbf{c}$ as $[hkl]$. It is easily seen that in cubic crystals, the direction $[hkl]$ is perpendicular to the plane (hkl). Fig. 2.4 shows some of the most common surfaces in a cubic unit cell crystal.

2.1.2 Band theory of semiconductors

The physics of semiconductor crystals can best be understood in terms of band theory, which describes energy levels of atoms in crystalline solids. According to quantum mechanics, electrons associated with a single atom occupy a

Crystal system	Bravais lattice	Unit cell shape
Triclinic	simple	$a \neq b \neq c$; $\alpha \neq \beta \neq \gamma \neq 90^\circ$
Monoclinic	simple base-centered	$a \neq b \neq c$; $\alpha = \beta = 90^\circ \neq \gamma$
Orthorhombic	simple base-centered body-centered face-centered	$a \neq b \neq c$; $\alpha = \beta = \gamma = 90^\circ$
Tetragonal	simple body-centered	$a = b \neq c$; $\alpha = \beta = \gamma = 90^\circ$
Trigonal	simple	$a = b = c$; $\alpha = \beta = \gamma \neq 90^\circ$
Hexagonal	simple	$a = b \neq c$; $\alpha = 120^\circ, \beta = \gamma = 90^\circ$
Cubic	simple body-centered face-centered	$a = b = c$; $\alpha = \beta = \gamma = 90^\circ$

Table 2.1: Seven distinct crystal systems and their corresponding fourteen Bravais lattices in three-dimensional space.

set of discrete energy levels, known as **atomic orbitals**. When two atoms are brought sufficiently close to each other, the electrons in their respective outermost atomic orbitals (**valence electrons**) interact with each other and rearrange themselves, so as to lower the total energy of the two-atom system, a process known as **bonding**. This process creates new energy levels in which the electrons can occupy. As more and more atoms are brought together, this bonding process would occur between any two neighboring atom pairs, resulting in a multitude of energy levels. A real-world macroscopic crystal sample consists of on the order of 1 mole of, or 6.02×10^{23} densely packed atoms. In such a sample, the vast amount of interatomic bondings generate innumerable energy levels such that individual energy levels cannot be clearly distinguished from

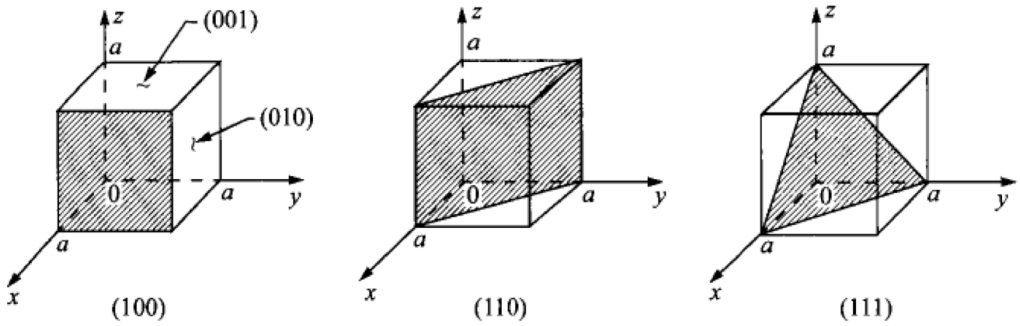

Figure 2.4: Miller indices of some important planes in a cubic crystal. Reproduced from Fig. 1.2, [4].

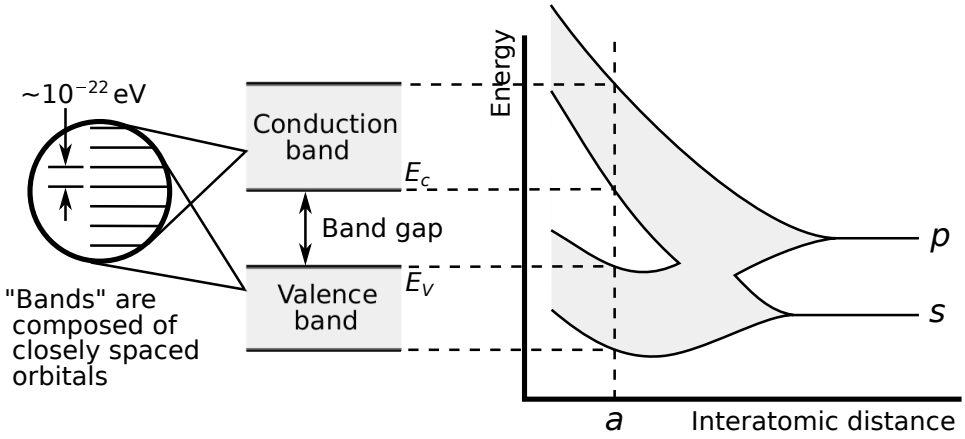

Figure 2.5: Origin of electronic energy bands in a crystal. Reproduced from [5].

adjacent ones. The thus-formed continuous energy levels of electrons are called **"energy bands"**. (Fig. 2.5)

When the atoms are placed on a crystalline lattice, their electrons fill up the energy bands according to a known probability distribution. At absolute zero temperature ($T = 0$K), all energy levels E below a certain threshold known as the **"Fermi level"** E_F are completely filled up with electrons, and all energy levels above E_F are empty. As the temperature increases to finite values, the electrons that occupy energy levels close to the Fermi level tends to disperse, so that there

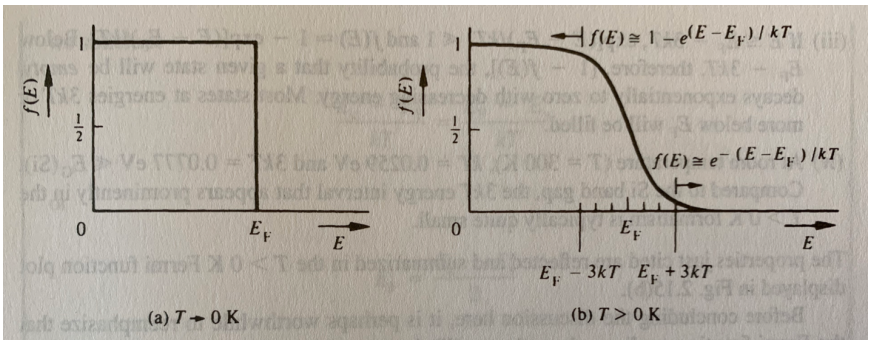

Figure 2.6: The Fermi function at (a) zero temperature ($T = 0$K), and (b) finite temperature. Reproduced from Fig. 2.15, [1].

is a finite probability that an energy level above the Fermi level is occupied by electron(s), and a finite probability that an energy level below the Fermi level is unoccupied by electron(s). For physical motivations that will become clear in subsequent sections, it is beneficial to introduce the notion of **"holes"**, which can be simply understood as the lack of an electron. When an electron with one negative unit of charge $-e$ is excited to a higher energy level, it leaves a hole with one positive unit of charge $+e$ at its original energy level; hence both electrons and holes are charge carriers of opposite signs. Under thermal equilibrium, the probability distribution for electrons, known as the **"Fermi function"**, is given by

$$f(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{k_B T}\right)} \tag{2.1}$$

where $k_B = 8.617 \times 10^{-5}$eV/K is the Boltzmann constant. As the material as a whole must remain charge neutral, the probability distribution for holes is given by

$$1 - f(E) = \frac{1}{1 + \exp\left(\frac{E_F - E}{k_B T}\right)}. \tag{2.2}$$

The plots of the Fermi function at zero temperature and finite temperature are shown in Fig. 2.6.

The Fermi function $f(E)$ tells us about the probability of occupation of electrons and holes as a function of energy at a given temperature under thermal equilibrium; it however does not contain any information on the number of electronic energy levels available at a given energy value E. This information is provided by the **density of states** $D(E)$. This is an intrinsic quantity of materials; each material has a different density of states. The density of electrons and holes in a material at a given energy E is therefore given by the product of the density of states at E and the probability for the carrier to occupy these states:

$$\begin{aligned} n(E) &= D(E)f(E); \\ p(E) &= D(E)(1 - f(E)). \end{aligned} \tag{2.3}$$

The density of states is a crucial indicator of the electronic properties of a material. In a material, the distribution of electron energy levels depend as a function of multiple factors, including the nature of interatomic bonding and lattice spacing. Some combinations of these parameters give a band structure that is continuous, while others give a band structure in which the energy bands are separated into two groups by a gap where no energy level exists; this gap is called the **"band gap"**. (Fig. 2.7) In the former case, the electrons can be continuously excited to gradually higher energy levels as the temperature increases; whereas in the latter case, the ability for an electron below the band gap to be excited to the top of the band gap depends on the size of the gap E_g. When E_g is large enough, it is impossible by means of increasing temperature alone to excite an electron across the gap, due to the very small probability of occupation by the Fermi function at the top of the gap. (Fig. 2.6) In this case, even at *finite temperature*, essentially all electrons will be confined in their energy levels below the gap (called the **"valence band"**), and no electrons will be able to flow freely inside the crystal at energy levels above the gap (called the **"conduction**

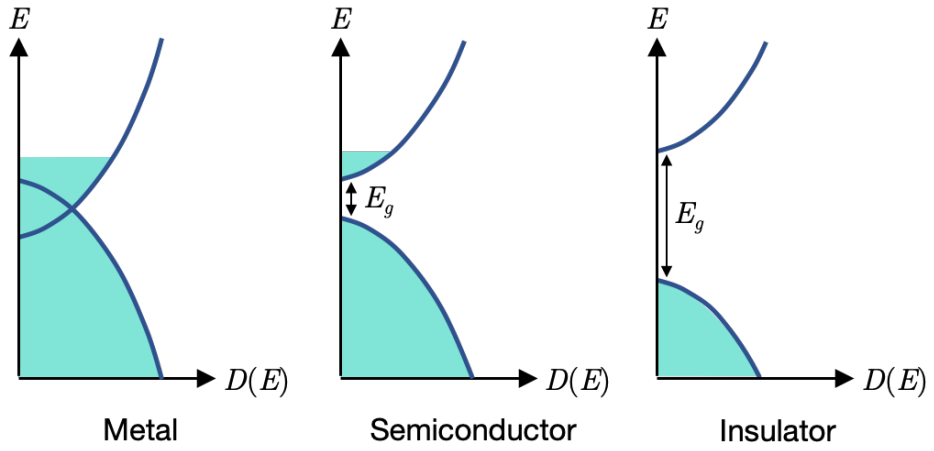

Figure 2.7: The schematics of density of states for three types of materials: metal, semiconductor, and insulator. Upward and downward curves represent conduction and valence bands respectively; the shades represent filling of electrons.

band"), hence electricity conduction is not possible. Such materials are properly named **insulators**. On the other hand, for sufficiently small E_g, at elevated temperatures a large number of electrons will be able to spill from the bottom of the gap into the top, giving rise to the so-called **"charge carriers"**. Such materials are called **semiconductors**.

Both density of states and Fermi function can provide crucial insight on the charge carrier distribution inside a material. (Sec. 2.1.3) The density of states can be obtained routinely from computer simulations (Chap. 3). With this information, we can perform further analysis such as calculating maximum charge carrier concentration at a particular dopant level (Chap. 4), or bond population analysis in a crystal (Chap. 7). These methods will be explained in detail in Chap. 3.

The best way to visualize the energy bands of a crystal is by plotting the **band structure** of the unit cell. In a band structure plot, the energy levels of elec-

trons are plotted as a function of **k-point**, corresponding to values of momentum of electron in the crystal (called **"crystal momentum"**). When an electron moves in free space (i.e. no atom is present), its wavefunctions are superpositions of plane waves with the functional form $\exp(i\mathbf{k}\cdot\mathbf{r})$ ($\mathbf{k}$ denotes the wavevector; $\mathbf{r}$ denotes the position). The electron's energy, which is purely kinetic, is described by quantum mechanics as

$$E(k) = E_{\text{kin}}(k) = \frac{\hbar^2 k^2}{2m_e}, \tag{2.4}$$

where $\hbar = 4.136 \times 10^{15}$eV·s is the reduced Planck constant, $m_e = 9.109 \times 10^{-31}$kg is the electron mass, and $k = |\mathbf{k}| = 2\pi/\lambda$ is the wavenumber of the electron with wavelength λ. In a periodic crystal, the electron's energy is inevitably affected by the presence of atoms. The nuclei potential gives rise to a potential energy term $U(\mathbf{k})$ to the electron, while the kinetic energy term of the electron becomes

$$E_{\text{kin}}(k) = \frac{\hbar^2 k^2}{2m_e^*}, \tag{2.5}$$

where the term m_e^* here represents the **"effective mass"** of an electron inside the crystal. This mass is not the true electron mass, as the mass of electron always stays constant; it purely reflects the *effect* that the crystal potential has on the electron's motion, *as if* the electron has changed its mass. In other words, the smaller the effective mass, the faster the electron moves inside a crystal. (This observation has important consequences for device performance; see for example Sec. 2.3.1.) This result also applies to holes. The effective masses of electrons and holes can be obtained from the band structure by

$$m^* = \hbar^2 \left(\frac{\partial^2 E}{\partial k^2} \right)^{-1}, \tag{2.6}$$

where the partial derivative is taken at the CBM/VBM, where the potential energy of electron/hole is zero. We see from this "effective mass theory" that the

conduction and valence bands should locally resemble parabola at their respective extrema (see, for example, the band structures of Si and GaAs shown in Fig. 2.8); moreover, the effective masses is related to the curvature of the band extrema (the greater the effective mass, the smaller the curvature, the flatter the band). Hence the effective masses is a direct measure of the density of states available at CBM/VBM. This very useful relation can be quantified by defining the "effective density of states" of CBM/VBM:

$$\begin{aligned} N_C &= 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2}; \\ N_V &= 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2}. \end{aligned} \tag{2.7}$$

In every band structure, the electronic energy levels are conventionally plotted as the k-point moves along a certain path connecting **high-symmetry points** inside the **(first) Brillouin zone** of the crystal. The Brillouin zone is the set of all possible unique wavevectors $\mathbf{k}$, corresponding to wavelengths λ greater than the lattice constants, that can describe any electronic wavefunction in a crystal. (Sec. 3.3.6) The Brillouin zone of the face-centered cubic (fcc), diamond, and zinc blende crystal structure is shown in Fig. 2.9, along with its associated high-symmetry points (Table 2.2) and directions. Band structures is a materials-specific property, and they provide a complete and unique picture of electron dynamics for each crystal.

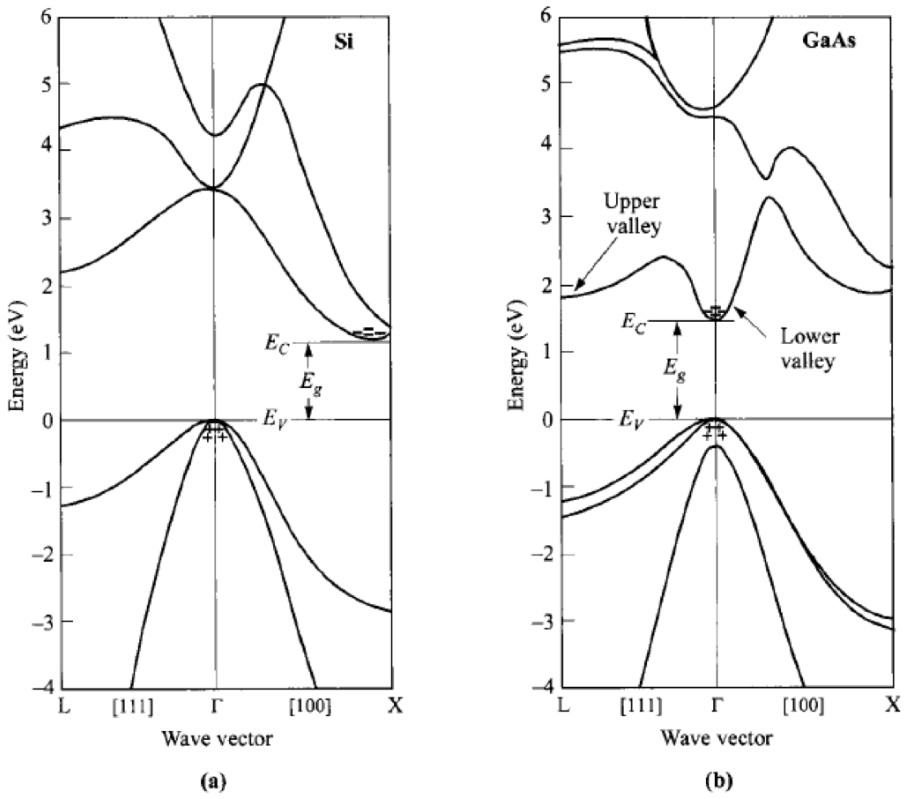

Figure 2.8: Energy-band structures of (a) Si and (b) GaAs, where E_g is the energy band gap. Plus signs (+) indicate holes in the valence bands and minus signs (-) indicate electrons in the conduction bands. Reproduced from Fig. 1.4, [4].

2.1.3 Impurities in semiconductors

Doping in semiconductors

It is clear from the previous section that for semiconductors, temperature can serve as a "knob" to control the degree of electricity conduction. However for the application of electronic devices, this is neither sufficient nor precise. For instance, many commonly used semiconductor materials have much smaller concentrations of charge carriers than typical metals at room temperature. (Table 2.3) Such **intrinsic semiconductors** contains relatively small concentration

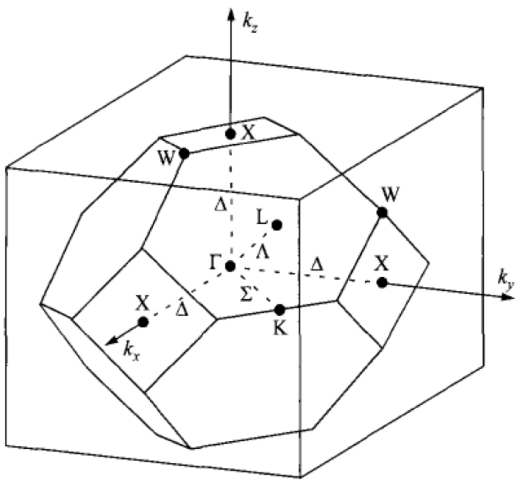

Figure 2.9: Brillouin zones for face-centered cubic (fcc), diamond and zinc blende crystal structures, showing high-symmetry points and lines. Reproduced from Fig. 1.3, [4].

High-symmetry point	Coordinates in Brillouin zone
Γ	$(0, 0, 0)$
X	$\frac{2\pi}{a}(0, 1, 0)$
L	$\frac{2\pi}{a}(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$
W	$\frac{2\pi}{a}(\frac{1}{2}, 1, 0)$
U	$\frac{2\pi}{a}(\frac{1}{4}, 1, \frac{1}{4})$
K	$\frac{2\pi}{a}(\frac{3}{4}, \frac{3}{4}, 0)$
High-symmetry line	**Direction**
Δ	[010] $(\Gamma - X)$
Σ	[110] $(\Gamma - K)$
Λ	[111] $(\Gamma - L)$

Table 2.2: Notation of high-symmetry points and lines in the fcc Brillouin zone. [6]

of impurities, and often do not possess sufficient amount of charge carriers required for standard device operations. In order to overcome the obstacle, another strategy, namely doping (see Sec. 2.1), is routinely used in semiconductor industry. Doping refers to the practice of adding external chemical elements called **"dopants"** to the crystal lattice of the semiconductor. The charge carrier

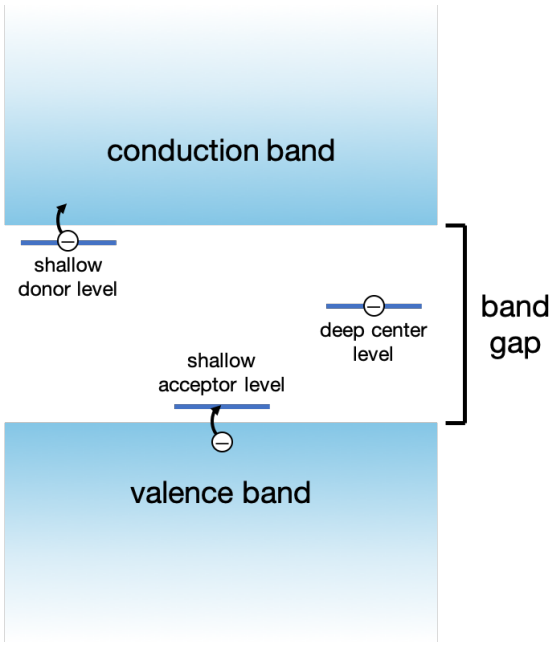

Figure 2.10: Schematic of the energy levels of three types of impurities (donor, acceptor, deep center) in a semiconductor band structure. The $\ominus$ symbol represents an electron.

concentration in the semiconductor depends on both the chemical nature and the amount of dopants added within a given range. This property of semiconductor materials allows effective and precise control over charge carrier concentrations, and hence the degree of electricity conduction (electrical conductivity), needed in a semiconductor-based device.

Semiconductor material	intrinsic carrier concentration at 300K
Silicon (Si)	$1 \times 10^{10}\text{cm}^{-3}$
Germanium (Ge)	$2 \times 10^{13}\text{cm}^{-3}$
Gallium Arsenide (GaAs)	$2 \times 10^{6}\text{cm}^{-3}$

Table 2.3: Common semiconductor materials used in industry and their intrinsic carrier concentration at $T = 300$K. [1]

To help understand the mechanism of carrier concentration modulation by doping, we again resort to the energy band picture of electrons in a semicon-

ductor. Fig. 2.10 shows the energy levels of electrons in a typical semiconductor material with impurities. As discussed in the previous section, the conduction band are the energy levels in which excited valence electrons can flow freely inside the crystal, thereby capable of conducting electricity; the valence band are the energy levels where valence electrons are bound to the atomic nucleus, and thus cannot conduct electricity. These energy bands are always present, regardless of whether impurities exist in the crystal. However, the presence of relatively large amount of impurities in a semiconductor can often modify the energy landscape and significantly alter the distribution of electrons. Depending on the electronic nature, an impurity belongs to one of the three categories: **donor**, **acceptor**, and **deep center.** A donor is a chemical species which can donate valence electrons to the conduction band. When a donor is properly incorporated into the crystal, it introduces an additional energy level within the band gap that is very close to the minimum point of the conduction band (**"conduction band minimum (CBM)"**); such level is called the **shallow donor level**. The proximity of this donor level to the CBM allows electrons occupying this level to be easily excited to the CBM. The consequence is an increase of electron concentration in the conduction band (and a decrease of hole concentration in the conduction band); hence incorporation of donors is known as **"n-type doping"** ("n" for negatively charged electrons). Similarly, when an acceptor is properly incorporated into the crystal, it introduces an additional energy level (acceptor level) within the band gap that is very close to the maximum point of the valence band (**"conduction band maximum (VBM)"**); such level is called the **shallow acceptor level**. The proximity of this acceptor level to the VBM allows electrons occupying the highest valence bands to be easily excited to the acceptor level. The consequence is an increase of hole concentration

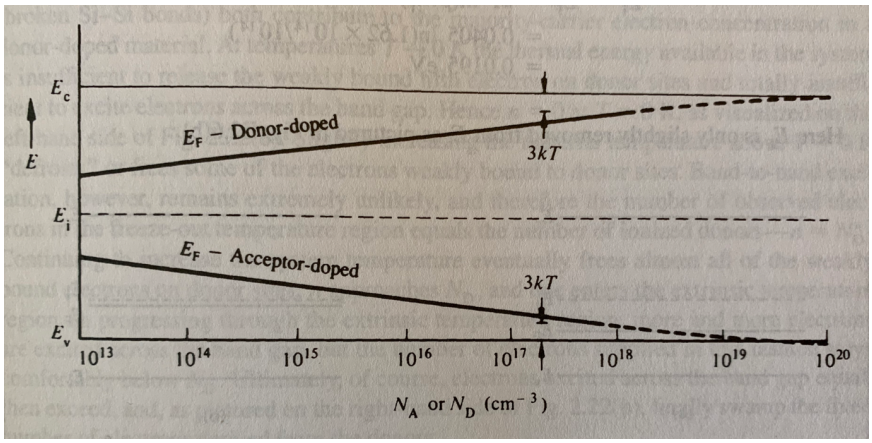

Figure 2.11: Fermi level positioning in Si at 300K as a function of the doping concentration. Reproduced from Fig. 2.21, [1].

in the valence band (and a decrease of electron concentration in the conduction band); hence incorporation of acceptors is known as **"p-type doping"** ("p" for positively charged holes). Hence we see that doping with donors increase free electron concentration; doping with acceptors increase free hole concentration. For materials with band gap (semiconductors and insulators), the Fermi level position within the band gap shifts as a function of carrier concentration (see Eq. (2.3)), and therefore depends in turn on the doping concentration. (Fig. 2.11) Finally, a deep center is a type of impurity which introduces at least one energy level within the band gap that is far away from both conduction band and valence band. Any electron or hole that occupies this level cannot be easily excited to either side of the band gap, and is thus localized ("trapped") at the defect level. Such impurities are typically considered undesirable for the purpose of doping, but they may useful in other applications. For instance, the nitrogen-vacancy (NV) center in diamond can potentially act as a quantum bit (qubit), which is the basic unit of information in a quantum computer. [7]

In order for a chemical element to behave as donor or acceptor in a particu-

Donors	$\lvert E_b \rvert$ (eV)	Acceptors	$\lvert E_b \rvert$ (eV)
Antimony (Sb)	0.039	Boron (B)	0.045
Phosphorous (P)	0.045	Aluminum (Al)	0.067
Arsenic (As)	0.054	Gallium (Ga)	0.072
		Indium (In)	0.16

Table 2.4: Common shallow donors and acceptors used for silicon, and their absolute binding energies. Reproduced from Table 2.3, [1].

lar semiconductor material, it must possess certain chemical properties. From an atomistic point of view, a donor-type or acceptor-type dopant is incorporated into the crystal lattice of a semiconductor by substituting atoms of the host material. As the host atoms form bonds in a crystal, such atomic substitution breaks existing bonds between the substituted host atoms and the neighboring host atoms, and forms new bonds between the dopant atom and the neighboring host atoms. As a chemical rule of thumb known as the "octet rule", an atom in a bonding environment is stabilized when eight electrons are present in its outermost shell. Such condition is satisfied with every host atom in a stable crystal. That means if the impurity species has the same number of valence electrons as the host atom it substitutes, then the bonding environment will be the same as that before the substitution, and no extra electrons will be donated to or accepted from the sea of free electrons in the crystal ("reservoir"). On the other hand, if the impurity species possess more valence electrons than the host atom species it substitutes, these extra valence electron(s) may not participate in the bonding process, and will become free electrons. Similarly, if the impurity species possess fewer valence electrons than the host atom species it substitutes, a same number of missing electrons are required from the sea of free electrons in order to form the stable octet for the impurity atom. (Fig. 2.12) From this reasoning, we see that shallow donors should have greater valence

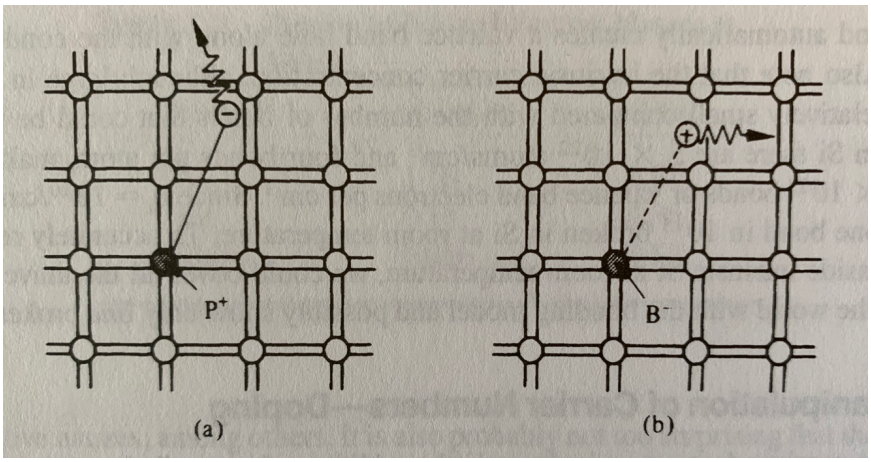

Figure 2.12: Visualization of (a) a donor and (b) acceptor action using the bonding model. In (a) the Column V element P is substituted for a Si atom; in (b) the Column III element B is substituted for a Si atom. Reproduced from Fig. 2.10, [1].

than the substituted species, and shallow acceptors less valence than the substituted species. It must to be emphasized that the reverse is not always true; namely, simply having a different number of valence electrons does not entail a particular chemical species is a shallow donor/acceptor, as it may be a deep center. In other words, the difference between the energy level of the impurity and that of the closest band edge (called the "binding energy" of the dopant) must be small enough (typically less than 0.1 eV) in order to qualify as shallow dopants. Table 2.4 lists commonly used shallow donors and acceptors in silicon and their binding energies.

Finally, it is worth noting that although doping is an effective way of controlling carrier concentration in a semiconductor, it does not always work as expected. Specifically, it is often found that as concentration of the dopant species in a crystal increases beyond a threshold value, the carrier concentration will saturate and stop increasing further, a phenomenon known as **"dopant deactivation"**. Such difficulty in doping may be traced to three possible origins. [8]

First, the dopant might have an energy level(s) that is not "shallow" enough, so that the carriers cannot be excited to the corresponding band edge at room temperature. Second, the dopant atoms may cluster and form a secondary phase beyond a concentration threshold, either due to insufficient solid solubility inside the host crystal or a tendency to segregate towards the epitaxial surface of the host crystal. In this case, the local bonding environment in this new phase becomes the same as that in the bulk crystal consisting of purely the dopant species, and thus no new charge carriers will be generated. Last but not least, as the Fermi level shifts towards the band edge as a consequence of increased carrier concentration due to doping, defects of opposite charge become more likely to form inside the crystal (See Sec. 2.1.3 for detailed discussion), causing the net charge carrier concentration to saturate. These bottlenecks of doping have become a major challenge in many semiconductor-based applications, as the performances of many semiconductor devices depend critically on the ability to increase charge carrier concentration. Two research works in this thesis (Chap. 4 and 7) are concerned mainly with the problem of dopant deactivation, through a comprehensive investigation of the physical and chemical mechanisms, as well as proposing possible strategies to overcome this performance-limiting obstacle.

Defects in semiconductors

Naturally-occuring crystals never comes as perfect arrangements of atoms in our simple mathematical model. Even without intentional doping, it is thermodynamically favorable for certain impurities and irregularities to form inside a crystal; such imperfections are called defects. Defects come in a great variety of types. In terms of geometry, three main types of defects can be found in a

crystal: **point defects**, **line defects**, and **planar defects**. Point defects are atom-sized modifications on a single lattice site. There are four main types of point defects: **vacancies**, **substitutionals**, **interstitials**, and **antisites**. (Fig. 2.13) A vacancy is the lack of an atom in a crystal; a substitutional is the occupation of a foreign species not part of the crystal itself; an interstitial is an atom situated in the void space between atoms in the crystal; and an ansisite is the occupation of a species that belongs to the crystal but on the wrong lattice site. Line defects refer to dislocations, where an entire line of atoms shift out of their original place on the lattice due to shear stress. Dislocations include edge dislocation (dislocation line parallel to stress direction) and screw dislocation (dislocation line perpendicular to stress direction). Planar defects are discontinuity of periodic atomic arrangements across a plane; these include surfaces (termination of atomic planes), interfaces (atomic plane where two different crystals meet), grain boundaries (atomic plane where two regions of the same crystal with different crystallographic orientations meet), stacking faults (missing of an atomic layer in a regular stacking sequence of atomic planes), and twin boundaries (atomic plane where two regions across the plane are mirror images of each other). As we shall see, the defects with most electronic importance are point defects, so we will focus solely on point defects from this point on.

Point defects can behave electrically as either shallow donors, shallow acceptors, or deep centers. For defects that behave as shallow donors or acceptors (the so-called "electrically active" defects), they would either assist or hinder n-type or p-type doping, depending on whether the defect contributes the same or opposite type of charge carriers compared to the dopant. On the other hand, defects that behave as deep centers trap carriers at their energy levels deep inside the gap, causing a bottleneck in doping efficiency at high concentrations. Note

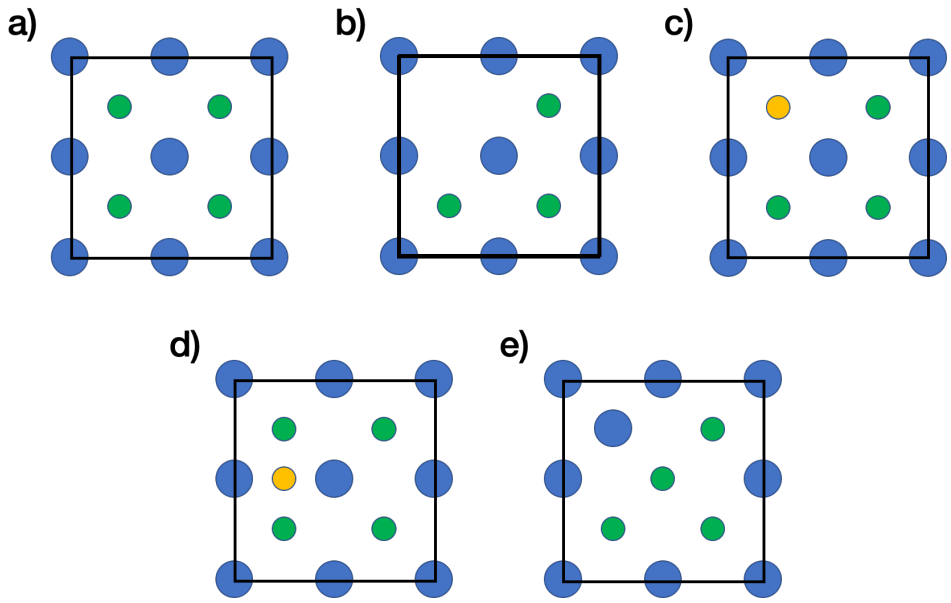

Figure 2.13: Schematic of various point defect configurations in reference to (a) the perfect crystal, including (b) vacancy; (c) substitutional; (d) interstitial; and (e) antisite.

that a deep center can act as a donor or an acceptor, depending on the position of the Fermi level. Unlike shallow donors or acceptors, many deep centers possess multiple **charge states** inside the band gap, meaning that the defect can retain different numbers of electrons or holes as the Fermi level varies; the Fermi levels at which the defect acquires or loses charge carrier are called **"charge transition levels"**. We shall see in Sec. 3.4.1 and 4.2.5 that these levels can be calculated using a quantitative formalism of point defect thermodynamics, and can provide information on the maximum carrier concentration achievable inside a particular semiconducting material at a given temperature, a critical knowledge in semiconductor device processing.

Point defects are inherently present in any crystal; such phenomenon can be understood from thermodynamic argument. Consider a piece of gallium arsenide (GaAs) crystal at room temperature. There are four Ga atoms and four

As atoms in the conventional cubic unit cell with side length $a = 5.65$Å, resulting in a density of Ga lattice sites $n_s = 2.22 \times 10^{22}\text{cm}^{-3}$. In bulk GaAs crystal, a Ga vacancy in -3 charge state (meaning three electrons are bound to the vacancy; denoted as V_{Ga}^{-3}) has a formation energy (energy required to remove a Ga atom from the lattice) $E^f = 0.24$eV for Fermi level at 0.3eV below the CBM, a condition corresponding to moderate n-type doping. [9] Based on these results, it is possible to deduce that at extremely high n-type doping levels, the thermal equilibrium concentration of V_{Ga}^{-3} in GaAs, given by the Arrhenius-type formula

$$n_V = n_s \exp\left(\frac{-E^f}{k_B T}\right), \tag{2.8}$$

is equal to $2.0\times10^{18}\text{cm}^{-3}$ at room temperature. At such non-negligible concentrations, these defects would often contribute charges with opposite sign to that of the dominant carrier (**"charge compensation"**). Furthermore, even intentional dopants may start to exhibit undesirable properties when located on the wrong lattice sites or under non-standard conditions. For example, Si is commonly used as donors in GaAs when incorporated on Ga sites, but can also behave as acceptors when substituting As atoms (such behavior is called **"amphoteric doping"**) [10], or even as deep centers when it adopts an alternative bonding geometry on Ga sites when under hydrostatic pressure greater than 2 GPa (e.g. "DX centers" in Si-doped GaAs) [11]. These challenges all could contribute to dopant deactivation at high doping concentrations, severely limiting the efficacy of doping. Hence, it is crucial to fully understand the types and behaviors of such doping-hindering defects and their interactions with dopants as a first step towards overcoming the doping bottleneck in semiconductor materials.

Finally, defects are prevalent on the surfaces and interfaces of materials. Common defects of this kind include dimers, dangling bonds (DBs), vacancies, antisites, and adatoms. These defects often contribute significantly to the char-

acteristics and performance of electronic devices. For example, presence of defects on semiconductor surfaces introduces **surface states** inside the band gap, causing the Fermi level near the surface to be pinned at a fixed position irrespective of the doping concentration and doping type in the bulk (known as **"Fermi level pinning"**). [12, 13, 14, 15] (Fig. 2.14) Fermi level pinning limits the carrier concentration near the surface, which in turn degrades the charge transfer characteristics across the interface between the semiconductor and the metal / oxide. Specifically, Fermi level pinning at metal-source/drain interface leads to the formation of Schottky barriers with height independent of the metal work function, which adds parasitic contact resistances to the device. Fermi level pinning at oxide-channel interface traps the minority carriers for positive gate voltage (inversion mode), thereby limiting the degree of gate modulation on the channel conductance. [11, 16] For these crucial reasons, removal of surface and interface defects has been one of the major challenges for III-V processing and integration in electronic devices.

Vibrational fingerprint of impurities in materials

For a doped semiconductor sample, it is often important to have a knowledge of the distribution of dopants. Especially when a particular semiconducting material approaches its doping limit, it is essential to be able to pinpoint the lattice locations of dopants and defects in order to obtain a fundamental understanding of the limiting factors. One of the most common experimental methods for detecting positions of impurities in a crystal is the spectroscopic methods. This method utilizes the fact that all atoms are in constant vibrational motion around their equilibrium positions, even under thermal equilibrium. Due to the global

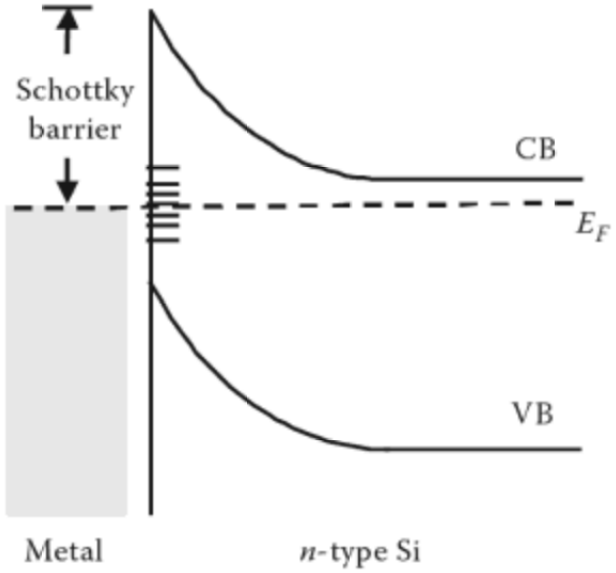

Figure 2.14: Real metal-semiconductor interface. Surface states, indicated by horizontal lines, pin the Fermi level at $\sim 1/3E_g$ above the valence band (VB) maximum. This results in a Schottky barrier of $\sim 2/3E_g$. Reproduced from Fig. 2.22, [11].

(space group) as well as local (point group) symmetries of the crystalline lattice (Sec. 2.1.1), atoms in the lattice exhibit patterns of collective vibrational motions at certain frequencies, known as **"phonon modes"**. Such phonon modes depend sensitively on the atomic mass and interatomic force constants, a measure of the mechanical strength of bonding analogous to macroscopic spring constants. Inclusion of defects in the lattice usually brings about modifications in atomic mass and local interatomic force constants, leading to a change in the phonon modes of the crystal. When the atomic mass of the impurity is much smaller than any of the host atom species, or the force constants are much higher compared to those of the host bonds, the vibrational pattern of the impurity and its close atoms will become separated form that of the host crystal, thereby creating **local vibrational modes (LVM)**. (Fig. 2.15; detailed discussion see Chap. 5) As the bond strengths depend on the local atomic environment,

phonon modes are an effective source of information that offers critical insights on the lattice locations of a particular atomic species. Historically, the doping behavior of technologically important semiconductors such as silicon and gallium arsenide have been thoroughly understood with the help of spectroscopic methods, through accurate identication of local vibrational modes of dopants, defects, and dopant-defect complexes. [17]

Experimentally, phonon modes are obtained by spectroscopic methods. Spectroscopic methods measure the **"phonon density of states" (PDOS)**, or **phonon spectrum** of a crystal, namely the intensity of vibration within a range of frequencies. Characteristic phonon modes of a system appear as peaks in the phonon spectrum. There are two main types of spectroscopic methods, namely infrared (IR) spectroscopy and Raman spectroscopy. [19] In both methods, a light source is required to shine a beam of light onto the material sample. In infrared spectroscopy, the quantized units of light, or **photons**, are absorbed by the atoms and turned into quantized units of vibrations, or **phonons**, with equal amount of energy inside the crystal. (Fig. 2.16(a)) In Raman spectroscopy, the photons are not absorbed by the atoms, but rather scattered by the atoms, as the photons involved provide much higher energies than that of vibrational modes; (Fig. 2.16(b)) in this case, the photons scattered by the atoms may lose or gain energy, depending on whether the atoms are excited from the ground state to a vibrational state or otherwise. [19] As the energy difference between each vibrational state and the ground state of a material stays constant independent of the spectroscopic method used, these two methods produce similar phonon spectra for a particular system; however, there exist important differences. Not all vibrational modes of a system can be detected by *both* IR (IR-active) or Raman (Raman-active) spectroscopy. As IR spectroscopy utilizes photon absorption as

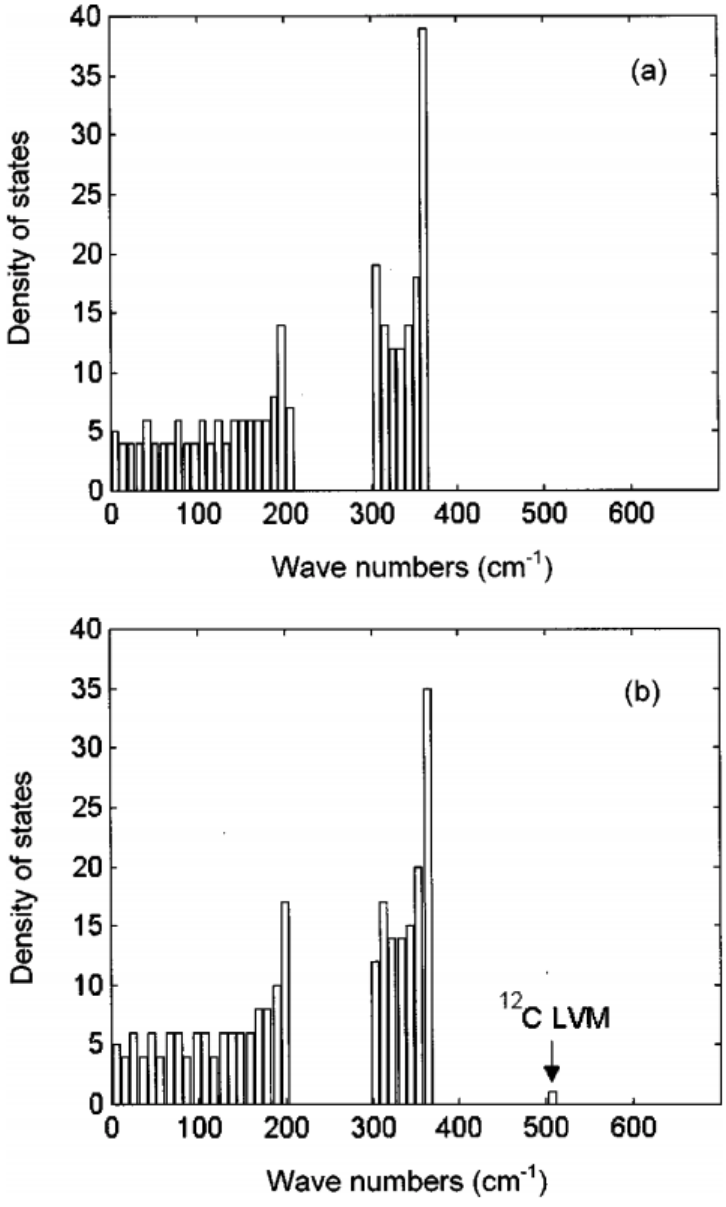

Figure 2.15: Density of vibrational states for (a) undoped and (b) ^{12}C-doped GaP, calculated numerically with the linear-chain model. A LVM due to ^{12}C has a calculated frequency of 510 cm^{-1}. Reproduced from Fig. 1, [18].

a means of exciting vibrations, a particular vibrational mode must change the electric dipole moment of the local group of atoms (or molecule) to signal the occurrence of such absorption. In contrast, as scattered photons originate from oscillations of electric dipoles, the intensity of scattered photons is proportional to the polarizability of the local atom group (or molecule); hence in Raman spectroscopy only vibrational modes that change the polarizability can be detected. These criteria are formally known as "selection rules". [19]

Due to the limitations stated above, it is not guaranteed that any one spectroscopic method will be able to uncover all phonon modes for a given system. Furthermore, in complicated material systems such as random alloys, the lack of long-range order will cause disruption in propagation of phonons in the crystal, resulting in a relatively large number of "background" local vibrational modes that blur the signal of vibration of dopants and defects. In Chap. 4, such phenomenon is investigated using advanced computational modeling techniques for the first time.

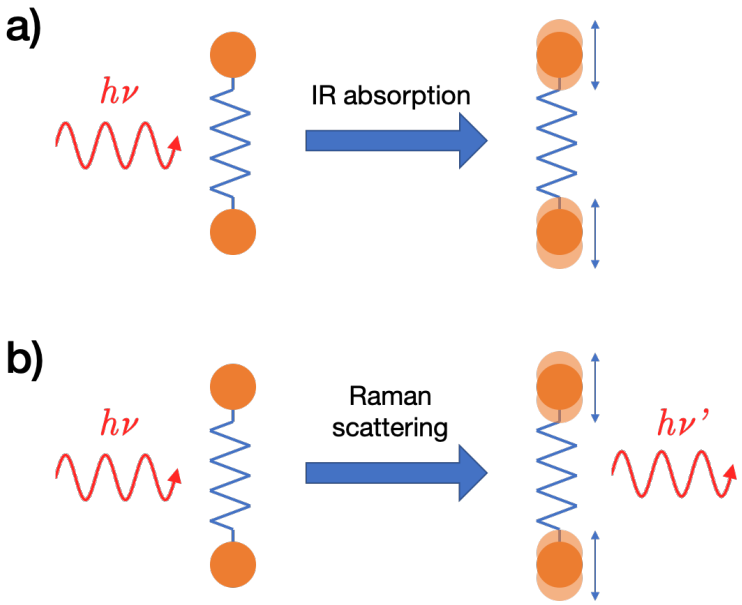

Figure 2.16: Schematics of mechanisms of (a) infrared (IR) spectroscopy, and (b) Raman spectroscopy. Details see texts of Sec. 2.1.3.

2.2 Basics of MOSFET

Metal-oxide-semiconductor field effect transistors (MOSFETs) are an important class of electronic devices that serves as the workhorse for today's extensive computing needs. Fig. 2.17 shows the structure of a planar MOSFET. In a MOSFET, the source and drain regions are semiconductors that are heavily-doped with the same polarity (n-type or p-type), while the substrate is lightly-doped with the opposite polarity. The contacts are metals that connects different parts of the device with external electric circuits. The operation of a MOSFET is controlled by the voltage applied to the gate V_G, which introduces a vertical electric field through the gate oxide. When a positive gate-to-source voltage V_{GS} is applied in an n-type MOSFET (n-MOSFET), excess holes under the gate are driven away from the vicinity of the gate, creating a **depletion region**. As V_{GS} keeps increasing, excess free electrons are attracted from source and drain to the de-

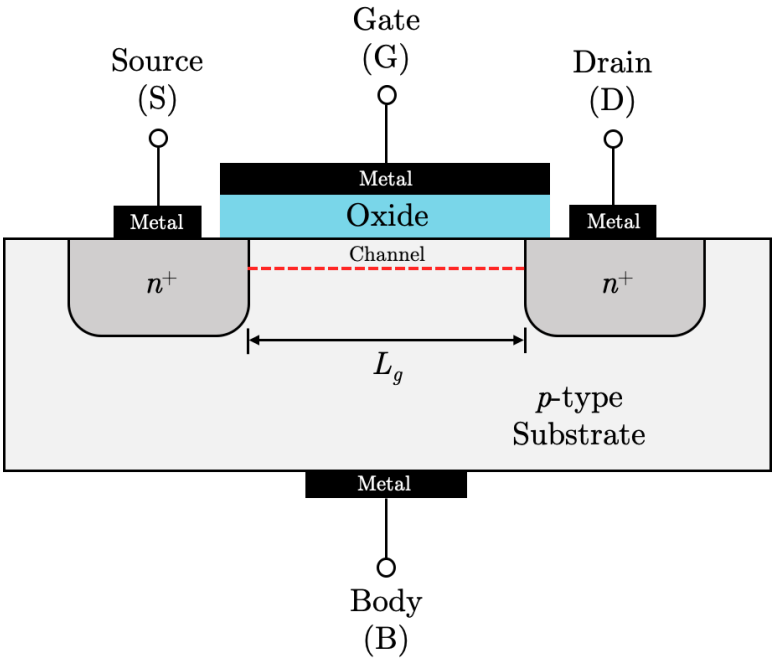

Figure 2.17: Schematic cross-section of a planar n-type metal-oxide-semiconductor field effect transistor (n-MOSFET). In an n-MOSFET, the source and drain regions are heavily n-type doped (denoted by "n^+") semiconductors, and the substrate are lightly p-type doped.

pletion region, creating a new "inversion layer" (channel) filled with electrons. Additionally, application of a small drain-to-source voltage V_{DS} would introduce a lateral electric field across the channel, forcing the electrons to flow from the drain to the source, thereby creating a current I_D in the opposite direction.

The magnitude of I_D varies with respect to both V_{GS} and V_{DS}. At a fixed positive value of V_{DS}, when V_{GS} is below a threshold voltage V_T, there is still not enough free electrons in the depletion region, therefore I_D remains close to zero (called the **"OFF-mode"** of MOSFET). The onset of I_D occurs when V_{GS} exceeds V_T, switching the MOSFET to textbf"ON-mode". (Fig. 2.18) A device like MOSFET that has two states (ON and OFF) belongs to the category of **"logic**

devices", as it can perform logic operations. On the other hand, at a fixed $V_{GS} > V_T$, I_D exhibits distinct growing patterns as V_{DS} increases. When V_{DS} is small ($V_{DS} \ll V_{GS} - V_T$), the lateral electric field causes the shape of the channel to taper towards the source, making it easier for electrons to flow out of the drain, thereby increasing I_D in a linear fashion (the "linear region"). When V_{DS} reaches a threshold value ($V_{DS} = V_{GS} - V_T$), the tapering of the channel becomes so severe that the channel is "pinched-off" from the drain. (Fig. 2.19) In this scenario, I_D would saturate at a maximum level, even if V_{DS} keeps increasing beyond the threshold value (the "saturation region"). Quantitatively, the dependence of I_D on V_{GS} and V_{DS} can be described by the following equations:

$$\begin{aligned} I_D &\approx 0 && (V_{GS} < V_T); \\ I_D &\approx \mu_{\text{inv}} C_{\text{ox}} \frac{W}{L} (V_{GS} - V_T) V_{DS} && (V_{GS} > V_T, V_{DS} \ll V_{GS} - V_T); \\ I_D &\approx \mu_{\text{inv}} C_{\text{ox}} \frac{W}{2L} (V_{GS} - V_T)^2 && (V_{GS} > V_T, V_{DS} \geq V_{GS} - V_T). \end{aligned} \tag{2.9}$$

From these equations, it is clear that as the width W and length L of the gate becomes smaller than ever, in order not to let I_D in ON-mode (I_{on}) degrade with the scaling, it is crucial to improve on materials-specific parameters such as the electron mobility in the inversion layer (μ_{inv}). Another equally important challenge of transistor scaling is that, in a realistic MOSFET, there exists a very small but finite current I_D even when no gate voltage is applied ($V_{GS} = 0$). This **leakage current** I_{off} is the main source of *static* power consumption for a MOSFET. Hence it is crucial to reduce I_{off} so that a chip containing billions of transistors would not waste exceeding amount of energy in standby mode. Finally, the subthreshold swing S, defined by the inverse slope of the $I_D - V_{GS}$ curve in the substreshold region ($V_{GS} < V_T$), is a key parameter that determines the switch-

ing speed from OFF to ON-mode for a MOSFET and the magnitude of V_T. The smaller the S, the steeper the slope, and the smaller the V_T which leads to less *dynamic* power consumption per transistor. S is given by

$$S = \ln(10)\left(\frac{k_B T}{q}\right)\left(1 + \frac{C_{\text{dep}}}{C_{\text{ox}}}\right), \tag{2.10}$$

where q is the elementary charge, C_{dep} and C_{ox} are the capacitances of the depletion layer and the oxide, respectively. It is clear from this equation that S can be lowered by increasing C_{ox} (through reducing thickness of oxide or using materials with high permittivity ("high-κ")) or decreasing C_D (through reducing density of interfacial defects between the oxide and the substrate); nonetheless, in a MOSFET S is always greater than the ""thermionic limit", which equals $\ln(10)(k_B T/q) = 60$mV per decade (10 times) increase in I_D at room temperature ($T = 300$K). Alternative devices such as tunneling field effect transistor (TFET) is capable of breaking this limit by taking advantage of band-to-band tunneling rather than thermal injection of carriers into the channel. Improvement on these three parameters (increase I_{on}, decrease I_{off}, decrease S) are the major objectives and challenges of transistor scaling.

2.2.1 metal-semiconductor contact

The (metal-semiconductor) contacts are an important component of all electronic devices. They serve as the connection between the device and the external circuit. As the transistor size keeps scaling down, so does the contact length L_{c}. Contact resistance, an inherent quantity of metal-semiconductor contact that measures the easiness of carrier transport across the junction, increases as $\coth(L_{\text{c}})$ at the limit of small L_{c}. [20] (Fig. 2.20) This phenomenon causes the

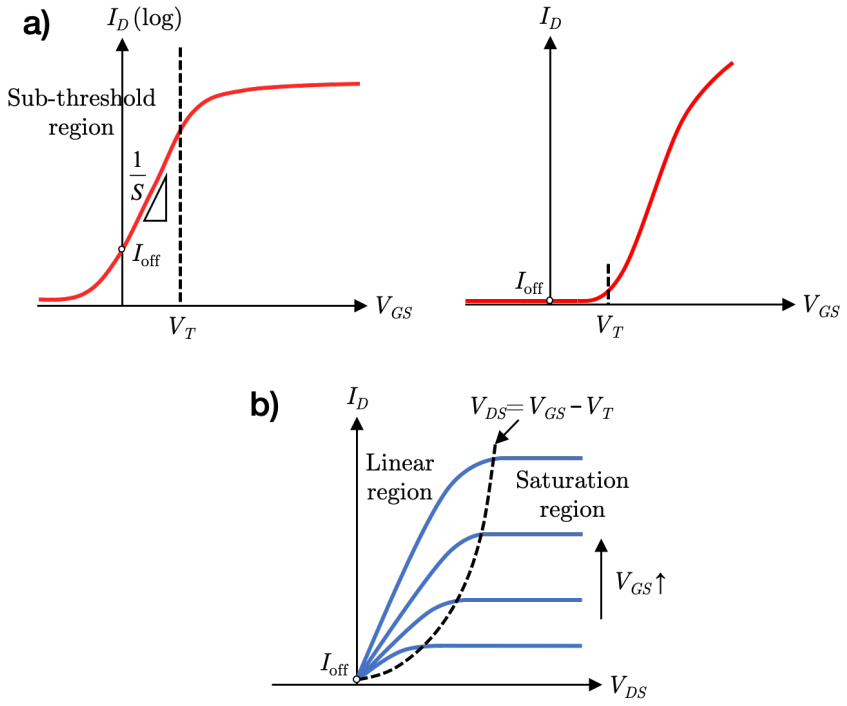

Figure 2.18: (a) $I_D - V_{GS}$ characteristic of a MOSFET (in log and linear scale for I_D); (b) $I_D - V_D$ characteristic of a MOSFET.

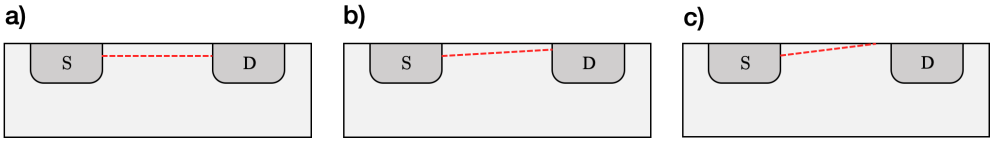

Figure 2.19: Various phases of MOSFET operations for $V_{GS} > V_T$. (a) $V_{DS} = 0$, where the channel region is flat; (b) $V_{DS} < V_{GS} - V_T$, where the channel region is tapered towards the source; (c) $V_{DS} > V_{GS} - V_T$, where the channel region is pinched off from the drain.

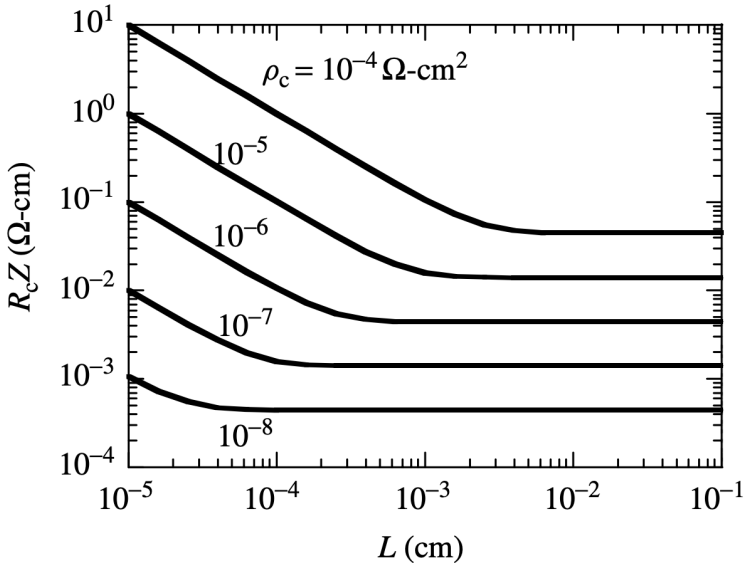

Figure 2.20: Front contact resistance-contact width product as a function of contact length and specific contact resistivity for sheet resistance R_{sh} = 20Ω/square and semiconductor resistance R_{sm} = 0. Reproduced from Fig. 3.18, [21].

contact resistance to take up an increasing portion of the overall device parasitic resistances (Fig. 2.21), hence becoming a major limiting factor of performance in ultra-scaled devices. It is critical for such devices to have contacts with low **(specific) contact resistivity** (contact resistance normalized by area).

Fundamentally, the contact resistivity originates from two physical mechanisms. First, according to the Schottky-Mott rule, [23, 24] the electronic properties of the contact is mainly determined by the relative magnitude of the work function of the metal ϕ_m and that of the semiconductor ϕ_s, both defined to be the difference between the vacuum level E_{vac} and their respective Fermi level E_{Fm} and E_{Fs}. When the metal and the semiconductor come into contact, the energy bands of the semiconductor would bend as the Fermi level must remain constant across the interface under thermal equilibrium. Fig. 2.22 shows three

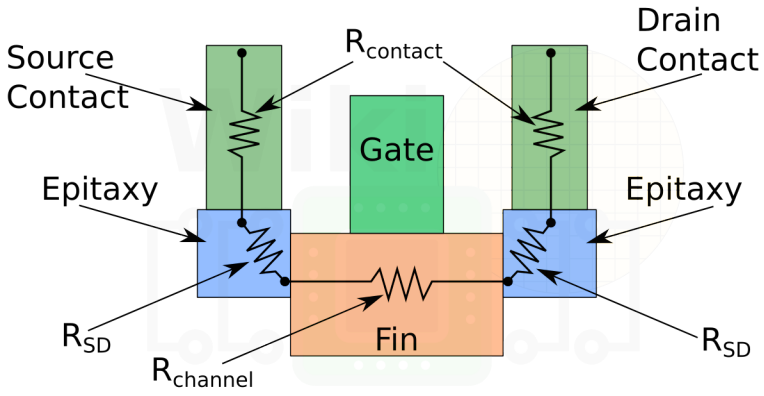

Figure 2.21: Components of the parasitic resistances in a MOSFET device: contact resistance (R_{contact}), source/drain resistance ($R_{\text{S/D}}$), and channel resistance (R_{channel}). Reproduced from [22].

possible scenarios of band bending for an n-type doped semiconductor, corresponding to $\phi_m < \phi_s$, $= \phi_s$, and $> \phi_s$ respectively. In reality, almost all metals have $\phi_m > \phi_s$ for common semiconductors, therefore a **Schottky barrier** exists at the metal-semiconductor interface, with height $\phi_B = \phi_m - \chi$ where χ is the electron affinity of the semiconductor. It is evident that the Schottky barrier height is a materials-specific property of the *bulk* metal and semiconductor, independent of the conditions of the interface. On the other hand, another important mechanism is the Fermi level pinning by defects occuring at the metal-semiconductor interface (Sec. 2.1.3). These defects trap the free electrons at their mid-gap energy levels, thereby limiting the ability for the electrons to move across the Schottky barrier. As the defects originate from the semiconductor surface, it is found experimentally that contrary to the Schottky model, the Schottky barrier height is largely independent of the work function of the metal for pinned semiconductors.

As expected, the contact resistivity of a metal-semiconductor junction depends positively on the Schottky barrier height. Another critical "knob" for

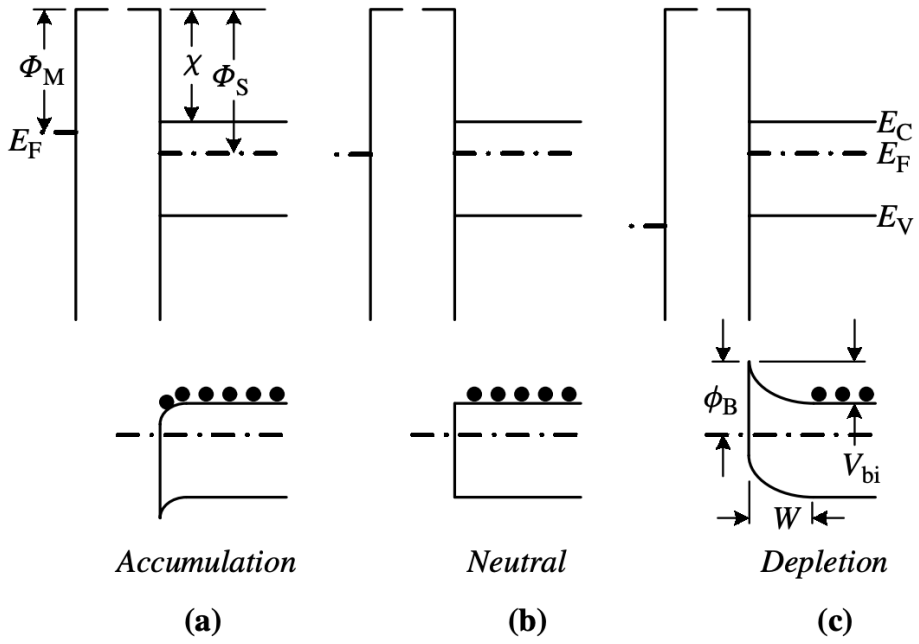

Figure 2.22: Metal-semiconductor contacts according to the simple Schottky model. The upper and lower parts of the figure show the metal-semiconductor system before and after contact, respectively. Reproduced from Fig. 3.1, [21].

tuning the contact resistivity is the level of doping in the semiconductor. The band diagrams of these three scenarios are illustrated in Fig. 2.23. Near the metal-semiconductor interface, there exists a region in the semiconductor that is depleted of free carriers (called **"depletion region"**) due to drift of carriers into the metal. The width of this depletion region is inversely proportional to the square root of doping level N_d in the semiconductor at a given ϕ_B. Therefore, depending on the semiconductor doping level, carrier transport across metal-semiconductor junction can manifest in one of three scenarios: thermionic emission (TE), thermionic-field emission (TFE), and field emission (FE). In thermionic emission, the depletion width is large due to light doping, hence the carriers must be thermally excited in order to cross the barrier, mak-

ing the contact a **rectifying junction** (diode). In field emission, the depletion width is sufficiently small due to heavy doping, so that the carriers can tunnel directly through the barrier without being thermally excited, making the contact an **Ohmic contact**. The onset of dependence of contact resistivity on doping concentration happens in the TFE regime, where tunneling starts to happen. Quantitatively, the contact resistivity in the FE regime is given by

$$\rho_c = \rho_{c0} \exp\left(\frac{2\phi_B}{\hbar}\sqrt{\frac{\epsilon_s m^*}{N_d}}\right), \tag{2.11}$$

where ρ_{c0} is a materials-specific constant, ϵ_s the dielectric constant of the semiconductor, and m^* the effective mass of the majority carrier in the semiconductor. It is evident from Eq. (2.11) that as the stringent scaling requirement demands very low contact resistivity in an ultra-scaled MOSFET device, it is critically important for a semiconductor material to be able to have sufficiently high concentrations of active dopants. This turns out to be one of the most serious bottlenecks for incorporation of III-V materials in next-generation MOSFET devices.

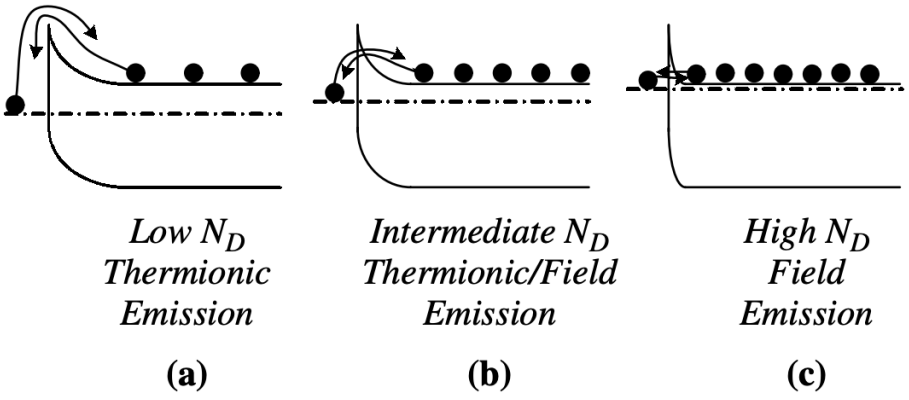

Figure 2.23: Depletion-type contacts to n-type substrates with increasing doping concentrations. The electron flow is schematically indicated by the electrons and their arrows. Reproduced from Fig. 3.3, [21].

2.3 Selected semiconductor materials of technological importance

Historically, semiconducting materials such as Si and GaAs have played major roles in the development of almost all kinds of electronic devices due to their exceptional properties. (Sec. 2.1.3) In recent years, several new semiconducting materials have come forward as candidate materials used in next-generation electronic devices, for their alleged superior properties compared to their already successful predecessors. This section will detail two kinds of such materials, which are the main focus of this thesis.

2.3.1 III-V compounds

Crystal structure

III-V compounds are an important class of semiconductors since the earliest days of electronics. The usage of gallium arsenide (GaAs) in transistors dates back to 1965. [25] Nowadays, GaAs finds widespread applications in electronic devices such as heterojunction bipolar transistors (HBTs), field effect transistors (FETs), and diodes, enabling high-speed and high-frequency electronics. [26] III-V compounds obtains its name from the fact that the family of compounds all consist of at least one group-IIIA species and one group-VA species. (Fig. 2.24) Two main crystal structures exist for III-V compounds: zinc blende and wurtzite. (Fig. 2.2) Such phenomenon is known as **"polymorphism"**, and individual structures **"polymorphs"**. Both polymorphs of binary III-V compounds share the same point group; namely, each atom is bonded covalently with four nearest-neighbor atoms of alternate species, forming a regular tetrahedra (known as **"tetrahedral coordination"**). The difference of these two polymorphs lies in their space groups: zinc blende structure belongs to space group $F4\bar{3}m$ in Herman-Mauguin notation (or number 216); wurtzite structure belongs to space group $P6_3mc$ in Herman-Mauguin notation (or number 186). It is well known that at room temperature, bulk arsenide crystals such as gallium arsenide (GaAs) and indium arsenide (InAs) are found in zinc blende structure, whereas bulk nitride crystals such as gallium nitride (GaN) and indium nitride (InN) are found in wurtzite structure. [27] The relative ground state energy of these two polymorphs depend only on the properties of free atoms of the constituent chemical species. [27]

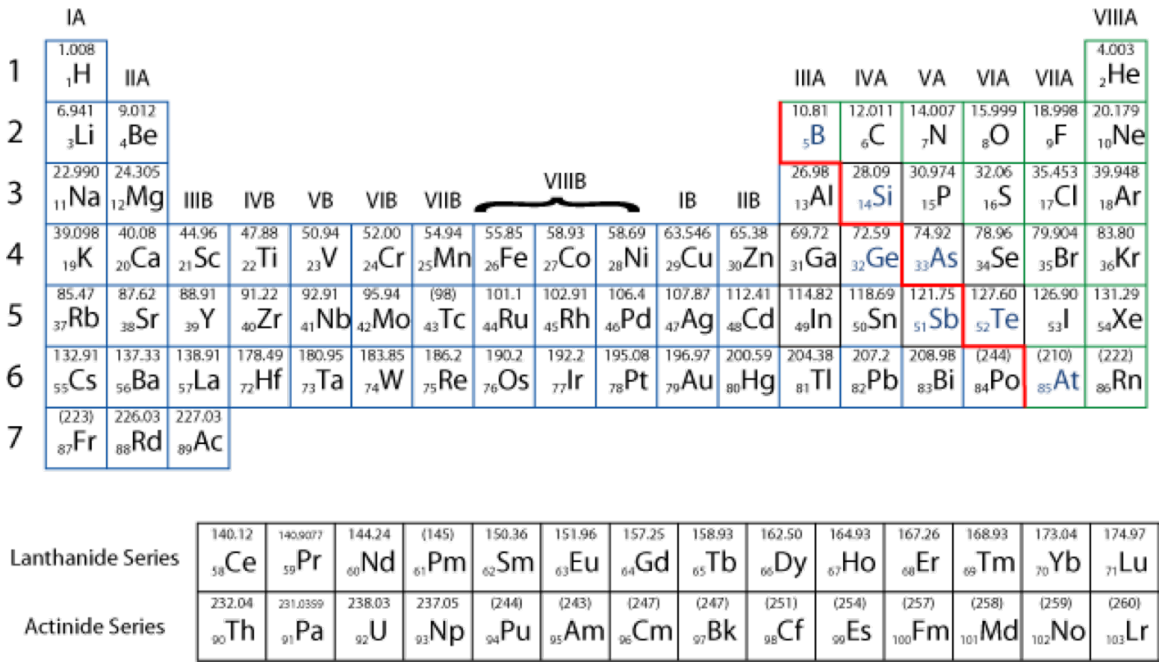

Figure 2.24: Periodic table of elements. Reproduced from [28].

Electronic properties

III-V semiconductors are long known for their superior electronic properties, thanks to their band structure. First, all III-V semiconductors are **direct band gap** semiconductors, meaning that the CBM and VBM are located at the same **k**-point. This allows the electrons to be excite directly from the VBM to the CBM without the assistance of phonons, as opposed to **indirect band gap** semiconductors such as silicon. This property is crucial for applications in optoelectronics such as photovoltaics, photodiodes, and laser diodes, as the electrons from the conduction band annihilates holes from the valence band through **radiative recombination**, emitting photon in the process. Another important advantage of III-V compounds is their significantly smaller effective electron masses m_e^* compared with silicon. (Table 2.5) This property has two major implications: (1)

higher electron injection velocity (v_{inj}) compared to silicon ($v_{\text{inj}} \sim \sqrt{1/m_e^*}$), and thus higher operating current (I_{on}) in transistors; (2) the electron mobility (μ_e) of III-V compounds is higher compared to silicon ($\mu_e \sim 1/m_e^*$), and hence faster switching speed of transistors. These superior electronic properties allow substantial improvement of device performance while keeping a moderate power consumption. This is the main reason and motivation that certain III-V materials are considered potential candidates for use as source, drain, and channel materials in next-generation field effect transistors (FETs) with gate length below 10nm.

Material/Property (T = 300K)	Si	Ge	GaAs	InAs	$In_{0.53}Ga_{0.47}As$	InSb
a (Å)	5.431	5.658	5.653	6.058	5.869	6.479
$m_e^*/m_e(\Gamma)$	0.19	0.08	0.063	0.023	0.041	0.014
N_C (cm^{-3})	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}	8.3×10^{16}	2.08×10^{17}	4.16×10^{16}
μ_e (cm^2/V-s)	1,450	3,900	9,200	33,000	12,000	77,000
E_g (eV)	1.12	0.66	1.42	0.35	0.74	0.17
ϵ_r	11.7	16.2	12.9	15.2	13.9	17.7

Table 2.5: Selected basic materials properties (lattice constant a, relative effective electron mass m_e^*/m_e at Γ-point, effective conduction band density of states N_C, electron mobility μ_e, band gap E_g, and dielectric constant ϵ_r) of some commonly used undoped semiconductors, measured at T = 300K. Data from [29].

Alloying and ordering

In semiconductor processing, it is often desirable to have a material with certain properties that are intermediate between those of two pure materials. For example, in applications such as metal-oxide-semiconductor field effect transistors (MOSFET), GaAs has a relatively large band gap (1.42 eV), which causes undesirable Fermi level pinning near CBM by As-As dimer antibonding (σ^*) states at the interface between GaAs and high-κ dielectric oxides such as hafnium ox-

ide (HfO_2) and aluminum oxide (Al_2O_3). [30, 31] On the other hand, InAs has a very small band gap (0.35 eV), allowing electrons to tunnel from the conduction band of the drain to the valence band of the substrate via band-to-band tunneling (BTBT), resulting in leakage current which degrades the performance of the MOSFET device. [32, 33] One way to mitigate these problems is to mix GaAs and InAs during growth phase to form an alloy compound $In_{1-x}Ga_xAs$, where indium and gallium atoms constitute $1 - x$ and x proportions of all the cations in the crystal respectively ($0 \leq x \leq 1$). Such a **pseudobinary** compound has intermediate band gap between that of GaAs and of InAs, which is large enough to reduce the magnitude of BTBT current, and small enough so that the interfacial defect states lies entirely within the conduction band. Furthermore, at a specific composition $In_{0.53}Ga_{0.47}As$, the lattice constant of the alloy (a = 5.869Å) matches exactly with that of InP, a commonly used III-V substrate, allowing perfect epitaxial growth without inducing strain and dislocations. These advantages make $In_{0.53}Ga_{0.47}As$ a very appealing candidate for future generation electronic devices such as MOSFETs.

In a real-world alloy such as InGaAs, the arrangement of different cation species (In and Ga in InGaAs) on their common sublattice deviates from that of a **random alloy**, where each cation site is randomly occupied by either an indium or a gallium atom, with respective probability $1-x$ and x. (Fig. 2.25) Under certain circumstances, atoms of same species tend to aggregate while atoms of different species tend to separate, which leads to **phase separation** if the constituent pure materials form distinct regions ("domains") inside the alloy. In other scenarios, atoms tend to arrange themselves in regular fashion, thereby exhibiting either **short-range order (SRO)**, where the number of A-B type bonds exceeds that in a random alloy; or **long-range order (LRO)**, where the atoms

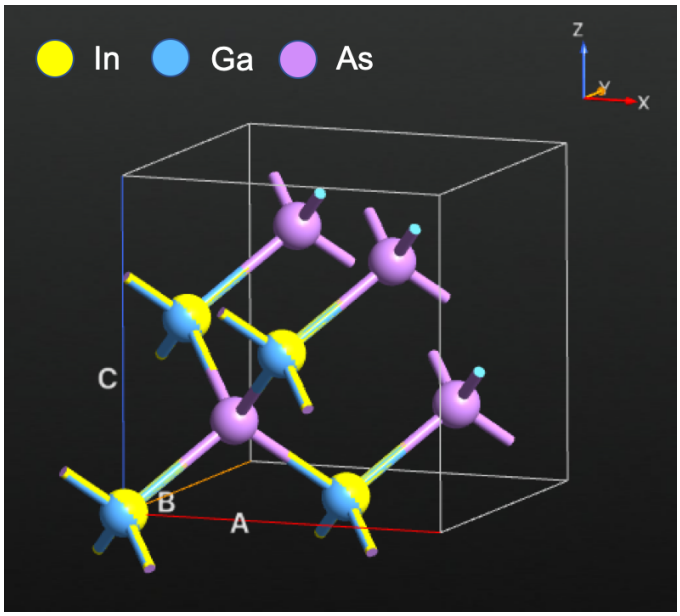

Figure 2.25: The schematic conventional unit cell of random $In_{0.5}Ga_{0.5}As$; each cation lattice site is equally likely to be occupied by either an indium (In) atom or a gallium (Ga) atom. Note that the "unit cell" depicted here only represents the repeatable unit of the underlying lattice, not the actual atomic arrangement within the cell.

form periodic superstructures called **superlattices** with periodicity greater than that of the constituent crystal in at least one direction. [1] It is observed in experiments that long-range ordering would spontaneously appear in almost all III-V alloys under certain growth conditions across the spectrum of epitaxial growth methods, such as liquid phase epitaxy (LPE), molecular-beam epitaxy (MBE), metal-organic vapour-phase epitaxy (MOVPE), and vapor-levitation epitaxy (VLE). [35] Fig 2.26 shows commonly observed types of superlattice structures for III-V alloys.

According to thermodynamics of bulk crystals, the tendency of atoms in an

[1] The degree of short-range order and long-range order in an alloy can be quantified; one of the methods is proposed by Cowley [34].

alloy to exhibit ordering is determined by three main factors: formation *enthalpy* of the random phase (ΔH_R), formation *entropy* of the random phase (ΔS_R), and the excess energy of the order phase of interest relative to its constituent pure phases (ΔE_O). To be precise, if the alloy is grown above the temperature $T_R = \Delta H_R/\Delta S_R$, the alloy is more stable in the random phase; on the other hand, if the alloy is grown below the temperature $T_O = (\Delta H_R + \Delta E_O)/\Delta S_R$, then the alloy is more stable in the ordered phase of interest. [36] Using high-accuracy computational modeling, it is found that the excess energy ΔE_O of most long-range ordered (LRO) bulk phases of III-V pseudobinary alloys are positive except for chalcopyrite and famatinite phases; nevertheless, the excess energy of all these metastable LRO bulk phases are very small (the largest value being 110 meV for $In_{0.5}Ga_{0.5}As$ with CuPt-B phase [37]). Combined with the experimental observation of various LRO phases, an important implication is that factors other than bulk thermodynamics must play an decisive role in the formation of ordering in the alloy. For example, it is found that if the alloy grown is *coherent* on an lattice-mismatched substrate (i.e. the interface is free of dislocations), then the ordered phases are much more likely to occur, especially at low growth temperatures. This is because the external strain on the alloy lattice imposed by the substrate is maximally relieved by increasing the fraction of certain types of atomic clusters while decreasing the fraction of others, thereby creating ordering that deviates from total randomness. [38] In addition, *surface* thermodynamics and growth kinetics may also influence the type of ordering of the alloy. [39] For example, it is shown that certain LRO phases of $In_{0.5}Ga_{0.5}P$, such as CuPt-B and CuAu-I, possess the lowest formation energy when such ordering occurs near the surface due to reconstruction, a trend in contrast with the results of the bulk alloy. [39] In terms of growth kinetics, it is found that growth temperature, growth

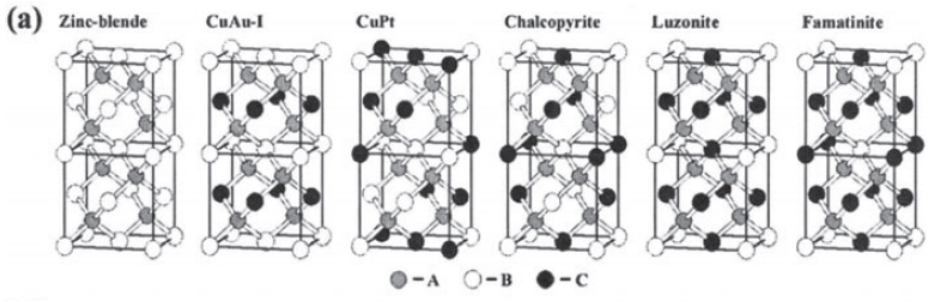

Figure 2.26: Atomic models for the random zinc blende structure and various possible superlattice structures for III-V alloys. Reproduced from Fig. 2.1, [41].

rate and substrate orientation can all influence the degree and type of LRO in near-surface $GaAs_{0.5}Sb_{0.5}$ grown by MOVPE. [40] Knowledge of ordering of a III-V alloy is very important, as the degree and type of LRO can alter many basic properties of the semiconductor alloy compared to the random case, such as reduction of band gap, splitting of valence bands, transition from indirect band gap to direct band gap, decrease of effective electron mass, and induction of internal electric field. [39]

Challenges

Despite the many advantages of III-V materials described above, widespread usage of III-V materials in field effect transistors has yet to become a reality. Several major obstacles must be overcome in order to truly unleash the promised potential of III-V materials as an alternative material for next-generation logic devices.

First, one of the key advantages of silicon is the high-quality interface it can form with its native oxide, SiO_2, which can be readily formed when sili-

con is oxidized. Such interface has a very low mid-gap interfacial defect density ($\sim 10^{10}\text{cm}^{-2}\text{eV}^{-1}$) [42]. In contrast, native oxides of III-V compounds such as gallium oxide (Ga_2O_3) do not possess such beneficial properties. For example, when exposed to air or low-vacuum environment, the oxidized surface of GaAs readily forms a wide variety of unwanted defects such as As and Ga dangling bonds, As-As dimers, and Ga vacancies, thereby introducing high interfacial defect density between the substrate and the native oxide. [31] Such consequence is undesirable for scaled-down transistors, due to Fermi level pinning effect which causes an increase in the subthreshold swing and renders the gate modulation ineffective. This has historically been one of the main bottlenecks of III-V-based FETs. Fortunately, in recent years important progress has been made that greatly reduces the interfacial defect density. Specifically, atomic layer deposition (ALD) of high-κ oxides such as HfO_2 and Al_2O_3 is used to suppress the spontaneous formation of poor-quality native oxide on GaAs surface [42]. Experiments have shown that the Fermi level is largely unpinned at the interface between GaAs and ALD-deposited oxide. [42] ALD-deposited oxides has been demonstrated on other III-V compounds as well. [31] It is found that surface processing techniques such as pre-deposition cleaning treatment, as well as using III-V compounds containing indium, significantly improves the quality of the oxide-substrate interface. The former method removes the III-V surface defects, while the latter method pushes the defect states above the conduction band due to the smaller band gap compared to GaAs.

Second, due to the immense success of silicon in driving Moore's law for five decades, currently existing semiconductor processing tools and infrastructure are designed and tuned to work with silicon substrates on which the chips are made. It is hence a great economical advantage if III-V compounds can be

integrated on Si substrates. However, this has met with numerous challenges, mostly due to the relatively large lattice mismatch between the two class of materials. (Table 2.5) This lattice mismatch creates a large number of defects such as dislocations and antiphase domain boundaries at the interface between the Si substrate and the epitaxially grown III-V compound, which leads to shorts and traps that severely degrades the device performance. [43] Various methods have been devised to tackle this problem with limited success, such as using graded ternary III-V buffer layers with continuously varying lattice constant, transfering the III-V thin film onto dielectric-covered Si substrate, and aspect ratio trapping (ART) where III-V material is grown inside trenches formed by SiO_2 on top of Si substrate. [44] One of the more recent and promising route is template-assisted selective-area epitaxy (TASE), [45] where the growth of III-V materials is confined by a "template" to a nanometer-thin cross section (with diameter < 100nm) on Si substrate. This small cross section completely eliminates the typical growth defects that result from bulk growth. Furthermore, this method allows growth of III-V material on Si substrate with arbitrary orientation, unlimited by the lattice spacings of Si substrate in a particular direction.

Last but not least, the parasitic resistances of a transistor have a great impact on the magnitude of I_{on} at a given V_{GS} and V_{DS}. Both the channel resistance R_{channel} and the contact resistance R_{contact} depend negatively on the doping level N_d in the semiconductor, hence it is critical to obtain a sufficiently high level of doping for extreme-scaled transistors. For silicon, n-type doping that results in a free electron concentration of above 10^{20}cm^{-3} is routinely achieved. [46, 43] In contrast, n-type doping in InGaAs by Si, one of the dopants with best activation and low diffusivity, is found to saturate at a free electron concentration of $\approx 1.5 \times 10^{19}\text{cm}^{-3}$ upon thermal annealing at high temperatures, a step required

for maximizing dopant activation and remove damage from dopant implantation. [47] (Fig. 2.27) This saturation level remains relatively constant regardless of implant conditions and anneal conditions. The root cause of this dopant activation limiting phenomenon was not perfectly understood. Several hypotheses have been proposed aiming for an explanation, including (1) insufficient solid solubility of dopants, (2) amphoteric nature of group IV dopants in III-V compounds, and (3) presence of charge-compensating defects such as vacancies. [43] Chap. 4 of this thesis presents a comprehensive computational modeling study confirming that the formation of negatively-charged cation (In/Ga) vacancies is responsible for the dopant deactivation. In addition, many strategies for breaking the limit of active dopant concentration have been proposed, including lowering growth temperature, using annealing techniques with shorter time duration such as rapid thermal annealing or laser spike annealing, [48] and co-implanation with other low diffusivity dopants. The motivation of the last strategy is that, due to the fact that group-VI elements such as sulfer (S), selenium (Se) or tellurium (Te) would preferentially occupy group-V (arsenic) sites in InGaAs, co-doping with group-VI dopants would likely result in reduced amphoteric doping of Si and hence increased carrier concentration. Experiments so far have shown limited degree of improvement in dopant activation by co-implanation. [43] Especially surprising is the fact that co-implantation of silicon with sulfer, which shows high percentage of activation when doped alone, exhibits only slight increase in the carrier concentration upon annealing to $1.7 \times 10^{19} \mathrm{cm}^{-3}$. [49] It is not clear whether most of the group-VI dopants truly occupy group-V lattice sites in these samples. Spectroscopic methods such as infrared (IR) or Raman spectroscopy are deployed to determine the local dopant configuration, but such methods are challenging to provide definitive answers

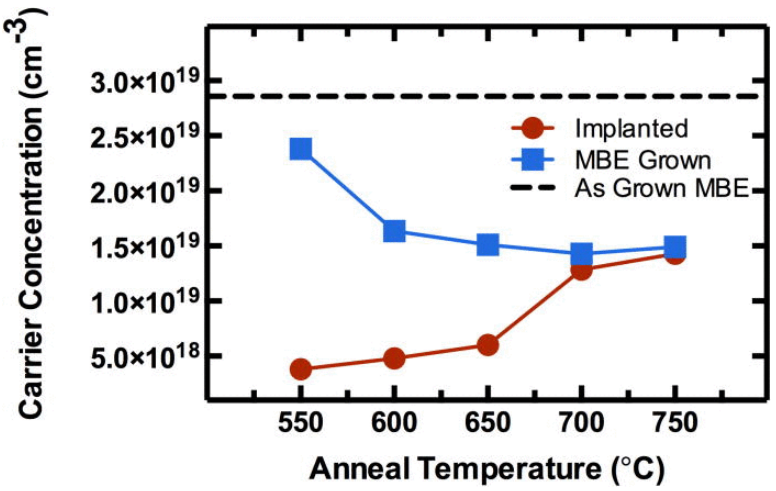

Figure 2.27: Calculated net donor concentration (N_d) of Si^+-implanted and MBE Si-doped $In_{0.53}Ga_{0.47}As$ specimens as a function of annealing temperature after 10m furnace anneals at 550, 600, 650, 700, and 750°C. Reproduced from Fig. 4, [47].

due to variations in the cation arrangement in a short-ranged or near-random alloy. Chap. 5 of this thesis aims to tackle this problem by a novel *computational* scheme which links particular local phonon modes to certain dopant / defect configurations in a random alloy, taking into full account the dependence on local atomic environment. Finally, the 2018 version of the International Roadmap for Devices and Systems (IRDS) [50] requires that for sub-10nm transistor devices, the contact resistivity must be less than $1 \times 10^{-9}\Omega$-cm^2. Current contact technologies for III-V semiconductors cannot yet achieve this ambitious goal, (Fig. 2.28) which requires a fundamental understanding of the metal-semiconductor interface on the atomic level. Chap. 6 of this thesis focuses on the impact of various material- and atomistic-level factors on the contact resistivity of Ni-In(Ga)As interface.

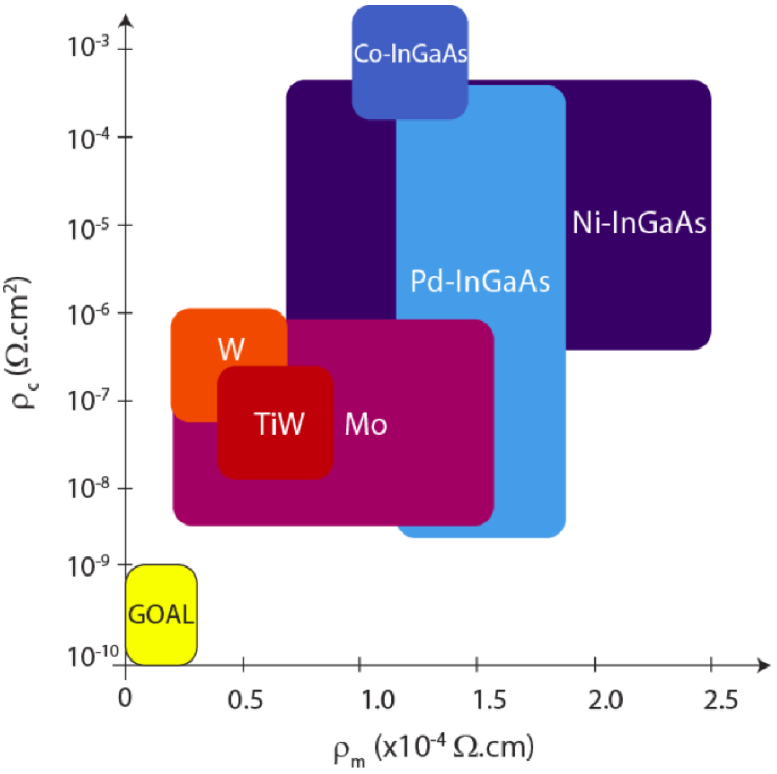

Figure 2.28: Landscape of contact resistivity vs. metal film resistivity of Si-compatible ohmic contacts to n-InGaAs [51]. The desired regime of operation is the bottom left corner. Reproduced from Fig. 4, [52].

2.3.2 Gallium oxide

Metal oxides, like III-V materials, is another class of semiconductor material with important electronic applications. Among them, gallium oxide (Ga_2O_3) has in recent years experienced a surge of interest due to its superior electronic properties compared to conventional semiconductors in power electronics, such as wide band gap (4.6 – 4.9eV), high breakdown field (6 – 8MV-cm^{-1}), and excellent chemical and thermal stability. [53, 54] (Table 2.6) These properties makes Ga_2O_3 an attractive candidates for a wide range of applications, including power electronics, solar blind UV detectors and gas sensors. [53] There are five identified

polymorphs of Ga_2O_3, namely corundum (α), monoclinic (β), defective spinel (γ), and orthorhombic (δ, ϵ). [53] Out of these polymorphs, β-Ga_2O_3 is the most stable phase at standard condition, therefore making it the most used in industry. Fig. 2.29 shows the crystal structure of β-Ga_2O_3, and two of the most commonly used surfaces of β-Ga_2O_3 in device applications. [55]

Material/Property (T = 300K)	Si	GaAs	4H-SiC	GaN	Ga_2O_3
E_g (eV)	1.12	1.42	3.25	3.4	4.85
ϵ_r	11.8	12.9	9.7	9.0	10.0
E_c (MV-cm^{-1})	0.3	0.4	2.5	3.3	8.0

Table 2.6: Selected basic materials properties (band gap E_g, dielectric constant ϵ_r, and breakdown field E_c) of some commonly used undoped semiconductors, measured at T = 300K. Data from [53].

Unlike InGaAs, which shows a bottleneck of n-type doping at $\sim 10^{19}$cm^{-3} (Sec. 2.3.1), Ga_2O_3 can be doped n-type up to $\sim 10^{20}$cm^{-3}. [54] Nonetheless, the doping level actually achieved varies greatly depending on the dopant species and growth condition. One of the issues with doping in Ga_2O_3 is that some species of dopants (for example, tin (Sn)) experience significant degree of segregation towards the surface when subjected to thermal treatment such as annealing; namely, most of the grown-in dopants to leave their activated bulk sites and form a separate contiguous solid phase above the Ga_2O_3 surface. This phenomenon severely hampers the performance of Sn-doped Ga_2O_3. [56] To date, the cause of dopant segregation in Ga_2O_3 has not been elucidated. In Chap. 7 of this thesis, we use computational modeling to reveal that both surface thermodynamics of dopants and surface defects (vacancies) play a big role in dopant segregation towards the Ga_2O_3(010) surface.

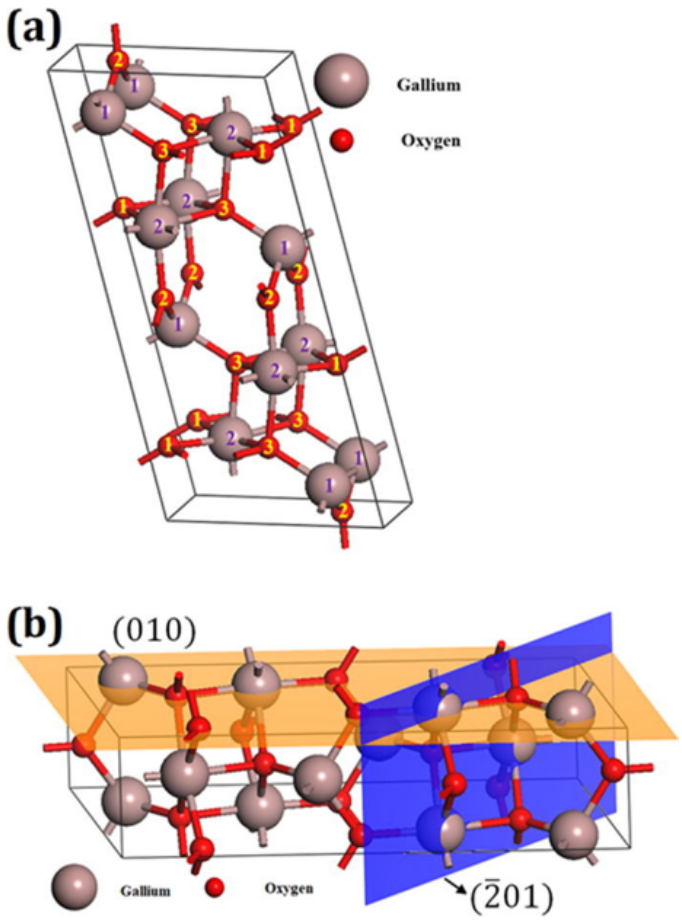

Figure 2.29: (a) β-Ga_2O_3 crystal structure; (b) (010) and ($\bar{2}$01) surfaces. Reproduced from Fig. 4, [55].

Bibliography

[1] Robert F Pierret. *Semiconductor device fundamentals*. Pearson Education India, 1996.

[2] Peter Y Yu and Manuel Cardona. *Fundamentals of semiconductors: physics and materials properties*. Springer, 2010.

[3] Robert F Pierret and Gerold W Neudeck. *Advanced semiconductor fundamentals*, volume 6. Addison-Wesley Reading, MA, 1987.

[4] Simon M Sze and Kwok K Ng. *Physics of semiconductor devices*. John wiley & sons, 2006.

[5] Wikipedia contributors. Band gap — Wikipedia, the free encyclopedia, 2019. [Online; accessed 21-August-2019].

[6] Peter Hadley. The first brillouin zone of a face centered cubic lattice, 2019. [Online; accessed 31-August-2019].

[7] Marcus W Doherty, Neil B Manson, Paul Delaney, Fedor Jelezko, Jörg Wrachtrup, and Lloyd CL Hollenberg. The nitrogen-vacancy colour centre in diamond. *Physics Reports*, 528(1):1–45, 2013.

[8] Su-Huai Wei. Overcoming the doping bottleneck in semiconductors. *Computational Materials Science*, 30(3-4):337–348, 2004.

[9] Fedwa El-Mellouhi and Normand Mousseau. Self-vacancies in gallium arsenide: An ab initio calculation. *Physical Review B*, 71(12):125207, 2005.

[10] Chenming Hu. *Modern semiconductor devices for integrated circuits*, volume 2. Prentice Hall Upper Saddle River, NJ, 2010.

[11] Matthew D McCluskey and Eugene E Haller. *Dopants and defects in semiconductors*. CRC Press, 2012.

[12] WE Spicer, PW Chye, CM Garner, I Lindau, and P Pianetta. The surface electronic structure of 3–5 compounds and the mechanism of fermi level pinning by oxygen (passivation) and metals (schottky barriers). *Surface Science*, 86:763–788, 1979.

[13] WE Spicer, PW Chye, PR Skeath, C Yu Su, and I Lindau. New and unified model for schottky barrier and iii–v insulator interface states formation. *Journal of Vacuum Science and Technology*, 16(5):1422–1433, 1979.

[14] Winfried Mönch. *Semiconductor surfaces and interfaces*, volume 26. Springer Science & Business Media, 2013.

[15] John Robertson. Model of interface states at iii-v oxide interfaces. *Applied Physics Letters*, 94(15):152104, 2009.

[16] Phaedon Avouris and Christos Dimitrakopoulos. Graphene: synthesis and applications. *Materials today*, 15(3):86–97, 2012.

[17] Eicke R Weber. *Imperfections in III/V materials*, volume 38. Elsevier, 1993.

[18] MD McCluskey. Local vibrational modes of impurities in semiconductors. *Journal of Applied Physics*, 87(8):3593–3617, 2000.

[19] Martin T Dove and Martin T Dove. *Introduction to lattice dynamics*, volume 4. Cambridge university press, 1993.

[20] HH Berger. Contact resistance and contact resistivity. *Journal of the Electrochemical Society*, 119(4):507–514, 1972.

[21] Dieter K Schroder. *Semiconductor material and device characterization*. John Wiley & Sons, 2015.

[22] Vlsi 2018: Globalfoundries 12nm leading-performance, 12lp, 2018. [Online; accessed 2-September-2019].

[23] W Schottky. Deviations from ohm's law in semiconductors. *Phys. Z*, 41:570, 1940.

[24] NF Mott. Note on the contact between a metal and an insulator or semiconductor. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 34, pages 568–572. Cambridge University Press, 1938.

[25] H Becke, R Hall, and J White. Gallium arsenide mos transistors. *Solid-State Electronics*, 8(10):813–823, 1965.

[26] Albert G Baca and Carol IH Ashby. *Fabrication of GaAs devices*. Number 6. IET, 2005.

[27] Chin-Yu Yeh, ZW Lu, S Froyen, and Alex Zunger. Zinc-blende–wurtzite polytypism in semiconductors. *Physical Review B*, 46(16):10086, 1992.

[28] Philip J Grandinetti. Periodic table, 2019. [Online; accessed 31-August-2019].

[29] Semiconductors on nsm. http://www.ioffe.ru/sva/nsm/semicond/, cited August 2019.

[30] L Lin and J Robertson. Defect states at iii-v semiconductor oxide interfaces. *Applied Physics Letters*, 98(8):082903, 2011.

[31] Jesús A Del Alamo. Nanometre-scale electronics with iii–v compound semiconductors. *Nature*, 479(7373):317, 2011.

[32] Jianqiang Lin, Dimitri A Antoniadis, and Jesús A del Alamo. Off-state leakage induced by band-to-band tunneling and floating-body bipolar effect in ingaas quantum-well mosfets. *IEEE Electron Device Letters*, 35(12):1203–1205, 2014.

[33] Qing Chen and Wenyuan Yang. The scaling-down and performance optimization of inas nanowire field effect transistors. *ECS Transactions*, 86(7):41–49, 2018.

[34] JM Cowley. Short-range order and long-range order parameters. *Physical Review*, 138(5A):A1384, 1965.

[35] GB Stringfellow and GS Chen. Atomic ordering in iii/v semiconductor alloys. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 9(4):2182–2188, 1991.

[36] GP Srivastava, José Luis Martins, and Alex Zunger. Atomic structure and ordering in semiconductor alloys. *Physical Review B*, 31(4):2561, 1985.

[37] JE Bernard, RG Dandrea, LG Ferreira, S Froyen, S-H Wei, and A Zunger. Ordering in semiconductor alloys. *Applied physics letters*, 56(8):731–733, 1990.

[38] Jeffrey Y Tsao. *Materials fundamentals of molecular beam epitaxy*. Academic Press, 2012.

[39] Alex Zunger. Spontaneous atomic ordering in semiconductor alloys: Causes, carriers, and consequences. *MRS Bulletin*, 22(7):20–26, 1997.

[40] HR Jen, MJ Jou, YT Cherng, and GB Stringfellow. The kinetic aspects of ordering in gaas1-xsbx grown by organometallic vapor phase epitaxy. *Journal of Crystal Growth*, 85(1-2):175–181, 1987.

[41] Angelo Mascarenhas. *Spontaneous ordering in semiconductor alloys*. Springer Science & Business Media, 2012.

[42] PD Ye, GD Wilk, J Kwo, BAYB Yang, H-JL Gossmann, MAFM Frei, SNG Chu, JP Mannaerts, MASM Sergent, MAHM Hong, et al. Gaas mosfet with oxide gate dielectric grown by atomic layer deposition. *IEEE Electron Device Letters*, 24(4):209–211, 2003.

[43] AG Lind, HL Aldridge, C Hatem, ME Law, and KS Jones. Dopant selection considerations and equilibrium thermal processing limits for n+-in0. 53ga0. 47as. *ECS Journal of Solid State Science and Technology*, 5(5):Q125–Q131, 2016.

[44] Heike Riel, Lars-Erik Wernersson, Minghwei Hong, and Jesus A Del Alamo. Iii–v compound semiconductor transistors—from planar to nanowire structures. *Mrs Bulletin*, 39(8):668–677, 2014.

[45] H Schmid, Mattias Borg, K Moselund, L Gignac, CM Breslin, J Bruley, D Cutaia, and H Riel. Template-assisted selective epitaxy of iii–v nanoscale devices for co-planar heterogeneous integration with si. *Applied Physics Letters*, 106(23):233101, 2015.

[46] Mao Wang, A Debernardi, Y Berencén, R Heller, Chi Xu, Ye Yuan, Yufang Xie, R Böttger, L Rebohle, W Skorupa, et al. Breaking the doping limit in silicon by deep impurities. *Physical Review Applied*, 11(5):054039, 2019.

[47] Aaron G Lind, Henry L Aldridge Jr, Cory C Bomberger, Christopher Hatem, Joshua MO Zide, and Kevin S Jones. Comparison of thermal annealing effects on electrical activation of mbe grown and ion implant si-doped in0. 53ga0. 47as. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, 33(2):021206, 2015.

[48] Victoria Carrie Sorg. Laser annealing and dopant activation in iii-v materials. 2017.

[49] Aaron G Lind, Henry L Aldridge Jr, Kevin S Jones, and Christopher Hatem. Co-implantation of al+, p+, and s+ with si+ implants into in0. 53ga0. 47as. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, 33(5):051217, 2015.

[50] International roadmap for devices and systems (irds™) 2018 edition. `https://irds.ieee.org/editions/2018`. Accessed: 2019-08-16.

[51] J Lin, DA Antoniadis, and JA del Alamo. Ingaas quantum-well mosfet arrays for nanometer-scale ohmic contact characterization. *IEEE Transactions on Electron Devices*, 63(3):1020–1026, 2016.

[52] Jesús A Del Alamo, Dimitri A Antoniadis, Jianqiang Lin, Wenjie Lu, Alon Vardi, and Xin Zhao. Nanometer-scale iii-v mosfets. *IEEE Journal of the Electron Devices Society*, 4(5):205–214, 2016.

[53] Michael A Mastro, Akito Kuramata, Jacob Calkins, Jihyun Kim, Fan Ren, and SJ Pearton. Perspective—opportunities and future directions for ga2o3. *ECS Journal of Solid State Science and Technology*, 6(5):P356–P359, 2017.

[54] Marko J Tadjer, John L Lyons, Neeraj Nepal, Jaime A Freitas, Andrew D Koehler, and Geoffrey M Foster. Review—theory and characterization of doping and defects in β-ga2o3. *ECS Journal of Solid State Science and Technology*, 8(7):Q3187–Q3194, 2019.

[55] SJ Pearton, Jiancheng Yang, Patrick H Cary IV, F Ren, Jihyun Kim, Marko J Tadjer, and Michael A Mastro. A review of ga2o3 materials, processing, and devices. *Applied Physics Reviews*, 5(1):011301, 2018.

[56] Kohei Sasaki, Masataka Higashiwaki, Akito Kuramata, Takekazu Masui, and Shigenobu Yamakoshi. Growth temperature dependences of structural and electrical properties of ga2o3 epitaxial films grown on β-ga2o3 (010) substrates by molecular beam epitaxy. *Journal of Crystal Growth*, 392:30–33, 2014.

CHAPTER 3

THEORIES AND METHODS

This chapter provides a detailed discussion of all the theories and computational methods used in materials modeling and simulation relevant to this thesis.

3.1 Computational materials science

Since the dawn of human history, the discovery and deployment of materials have always been a story of trial and error. Successful application of a new material typically involves thousands of failures and long periods of research and development, a fact vividly captured by the quote of Thomas A. Edison, the famous inventor who discovered successful usage of tungsten filament in incandescent light bulbs: "I have not failed. I just found 10,000 ways that won't work." While this strenuous method of materials discovery and experimentation had its glory days in the past, it can no longer keep up with the exponentially-increasing need of today's lightning fast development of human civilization. A cheaper, faster and more effective way of materials research and development is currently in critical need.

With the rapid increase of computing power enabled by Moore's law, nowadays we have more computational resources and capacity than ever before. Computational science, "a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problems", [1] has become increasingly important in the 21st century that it has been termed the "third pillar" of science, alongside theory and experiment. [1] Computa-

tional science achieves this end by making use of computational models, which can be regarded as a approximated version of reality that is simple enough yet captures the essence of the problem at hand. This new method of investigation has many advantages over the other paradigms; it is mostly suitable when the problem is too complicated to be analyzed by theory, or too costly, takes too long, or too unrealistic to be studied by experiments. [1] It is with these advantages that computational sciences has found wide applications in various traditional areas such as social sciences, physical sciences, biological sciences and medicine, climate sciences, geosciences, national security, and engineering and manufacturing. [1]

Computational materials science, an intersection of materials science and computation, aims to understand the properties and phenomena of materials and help design novel, better materials. [2] Computational materials science has the potential to guide and complement costly and time-consuming experiments on an unprecedented scale, the importance of which has been fully recognized in governmental programs such as "Materials Genome Initiative" of the United States and "Horizon 2020" of the European Union. Computational materials science covers both static and dynamic phenomena of materials. Depending on the length scale and time scale, different computational methods are deployed to best capture the most relevant physical mechanisms of the materials system. (Fig. 3.1) Typically, methods that are suitable for length scales closer to the atomic scale make fewer empirical assumptions and rely more on the fundamental theory of atomic interactions, thereby achieving higher accuracy; the downside is that the size of the system that can be simulated is typically much smaller than the higher-level phenomenological methods. At the bottom rung of the ladder of all methods lies **first principle** (***ab initio***) methods, which make

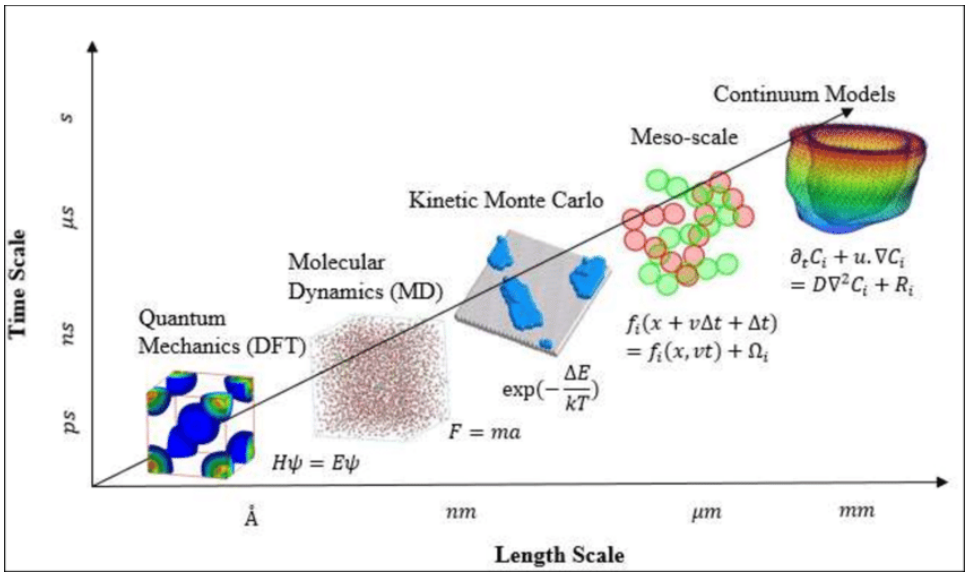

Figure 3.1: Typical methods in computational materials science in terms of size and time. Reproduced from [3].

(in principle) no additional assumptions than quantum mechanics itself. In this thesis, we use first principles (*ab initio*) density functional theory (DFT) in all of our works. This level of theory is required because the physics and chemistry of dopants and defects depends sensitively on the details of the atomic interaction, which requires a sufficiently accurate treatment.

3.2 Foundations of *ab initio* computational methods

3.2.1 Schrödinger's equation

All matters in nature are composed of atoms. With this basic fact comes an important insight: all materials follow the laws of nature that governs the interactions of atoms. It follows that as long as we have the knowledge of these

laws of nature, we can, at least in principle, know the properties of materials from the theory, just as we can know the motions of celestial bodies with the help of Newton's law of gravity and Einstein's theory of relativity. Fortunately, the fundamental laws of matter on the atomic scale have already been found, nearly a century ago: quantum mechanics, one of the pillars of modern physics and chemistry. At the core of quantum mechanics lies Schrödinger equation, the centural law that governs interactions of electrons and nucleus, the basic building blocks of everyday matters.

The **(time-independent) Schrödinger equation** can be expressed succinctly as:

$$\hat{H}\Psi = E\Psi. \tag{3.1}$$

$\hat{H}$ is called the **Hamiltonian operator** or simply the **Hamiltonian** of a system; it is a function which acts on another function (Ψ) to give the total energy E of all particles in the system, which is the sum of kinetic energy and potential energy. Ψ, called the **wavefunction**, is the fundamental quantity of quantum mechanics. It describes the physical state of all particles in the system, in the sense that $|\Psi|^2 = \Psi^*\Psi$ (* denotes complex conjugate) equals the *amplitude of probability* of the system in a given spatial configuration at a particular time. Note that the word "state" here assumes a different meaning from that in classical mechanics, where both a particle's position **r** and momentum **p** are needed to fully describe the state of a particle. In quantum mechanics, thanks to the Heisenberg uncertainty principle, we simply cannot know for certain both **r** and **p** at the same time, even in theory. Rather, what we *do* know for certain is that, when a measurement is made *on a particular physical quantity* (called an **"observable"**), such as position or energy, the wavefunction Ψ **collapses** from a **superposition** of eigenstates to a single **eigenstate** which yields a determinate value (one of the

eigenvalues) of *that observable only*. In other words, we cannot, even in theory, know what state of the system is in until we take a measurement. (This is called the "Copenhagen interpretation" of quantum mechanics.)

Any molecule or material other than a single hydrogen atom consists of more than two elementary particles. In a many-particle system containing M atoms and N electrons, the Schrödinger equation assumes the form

$$\hat{H}\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M) = E\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M), \tag{3.2}$$

where the wavefunction Ψ is a function of all electron positions and spins ($\mathbf{x}_i = \{\mathbf{r}_i, \sigma_i\}$) and all nucleus positions ($\mathbf{R}_I$). The many-body Hamiltonian is given by: [4]

$$\begin{aligned}
\hat{H} &= \hat{T}_e + \hat{T}_n + \hat{V}_{e-e} + \hat{V}_{n-e} + \hat{V}_{n-n}; \\
\hat{T}_e &= -\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2; \\
\hat{T}_n &= -\sum_{I=1}^{M}\frac{\hbar^2}{2M_I}\nabla_I^2; \\
\hat{V}_{e-e} &= \frac{1}{2}\sum_{j\neq i}^{N}\sum_{i=1}^{N}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}; \\
\hat{V}_{n-e} &= -\sum_{i=1}^{N}\sum_{I=1}^{M}\frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|}; \\
\hat{V}_{n-n} &= \frac{1}{2}\sum_{J\neq I}^{M}\sum_{I=1}^{M}\frac{Z_I Z_J e^2}{|\mathbf{R}_I - \mathbf{R}_J|},
\end{aligned} \tag{3.3}$$

where $\hat{T}_e$, $\hat{T}_n$, $\hat{V}_{e-e}$, $\hat{V}_{n-e}$, and $\hat{V}_{n-n}$ denote respectively the electron kinetic energy, nucleus kinetic energy, electron-electron Coulomb interaction energy, nucleus-electron Coulomb interaction energy, and nucleus-nucleus Coulomb interaction energy. It is important to fully grasp the complexity of the many-body wavefunction Ψ, which is a function of $3(M + N)$ variables. Indeed, it was not long

after the birth of quantum mechanics that scientists found solving the many-body Schödinger equation (3.3) is an unattainable task for all but the smallest system, the hydrogen atom. Nonetheless, a typical macroscopic piece of materials consists of roughly Avogadro's number (6.02×10^{23}) of atoms. It therefore seemed to many physicists and chemists in the early 20th century that it woule be very difficult, if not entirely hopeless, to understand materials and molecules entirely from the very law that governs them, i.e. quantum mechanics. As Paul A. M. Dirac, a founding father of quantum mechanics, famously put it in 1929: "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."

3.2.2 Born-Oppenheimer approximation

One of the first steps of approximation towards solving the "materials problem" was proposed by Max Born and J. Robert Oppenheimer in 1927. [5] They argued that as the mass of electrons (9.11×10^{-31} kg) is significantly smaller than the masses of atomic nuclei (the lightest nucleus is the proton, with the mass 1836 times the electron mass), the motion of an electron should be significantly faster than that of a nucleus. Therefore, from the electron point of view, the nuclei can be *regarded* as stationary. This insight makes it a reasonable approximation to separate the wavefunction of the system into an electron part and a nuclei part,

namely

$$\Psi(\mathbf{x}_{1,2,\ldots,N};\mathbf{R}_{1,2,\ldots,M}) = \Psi_e(\mathbf{x}_{1,2,\ldots,N};\mathbf{R}_{1,2,\ldots,M})\Psi_n(\mathbf{R}_{1,2,\ldots,M}). \quad (3.4)$$

In this expression, the electronic wavefunction Ψ_e depends only parametrically on the nuclei positions $\mathbf{R}_{1,2,\ldots,M}$, meaning that it can be regarded as the solution of an "electronic Hamiltonian" $\hat{H}_e$ where the nuclei kinetic energy $\hat{T}_n$ and the nucleus-nucleus Coulomb interaction energy $\hat{V}_{n-n}$ in the system Hamiltonian $\hat{H}$ vanish, and the nucleus-electron Coulomb interaction energy $\hat{V}_{n-e}$ becomes an additive constant. The total energy of the system can then be expressed as the sum of the electronic energy E_e = and the nuclei energy E_n. As nuclei do not participate in chemical reactions, it is thus very valuable to be able to focus on the wavefunction of only electrons, which now is a function of effectively only $3N$ variables.

3.2.3 Hartree-Fock method

One of the earliest attempts to approximately solve many-body Schrödinger's equation was the Hartree-Fock method. The basic assumption of Hartree-Fock method is that each electron in the system interact with all other electrons not in pair-wise fashion, but only through an effective mean-field Coulomb potential; this is known as the **independent electron approximation**. Douglas Hartree proposed in 1928 [6] that as a direct consequence of this approximation, the many-electron wavefunction Ψ_e can be written as the product of single-electron wavefunctions ψ_i:

$$\Psi_e(\mathbf{x}_{1,2,\ldots,N}) = \prod_{i=1}^{N} \psi_i(\mathbf{x}_i), \quad (3.5)$$

and the many-electron Hamiltonian $\hat{H}_e$ can be decomposed as a sum of single-electron Hamiltonians $\hat{h}_i$:

$$\begin{aligned}
\hat{H}_e &= \sum_{i=1}^{N} \hat{h}_i = \sum_{i=1}^{N} [\hat{t}_i + (\hat{v}_{n-e})_i + (\hat{v}_{e-e})_i]; \\
\hat{t}_i &= \frac{\hbar^2}{2m_e} \nabla_i^2; \\
(\hat{v}_{n-e})_i &= -\sum_{I=1}^{M} \frac{Z_I e^2}{|\mathbf{r}_i - \mathbf{R}_I|}; \\
(\hat{v}_{e-e})_i &= -e \int \mathrm{d}\mathbf{r}' \, \frac{n(\mathbf{r}')}{|\mathbf{r}_i - \mathbf{r}'|}
\end{aligned} \tag{3.6}$$

where $\hat{t}_i$, $(\hat{v}_{n-e})_i$ and $(\hat{v}_{e-e})_i$ are the one-electron kinetic energy, nuclear-electron potential energy, and electron-electron potential energy, respectively. The electron density $n(\mathbf{r})$ is approximated by

$$n(\mathbf{r}) = \sum_{\sigma_i} \sum_{i=1}^{N/2} |\psi_i(\mathbf{r})|^2. \tag{3.7}$$

The one-electron Schrödinger-like equation $[\hat{t}_i + (\hat{v}_{n-e})_i + (\hat{v}_{e-e})_i]\psi_i = \epsilon_i \psi_i$ is called the "Hartree equation".

The apparent advantage of Hartree's approximation is the extremely simple form. Nonetheless, it has a fatal flaw: the product form of the many-electron wavefunction (Eq. (3.5)) fails to satisfy the antisymmetry principle, namely exchanging the position-spin index $\mathbf{x}$ of any two electrons changes the sign of Ψ_e. Vladimir Fock refined the formalism in 1930, [7, 8] by recognizing that Ψ_e must be a particular linear combination of the product of ψ_i called the "Slater's determinant" in order to satisfy the antisymmetry principle. In this case, Ψ_e assumes the form

$$\Psi_e(\mathbf{x}_{1,2,\ldots,N}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_N(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) & \cdots & \psi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{x}_N) & \psi_2(\mathbf{x}_N) & \cdots & \psi_N(\mathbf{x}_N) \end{vmatrix}. \tag{3.8}$$

With this form of Ψ_e, it can be shown that an additional "exchange potential" term must be added to the Hartree equation to take into full account the anti-symmetry principle; namely,

$$[\hat{t}_i + (\hat{v}_{n-e})_i + (\hat{v}_{e-e})_i]\psi_i + \int d\mathbf{r}' \hat{v}_x(\mathbf{r}, \mathbf{r}')\psi_i(\mathbf{r}') = \tilde{\epsilon}_i \psi_i, \tag{3.9}$$

with

$$\hat{v}_x(\mathbf{r}, \mathbf{r}') = -\sum_{\sigma_j} \sum_{j}^{N/2} \frac{\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|}. \tag{3.10}$$

Equation (3.9) is called the "Hartree-Fock equation".

3.3 Density functional theory

Density functional theory is arguably one of the most successful theory in the physical sciences. It has become the workhorse for today's materials simulation and modeling in various disciplines, judging from the exponential increase in both the number of papers published and number of patents issued containing the keyword "density functional theory". ((Fig. 3.2)) One of the inventors of density functional theory, Walter Kohn, was awarded the Nobel Prize in Chemistry for his work in 1998. This section serves as an introduction to this remarkable method.

3.3.1 The Hohenberg-Kohn theorems

Although Hartree-Fock method helped greatly reduce the complexity of the problem, it is still based on the idea that knowledge of the wavefunction of the

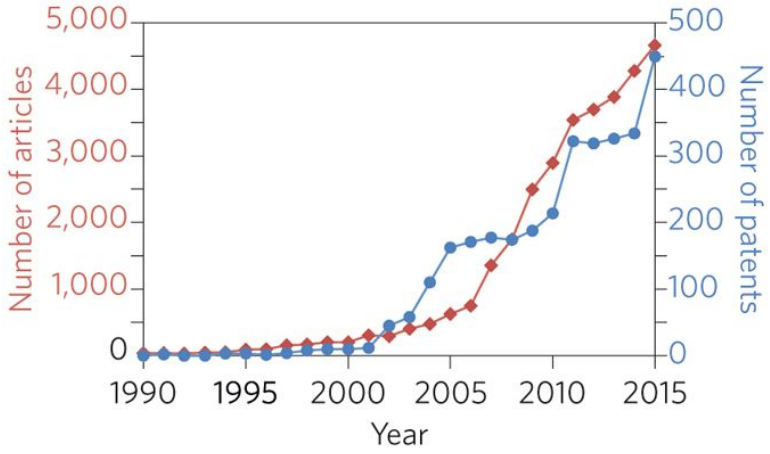

Figure 3.2: Number of articles and patents in materials science including the term "density functional theory" published per year during the past 25 years. Reproduced from [3].

system is required to solve for the energy of the system. This idea was the foundation of a collection of computational methods known as "wavefunction-based methods" including Hartree-Fock (HF) method, Møller-Plesset (MP) perturation theory, coupled cluster (CC) methods, multi-configurational self-consistent field (MCSCF), configuration interaction (CI), and so forth, each with various degree and type of approximations and accuracy. These methods all rely on solving for the electronic wavefunction, and therefore are typically restricted to very small systems containing < 100 electrons. This makes solution of many realistic materials and molecules intractable.

An alternative path towards solving the "materials problem" was first suggested independently by Llewellyn Thomas and Enrico Fermi in 1927. They proposed that the kinetic energy and the potential energy of *non-interacting* electrons in an infinite homogeneous electron gas (HEG) can be expressed as functionals (function acting on a function) of only the electron density $n(\mathbf{r})$ (hence

the name "density functional theory"), defined by

$$n(\mathbf{r}) = N_e \int \mathrm{d}\mathbf{r}_2 \ldots \int \mathrm{d}\mathbf{r}_N \sum_{\sigma_1} |\Psi_e(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)|^2, \tag{3.11}$$

i.e. the probability of finding *any* electron of any spin σ at a given position $\mathbf{r}$. In particular, the kinetic energy functional $T_e[n(\mathbf{r})]$, called the "Thomas-Fermi model", is given by

$$T_e[n(\mathbf{r})] = \frac{3}{10}(3\pi^2)^{2/3} \int \mathrm{d}\mathbf{r}\, n(\mathbf{r})^{5/3}. \tag{3.12}$$

This original idea was later refined by Dirac in 1930, attempting to take into account the interaction of electrons, namely **exchange** (the quantum-mechanical repulsion of same-spin electrons due to Pauli exclusion principle) and **correlation** (the non-independent spatial distribution of electrons due to Coulomb repulsion). Despite the novelty of these attempts, the results were not very accurate. This raises a fundamental question: can the state of the interacting many-body system be obtained uniquely by the electron density of the system?

This question was finally put to definitive rest in 1964, when a paper [9] containing two ingenious and elegant theorems named after Pierre Hohenberg and Walter Kohn profoundly changed the way scientists think about the relation between the electron density and the energy of a many-particles system. In essence, the Hohenberg-Kohn theorems establish an *exact* one-to-one correspondence between the ground state energy and a particular electron density (the ground state electron density) of *any* interacting many-particle system. This has drastically reduced the dimension of the problem from $3N$ to only 3, namely the spatial dimensions of the electron density $n(\mathbf{r}) = n(x, y, z)$. This astounding result makes it possible, at least in theory, to simulate realistic materials systems with size much larger than what is possible with the wavefunction-based methods. The first theorem states:

Theorem 1. *The external potential $\hat{V}_{ext}(\mathbf{r})$ of a system of interacting particles is an unique functional (up to an additive constant) of the ground state electron density $n_{GS}(\mathbf{r})$ of the system.*

In the case of many-particle systems, the external potential $\hat{V}_{\text{ext}}(\mathbf{r})$ of the electrons refers to the nuclear-electron repulsion term $\hat{V}_{n-e}$. The first theorem implies that, since the electronic Hamiltonian $\hat{H}_e$ is completely determined by fixing $\hat{V}_{n-e}$, the eigenstates (wavefunctions) of the system is also completely determined. Therefore, the electronic kinetic energy $\hat{T}_e$ and the electron-electron repulsion functional $\hat{V}_{e-e}$ are functionals of the ground state electron density as well. The total energy of the electrons can then be expressed as

$$\begin{aligned} E_e[n(\mathbf{r})] &= \underbrace{N_e \int \mathrm{d}\mathbf{r}_2 \ldots \int \mathrm{d}\mathbf{r}_N \sum_{\sigma_1} \Psi_e^* \hat{T}_e \Psi_e + N_e \int \mathrm{d}\mathbf{r}_2 \ldots \int \mathrm{d}\mathbf{r}_N \sum_{\sigma_1} \Psi_e^* \hat{V}_{e-e} \Psi_e}_{F_{\text{HK}}[n]} \\ &\quad + N_e \int \mathrm{d}\mathbf{r}_2 \ldots \int \mathrm{d}\mathbf{r}_N \sum_{\sigma_1} \Psi_e^* \hat{V}_{n-e} \Psi_e \\ &= F_{\text{HK}}[n(\mathbf{r})] + \int \mathrm{d}\mathbf{r}\, \hat{V}_{\text{ext}}(\mathbf{r})\, n(\mathbf{r}) \end{aligned} \tag{3.13}$$

where $F_{\text{HK}}[n(\mathbf{r})]$ is called the "Hohenberg-Kohn (HK) functional"; it is independent of the external potential V_{ext}, and is thus a "universal" functional.

The second theorem states:

Theorem 2. *The total electronic energy functional $E_e[n(\mathbf{r})]$ gives the lowest energy value E_{GS} for a given $\hat{V}_{ext}(\mathbf{r})$ if and only if the electron density $n(\mathbf{r}) = n_{GS}(\mathbf{r})$.*

Combined with the first theorem, we can draw a very powerful conclusion: the ground state properties (including energy) of *any* system is dependent on

the electron density *only*. In other words, as long as $F_{\mathrm{HK}}[n]$ is known, no wavefunction is required at all in order to know *exactly* all the information related to the ground state of an interacting many-particle system.

3.3.2 The Kohn-Sham equations

The universal functional $F_{\mathrm{HK}}[n]$ is the core concept of the Hohenberg-Kohn theorems. From Eq. (3.6), it is evident that $F_{\mathrm{HK}}[n]$ is an *implicit* functional of the electron density $n(\mathbf{r})$. Hohenberg-Kohn theorems, despite their conceptual significance, are *existence* theorems; namely, they prove that a universal functional capable of yielding the ground state energy for any system exists, but do not provide the actual form of dependence of the functional on $n(\mathbf{r})$.

The Kohn-Sham equations, proposed by Walter Kohn and Lu Jeu Sham in 1965, [10] concretize the form of $F_{\mathrm{HK}}[n]$ by making an important simplification, thereby marking the starting point of the practical density functional theory as we know it today. Kohn and Sham proposed that formally, $F_{\mathrm{HK}}[n]$ can be decomposed as

$$\begin{aligned} F_{\mathrm{HK}}[n(\mathbf{r})] &= T_s[n(\mathbf{r})] + J[n(\mathbf{r})] + E_{\mathrm{xc}}[n(\mathbf{r})] \\ &= T_s[n(\mathbf{r})] + \frac{1}{2}\int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\mathrm{xc}}[n(\mathbf{r})] \end{aligned} \tag{3.14}$$

where T_s is the kinetic energy functional of the *noninteracting* electron gas, J is the classical Coulomb self-interaction energy, and E_{xc}, named the **"exchange-correlation functional"**, contains all energy contributions due to exchange and correlation effects of the electron density not captured by T_s and J. This decomposition is meritable since T_s can be written down exactly (see next paragraph), and E_{xc}, while unknown, has a much smaller magnitude (typically accounts for

less than 10% of the total energy [2]) compared to the other two terms and hence does not introduce significant error in the total energy.

The crucial insight of Kohn and Sham is that they showed it is possible to construct a *fictitious* ground state density of the *noninteracting* electrons, such that it is equal to the ground state density n_{GS} of the real *interacting* electrons. The procedure is by substituting the many-electron problem by an *one-electron* problem, where each electron in the system is seen as moving in an noninteracting effective medium resulted from all other electrons. This way, the total electron density $n(\mathbf{r})$ can be written as the sum of one-electron effective densities $\tilde{n}_i(\mathbf{r})$:

$$n(\mathbf{r}) = \sum_{\sigma} \sum_{i=1}^{N/2} \tilde{n}_i(\mathbf{r}) = \sum_{\sigma} \sum_{i=1}^{N/2} |\phi_i(\mathbf{r})|^2 \tag{3.15}$$

where $\phi_i(\mathbf{r})$ are the fictitious Kohn-Sham wavefunctions of a single electron in the system. The index i runs through all orbitals (number of electrons / number of possible spins (2)). It is important to note that the purpose of ϕ_i is purely to construct the correct total density $n(\mathbf{r})$ of the interacting electron system, not to provide an accurate description of the actual state of any real electron in the system. The kinetic energy functional can then be expressed as

$$T_s[n(\mathbf{r})] = \frac{1}{2} \sum_{\sigma} \sum_{i=1}^{N/2} \int \mathrm{d}\mathbf{r}\, |\nabla \phi_i(\mathbf{r})|^2 \tag{3.16}$$

In this form, the kinetic energy T_s is not an explicit functional of $n(\mathbf{r})$, but rather an explicit functional of the one-electron Kohn-Sham wavefunctions $\phi_i(\mathbf{r})$. As the kinetic energy of the real interacting electron system T_e is in general different from T_s, the difference is entirely captured in the exchange-correlation functional E_{xc}. With the ansatz of Kohn-Sham wavefunctions, it can be shown that the electronic Hamiltonian (3.6) can be written as a sum of one-electron operators. This leads to a Schrödinger-like equation, called the **Kohn-Sham (K-S)**

equation, which acts on a single Kohn-Sham wavefunction $\phi_i(\mathbf{r})$ and yields the ground state density $n_{\mathrm{GS}}(\mathbf{r})$ and energy E_{GS} of the entire system:

$$\begin{aligned} \hat{H}_{\mathrm{KS}}\,\phi_i(\mathbf{r}) &= \varepsilon_i^{\mathrm{KS}}\,\phi_i(\mathbf{r}), \\ \hat{H}_{\mathrm{KS}} = \frac{\delta E_e[n]}{\delta n(\mathbf{r})} &= \frac{\hbar^2}{2m_e}\nabla_i^2 + \underbrace{\int \mathrm{d}\mathbf{r}'\,\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta E_{\mathrm{xc}}[n]}{\delta n(\mathbf{r})} + V_{\mathrm{ext}}}_{V_{\mathrm{eff}}[n]}, \end{aligned} \tag{3.17}$$

where δ denotes functional derivative (derivative with respect to a function). It is worthy of note that although the K-S equation gives the correct ground state density $n_{\mathrm{GS}}(\mathbf{r})$ and energy E_{GS}, it does *not* give the correct eigenenergies ε_i of the real electrons. Indeed, it is recognized that the Kohn-Sham eigenenergies $\varepsilon_i^{\mathrm{KS}}$ does not in general correspond to any physical quantity of the system; [10, 11] the notable exceptions being that the energy of the highest occupied K-S orbital in a metal gives the negative work function, [12] and the average position of the K-S CBM and VBM in a semiconductor (insulator) is the exact position relative to the vacuum. [13] Nonetheless, in practice it is found that the collection of K-S eigenenergies ε_i give a reasonably accurate representation of the real band structure of the materials system. The most prominent inaccuracy lies in the band gap of semiconductors/insulators, whose real value is typically severely underestimated with Kohn-Sham eigenenergies. The underestimation of band gap is one of the main known deficiencies and challenges of density functional theory. In Sec. 3.3.5, two state-of-the-art methods used in this thesis for reducing this discrepancy are briefly introduced.

Solving the Kohn-Sham equation, although much easier than solving the many-body Schrödinger equation, is no straightforward task. Over the years, many clever schemes and techniques are employed in obtaining accurate solutions of K-S equation. One of them is the **"self-consistent field" (SCF)** method. This method originates from the observation that the K-S Hamiltonian $\hat{H}_{\mathrm{KS}}$ it-

self is a functional of the electron density $n(\mathbf{r})$ of the system, therefore it depends on K-S wavefunctions ϕ_i, the very quantity we seek to solve. The SCF method uses an iterative procedure to find the solution, starting from an adequate initial guess $n_0(\mathbf{r})$ (usually the superposition of atomic electron densities). At each SCF iteration, the K-S equation is solved with the effective potential V_{eff} of the current density n_k, yielding a new set of K-S wavefunctions ϕ_i and the new electron density n_{k+1}. The electron density is optimized towards convergence by a procedure called "charge mixing", where the input electron density of each iteration k is calculated as a function of a selected number of previous input and output electron densities. When the residual density $R[n]$ becomes smaller than a predefined tolerance, the K-S equation is considered to be solved, with the final electron density being the correct ground state density n_{GS} of the system of interest. Fig. 3.3 summarizes the procedure of a Kohn-Sham SCF calculation.

3.3.3 Exchange-correlation functional

As seen in Sec. 3.3.2, in the Kohn-Sham formalism, the only unknown quantity in the total energy functional $E_e[n]$ is the exchange-correlation functional $E_{\text{xc}}[n]$. As the problem of solving the Kohn-Sham equation reduces to knowing the functional form of $E_{\text{xc}}[n]$, $E_{\text{xc}}[n]$ plays a central role in density functional theory. In theory, the exact form of $E_{\text{xc}}[n]$ contains all the exchange and correlation effects of any real many-electron system, and therefore makes solving the Kohn-Sham equation fully equivalent to solving the many-body Schödinger equation, except with tremendously less complexity. This grand promise has sparked and sustained the five-decade-long quest for $E_{\text{xc}}[n]$ since the invention of density functional theory. Nonetheless, this goal has proven to be far more

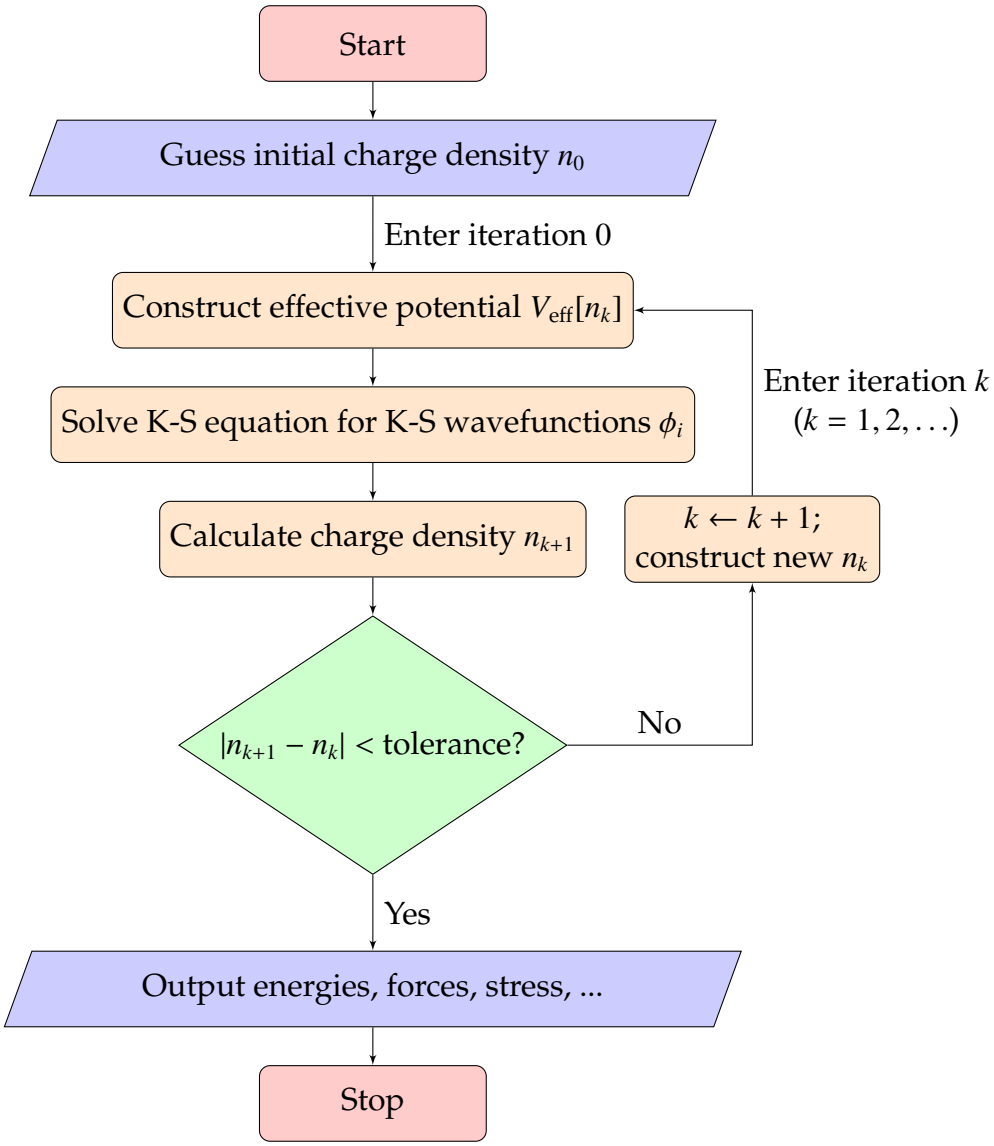

Figure 3.3: Flowchart of a self-consistent Kohn-Sham DFT calculation.

elusive, due to the extremely complicated nature of $E_{\text{xc}}[n]$ and the lack of a systematic approach to improve $E_{\text{xc}}[n]$. To date, various approximations of $E_{\text{xc}}[n]$ have been proposed, with varying degree of success and deficiency in modeling properties of molecules and materials.

Local density approximation (LDA)

The earliest and simplest exchange-correlation (xc) functional is the **local (spin) density approximation (L(S)DA)**, where the value of $E_{\mathrm{xc}}[n]$ at any location $\mathbf{r}$ is assumed to depend only on the *local* values of n($\mathbf{r}$). The general functional forms of spin-independent and spin-dependent local density approximation are given by:

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{LDA}}[n] &= \int \mathrm{d}\mathbf{r}\, n(\mathbf{r})\, \varepsilon_{\mathrm{xc}}[n(\mathbf{r})]; \\ E_{\mathrm{xc}}^{\mathrm{LSDA}}[n_\uparrow, n_\downarrow] &= \int \mathrm{d}\mathbf{r}\, n(\mathbf{r})\, \varepsilon_{\mathrm{xc}}[n_\uparrow(\mathbf{r}), n_\downarrow(\mathbf{r})], \end{aligned} \tag{3.18}$$

where $\varepsilon_{\mathrm{xc}}$ is the exchange-correlation energy per particle. In the original formulation of L(S)DA put forward by Kohn and Sham in the same seminal paper, [10] the electrons are assumed to form a homogeneous electron gas (HEG) with uniform charge density n everywhere in space. In this limit, it was shown that the assumption of locality of electronic exchange and correlation holds exactly, and $\varepsilon_{\mathrm{xc}}$ can be decomposed into an exchange part ε_{x} and a correlation part ε_{c}. Specifically, ε_{x} is given by an extension of the Dirac's formula of exchange energy $\varepsilon_{\mathrm{x}}[n] = -(9\alpha/8)(3/\pi)^{1/3} n^{1/3}(\mathbf{r})$ (α is a parameter) as in Thomas-Fermi-Dirac equation (Sec. 3.3.1), and ε_{c} is given by an interpolation formula between the known correlation energies in the high- and low-density limit of the HEG. A more recent version of LDA uses the values of $\varepsilon_{\mathrm{xc}}$ from highly accurate quantum Monte Carlo simulation of HEG. [14]

Despite the very crude and simplistic nature of L(S)DA, numerous calculations have confirmed that in real materials systems, L(S)DA already gives reasonably accurate structural parameters such as lattice constants (often within several percent of the experimental value) for simple metals and semiconductors, and tends to faithfully describe various molecular properties such as

equilibrium structures, harmonic frequencies, and charge moments, and crystal properties such as band structure (except for band gap). [15] Such unexpected accuracy are mainly due to three reasons: (1) the overestimating error by L(S)DA in the exchange energy E_x largely cancels out the underestimating error in the correlation energy E_c for simple electronic systems; (2) the exchange-correlation hole $n_{xc}(\mathbf{x}, \mathbf{x}')$, defined by the difference between the pair correlation function of electrons $g(\mathbf{x}, \mathbf{x}')$ and the electron density $n(\mathbf{r}')$, satisfies the exact sum rules under the description of L(S)DA; (3) the electron-electron Coulomb repulsion depends on the spherical average of the XC hole, which is reasonably well reproduced by L(S)DA. On the other hand, not surprisingly, L(S)DA has a series of drawbacks. For instance, L(S)DA is well known to underestimate lattice constants and bond lengths, while severely overestimating cohesive energies, phonon frequencies, and elastic moduli. [16] Furthermore, L(S)DA poorly predicts parameters related to chemical reactions, such as reaction enthalpies and activation energy barriers. These errors are a manifestation of L(S)DA to overbind atoms (by $\sim$ 1eV/atom), due to the incorrect *local* features of the L(S)DA exchange-correlation hole. This overestimation of binding energies is significantly larger than required by the so-called "chemical accuracy", namely better than 1 kcal/mol or 50 meV/atom. Despite these deficiencies, the local (spin) density approximation still finds its use today, mainly for its moderate computational cost. Most commonly used variations of LDA functionals include VWN [17] and Perdew-Zunger [18].

Generalized gradient approximation (GGA)

Many shortcomings of the local density approximation are to some extent alleviated by the next generation exchange-correlation functional: the **generalized gradient approximation (GGA)**, which has become one of the standard functionals in today's simulation of solids. GGA stems from the observation that in many real molecules and materials, the electron density varies rapidly in space, hence the homogeneous electron gas (HEG) is fundamentally not a truthful representation of the real system. The main modification of GGA over LDA is the inclusion of the dependence on the gradient of electron density $\nabla n(\mathbf{r})$ (hence GGA is also known as a "semilocal functional"). The general functional forms of GGA are given by: [19]

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{GGA}}[n] &= \int \mathrm{d}\mathbf{r}\, f[n(\mathbf{r}), \nabla n(\mathbf{r})]; \\ E_{\mathrm{xc}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow] &= \int \mathrm{d}\mathbf{r}\, f[n_\uparrow(\mathbf{r}), n_\downarrow(\mathbf{r}), \nabla n_\uparrow(\mathbf{r}), \nabla n_\downarrow(\mathbf{r})]. \end{aligned} \tag{3.19}$$

Unlike in L(S)DA, the choice for f is not unique. One of the most popular choices to this day, the PBE (Perdew-Burke-Ernzerhof) functional, [20] gives $E_{\mathrm{xc}}^{\mathrm{GGA}}$ as

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow] &= E_{\mathrm{x}}^{\mathrm{GGA}}[n] + E_{\mathrm{c}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow]; \\ E_{\mathrm{x}}^{\mathrm{GGA}}[n] &= \int \mathrm{d}\mathbf{r}\, n(\mathbf{r})\, \varepsilon_{\mathrm{x}}[n]\, F_{\mathrm{x}}[s(\mathbf{r})]; \\ E_{\mathrm{c}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow] &= \int \mathrm{d}\mathbf{r}\, n(\mathbf{r})\, \varepsilon_{\mathrm{c}}[n]\, F_{\mathrm{c}}[r_s, \zeta, t(\mathbf{r})]. \end{aligned} \tag{3.20}$$

where F_{x} and F_{c} are called the exchange and correlation "enhancement factor" which contain the dependence on the density gradient, through two dimensionless terms $s(\mathbf{r}) \propto |\nabla n|/n^{4/3}$ and $t(\mathbf{r}) \propto |\nabla n|/n^{7/6}$; $r_s = (4\pi n/3)^{-1/3}$ is the Wigner-Seitz radius (the radius of average spherical volume taken up by an electron in the material), and $\zeta = (n_\uparrow - n_\downarrow)/n$ is the relative spin polarization. Like L(S)DA

functionals, the construction of the PBE functional is entirely from physical constraints on the exchange-correlation hole and the energy, thus PBE belongs to non-empirical functionals (i.e. without empirical parameters to fit to experimental results). Other formulations of GGA functionals include PW91 [21], Lee-Yang-Parr (LYP) [22], Perdew86 (P86) [23], and revised PBE (rPBE) [24].

In practice, GGA has several advantages over L(S)DA. For crystals, GGA is shown to produce lattice constants with generally better accuracy. For molecules, GGA greatly improves on atomization and bond dissociation energies, bond lengths, and transition barriers, typically within 10% of the true values (while errors of L(S)DA can be as large as 100%). However, like with any approximated functionals, there is still room for improvement for GGA, the most notable being that GGA tends to weakly underbind atoms in solids, showing the opposite trend to L(S)DA. Generally, PBE performs much better for molecules than for solids. A later variation of the PBE functional, named PBEsol [25], is designed for solids and surfaces; it tends to yield better accuracy for solid-phase properties such as lattice constant, bulk moduli, phonon frequencies, and surface energies.

3.3.4 Jacob's ladder in DFT

Although it is well recognized that there is no magic recipe for systematic improvement on the accuracy of XC functional for all properties of chemical systems, a formal hierarchy that resembles the Jacob's ladder can be constructed for different levels of XC functionals, [26] with higher level XCs possessing greater complexity and (in principle) accuracy than lower level ones. (Fig. 3.4) Not

surprisingly, local (spin) density approximation (L(S)DA) and general gradient approximation (GGA) occupy respectively the lowest (first) and the second rung on the Jacob's ladder in DFT. The third rung corresponds to the so-called "meta-GGA" functionals, which depends on the second derivative of electron density $\nabla^2 n$ and/or electronic kinetic energy densities $\tau_\sigma = \sum_i^{N/2} |\nabla \phi_{i,\sigma}|^2/2$, in addition to electron density $n(\mathbf{r})$ and density gradient ∇n. Starting from the fourth rung, the XC functionals become significantly more complex as they make use of the Kohn-Sham orbitals. Specifically, hybrid functionals, the fourth rung of the ladder, includes a dependence on all the occupied K-S orbitals; whereas random-phase-approximation (RPA)-like functionals, the fifth rung of the ladder, depends on all (occupied and unoccupied) K-S orbitals.

As a rule of thumb, the higher the level of a functional, the more expensive the computational cost. As this thesis focuses on various properties in solids, with the exception of Chap. 4 (which used the LDA functional due to constrained computational resources), all other research works presented in this thesis (Chap. 5-7) use PBEsol functional (except for calculating band gaps), as it strikes a good balance between computational cost and accuracy.

3.3.5 Band gap correction: beyond conventional DFT

For gapped materials such as semiconductors and insulators, the magnitude of the band gap is a critical quantity that determines the electronic and optical properties of the material. (Sec. 2.1.2) Unfortunately, one of the paramount difficulties of conventional (local and semilocal) density functionals (LDA, GGA, and meta-GGA) is their inability to predict the correct band gap value. Due to

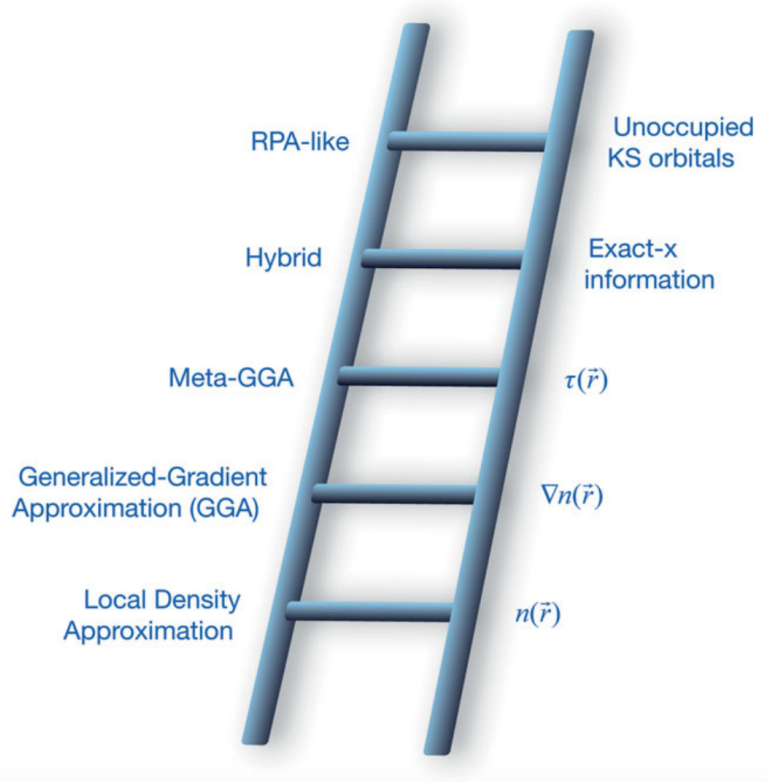

Figure 3.4: Jacob's ladder of density functional approximations. The rungs are labeled on the left, and their added ingredients are shown on the right. Reproduced from [3].

the "delocalization error" of electrons and lack of derivative discontinuity, conventional density functionals typically underestimate the band gap by about 50% on average, thereby severely hindering the predictive ability of DFT. Alternative methods beyond conventional DFT have been proposed aiming to remedy this problem, with two methods – hybrid functional and *GW* method – being two of the most satisfactory state-of-the-art methods in this regard.

Hybrid functional

Hybrid functionals stem from the observation that the exchange energy E_x of same-spin electrons account for a major proportion ($\sim 85-95\%$) of the exchange-

correlation energy of the many-electron systems. [27] It is therefore possible to improve on GGA functionals by making use of the exact exchange term of the Hartree Fock method (Sec. 3.2.3):

$$E_{\mathrm{x}}^{\mathrm{HF}}[n] = -\frac{1}{2}\sum_{\sigma}\sum_{i}^{N/2}\sum_{j}^{N/2}\int \mathrm{d}\mathbf{r}\int \mathrm{d}\mathbf{r}'\,\frac{\phi_{i,\sigma}^{*}(\mathbf{r})\phi_{j,\sigma}(\mathbf{r})\phi_{j,\sigma}^{*}(\mathbf{r}')\phi_{i,\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{3.21}$$

Note that $E_{\mathrm{x}}^{\mathrm{HF}}[n]$ is evaluated with the Kohn-Sham orbitals ϕ_i instead of the Hartree-Fock orbitals ψ_i. It is evident from Eq. (3.21) that $E_{\mathrm{x}}^{\mathrm{HF}}$ is a non-local density functional, thus it has the advantage of completely eliminating the unphysical many-electron self-interaction error in the approximated exchange energy of GGA which tends to over-delocalize electrons. Nonetheless, when applied to molecules and solids, the XC functional $E_{\mathrm{xc}} = E_{\mathrm{x}}^{\mathrm{HF}} + E_{\mathrm{c}}^{\mathrm{GGA}}$ gives unrealistic results, [28] due to the fact that $E_{\mathrm{x}}^{\mathrm{HF}}$ is an exact result for non-interacting electron systems. As electrons in most realistic material systems lie between the non-interacting limit and the fully-interacting limit (where $E_{\mathrm{x}}^{\mathrm{LDA/GGA}}$ performs reasonably well due to error-cancellation effects), practical hybrid functionals define the exchange energy E_{x} as a mixture of $E_{\mathrm{x}}^{\mathrm{HF}}$ and $E_{\mathrm{x}}^{\mathrm{GGA}}$ to mimic the real static correlation effect. (This formalism is called the *generalized* Kohn-Sham DFT.) Common variants of hybrid functionals include B3LYP, [17, 22, 29, 30] B3PW91, [29] PBE0, [31, 32] and HSE, [33, 34] with their XC functionals defined by

$$\begin{aligned}
E_{\mathrm{xc}}^{\mathrm{B3LYP}} &= 0.08E_{\mathrm{x}}^{\mathrm{LSDA}} + 0.20E_{\mathrm{x}}^{\mathrm{HF}} + 0.72E_{\mathrm{x}}^{\mathrm{B88}} + 0.19E_{\mathrm{c}}^{\mathrm{VMN}} + 0.81E_{\mathrm{c}}^{\mathrm{LYP}};\\
E_{\mathrm{xc}}^{\mathrm{B3PW91}} &= 0.08E_{\mathrm{x}}^{\mathrm{LSDA}} + 0.20E_{\mathrm{x}}^{\mathrm{HF}} + 0.72E_{\mathrm{x}}^{\mathrm{B88}} + 0.19E_{\mathrm{c}}^{\mathrm{VMN}} + 0.81E_{\mathrm{c}}^{\mathrm{PW91}};\\
E_{\mathrm{xc}}^{\mathrm{PBE0}} &= \frac{1}{4}E_{\mathrm{x}}^{\mathrm{HF}} + \frac{3}{4}E_{\mathrm{x}}^{\mathrm{PBE}} + E_{\mathrm{c}}^{\mathrm{PBE}};\\
E_{\mathrm{xc}}^{\mathrm{HSE}} &= aE_{\mathrm{x}}^{\mathrm{HF,\,SR}}(\omega) + (1-a)E_{\mathrm{x}}^{\mathrm{PBE,\,SR}}(\omega) + E_{\mathrm{x}}^{\mathrm{PBE,\,LR}}(\omega) + E_{\mathrm{c}}^{\mathrm{PBE}}.
\end{aligned} \tag{3.22}$$

where a is an empirical parameter for tuning the band gap (see next paragraph). In particular, the HSE functional splits the exchange parts of Hartree-Fock and

PBE into a short-range part and a long-range part, by decomposing the Coulombic $1/r$ term in the exchange energies:

$$\frac{1}{r} = \frac{\mathrm{erfc}(\omega r)}{r} + \frac{\mathrm{erf}(\omega r)}{r}, \tag{3.23}$$

where ω is the range-separation parameter characterizing the spatial extent of the short-ranged screening effect of the electrons, such that the electron-electron Coulomb interaction becomes negligibly small beyond a distance of $\sim 2/\omega$. The HSE functional is reduced to PBE0 at $\omega = 0$, and asymptotically approaches PBE as $\omega \to \infty$. By including the electronic screening effect, the HSE functional rectifies the divergence behavior of orbital energy derivatives in metals and small-gapped semiconductors and reduces the high computational cost associated with the long-ranged Coulomb interaction in $E_{\mathrm{x}}^{\mathrm{HF}}$; [33] therefore, HSE achieves considerable improvement on properties of solids compared to other hybrid functionals.

The hybrid functionals are shown to be able to greatly improve on many molecular and material properties, such as atomization energies, bond lengths, and vibrational frequencies. One of the most striking results of hybrid functionals is the great improvement of predicted band gap of gapped materials over conventional functionals, as is clearly shown in Fig. 3.5. The main reasons of this success are two-fold. First, within the framework of generalized Kohn-Sham DFT, HSE naturally contains an approximate derivative discontinuity ($\Delta_{\mathrm{xc}} = \partial E_e/\partial N_e|_{N_e \to N^+} - \partial E_e/\partial N_e|_{N_e \to N^-}$, where N_e is the (variable) number of electrons in the system), which can be shown to be the difference between the true (fundamental) energy gap G and the Kohn-Sham band gap g. Second, HSE includes a portion of the exact HF exchange energy, which is free of self-interaction error (SIE) for single atoms; this has the effect of correcting the effect of SIE on K-S eigenenergies, a signature defect of semi-local (GGA) XC func-

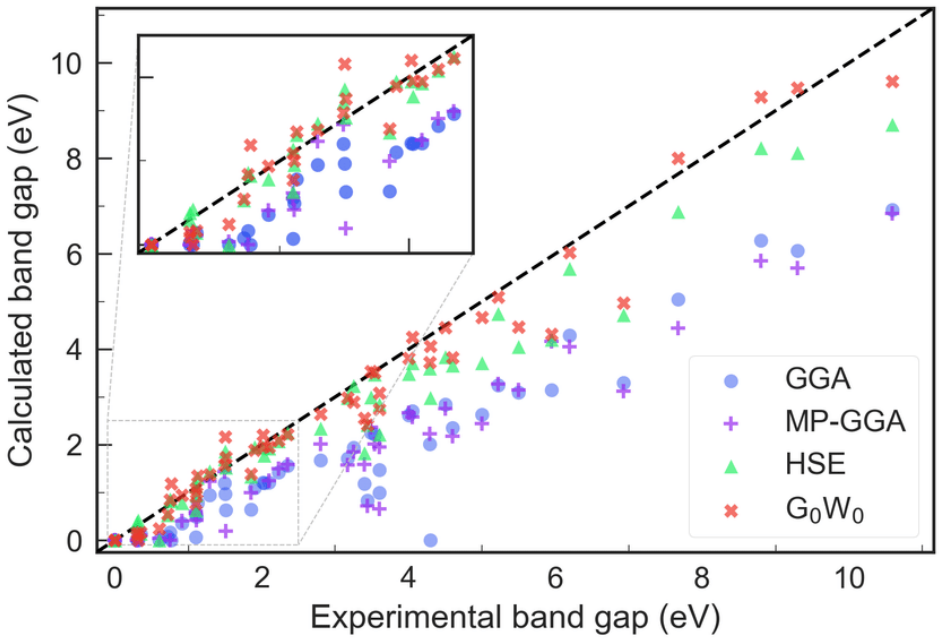

Figure 3.5: Comparative plot of the calculated and experimentally available values for all the electronic band gaps obtained in the current work. Legend: GGA, HSE, and G_0W_0 denote the results of this work for the corresponding level of theory. MP-GGA denote the results of Materials Project. Reproduced from [36].

tionals. [35]

One of the few challenges currently facing hybrid functionals is that, as the exact HF exchange energy depends on all occupied K-S orbitals, the computational expense is typically considerably higher (by about a factor of $\sim$ 100) than that of the local and semilocal functionals. Simulation of large systems using hybrid functionals is still considered a challenge in DFT. In this thesis, due to constraints on computational resources, we use hybrid functional only for the purpose of correcting the band gaps of semiconductors, which does not require large supercells (Sec. 3.3.6), unlike in calculation of total energy of defect-containing crystals).

GW method

Density functional theory is, by definition, a theory of the ground-state; this implies that properties related to the excited states of electrons, including the band gap (the energy difference between the lowest excited state and the highest ground state), cannot be accurately calculated using DFT, *even if the exact universal functional is known*. Hybrid functionals such as HSE gives satisfactory band gaps in many cases, but they still fundamentally belong within the framework of DFT. To treat such problems in a rigorous manner, higher-level methods must be employed. One of such methods is the so-called (quasiparticle) *GW* method, where the system is treated on the footing of many-body perturbation theory. The term "quasiparticle" refers to the collection of a single bare electron and its surrounding positive polarization charge resulted from Coulomb repulsion of other electrons. The critical insight of the *GW* method is that, similar to the spirit of the Hartree approximation, a system of interacting electrons can be regarded as a system of *weakly interacting* quasiparticles. *GW* method seeks an exact form of many-electron exchange and correlation of this system of quasiparticles, by replacing the Kohn-Sham XC functional with the so-called "(electron) self-energy" Σ:

$$V_{\mathrm{xc}}[n(\mathbf{r})]\phi_i(\mathbf{r}) \to \int \mathrm{d}\mathbf{r}'\, \Sigma(\mathbf{r}, \mathbf{r}'; \varepsilon_i)\phi_i(\mathbf{r}'). \tag{3.24}$$

The effect of replacing XC functional with the self-energy on the Kohn-Sham band structure is to perturb each eigenenergy ε_i by a correction term

$$\Delta\varepsilon_i = \mathrm{Re}\left[\int \mathrm{d}\mathbf{r}'\, \phi_i^*(\mathbf{r})\Sigma(\mathbf{r}, \mathbf{r}'; \varepsilon_i)\phi_i(\mathbf{r}') - \int \mathrm{d}\mathbf{r}\, \phi_i^*(\mathbf{r})V_{\mathrm{xc}}[n(\mathbf{r})]\phi_i(\mathbf{r})\right], \tag{3.25}$$

with Re denoting the real part of a complex quantity. The self-energy can be calculated (in time domain) as the product of the Green's function $G(\mathbf{r}t, \mathbf{r}'t')$ and

the dynamically screened Coulomb interaction $W(\mathbf{r}t, \mathbf{r}'t')$. It captures the energy contributions of all electron-electron interactions in a system. Unlike XC functionals, Σ is an energy-dependent quantity; furthermore, it is truly non-local. Furthermore, Σ contains the exact derivative discontinuity Δ_{xc} which is crucial in reproducing the correct band gaps. Fig. 3.5 shows the clear advantage of *GW* method over all DFT methods (including the hybrid functional HSE). Nonetheless, since calculation of Σ requires a large number of *unoccupied* electronic orbitals, *GW* is even more expensive computationally than HSE. In practice, *GW* is typically performed by perturbing the converged Kohn-Sham orbitals produced by a preceding DFT calculation.

GW methods come in many different flavors. The most commonly-used and computationally cheapest flavor is the so-called "one-shot *GW*", or G_0W_0 approximation. This method is a first-order perturbation based on DFT results; namely, eqn (3.25) is solved by performing a Taylor expansion on the right hand side and taking the first order correction term

$$Z_i \left[\int \mathrm{d}\mathbf{r}\, \phi_i^*(\mathbf{r})(\Sigma(\varepsilon_i) - V_{\mathrm{xc}})\phi_i(\mathbf{r}) \right] \tag{3.26}$$

with the renormalization factor

$$Z_i = \left[1 - \int \mathrm{d}\mathbf{r} \phi_i^*(\mathbf{r}) \frac{\partial \Sigma(\omega)}{\partial \omega}\Big|_{\omega=\varepsilon_i} \phi_i(\mathbf{r}) \right]^{-1}, \tag{3.27}$$

instead of being solved self-consistently. It is hence expected that G_0W_0 approximation would depend on the accuracy of the DFT starting point, yet for many common semiconductors, the G_0W_0 band gap agrees remarkably well with experiment even for LDA as the starting point. In Chap. 4, we use G_0W_0 to calculate the shifts in CBM and VBM of the CuAu-I-ordered $In_{0.5}Ga_{0.5}As$.

3.3.6 Practical aspects of DFT simulation of crystals

In practical DFT calculations, several importact points warrant special care. This section includes some important practical aspects of DFT calculations, specifically applied to modeling of crystals.

Periodic boundary conditions

In Sec. 2.1.1, we have seen that all crystals are solids with essentially an infinite copies of a basic structure (unit cell) repeated periodically in space. It is apparent that no practical computers can handle near-infinite (Avogadro's number) amount of atoms. In many mainstream DFT codes, the **periodic boundary conditions (PBC)** are imposed, which means that all the atoms inside a simulation box (called the "cell") is repeated periodically in all three lattice dimensions ($\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$). Thanks to the periodic nature of crystals, the PBC scheme allows simulation of the entire crystal with only a very small number of atoms.

It is very important to note that PBC is the natural condition satisfied by the bulk region of pristine crystals without any defects; therefore, great care must be taken to simulate non-bulk-like or defective materials. In this thesis, we mainly focus on two kinds of materials systems: (1) bulk crystal with point defect(s); (2) crystal surface. For bulk crystal with point defect(s), the PBC scheme implies that the defect(s) are repeated with the same periodicity as that of the host crystal, forming a periodic regular array in three-dimensional space. Although this picture is far from physical per se, the most important point from the simulation perspective is that the distance between the defect(s) themselves and their *nearest periodic images* due to PBC is sufficiently large. Using cells with small lat-

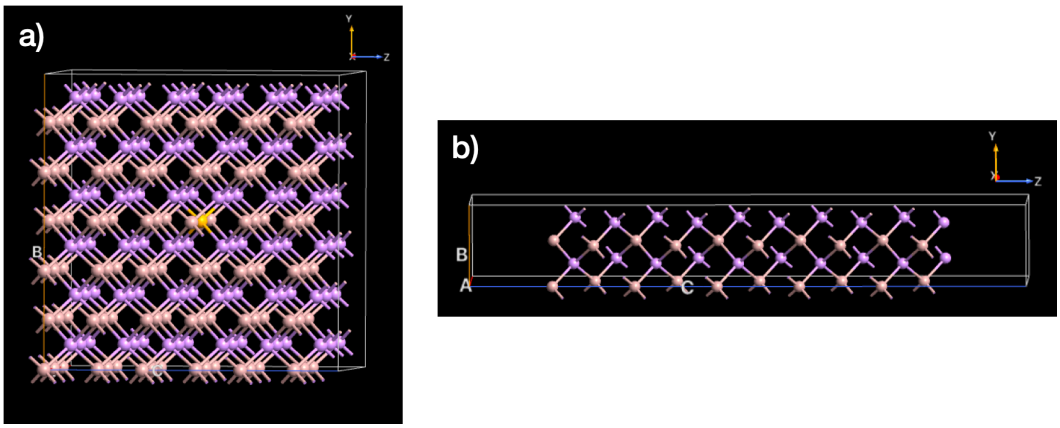

Figure 3.6: Cell representation of (a) a bulk crystal with point defect(s); (b) a pristine crystal surface.

tice constants often causes spurious interactions between the defect(s) and their images, which introduces various errors to the simulation results of defective materials. Therefore, it is critically important to use large-sized simulation cells that contain many replica of the crystal unit cell (called an **"supercell"**) in defect simulations. (Fig. 3.6(a)) (This important point is covered in detail in Sec. 3.4) For crystal surface, it is important to incorporate "vacuum" (in the sense of a region lack of atoms) into the simulation cell, as every real surface is interfaced with free space. (Fig. 3.6(b)) The way to define such a system in the PBC scheme is to create a **slab** composed of crystal unit cells stacked in the direction along the surface orientation, which is terminated on its two ends by vacuum in that direction. Like with the defective crystals, the thickness of both the slab and the vacuum must be large enough, so that one end of the slab and the opposite end (or its nearest periodic image) have minimal spurious interactions. This aspect is especially important if there exist impurities on or near the surface. (Chap. 7)

Plane wave basis set

We have indicated in Sec. 2.1.2 that plane waves are natural mathematical descriptors for electronic wavefunctions in a crystal. The underlying reason is the Bloch's theorem, which states that for an electron in an external potential $V(\mathbf{r})$ with periodic translational symmetry, the one-particle wavefunction $\psi(\mathbf{r})$ can be expressed as the product of a plane wave and a periodic function $u_{\mathbf{k}}(\mathbf{r})$ with the same periodicity as $V(\mathbf{r})$: (Fig. 3.7)

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{3.28}$$

Such a function can be Fourier expanded into a sum of plane waves:

$$u_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} C_{\mathbf{k}}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{3.29}$$

where Ω is the volume of the simulation cell. In this equation, G belongs to a particular set of vectors in the **reciprocal space** (the space of all wavevectors k). This set of vectors (the "reciprocal lattice vectors") are the Fourier-transformed Bravais lattice vectors $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$ (n_i = integer); namely, $\mathbf{G} = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + m_3\mathbf{b}_3$ (m_i = integer), such that

$$\begin{aligned}
\mathbf{b}_1 &= 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\Omega} = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}; \\
\mathbf{b}_2 &= 2\pi \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\Omega} = 2\pi \frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_2 \cdot (\mathbf{a}_3 \times \mathbf{a}_1)}; \\
\mathbf{b}_3 &= 2\pi \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\Omega} = 2\pi \frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_3 \cdot (\mathbf{a}_1 \times \mathbf{a}_2)}.
\end{aligned} \tag{3.30}$$

These expressions satisfy the orthonormality relations $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$, which implies $e^{i\mathbf{G}\cdot\mathbf{R}} = 1$. It follows that the electron wavefunctions $\psi_k(\mathbf{r})$ are "periodic" in

the *reciprocal* space as well:

$$\begin{aligned}\psi_{n,\mathbf{k}+\mathbf{G}}(\mathbf{r}) &= e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} u_{n,\mathbf{k}+\mathbf{G}}(\mathbf{r}) \\ &= e^{i\mathbf{k}\cdot\mathbf{r}}(e^{i\mathbf{G}\cdot\mathbf{r}} u_{n,\mathbf{k}+\mathbf{G}}(\mathbf{r})) \\ &= e^{i\mathbf{k}\cdot\mathbf{r}} \tilde{u}_{n,\mathbf{k}}(\mathbf{r}) \\ &= \psi_{\tilde{n},\mathbf{k}}(\mathbf{r}),\end{aligned} \tag{3.31}$$

since $\tilde{u}_{\mathbf{k}}(\mathbf{r})$ is by construction a periodic function in $\mathbf{R}$ like $u_{\mathbf{k}}(\mathbf{r})$. Strictly speaking, this equation tells us that the electronic wavefunction with band index n at wavevector $\mathbf{k} + \mathbf{G}$ is equivalent to the wavefunction with band index $\tilde{n}$ at wavevector $\mathbf{k}$. (Fig. 3.8) This result has a significant implication, namely in a periodic crystal, any wavefunction can be described equivalently by $\mathbf{k}$-points within the Brillouin zone (Sec. 2.1.2) only, thereby reducing an infinite-space problem to a finite-space one without losing any information. However, Eq. (3.27) still contains an infinite *number* of plane waves. In practice, a finite number of plane-waves must be used. The number of plane-waves included is controlled by the (kinetic) energy cutoff E_{cut},

$$E_{\text{cut}} = \frac{\hbar^2}{2m_e} G_{\text{cut}}^2 \tag{3.32}$$

which corresponds to the kinetic energy of an (non-interacting) electron with wavenumber G_{cut}. This means that all the wavefunctions with $|\mathbf{k} + \mathbf{G}| \leq G_{\text{cut}}$ are included in the plane-wave expansion of $\psi_{\mathbf{k}}$. Similarly, the electron density $n(\mathbf{r})$ can also be expanded exactly in the plane-waves, with the density cutoff n_{cut} equal to $4E_{\text{cut}}$, corresponding to eight times the number of plane-waves with $|\mathbf{k}+\mathbf{G}| \leq G_{\text{cut}}$. The plane-waves form a complete (when $E_{\text{cut}} \to \infty$) and orthonormal **basis set** for expanding the electronic wavefunctions and density in periodic systems, and the resulting energies E_e and ε_i can be systematically improved by simply increasing E_{cut}. It can be shown that the entire Kohn-Sham equation can

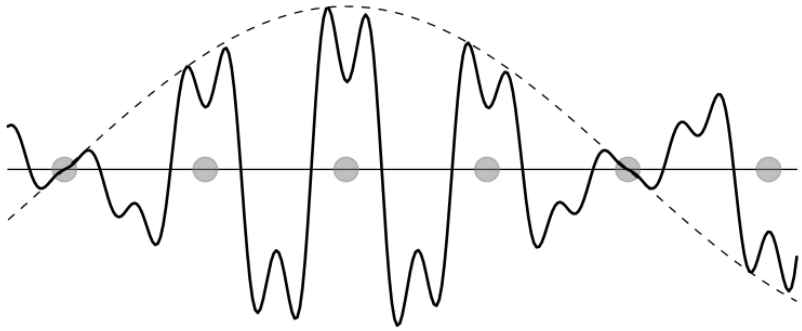

Figure 3.7: Solid line: A schematic of a typical Bloch wave in one dimension. (The actual wave is complex; this is the real part.) The dotted line is from the $e^{i\mathbf{k}\cdot\mathbf{r}}$ factor. The light circles represent atoms. Reproduced from [37].

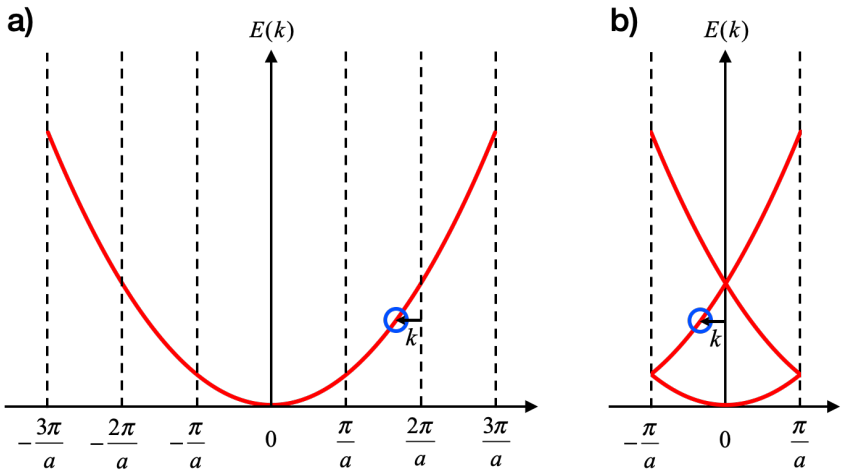

Figure 3.8: Schematic of band structure of an one-dimensional crystal in (a) extended scheme; (b) reduced scheme. The blue circles in both plots denotes the same point in the band structure, corresponding to the same wavefunction $\psi_{n=1,k+2\pi/a} = \psi_{\tilde{n}=2,k}$.

be expressed in terms of plane waves, making it easier to compute with efficient fast Fourier transform (FFT) operations in the reciprocal space.

k-point sampling

As we have seen in the previous section, the Kohn-Sham wavefunction $\psi_{\mathbf{k}}(\mathbf{r})$ is uniquely determined by the **k**-points inside the Brillouin zone (BZ). This means that many physical quantities such as total energy, electron density and density of states can be expressed as an integral over all **k**-points in the Brillouin zone (BZ):

$$\bar{f} = \frac{1}{\Omega_{\mathrm{BZ}}} \int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}\, f(\mathbf{k}), \tag{3.33}$$

where $\Omega_{\mathrm{BZ}} = (2\pi)^3/\Omega$ is the volume of the BZ. In numerical calculations, such a continuous three-dimensional integral must be approximated by a discrete weighted summation

$$\bar{f} \approx \sum_{\mathbf{k}_j \in \mathrm{IBZ}} w_{\mathbf{k}_j} f(\mathbf{k}_j), \tag{3.34}$$

where IBZ denotes the "irreducible Brillouin zone", namely the subspace of BZ which cannot be further reduced by the point group symmetry of the crystal. In principle, the denser the **k**-points used in the summation, the more accurate the approximation will be. However, the number of **k**-points used is limited by the computational resources. Hence it is very important to choose the appropriate **k**-points included in the summation that most accurately approximate the integral (3.31) while maintaining feasible computational cost.

The most straightforward and widely used **k**-point sampling scheme was proposed by Monkhorst and Pack in 1976. [38] In this scheme, an equally-spaced three-dimensional mesh of **k**-points are sampled in the Brillouin zone, with number of **k**-points equal to N_1, N_2, N_3 along $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$, respectively. In general, from the expression of the Brillouin zone volume Ω_{BZ} we see that larger simulation cell leads to smaller reciprocal cell, which means fewer **k**-points need to be sampled in the BZ if the **k**-point density (hence the numerical accuracy) is

kept constant.

Another common method for **k**-point sampling is by Chadi and Cohen in 1976. [39] They construct a set of "special **k**-points" with the following procedure: expand the **k**-point dependent function $f(\mathbf{k})$ in Fourier series:

$$f(\mathbf{k}) = f_0 + \sum_{m=1}^{\infty} f_m A_m(\mathbf{k}) \tag{3.35}$$

where

$$A_m(\mathbf{k}) = \frac{1}{n_T} \sum_{|\mathbf{R}|<R_m} e^{i\mathbf{k}\cdot\mathbf{R}} \tag{3.36}$$

with n_T is the number of point group operations of the crystal lattice, and R_m the radius of the m-th "shell" of lattice vectors. From this expression, we see that if there exists a "special point" $\mathbf{k}_0$ such that $A_m(\mathbf{k}_0) = 0$ for all m, then $\bar{f} = f_0 = f(\mathbf{k}_0)$. However, such a point does not exist. However, for some crystal systems it is possible to find a point $\mathbf{k}_0$ such that $A_m(\mathbf{k}_0) = 0$ for $1 \le m \le N$ for some finite N. (Such $\mathbf{k}_0$ is called the "mean-value point". [40]) It is found that in simple cubic crystals, $\mathbf{k}_0 = \frac{2\pi}{a}(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ for $N = 3$. Chadi and Cohen expanded this idea by defining the set of special **k**-points $\mathbf{k}_i$ $(i = 1, 2, \cdots, n)$ using the criteria:

$$\begin{gathered} \sum_{i=1}^{n} w_{\mathbf{k}_i} A_m(\mathbf{k}_i) = 0 \qquad (m = 1, 2, \cdots, N); \\ \sum_{i=1}^{n} w_{\mathbf{k}_i} = 1. \end{gathered} \tag{3.37}$$

This way, for large enough N, we have

$$\bar{f} = f_0 \approx \sum_{i=1}^{n} w_{\mathbf{k}_i} f(\mathbf{k}_i). \tag{3.38}$$

Pseudopotential

As is made clear in Sec. 3.3, the central quantity of density functional theory is the electron density. An all-electron description of the system produces the ex-

act electron density of the system; such an description has, nevertheless, several shortcomings from a computational point of view. First, due to the requirement of orthogonality of wavefunctions, the wavefunctions of valence electrons exhibit high-frequency oscillations at small distance r to the nucleus ("core region"), which require an infeasible number of plane wave basis (scale as Z^3 with Z = atomic number) to accurately represent. Second, Coulomb potential between bare electrons diverges as $1/r$ as r approaches 0. These shortcomings lead to exceeding large total energies of the system, which is prone to large numerical errors.

The idea of pseudopotentials is motivated by the fact that the core electrons of an atom usually do not participate in chemical bonding, and therefore play a much less important role than the valence electrons in determining the chemical and physical properties of molecules and materials. It is therefore advantageous to construct **"pseudopotentials" (PPs)** where the atomic core (nucleus + core electrons) is "frozen", meaning that they are not represented by actual electrons but by an effective potential equal to the screened Coulomb potential of a single fixed, non-polarizable charge. This approximation has the effect of removing the divergence of Coulomb potential near $r = 0$. In addition, a "pseudo-wavefunction" can be constructed such that the rapid oscillations of the valence electron wavefunctions are replaced by artificial smooth ones; this approximation has the effect of reducing the size of the plane-wave basis set. To construct an accurate pseudopotential, a set of parameters (including core region radius r_c, XC functional, and pseudopotential type) must be optimized so that the resulting atomic energy levels and valence region wavefunctions must match those of an all-electron calculation. Moreover, the pseudo-wavefunctions/potentials matches the all-electron wavefunctions/potentials in

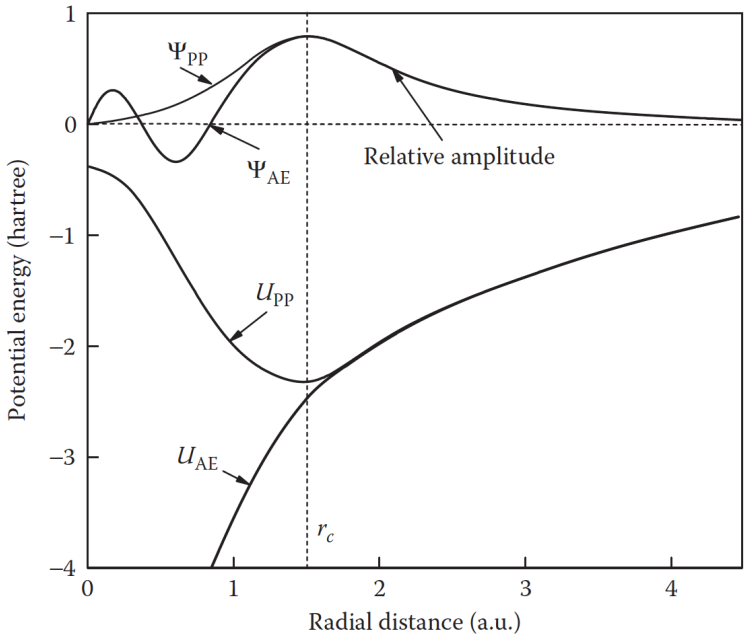

Figure 3.9: Schematic illustration of a pseudo wave function pseudized from a 3s wave function (showing the relative amplitude in arbitrary unit) and the corresponding pseudo- and all-electron (AE) potentials. Reproduced from [2].

the valence region ($r > r_c$). (Fig. 3.9)

There exists three main types of pseudopotentials: norm-conserving, ultrasoft, and projected augmented wave (PAW). In norm-conserving PPs, the integrated electron density within the core region must be equal to that of an all-electron calculation (hence the name "norm-conserving"). This construction has the advantage of faithful representation of valence electrons, but relatively large basis set is still required. Ultrasoft PPs substantially alleviates this issue while maintaining accuracy, by relaxing the charge conservation constraint of the norm-conserving PPs. This is achieved by dividing the total valence electron density into a very smooth ("ultrasoft") part and a more oscillatory part, treated as core augmentation charge. A third approach, PAW combines the ac-

curacy and transferability of all-electron calculations with the efficiency of the pseudopotential approach. In PAW, a set of one-electron pseudo-wavefunctions $\tilde{\psi}_i$ are defined such that the all electron wavefunctions ψ_i^{AE} can be obtained by a linear transformation operation on $\{\tilde{\psi}_i\}$. Furthermore, it utilizes atomic orbitals (defined on a atom-centered radial grid) to represent part of valence electron wavefunctions in the core region, and plane waves to represent those in the valence region. In this thesis, norm-conserving and ultrasoft PPs have been used in structural optimization and total energy calculations of crystals, whereas PAW PPs are used solely for population analysis due to their generally higher consumption of computer memory and their ability to reproduce all-electron density.

3.4 Modeling of point defects in semiconductors

In Sec. 2.1.3., it has been shown that the impact of dopants and defects on the properties of semiconductor cannot be overstated. Hence a thorough, atomic-level understanding of these impurities is of unprecedented importance in providing a guidence for improving the quality and performance of semiconducting materials in future devices. In a crystalline solid, two critical aspects of defects and dopants control the properties of a semiconductor: formation and migration. The former is fundamentally governed by thermodynamics, the latter by kinetics. The formation energy of an impurity tells us about the likelihood of *presence* of the impurity, whereas the migration energy of an impurity tells us about the likelihood of *diffusion* of the impurity. As dopant activation entails the dopant and defect atoms being located on particular lattice sites, both the formation and migration energies are essential in explaining this complex phe-

nomena. Nevertheless, as the impurity migration is a dynamic process, it is more difficult to be accurately modelled in density functional theory, as in a real crystal, the typical diffusion process occurs on a spatial and temporal scale far exceeding the capabilities of DFT. Even so, the formation energy itself provides critical information about the dopant/defect distribution in a crystal *under thermal equilibrium*. Hence, we focus on only the defect formation energies in all the research works presented in this thesis.

3.4.1 Defect formation energy

Under thermal equilibrium, the defect concentration of defect species D is given by the Arrhenius equation (Eq. (2.8)):

$$n_D = n_s \exp\left(\frac{-E^f(D)}{k_B T}\right), \tag{3.39}$$

which depends exponentially on the **defect formation energy** $E^f(D)$. Hence an accurate evaluation of $E^f(D)$ is critical. In the fully rigorous treatment, the defect formation energy under a constant finite temperature T and pressure P is the Gibbs *free* energy of formation

$$\Delta G^f(D) = \Delta E^f(D) - T\Delta S^f(D) + P\Delta V^f(D), \tag{3.40}$$

where $\Delta E^f(D)$ is the change in internal energy due to defect formation, $\Delta S^f(D)$ the defect formation *entropy* (consisting of configurational entropy $\Delta S^f_{\text{conf}}(D)$ and vibrational entropy $\Delta S^f_{\text{vib}}(D)$), and $\Delta V^f(D)$ the defect formation volume. For a typical defect formed at standard conditions, the latter two terms are typically much smaller than the formation enthalpy. [41, 42, 43] In addition, the entropy term requires expensive phonon calculations, and the defect formation

volume $\Delta V^f(D)$ is not well-defined in periodic DFT calculations. Hence, the most common practice in DFT literature is to consider only the internal energy term $\Delta E^f(D)$, which composes the bulk of $\Delta G^f(D)$. We will denote the defect formation energy $E^f(D) = \Delta E^f(D)$ from this point on.

The defect formation energy of a defect D in the charge state q in a gapped material (semiconductor or insulator) is given by the expression

$$E^f(D^q) = E_{\text{tot}}(D^q) - E_{\text{tot}}(\text{pure}) - \sum_i \Delta n_i \mu_i + q(E_F - E_{\text{VBM}}) + E_{\text{corr}}. \tag{3.41}$$

In this expression, $E_{\text{tot}}(D^q)$ is the total energy of the supercell containing a copy of defect D; $E_{\text{tot}}(\text{pure})$ is the total energy of the pristine supercell without any defect. Δn_i is the number of atoms of species i added to ($\Delta n_i > 0$) or removed from ($\Delta n_i < 0$) the supercell as a result of defect formation; μ_i is the atomic chemical potential of species i as in the simulated material. q is the charge state of the defect; E_F is the Fermi level of the system; E_{VBM} is the energy level of the valence band maximum of the pristine supercell. Finally, E_{corr} is a correction term which accounts for the spurious Coulombic interactions between a charged defect and its nearest neighboring images due to PBC. In the following subsections, we focus on three most important aspect of defect formation energy: atomic chemical potential, defect charge state, and finite-size corrections.

Atomic chemical potential

The atomic chemical potential μ_i indicates the growth conditions of the material. Typically, epitaxial growth of crystalline materials involves sources of elemental materials known as "atomic reservoirs", with which the material can exchange atoms. [44] In creating a defect, atoms from the material has to be either added

to the material from the reservoir, or removed from the material to the reservoir. The chemical potential μ_i of element i can be understood as the energetic cost of exchanging one atom of element i between the reservoir and the grown material, which are in constant thermal equilibrium. Large values of μ_i correspond to i-rich growth conditions, namely the supply of element i atoms from the reservoir are high compared to other elements. Conversely, small values of μ_i correspond to i-poor growth conditions, where the supply of element i atoms are relatively low.

In any particular materials system, the values of μ_i cannot be arbitrary. They are strictly bounded by several constraints: (1) μ_i cannot be greater than the energy per atom in the corresponding bulk elemental phase ($\mu_i < E_i$), meaning that no elemental phase inside the material can form; (2) the stoichiometric sum $m(\mu_A - E_A) + n(\mu_B - E_B) + \cdots$ composed of any *subset* of elements $A, B, \ldots$ present in the material cannot be greater than the formation enthalpy of their compound $A_m B_n \ldots$, meaning that no secondary phase inside the material can form; (3) the stoichiometric sum $m(\mu_A - E_A) + n(\mu_B - E_B) + \cdots$ composed of all elements $A, B, \ldots$ present in the material must be equal to the formation enthalpy of the material itself. The set of allowed values of μ_i for all elements i in the material (including the defect) is the intersection of volumes in the chemical potential space that satisfy all three constraints for every element i.

Defect charge state

From Sec. 2.1.3, we have seen that the charge state of a defect characterizes the defect's electronic properties within a gapped material. As the Fermi level varies with the free carrier concentration in the material, we will consider it as

an independent variable which can assume any value within the band gap. [1] As the Fermi level varies within the band gap, the defect may either assume the same charge state or switch between different charge states, depending on the number of defect energy levels within the gap. The charge transition level of a defect defines the position of Fermi level where the switch from one charge state to another occurs:

$$\varepsilon(q_1/q_2) = \frac{E^f(D^{q_1}, E_F = 0) - E^f(D^{q_2}, E_F = 0)}{q_2 - q_1} \tag{3.42}$$

where $E_F = 0$ means the Fermi level is located at the VBM. This equation tells us that at the transition level $E_F = \varepsilon(q_1/q_2)$, the formation energies $E^f(D^{q_1} = E^f(D^{q_2}$; the charge state transition from q_1 to q_2 therefore corresponds to a point where for $E_F < \varepsilon(q_1/q_2)$, $E^f(D^{q_1}) < E^f(D^{q_2})$; and for $E_F > \varepsilon(q_1/q_2)$, $E^f(D^{q_1}) > E^f(D^{q_2})$. (Fig. 3.10)

As discussed in Sec. 3.3.5, conventional XC functionals severely underestimate the band gap for semiconductors and insulators. Advanced computational methods such as hybrid functionals and *GW* method are capable of improving the band gap (and also the charge transition levels); [45, 46] nonetheless, the expensive nature of these methods prohibit large-scale calculations of supercells with limited computational resources. A compromise is to impose an *ex post facto* corrections on the band edge (CBM and VBM) positions. This procedure, known as "scissor operation", simply shifts the positions of CBM and VBM by ΔE_C and ΔE_V respectively, which are the differences between the conventional DFT results and the advanced method results. Despite being able to reproduce the correct band gap value, this *ad hoc* procedure presents an addi-

[1] This statement is not strictly true; for example, the Fermi level can locate inside the conduction band as well as the valence band for very heavily-doped narrow-gap semiconductors (such as InAs and InSb). We do not consider those cases here.

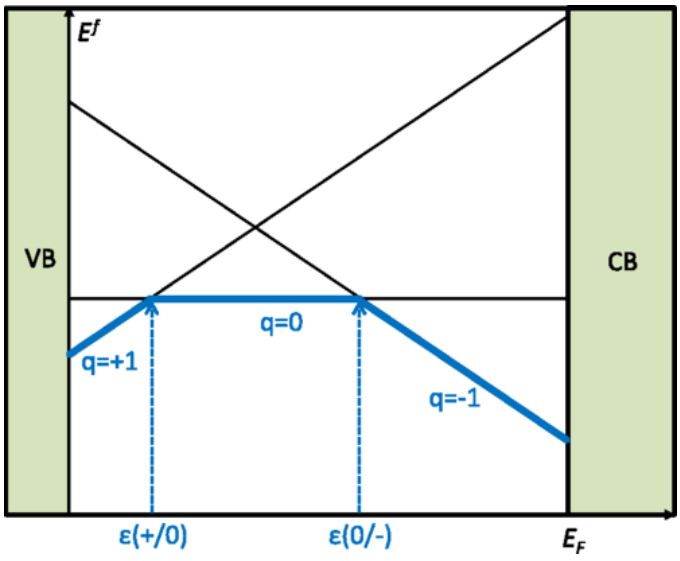

Figure 3.10: Schematic illustration of formation energy E^f vs Fermi level E_F for an amphoteric defect that can occur in three charge states q: +1, 0, and −1. Solid lines correspond to the formation energy as defined by Eq. (3.39). The defect exhibits two charge-state transition levels: a deep donor level $\varepsilon(+/0)$ and a deep acceptor level $\varepsilon(0/-)$. The thick solid lines indicate the energetically most favorable charge state for a given Fermi level. Reproduced from [41].

tional problem to the defect formation energies and calculated charge transition levels. Namely, it is not clear whether the charge transition level should stay at its absolute energy level (scheme (1) in Fig. 3.11), or keep constant its relative position to the nearest band edge (scheme (1) in Fig. 3.11). Lany and Zunger [47] proposed a general scheme for correcting the charge transition level with band edges. Specifically, they considered three different scenarios for defects: (1) truly shallow dopants (both before and after correction); (2) truly deep defects (both before and after correction); and (3) "pseudo-shallow" defects (shallow before correction, deep after correction). (Fig. 3.12) For truly shallow dopants, the charge transition level should follow the band edge position upon correction, as the delocalized character of the defect charge density belongs to that of a "per-

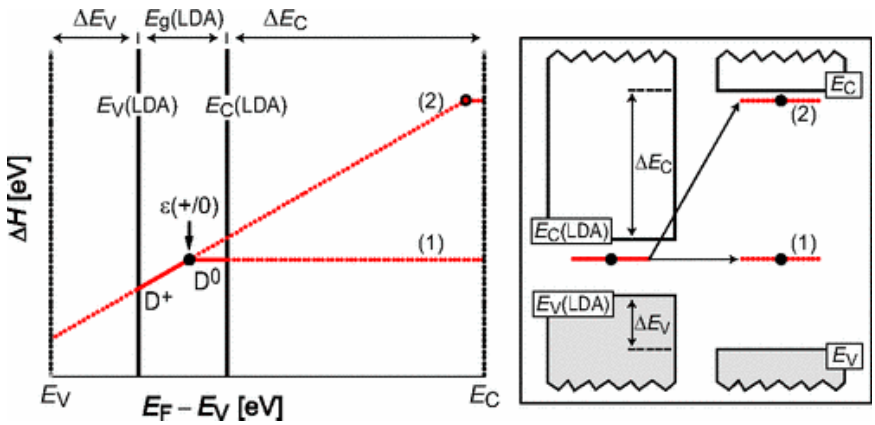

Figure 3.11: Schematic illustration of the effect on the formation energies (left) and single-particle energies (right) when the valence- and conduction-band edges in LDA are corrected by ΔE_V and ΔE_C toward the experimental gap $E_g = E_C - E_V$. Solid lines correspond to situation before correction and dotted lines to the situation after correction. The scale is chosen so as to illustrate the magnitude of corrections needed in ZnO. In general, the defect levels can be affected by the correction in varying degrees, as illustrated by examples (1) and (2). Reproduced from [47].

turbed host state"; for truly deep defects, the charge transition level should stay at its absolute energy level, as the localized character of the defect charge density belongs to that of a "defect-localized state". For "pseudo-shallow" defects (such as singly charged oxygen vacancy V_{O}^{+} in ZnO), since the nature of defect charge density is incorrectly described by the conventional XC functional, a much more complex correction scheme must be used, where the band structure of the host crystal must be modified *during the self-consistent calculation*. [47] For the second case, it is important to have a correct value of ΔE_C and ΔE_V, by having a common reference energy level of the two computation schemes. A natural choice is by aligning the average electrostatic potential $\bar{V}$ of the pristine supercell obtained using the two schemes. (Fig. 3.13)

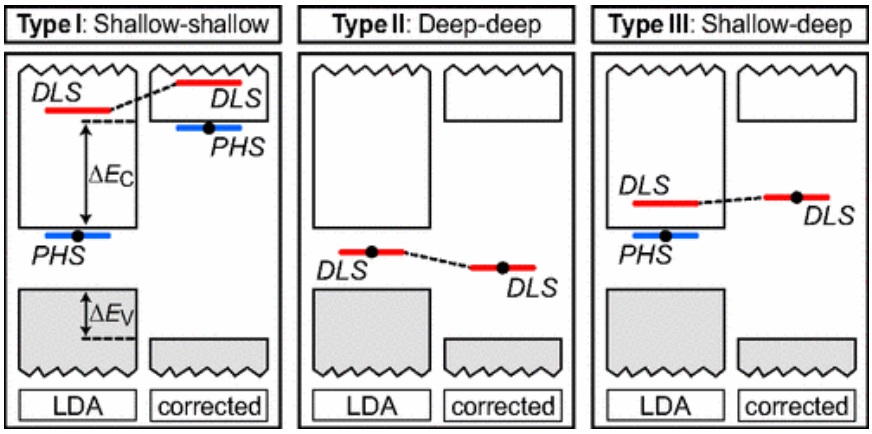

Figure 3.12: Schematic illustration of three qualitatively different behaviors of defect level during band-gap correction. If the primary defect level, i.e., the defect-localized state (DLS, red) is resonant inside the conduction band, the electron is released to a secondary, conduction-band-like perturbed-host state (PHS, blue). In this case, the defect exhibits shallow behavior. If the DLS lies inside the gap, the defect exhibits deep behavior. Type I: Shallow in LDA, shallow after correction. Type II: Deep in LDA, deep after correction. Type III: Shallow in LDA, deep after correction. Reproduced from [47].

Finite-size corrections

The finite-size correction term E_{corr} takes into account the energy contributions arisen from modeling a charged, defective system under periodic boundary conditions. There are two main contributions to E_{corr}. First, when a charged defect is introduced in a supercell, the $\mathbf{G} = 0$ terms of the Fourier-transformed classical Coulomb (Hartree) potential V_{H} and ionic potential V_{ext} of an infinite periodic array of charges would diverge to infinity. [49] To remedy this problem, these terms are set arbitrarily to zero, corresponding to imposing a uniform compensating charge background to obtain overall charge neutrality in the supercell. [47] This procedure leads to the unintended consequence of an arbitrary shift in the average electrostatic potential in the supercell, compared to the pris-

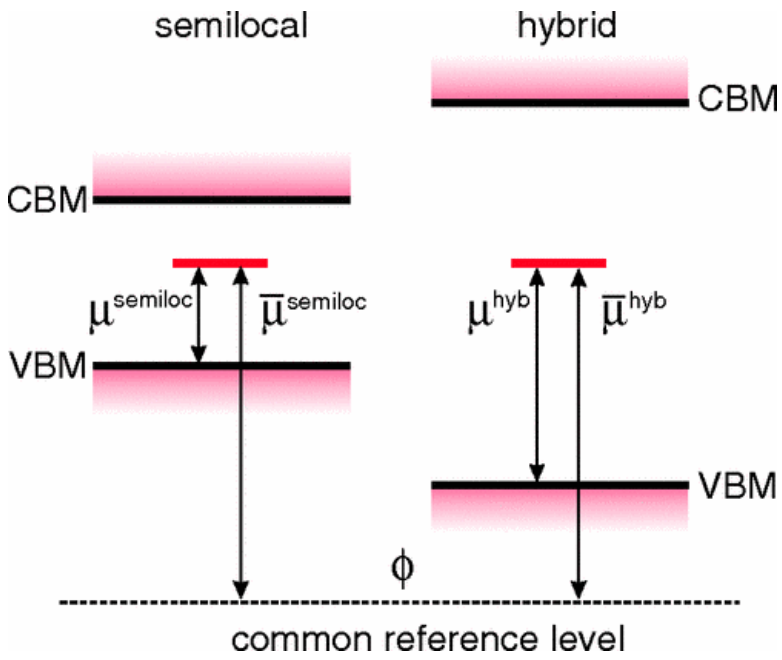

Figure 3.13: Schematic illustration of the alignment between energy levels obtained with a semilocal and a hybrid density functional. The charge transition levels μ and $\bar{\mu}$ are referred to the respective valence band maxima (VBM) and to a common reference level, respectively. The conduction band minima (CBM) are also shown. Reproduced from [48].

tine bulk. Second, for supercells that are not sufficiently large in all dimensions, the Coulomb interactions between the defect charge and its nearest periodic images do not vanish, (Fig. 3.14) which adds a spurious contribution to the total energy of the defective system. These two artifacts cause errors in defect formation energies, which in turn affect the accuracy of related quantities such as net carrier concentrations.

Fortunately, these sources of errors can be corrected with their respective means. For the error resulted from arbitrary potential shift, a procedure called "potential alignment" is needed; this procedure involves aligning the electrostatic potential of the pristine bulk system and that of the neutral defective sys-

tem, at a location **R** suffiently far from the defect:

$$\Delta E_{\mathrm{PA}}(D^q) = q\Delta V = q(\bar{V}_{\mathbf{R}}(D^q) - \bar{V}_{\mathbf{R}}(\mathrm{bulk})), \tag{3.43}$$

where $\bar{V}_{\mathbf{R}}(D^q)$ and $\bar{V}_{\mathbf{R}}(\mathrm{bulk})$ are the (local) atomic-sphere averaged electrostatic potential at **R** for defective and pristine bulk systems, respectively. This aligned potential provides a common energy reference for all pristine as well as defective systems. The second error is more difficult to correct, largely due to the complex nature of the spatial distribution of defect charges.

For the error resulted from artificial Coulomb interactions, several correction schemes have been proposed. Leslie and Gillan [50] argued and Makov and Payne [51] proved that, for an infinitely periodic array of localized charges q with a neutralizing background charge in a structureless dielectric with permittivity α, the correction energy ΔE_{Coul} can be expressed as the dipole term in the multipole expansion of defect charge density:

$$\Delta E_{\mathrm{Coul}}(D^q) = \frac{q^2\alpha_M}{2\varepsilon L} + \frac{2\pi qQ}{2\varepsilon L^3} + O(L^{-5}), \tag{3.44}$$

where α_M is the lattice-dependent Madelung constant, Q is the second radial moment of defect charge density, and $L = \Omega^{-1/3}$ is the characteristic lattice dimension of the supercell. This corretion scheme contains no emphirical fitting parameters; nonetheless, it requires a series of calculations with increasing L. This would lose its effectiveness if one of the dimensions of the unit cell is already large. Besides, the Makov-Payne correction scheme tends to overestimate the correction energy E_{Coul} for charged defects in semiconductors. [52]

An alternative scheme of correction is the so-called "FNV scheme" [53, 54]. In this method, the defect charge is approximated by a model charge q_{model} with the shape of a localized Gaussian function plus a delocalized exponential tail.

This model charge induces an electrostatic potential V, which corresponds to the difference between the electrostatic potential of the charged supercell and that of the neutral supercell. The long-range part of the potential, V_{lr}, characterizes all artificial Coulomb interactions between the model charge and its periodic images. The short-range part, $V_{\mathrm{sr}} = V - V_{\mathrm{lr}}$, is hence the real potential of the isolated defect charge. Then, assuming that the defect charge is localized, the short range potential should decay to zero in regions far from the defect. If it is not zero, then that nonzero constant should be equal to ΔV. The correction term, E_{corr}, is determined by the expression:

$$E_{\mathrm{corr}} = E_{\mathrm{iso}} - E_{\mathrm{per}} - q\Delta V; \tag{3.45}$$

where E_{iso} is the self-energy of the isolated model charge, and E_{per} refers to the artificial Coulomb energy of the periodic array of model charges. The FNV scheme has several advantages compared with Makov-Payne scheme, including: (1) convergence with respect to lattice parameter L is not necessary; and (2) the value of E_{corr} is largely independent of the shape details of the model charge. The downside of FNV scheme is that it cannot be used to correct formation energies of fully delocalized defects (such as shallow dopants). For practical purposes, we would assume (reasonably) that in this case, E_{corr} is approximately zero.

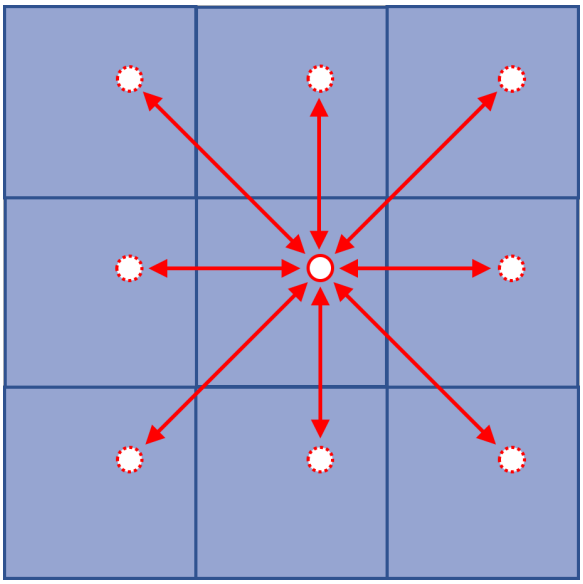

Figure 3.14: Schematic illustration of spurious Coulomb interactions between a charged defect in the supercell and its neighboring images.

3.5 Miscellaneous techniques in materials modeling

3.5.1 Special quasirandom structure

Substitutionally disordered solid solutions (alloys) are prevalent in nature, [55] and they have found many important applications in electronic devices. In order to have a thorough understanding of near-random alloys, it is critical to have an accurate model of a random crystal structure. Nevertheless, in a supercell with periodic boundary conditions, by construction no configuration can be truly random in the most strict sense, namely every (sub)lattice site may be occupied by an atom of either one of a set of elements (therefore almost no long-range order can occur in the crystal). One of the earliest and simplest attempts to model a random alloy is called the **"virtual crystal approximation" (VCA)**. [56] VCA is a mean-field approximation, meaning that it treats every

atom in a random crystal as the same species, namely a virtual atom whose potential $\langle V \rangle$ is the stoichiometric average of the potentials of all the elements that could occupy its lattice site. The main advantage of VCA is its low computational cost, as all the random mixing effects are included in the effective virtual atom. However, while VCA typically gives a qualitatively correct description of structural and electronic properties for simple semiconducting and ferromagnetic alloys, [57, 58, 59, 60] it suffers from several major limitations: (1) VCA is effective only when the perturbation in electron density, resulted from different atomic potentials of the alloying elements, are sufficiently small; (2) VCA completely neglects all chemical effects due to variations of local atomic environment. An alternative method called the **"coherent potential approximation" (CPA)** [61, 62] improves on VCA by taking into account the single-site scattering effect from randomly distributed elements. It is done by replacing the atomic potential at each lattice site with an effective potential V_0, which is determined *self-consistently* by requiring that within this potential, the average scattering from all the alloying elements at each lattice site is zero. Nonetheless, as a mean-field theory, CPA still fails to completely resolve all the disadvantages of VCA.

Despite these seemingly insurmountable difficulties in modeling random alloys, Walter Kohn's principle of nearsightedness [63], which states that any change at places far away has little impact on the local electronic structure, gives us hope in creating a model of random alloy that accurately reproduces the chemical properties of a real random alloy, even within a relatively small finite volume. The formalism of special quasirandom structure, proposed by Alex Zunger [64], is a theoretical framework of random alloy modeling aligned with this insight. The core idea is based on a powerful formalism called **"clus-**

ter expansion", which states that any macroscopic observable P of a crystal with any given atomic arrangement (configuration) σ can be expanded into a sum of contributions from particular groups of atoms called *clusters*:

$$P(\boldsymbol{\sigma}) = \sum_{\alpha} m_\alpha J_\alpha \rho_\alpha(\boldsymbol{\sigma}) \tag{3.46}$$

where m_α is the degeneracy of cluster α (i.e. the number of equivalent clusters to α in the crystal); J_α, called "effective cluster interaction" (ECI), is the expansion coefficient in the sum (it can be regarded as the contribution to P from cluster α); and $\rho_\alpha(\sigma)$ is the cluster correlation function of α (see below) in a crystal with configuration $\boldsymbol{\sigma}$. While Eq. (3.29) is exact if *all* clusters in the crystal are included, the true advantage of cluster expansion is that only a small number of short-ranged, low-ordered (having a few atoms) clusters (see Fig. 3.15 for example) are needed for the expansion to be sufficiently close to the true value $P(\boldsymbol{\sigma})$. For a given set of clusters $\{\alpha\}$, the ECIs $\{J_\alpha\}$ can be found by expressing Eq. (3.44) in matrix form $\mathbf{P} = \mathbf{\Pi J}$ and solve by matrix inversion $\mathbf{J} = \mathbf{\Pi}^{-1}\mathbf{P}$.

It is clear from Eq. (3.44) that cluster correlation functions $\{\rho_\alpha(\boldsymbol{\sigma})\}$ serve as the basis of the cluster expansion for a given configuration $\boldsymbol{\sigma}$. The cluster correlation function $\rho_\alpha(\boldsymbol{\sigma})$ is the average value of cluster functions $\Gamma_\alpha^{\mathbf{n}}$ over all clusters α' symmetrically equivalent to α in the crystal. The cluster functions, which capture both the geometry of and the atomic species in the cluster, are defined as

$$\Gamma_\alpha^{\mathbf{n}} = \prod_{i \in \alpha} \Theta_i^{n_i}(\sigma_i), \tag{3.47}$$

where σ_i denotes the species of atom i in cluster α, and $\Theta_i^{n_i}$ is the basis function with index n_i, which must satisfy the orthogonality condition. In binary random alloys, $\Theta_i^{n_i}(\sigma_i) = \sigma_i$.

In a completely random structure, by definition there is no correlation in

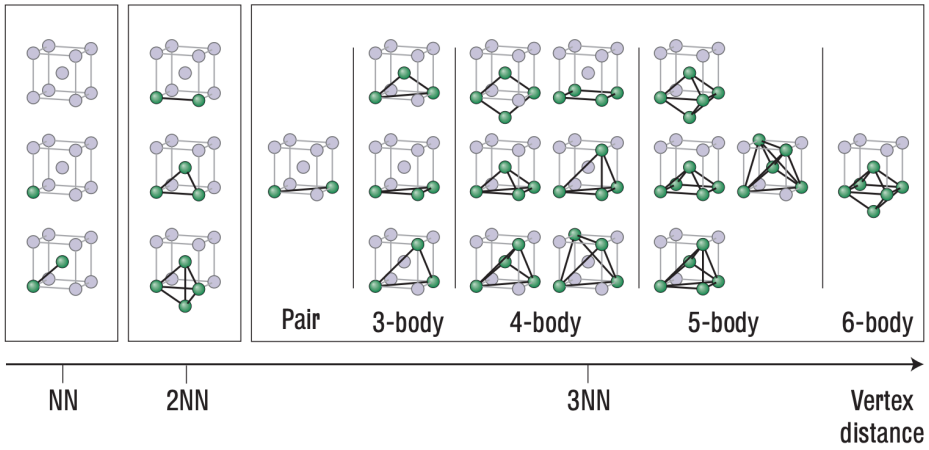

Figure 3.15: Geometrically unique clusters of fcc lattice sites. The average distance from the center of mass of the cluster increases moving left to right. Reproduced from [65].

atomic species arrangements between different clusters. Therefore, an accurate model of a random alloy must aim to satisfy this constraint as much as possible. Quantitatively, this means that the absolute difference

$$\Delta\rho_\alpha(\boldsymbol{\sigma}) = |\rho_\alpha(\boldsymbol{\sigma}) - \rho_\alpha(\boldsymbol{\sigma}_{\text{rand}})| \tag{3.48}$$

between the cluster correlation function of the model configuration $\boldsymbol{\sigma}$ and of the completely random configuration $\boldsymbol{\sigma}_{\text{rand}}$ must be minimized for as many symmetrically distinct clusters α as possible; any model that satisfies this criteria for a selected number of short-ranged clusters is called a **"special quasirandom structure" (SQS)**.

In the statistical sense, $\boldsymbol{\sigma}_{\text{rand}}$ can be calculated exactly; in a binary random alloy, for clusters containing k lattice sites, $\rho_\alpha(\boldsymbol{\sigma}_{\text{rand}}) = (2x - 1)^k$. In Chap. 5, as we consider vibrational signatures of dopants and defects in a random $In_{0.5}Ga_{0.5}As$, we need to have a 216-atom special quasirandom structure of $In_{0.5}Ga_{0.5}As$ which minimizes the target function, namely the sum of correla-

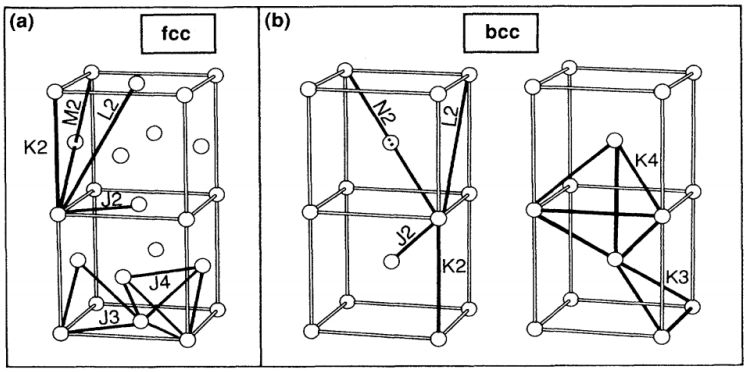

Figure 3.16: Schematics of selected short-ranged low-ordered atomic clusters in (a) face centered cubic (fcc) lattice; (b) body centered cubic (bcc) lattice. Reproduced from [66].

tion functions $\tilde{\rho} = \sum_{\alpha \in \Lambda} \Delta\rho_\alpha(\sigma)$ for a selected set of short-ranged, low-ordered clusters Λ. In this case, we choose $\Lambda = \{J_2, K_2, L_2, M_2, J_3, J_4\}$ on the cation sublattice which assumes fcc crystal structure. (Fig. 3.16) A version of simulated annealing, a meta-heuristic global optimization algorithm, is used to optimize the target function below the tolerance 10^{-2}. () The distribution of $\tilde{\rho}$ are shown in Fig. 3.17. We see that the distribution shows many local minima whose value exceeds the tolerance, hence it is crucial to use a global optimizer such as simulated annealing in order to find the required SQS.

3.5.2 Phonon calculation

Lattice dynamics

Atoms are in constant random oscillatory motion about their local energy minima, even under thermal equilibrium. In the harmonic approximation, the local potential energy landscape of a system of atoms can be described by a Taylor

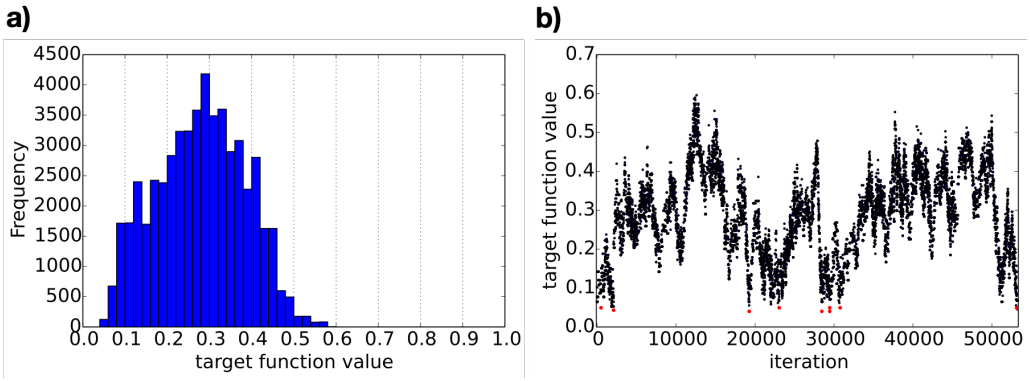

Figure 3.17: Distribution of the target function $\tilde{\rho}$ for configurations of 216-atom $In_{0.5}Ga_{0.5}As$ supercell traversed in the simulated annealing algorithm, shown as (a) histogram; (b) individual values, where the red dots represents the configurations with $\tilde{\rho} < 0.05$.

Algorithm 1: Algorithm for simulated annealing in SQS search.

Initialize:
random model configuration $\boldsymbol{\sigma}$;
target function value $\tilde{\rho}$ of $\boldsymbol{\sigma}$;
initial temperature $T_{\max}$;
final temperature $T_{\min} \ll T_{\max}$
while $T > T_{\min}$ **do**
  Swap two atoms of distinct species in $\boldsymbol{\sigma}$
  Calculate new target function value of $\boldsymbol{\sigma}$, $\tilde{\rho}'$
  if $\tilde{\rho}' < \tilde{\rho}$ **then**
    Accept the swap and update $\tilde{\rho} \leftarrow \tilde{\rho}'$
  else
    Accept the swap and update $\tilde{\rho} \leftarrow \tilde{\rho}'$ with probability $P = \frac{1}{1+\exp((\tilde{\rho}'-\tilde{\rho})/T)}$
  end if
  if $\tilde{\rho}'$ <tolerance **then**
    An SQS is found; break
  end if
  $T \leftarrow T - \Delta T$
end while

---

series about their equilibrium positions $\mathbf{r}_0$:

$$V(\mathbf{r}) = V(\mathbf{r}_0) + \frac{1}{2}\sum_{\alpha,\beta=1}^{N}\sum_{j,k=1}^{3}\frac{\partial^2 V(\mathbf{r})}{\partial r_{\alpha j}\partial r_{\beta k}}\Big|_{\mathbf{r}=\mathbf{r}_0} u_{\alpha j}u_{\beta k}, \tag{3.49}$$

where α,β are the atom indices, j,k are the Cartesian directions, and u is the (small) displacement of an atom along a particular direction. Taking derivative of the corresponding Hamiltonian with respect to $u_{\alpha j}$, we obtain the classical equation of motion for lattice dynamics

$$f_{\alpha j} = M_\alpha \ddot{u}_{\alpha j} = -\sum_{\beta=1}^{N}\sum_{k=1}^{3} F_{\alpha j,\beta k}u_{\beta k}. \tag{3.50}$$

where $f_{\alpha j}$ is the net force component on atom α in direction j, and $F_{\alpha j,\beta k} = \partial^2 V(\mathbf{r})/\partial r_{\alpha j}\partial r_{\beta k}\Big|_{\mathbf{r}=\mathbf{r}_0}$ is the $(\alpha j,\beta k)$-th entry of the **force constant matrix F** of the system. We will see that all the information of the system's vibrational properties are encoded in the **dynamical matrix D**, with $D_{\alpha j,\beta k} = F_{\alpha j,\beta k}/\sqrt{m_\alpha m_\beta}$.

In practical computational simulations, **D** can be calculated with two main methods: frozen phonon method, and density functional perturbation method. The frozen phonon method makes use of Eq. (3.48); namely, the dynamical matrix element $D_{\alpha j,\beta k}$ is obtained as follows: in a supercell with all the atoms at their equilibrium positions, displace the atom β in directions $+k$ and $-k$ (one at a time) by a small amount u, and calculate the corresponding force components $f^+_{\alpha j}$ and $f^-_{\alpha j}$; then $D_{\alpha j,\beta k}$ is given by

$$D_{\alpha j,\beta k} = \frac{f^+_{\alpha j} - f^-_{\alpha j}}{2u}. \tag{3.51}$$

The density functional perturbation method, on the other hand, operates in the reciprocal space, where **D** can be obtained by self-consistently calculating the charge density induced by a small perturbation $\partial n/\partial u$. This formalism is more evolved mathematically, and due to its complexity (scale as $O(N_e^4)$), it cannot yet

be used for large scale calculations in real space, e.g. supercells with defect(s). In Chap. 5, we use the frozen phonon method in our calculations of vibrational signatures of impurity modes in random InGaAs. We must note that, one of the limitations of the frozen phonon method is that it cannot deal with polar solids, where the internal electric field results in a non-analytical term at Γ point which induces splitting of high-frequency longitudinal and transverse optical (LO and TO) modes. This contribution can be naturally accounted for in the DFPT framework.

Finally, the dynamical matrix **D** for any system must satisfy two generic constraints. [67, 68] First, by definition the force constants between two atoms should not depend on the relative order of atom-direction $(\alpha j, \beta k)$ indices; therefore, **D** must be diagonally symmetric ($D_{\alpha j,\beta k} = D_{\beta k,\alpha j}$). Second, due to the translational symmetry of crystals, the forces acting on each atom cannot be changed upon a uniform displacement on all atoms. This implies that $\sum_k D_{\alpha j,\beta k} = 0$, which is known as the "acoustic sum rule". Both constraints must be applied iteratively in order for **D** to satisfy them simultaneously.

Local phonon density of states

The force constant $D_{\alpha j,\beta k}$ is a quantitative measure of the mechanical stiffness of interatomic bonds, which dictates the characters of vibration between these two atoms; therefore, the dynamical matrix **D** contains the complete vibrational information of the system. Nonetheless, the dynamical matrix itself does not correspond to any *experimentally* measurable quantity; instead, such quantities are stored in the spectrum of the dynamical matrix. In a periodic crystal, vibration of native atoms propagate throughout the lattice, thereby assuming the

form of plane waves with certain frequencies ω and wavevectors $\mathbf{k}$ The equation of motion (3.48) then becomes

$$\omega^2 \mathbf{e} = \mathbf{D} \cdot \mathbf{e}, \tag{3.52}$$

where $\mathbf{e}$ is the displacement vector of all the atoms in the system. Solving this eigenvalue problem gives the eigenfrequencies ω_p and eigen-displacement $\mathbf{e}_p$ of the vibrational **normal modes** of the system. Mature experimental techniques such as spectroscopic methods are capable of producing **phonon spectrum** of a materials system, which contains the intensities as a function of frequencies of vibration. Normal modes of vibration of a given materials system appear as peaks in its phonon spectrum. The frequency of a particular vibrational mode in the phonon spectrum directly indicates the strengths of the force constants, which depends on the underlying local lattice geometry and atomic configuration. Moreover, many important thermodynamic quantities, such as vibrational entropy and free energy, can be directly obtained from mathematical operations on the phonon spectrum.

The phonon spectrum can be approximately calculated within the *ab initio* framework; in this context, it is known as the **phonon density of states (PDOS)**, analogous to the electron density of states. Formally, the PDOS can be expressed as

$$g(\omega) = \frac{1}{N} \sum_{l=1}^{3N} \delta(\omega - \omega_l). \tag{3.53}$$

The PDOS describes the *global* vibrational patterns of the whole system; it is also possible to obtain the *local* phonon density of states (LPDOS) by projecting the PDOS onto each atom:

$$g(\omega, \mathbf{r}_\alpha) = \sum_{l=1}^{3N} \delta(\omega - \omega_l) |\mathbf{e}_\alpha^l|^2. \tag{3.54}$$

It is clear that the expressions (3.51) for PDOS and (3.52) for LPDOS both require knowledge of all the eigenfrequencies ω_p of the dynamical matrix. However, for very large systems, solving Eq. (3.50) exactly is impractical, as the direct matrix diagonalization typically scales as $O(N^3)$ with N = number of row/columns of the matrix. Fortunately, there exists an approach that can solve for $g(\omega, \mathbf{r}_\alpha)$ without solving for the eigen-displacement vector $\mathbf{e}$, and therefore is very efficient for large systems. To do so, we first recognize that the LPDOS can be expressed in terms of the lattice static Green's function $\mathbf{G}$, defined as: [69]

$$G_{\alpha j,\beta k}(\omega^2) = [(\omega^2\mathbf{I} - \mathbf{D})^{-1}]_{3\alpha-3+j,3\beta-3+k} = \sum_{l=1}^{3N} \frac{e^l_{\alpha j} e^{l\dagger}_{\beta k}}{\omega^2 - \omega_l^2}, \tag{3.55}$$

by

$$g(\omega, \mathbf{r}_\alpha) = 2\omega\left(\frac{1}{\pi}\lim_{\eta\to 0^+} \mathrm{Im}\, G_{\alpha j,\alpha j}(\omega^2 + i\eta)\right). \tag{3.56}$$

It is clear from Eq. (3.54) that the diagonal terms $G_{\alpha j,\alpha j}$ of the Green's function is all we need in order to obtain the LPDOS. $G_{\alpha j,\alpha j}$ of a large symmetric matrix $\mathbf{D}$ can be conveniently calculated using the recursion method first proposed by Haydock *et al.* [70, 71, 72] In this method, the dynamical matrix $\mathbf{D}$ is first transformed into a tridiagonal form (possible for all symmetric matrices):

$$\tilde{\mathbf{D}} = \begin{bmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & b_{3N-1} & a_{3N-1} & b_{3N} \\ 0 & \cdots & 0 & b_{3N} & a_{3N} \end{bmatrix}. \tag{3.57}$$

where a_l and b_l are determined with the Lanczos recursion algorithm [73] (Algorithm 2). This procedure is repeated until either the maximum number of iterations has been reached, or b_l converges to zero, whichever comes first. Typically, for systems with large $3N$, it takes far fewer steps n than $3N$ for a_l and b_l

to converge to their respective stable values, thereby making the Haydock recursion method a very efficient algorithm for large systems. After all of a_l and b_l in $\tilde{\mathbf{D}}$ and $\boldsymbol{\phi}_l$ have been obtained, the Green's function $\mathbf{G}$ in matrix representation can be calculated from the Green's function $\tilde{\mathbf{G}} = (z\mathbf{I} - \tilde{\mathbf{D}})^{-1}$: ($z \equiv \omega^2 + i\eta$)

$$\mathbf{G}(z) = \mathbf{L}\tilde{\mathbf{G}}(z)\mathbf{L}^T, \tag{3.58}$$

where $\mathbf{L} = [\boldsymbol{\phi}_1\, \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n \cdots]$ is the transformation matrix. From Eq. (3.56) the following relation can be deduced:

$$G_{\alpha j,\alpha j}(z) = \tilde{G}_{11}(z), \tag{3.59}$$

where $\tilde{G}_{11}$ is the entry on the first row and first column of $\tilde{\mathbf{G}}$. This equation reveals that the LPDOS can be obtained only if we know $\tilde{G}_{11}(z)$, which can be calculated as a continued fraction involving only a_l's and b_l's:

$$\tilde{G}_{11}(z) = \cfrac{1}{z - a_1 - \cfrac{b_2^2}{z - a_2 - \cfrac{b_3^2}{\ddots\ \cfrac{b_n^2}{z - a_n - b_{n+1}^2 t(z)}}}} \tag{3.60}$$

where $t(z) = [z - a_\infty - \sqrt{(z - a_\infty)^2 - 4\beta_\infty^2}]/(4\beta_\infty^2)$, with a_∞ and b_∞ being the converged values of a_l and b_l respectively. We find that for a typical 1000-atom system of dopant-containing random $In_{0.5}Ga_{0.5}As$, $n = 300$ yields an converged LPDOS for both the dopant atom and the host atoms.

Algorithm 2: Lanczos recursion algorithm.

Initialize:
$l \leftarrow 0$;
$\phi_0 \leftarrow \mathbf{0}$;
$\phi_1 \leftarrow (\ldots, 0, 1, 0, \ldots)^T$ ($(3\alpha - 3 + j)$-th entry = 1, other entries = 0);
$a_1 \leftarrow \boldsymbol{\phi}_1^T \mathbf{D} \boldsymbol{\phi}_1$;
$b_1 \leftarrow 0$;
Maximum number of iterations N_{maxiter}
while $l < N_{\text{maxiter}}$ or $b_l \neq 0$ **do**
  $\tilde{\phi}_{l+1} \leftarrow (\mathbf{D} - a_l \mathbf{I})\phi_l - b_l \phi_{l-1}$
  $a_{l+1} \leftarrow \tilde{\boldsymbol{\phi}}_{l+1}^T \mathbf{D} \tilde{\boldsymbol{\phi}}_{l+1}$
  $b_{l+1} \leftarrow \sqrt{\tilde{\boldsymbol{\phi}}_{l+1}^T \tilde{\boldsymbol{\phi}}_{l+1}}$
  $\boldsymbol{\phi}_l \leftarrow \frac{\tilde{\phi}_l}{b_l}$
  $l \leftarrow l + 1$
end while

3.5.3 Quantum transport

Charge transport in the nanoscale

Electron transport governs the electrical properties and performance of devices under operation. A two-terminal device can be schematically represented by Fig. 17. In the device, the two terminals, corresponding to the left electrode (LE) and right electrode (RE), are electron reservoirs which are respectively fixed at a given electron chemical potential (or voltage); the transport region ("sample"), which lies in between the electrodes, determines the transport behavior of electrons. On the most fundamental level, the transport characteristics of a device at zero bias (voltage difference between LE and RE) is governed by the interaction between electron and other particles (such as electrons and atoms) and quasiparticles (such as phonons). Upon encountering such entities, the electron wavefunction will be scattered either elastically or inelastically; such scatterings are the main source of resistance for electrons in a device. If the characteristic

dimensions of a device in the transport direction L_c is greater than the average distance between two elastic scattering events for an electron ("mean free path" L_m), it is then very likely for an electron to experience elastic scattering during its travel across the device; such regime is called "diffusive transport" (Fig. 3.18(a)). If $L_c < L_m$, than it is unlikely for an electron to experience resistance from elastic scattering; such regime is called "ballistic transport". Analogously, electronic transport without experiencing any inelastic scattering is called "coherent transport" (Fig. 3.18(b)), as it preserves the phase of the electronic wavefunction. Diffusive transport can be well described by semiclassical transport formalism based on Boltzmann's and Drude's theory, which leads to the familiar Ohm's law ($R = \rho L/A$, where ρ = resistivity of the material, A is the cross-sectional area in the transverse direction); in contrast, coherent ballistic transport is best described by the **Landauer-Büttiker formalism** [74, 75], which violates Ohm's law in that the resistance of a device is independent of the length of the device. For devices with gate length on the order of 1-10 nm, the most relevant transport regime is the ballistic regime. Fortunately, the essential transport characteristics of such devices can be simulated in practice via non-equilibrium Green's function (NEGF) method, combined with either *ab initio* DFT or empirical calculation methods such as tight binding.

Landauer-Büttiker formalism

For a nanoscale device, the electron is confined within a very small space with the dimension $\sim W$ in the transverse direction. This, rather than scattering, is the main source of resistance for electrons in the ballistic regime. To motivate our discussion, we first consider an idealistic one-dimensional device whose width

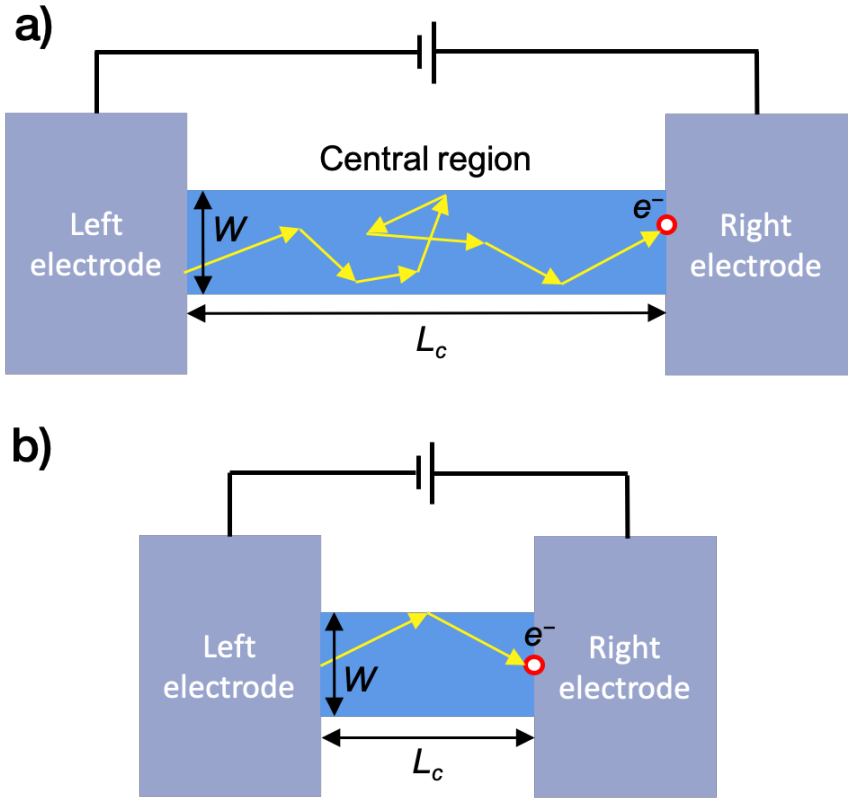

Figure 3.18: Schematics of (a) diffusive transport; (b) ballistic transport in a two-terminal device, where the characteristic length in the transport direction is L_c.

W is so thin, such that the electron wavefunction is confined to *one* conduction channel with energy level $E_0(k) = E_0(k = 0)+\hbar^2k^2/2m_e^*$ (k = longitudinal wavevector). When a forward bias $V = \frac{\mu_R-\mu_L}{e}$ ($\mu_L > \mu_R$) is applied between the electrodes, a current I is induced in the device. I consists of a right-going component I_+ (corresponding to $k > 0$) and a left-going component I_- (corresponding to $k < 0$).

Under finite temperature,

$$\begin{aligned} I &= I_+ + I_- \\ &= n_+ e v_+ + n_- e v_- \\ &= \left[\frac{e}{L}\sum_{k>0} T_k(E) v_k(E) f_L(E)\right] + \left[\frac{e}{L}\sum_{k'<0} T_{k'}(E) v_{k'}(E) f_R(E)\right] \\ &= \frac{e}{L}\left[\sum_k T_k(E) v_k(E)(f_L(E) - f_R(E))\right] \end{aligned} \tag{3.61}$$

In these equations, v represents the group velocity of the electron wavefunctions. The third equality comes from the fact that the number density of electrons transmitted $n = \sum_k T_k(E) f(E)$, where $T_k(E)$ is the transmission coefficient (ratio of transmitted current vs. total current), and $f(E)$ is the Fermi function, corresponding to the left or right electrode ($f_{L/R}(E) = [1 + \exp((E - \mu_{L/R})/k_B T)]^{-1}$). The last equality comes from the fact that the right-going transmission coefficient $T_+ = T_{k>0}$ should be equal to the left-going transmission coefficient $T_- = T_{k<0}$, and $v_- = -v_+$, due to time-reversal symmetry. Replacing the sum $\sum_k$ by $2(L/(2\pi))\int \mathrm{d}k$ (the factor 2 accounts for spin degeneracy) and $v(k)$ by $(1/\hbar)(\mathrm{d}E/\,\mathrm{d}k)$, we obtain

$$I = \frac{2e}{h}\int_{-\infty}^{\infty} \mathrm{d}E\, T(E)(f_L(E) - f_R(E)). \tag{3.62}$$

In the linear response regime ($|\mu_L - \mu_R|$ is small), we can Taylor expand the right hand side and obtain in the first-order approximation

$$I = \frac{2e}{h} T(E)(\mu_L - \mu_R). \tag{3.63}$$

Hence, in the $T = 0$K and full transmission ($T(E) = 1$) limit, the conductance $G = \frac{I}{V} = \frac{2e^2}{h}$; this is called the quantum of conductance G_0. For a realistic device with a finite width W, many such channels are available to the electrons; furthermore, the transmission probability of electronic wavefunctions in each

channel is always less than one due to existence of scattering in the sample. Therefore, for a device with M channels and transmission probabilities T_i, the total conductance is

$$G = \frac{2e^2}{h} \sum_{i=1}^{M} T_i. \tag{3.64}$$

This formula is called the Landauer-Büttiker formalism; it is a general formula for describing coherent ballistic transport in electronic devices.

Non-equilibrium Green's function (NEGF)

The non-equilibrium Green's function (NEGF) method [76, 77, 78] is a practical scheme to calculate transport characteristics of nanoscale devices. It is capable of treating devices with either weakly interacting ballistic contacts (as in the Landauer limit) or strongly interacting contacts. Within the NEGF framework, a device is modeled as an open-boundary configuration, where the outer parts of left/right electrodes are treated as the corresponding bulk material with periodic boundary conditions. Formally, the Hamiltonian of the system $\mathbf{H}$ can be expressed as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_L & \mathbf{H}_{L,C} & \mathbf{0} \\ \mathbf{H}_{L,C}^{\dagger} & \mathbf{H}_C & \mathbf{H}_{R,C} \\ \mathbf{0} & \mathbf{H}_{R,C}^{\dagger} & \mathbf{H}_R \end{bmatrix}, \tag{3.65}$$

where the subscript X ($X = L$ (left electrode), C (central region), R (right electrode)) denotes the Hamiltonian term for subspace X, and the subscript X, Y denotes the coupling Hamiltonian between region X and region Y. The elements of $\mathbf{H}$ are given by

$$H_{ij} = \int \mathrm{d}\mathbf{r} \phi_i^*(\mathbf{r} - \mathbf{R}_I) \mathbf{H}(\mathbf{r}) \phi_j(\mathbf{r} - \mathbf{R}_J), \tag{3.66}$$

where ϕ_i and ϕ_j are atomic orbitals centered around atom I and J respectively. Note that here, instead of the plane wave basis, a LCAO (linear combination of atomic orbitals) basis is used. This basis, unlike the plane wave basis, is in general non-orthogonal; hence the Schrödinger equation of the system assumes the matrix form

$$\mathbf{Hc} = E\mathbf{Sc}, \tag{3.67}$$

where $S_{ij} = \int \mathrm{d}\mathbf{r}\phi_i^*(\mathbf{r}-\mathbf{R}_I)\phi_j(\mathbf{r}-\mathbf{R}_J)$ is the overlap integral between orbitals i and j (nonzero for nonorthogonal orbitals). The overlap matrix $\mathbf{S}$ should assume the same form (Eq. (3.63)) as $\mathbf{H}$. We recognize that directly solving Eq. (3.63) is not feasible, as the Hamiltonian and overlap submatrices containing orbital(s) from left/right electrode must be infinite-dimensional due to the open boundary conditions. To solve this problem, we write Eq. (3.65) in another form:

$$(E\mathbf{S}-\mathbf{H})\mathbf{G} = \mathbf{I}, \tag{3.68}$$

with $\mathbf{I}$ = identity matrix. Eq. (3.66) can be considered as the definition of the **Green's function G**. This form is more advantageous to work with than Eq. (3.65), as it is possible to reduce the infinite-dimensional problem to a finite one without losing any useful information. To see this, we first note that the submatrices $\tilde{\mathbf{H}}_{LC}$ and $\tilde{\mathbf{H}}_{RC}$ of the "reduced device Hamiltonian" $\tilde{\mathbf{H}} \equiv (E+i\eta)\mathbf{S}-\mathbf{H}$ (*eta* is an infinitesimally small positive number, introduced to avoid the pole of $\mathbf{G}$) can be eliminated via block Gaussian elimination; this procedure would transform $\tilde{\mathbf{H}}_C$ into the form

$$\begin{aligned}\tilde{\mathbf{H}}_C' &= \tilde{\mathbf{H}}_C - (\tilde{\mathbf{H}}_{LC}^\dagger \tilde{\mathbf{H}}_L^{-1}\tilde{\mathbf{H}}_{LC}) - (\tilde{\mathbf{H}}_{RC}^\dagger \tilde{\mathbf{H}}_R^{-1}\tilde{\mathbf{H}}_{RC}) \\ &= \tilde{\mathbf{H}}_C - \mathbf{\Sigma}_L - \mathbf{\Sigma}_R,\end{aligned} \tag{3.69}$$

where $\mathbf{\Sigma}_L = \mathbf{H}_{L,C}^\dagger \mathbf{H}_L^{-1}\mathbf{H}_{L,C}$ and $\mathbf{\Sigma}_R = \mathbf{H}_{R,C}^\dagger \mathbf{H}_R^{-1}\mathbf{H}_{R,C}$ are the self-energy matrices of the left/right electrode respectively, which contain all the information of the inter-

action between the electrode and the central region. In order to facilitate calculation, the whole device with semi-infinite electrodes is divided into "principal layers" (Fig. 3.19), where the atomic orbitals in each layer only interacts with orbitals within the range covering nearest-neighbor layers. This decomposition is possible when there is a lack of long-range Coulomb interaction between the atoms, due to the nearsightedness principle of electron density [63]. The original infinite-dimensional reduced Hamiltonian $\tilde{\mathbf{H}}$ thus reads

$$\tilde{\mathbf{H}} = \begin{bmatrix} \ddots & \ddots & & & & \\ \ddots & \tilde{\mathbf{h}}_L & \tilde{\mathbf{h}}_{LL} & \mathbf{0} & & \\ & \tilde{\mathbf{h}}_{LL}^{\dagger} & \left[\tilde{\mathbf{H}}_C\right] & \tilde{\mathbf{h}}_{RR} & & \\ & & \mathbf{0} & \tilde{\mathbf{h}}_{RR}^{\dagger} & \tilde{\mathbf{h}}_R & \ddots \\ & & & & \ddots & \ddots \end{bmatrix}, \tag{3.70}$$

where

$$\tilde{\mathbf{H}}_C = \begin{bmatrix} \tilde{\mathbf{h}}_1 & \tilde{\mathbf{h}}_{1,2} & & \\ \tilde{\mathbf{h}}_{1,2}^{\dagger} & \ddots & \ddots & \\ & \ddots & \ddots & \tilde{\mathbf{h}}_{n-1,n} \\ & & \tilde{\mathbf{h}}_{n-1,n}^{\dagger} & \tilde{\mathbf{h}}_n \end{bmatrix}. \tag{3.71}$$

It can be seen from Eq. (3.63) and (3.68) that the semi-infinite submatrices $\tilde{\mathbf{H}}_{LC}$ and $\tilde{\mathbf{H}}_{RC}$ are non-zero only in the block $\tilde{\mathbf{H}}_{LL}$ and $\tilde{\mathbf{H}}_{RR}$ respectively. This means that the self-energy matrices $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ can be reduced to finite blocks with the same size as $\tilde{\mathbf{H}}_{LL}$ and $\tilde{\mathbf{H}}_{RR}$ respectively, without losing any information of electrode-central region interaction. After a series of matrix algebra, we can arrive at the following expression for transmission coefficient:

$$T(E) = \mathrm{Tr}(\mathbf{\Gamma}_L \mathbf{G}_C^{\dagger} \mathbf{\Gamma}_R \mathbf{G}_C) \tag{3.72}$$

where $\mathbf{\Gamma}_{L/R} = i(\mathbf{\Sigma}_{L/R} - \mathbf{\Sigma}_{L/R}^{\dagger})$ are the so-called "broadening functions", and $\mathbf{G}_C = \tilde{\mathbf{H}}_C^{-1}$ is the Green's function of the central region. Combined with the Landauer

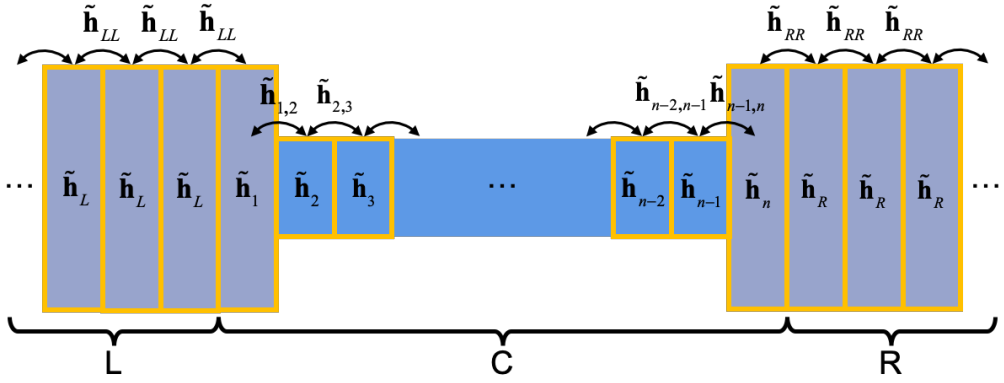

Figure 3.19: Schematics of a infinite two-electrode device, divided into principal layers; interaction within each layer is described by the block reduced Hamiltonian $\tilde{\mathbf{h}}_I$, whereas interaction of nearest-neighbor layers is described by $\tilde{\mathbf{h}}_{IJ}$.

current formula (Eq. (3.60)), we now have a computational recipe for calculating the current in a physical device model.

3.5.4 Population analysis

Chemists' intuition suggests that there exists an intrinsic connection between the chemical bonds and the electron density. From a quantum mechanical point of view, chemical bonds are special regions of the electron density, formally known as "bond critical points (CP)". A bond CP is defined mathematically as locations where the spatial gradient of the electron density $\nabla n(\mathbf{r}) = 0$, and the electron density $n(\mathbf{r})$ shows a local minimum along the axis connecting the two bonding atoms, and a local maximum along two orthogonal axes. With this rigorous definition of chemical bonds, it becomes possible to quantitatively analyze chemical bonding in material systems (called "population analysis"). Two of the many methods of population analysis are the "crystal orbital overlap

population" (COOP) [79] and "crystal orbital Hamilton population" (COHP) [80]. COOP and COHP are based on the idea that chemical bonding information can be derived from partitioning either the total electron number (COOP) or the electronic band structure (COHP) in an energy-resolved manner. Specifically, COOP is defined as the density of states weighted by the overlap matrix **S**:

$$\mathrm{COOP}_{ij}(E) = S_{ij} \sum_n f_J c_{ni}^* c_{nj} \delta(E - E_n), \tag{3.73}$$

where c_{ni} are the coefficients of expansion of electron density in linear combinations of atomic orbitals (LCAO), S_{ij} is the overlap between atomic orbitals φ_i and φ_j; f_J is the occupation number of each band J. Similarly, COHP is defined as the density of states weighted by the Hamilton matrix **H**:

$$\mathrm{COHP}_{ij}(E) = H_{ij} \sum_n f_J c_{ni}^* c_{nj} \delta(E - E_n), \tag{3.74}$$

where H_{ij} is the Hamiltonian matrix element of orbitals i and j. COOP is an indication of bond order, while COHP reflects bond strength between two neighboring atoms. When plotted as a function of one-electron energy E, both COOP and COHP can show different bonding characters within a specified energy range: for COOP, positive values indicate bonding (stabilizing) interaction, and negative values indicate anti-bonding (destabilizing) interaction; for COHP, vice versa. Another useful quantity is the negative integrated COHP (−ICOHP), defined as $-\int_{-\infty}^{E_F} \mathrm{d}E\, \mathrm{COHP}(E)$. The magnitude of −ICOHP is an indication of the strength of the covalent bonds between two atoms; the greater the |−ICOHP|, the stronger the covalent bonding interaction. In Chap. 7, we adopt the plane-wave formalism of COHP (projected COHP, or pCOHP) [81] to study the bond strength between surface dopant and nearest-neighbor vacancy on Ga_2O_3(010) surface.

Bibliography

[1] Daniel A Reed, Ruzena Bajcsy, Manuel A Fernandez, Jose-Marie Griffiths, Randall D Mott, Jack Dongarra, Chris R Johnson, Alan S Inouye, William Miner, Martha K Matzke, et al. Computational science: ensuring america's competitiveness. Technical report, PRESIDENT'S INFORMATION TECHNOLOGY ADVISORY COMMITTEE ARLINGTON VA, 2005.

[2] June Gunn Lee. *Computational materials science: an introduction*. Crc Press, 2016.

[3] Berna Akgenc. *A study of the relation of piezoelectric properties and nano structures through methods in computational physics*. PhD thesis, Yildiz Technical University, 2016.

[4] Richard M Martin and Richard Milton Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2004.

[5] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der physik*, 389(20):457–484, 1927.

[6] Douglas R Hartree. The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 89–110. Cambridge University Press, 1928.

[7] Vladimir Fock. Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems. *Zeitschrift für Physik*, 61(1-2):126–148, 1930.

[8] V Fock. "selfconsistent field" mit austausch für natrium. *Zeitschrift für Physik*, 62(11-12):795–805, 1930.

[9] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

[10] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[11] N David Mermin. Thermal properties of the inhomogeneous electron gas. *Physical Review*, 137(5A):A1441, 1965.

[12] JF Janak. Proof that $\partial e/\partial n_i = \varepsilon$ in density-functional theory. *Physical Review B*, 18(12):7165, 1978.

[13] John P Perdew and Mel Levy. Physical content of the exact kohn-sham orbital energies: band gaps and derivative discontinuities. *Physical Review Letters*, 51(20):1884, 1983.

[14] David M Ceperley and BJ Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 45(7):566, 1980.

[15] Wolfram Koch and Max C Holthausen. *A chemist's guide to density functional theory*. John Wiley & Sons, 2015.

[16] A Van de Walle and G Ceder. Correcting overbinding in local-density-approximation calculations. *Physical Review B*, 59(23):14992, 1999.

[17] Seymour H Vosko, Leslie Wilk, and Marwan Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of physics*, 58(8):1200–1211, 1980.

[18] John P Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B*, 23(10):5048, 1981.

[19] John P Perdew, John A Chevary, Sy H Vosko, Koblar A Jackson, Mark R Pederson, Dig J Singh, and Carlos Fiolhais. Atoms, molecules, solids, and

surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Physical review B*, 46(11):6671, 1992.

[20] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[21] John P Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244, 1992.

[22] Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2):785, 1988.

[23] John P Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, 33(12):8822, 1986.

[24] BHLB Hammer, Lars Bruno Hansen, and Jens Kehlet Nørskov. Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals. *Physical Review B*, 59(11):7413, 1999.

[25] John P Perdew, Adrienn Ruzsinszky, Gábor I Csonka, Oleg A Vydrov, Gustavo E Scuseria, Lucian A Constantin, Xiaolan Zhou, and Kieron Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical review letters*, 100(13):136406, 2008.

[26] John P Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*, volume 577, pages 1–20. AIP, 2001.

[27] AV Arbuznikov. Hybrid exchange correlation functionals and potentials: Concept elaboration. *Journal of Structural Chemistry*, 48(1):S1–S31, 2007.

[28] Enrico Clementi and Subhas J Chakravorty. A comparative study of density functional models to estimate molecular atomization energies. *The Journal of Chemical Physics*, 93(4):2591–2602, 1990.

[29] Axel D Becke. Density-functional thermochemistry. i. the effect of the exchange-only gradient correction. *The Journal of chemical physics*, 96(3):2155–2160, 1992.

[30] Philip J Stephens, FJ Devlin, CFN Chabalowski, and Michael J Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry*, 98(45):11623–11627, 1994.

[31] 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.

[32] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *The Journal of chemical physics*, 110(13):6158–6170, 1999.

[33] 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.

[34] Jochen Heyd, Gustavo E Scuseria, and Matthias Ernzerhof. Erratum: "hybrid functionals based on a screened coulomb potential" [j. chem. phys. 118, 8207 (2003)]. *The Journal of Chemical Physics*, 124(21):219906, 2006.

[35] Benjamin G Janesko, Thomas M Henderson, and Gustavo E Scuseria. Screened hybrid density functionals for solid-state chemistry and physics. *Physical Chemistry Chemical Physics*, 11(3):443–454, 2009.

[36] Protik Das, Mohammad Mohammadi, and Timur Bazhirov. Accessible computational materials design with high fidelity and high throughput. *arXiv preprint arXiv:1807.05623*, 2018.

[37] Wikipedia contributors. Bloch wave — Wikipedia, the free encyclopedia, 2019. [Online; accessed 12-September-2019].

[38] Hendrik J Monkhorst and James D Pack. Special points for brillouin-zone integrations. *Physical review B*, 13(12):5188, 1976.

[39] DJ Chadi and Marvin L Cohen. Special points in the brillouin zone. *Physical Review B*, 8(12):5747, 1973.

[40] Alfonso Baldereschi. Mean-value point in the brillouin zone. *Physical Review B*, 7(12):5212, 1973.

[41] Christoph Freysoldt, Blazej Grabowski, Tilmann Hickel, Jörg Neugebauer, Georg Kresse, Anderson Janotti, and Chris G Van de Walle. First-principles calculations for point defects in solids. *Reviews of modern physics*, 86(1):253, 2014.

[42] Anuj Goyal, Kiran Mathew, Richard G Hennig, Aleksandr Chernatynskiy, Christopher R Stank, Samuel T Murphy, David A Andersson, Simon R Phillpot, and Blas P Uberuaga. The conundrum of relaxation volumes in first-principles calculations of charge defects. *arXiv preprint arXiv:1704.04044*, 2017.

[43] Elif Ertekin, Varadharajan Srinivasan, Jayakanth Ravichandran, Pim B Rossen, Wolter Siemons, Arun Majumdar, Ramamoorthy Ramesh, and Jeffrey C Grossman. Interplay between intrinsic defects, doping, and free carrier concentration in srtio3 thin films. *Physical Review B*, 85(19):195460, 2012.

[44] Guo-Xin Qian, Richard M Martin, and DJ Chadi. First-principles study of the atomic reconstructions and energies of ga-and as-stabilized gaas (100) surfaces. *Physical Review B*, 38(11):7649, 1988.

[45] Hannu-Pekka Komsa and Alfredo Pasquarello. Assessing the accuracy of hybrid functionals in the determination of defect levels: Application to the as antisite in gaas. *Physical Review B*, 84(7):075207, 2011.

[46] Wei Chen and Alfredo Pasquarello. Accuracy of gw for calculating defect energy levels in solids. *Physical Review B*, 96(2):020101, 2017.

[47] 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*, 78(23):235104, 2008.

[48] Audrius Alkauskas, Peter Broqvist, and Alfredo Pasquarello. Defect energy levels in density functional calculations: Alignment and band gap problem. *Physical review letters*, 101(4):046405, 2008.

[49] J Ihm, Alex Zunger, and Marvin L Cohen. Momentum-space formalism for the total energy of solids. *Journal of Physics C: Solid State Physics*, 12(21):4409, 1979.

[50] M Leslie and NJ Gillan. The energy and elastic dipole tensor of defects in ionic crystals calculated by the supercell method. *Journal of Physics C: Solid State Physics*, 18(5):973, 1985.

[51] G Makov and MC Payne. Periodic boundary conditions in ab initio calculations. *Physical Review B*, 51(7):4014, 1995.

[52] Chris G Van de Walle and Jörg Neugebauer. First-principles calculations for defects and impurities: Applications to iii-nitrides. *Journal of applied physics*, 95(8):3851–3879, 2004.

[53] Christoph Freysoldt, Jörg Neugebauer, and Chris G Van de Walle. Fully ab initio finite-size corrections for charged-defect supercell calculations. *Physical review letters*, 102(1):016402, 2009.

[54] Christoph Freysoldt, Jörg Neugebauer, and Chris G Van de Walle. Electrostatic interactions between charged defects in supercells. *physica status solidi (b)*, 248(5):1067–1076, 2011.

[55] James A. Krumhansl. It's a random world. In Henry O. Hooper and Adriaan M. de Graaf, editors, *Amorphous Magnetism: Proceedings of the International Symposium on Amorphous Magnetism, August 17–18, 1972, Detroit, Michigan*, chapter 2, pages 15–25. Plenum Press, New York, 1973.

[56] Lothar Nordheim. Zur elektronentheorie der metalle. i. *Annalen der Physik*, 401(5):607–640, 1931.

[57] Stefano De Gironcoli, Paolo Giannozzi, and Stefano Baroni. Structure and thermodynamics of si x ge 1- x alloys from ab initio monte carlo simulations. *Physical review letters*, 66(16):2116, 1991.

[58] Nicola Marzari, Stefano de Gironcoli, and Stefano Baroni. Structure and phase stability of ga x in 1- x p solid solutions from computational alchemy. *Physical review letters*, 72(25):4001, 1994.

[59] DA Papaconstantopoulos and WE Pickett. Tight-binding coherent potential approximation study of ferromagnetic la 2/3 ba 1/3 mno 3. *Physical Review B*, 57(20):12751, 1998.

[60] P Slavenburg. Tife 1- x co x salloys and the influence of antistructural atoms. *Physical Review B*, 55(24):16110, 1997.

[61] Paul Soven. Coherent-potential model of substitutional disordered alloys. *Physical Review*, 156(3):809, 1967.

[62] B Velickỳ, S Kirkpatrick, and H Ehrenreich. Single-site approximations in the electronic theory of simple binary alloys. *Physical Review*, 175(3):747, 1968.

[63] Emil Prodan and Walter Kohn. Nearsightedness of electronic matter. *Proceedings of the National Academy of Sciences*, 102(33):11635–11638, 2005.

[64] Alex Zunger, S-H Wei, LG Ferreira, and James E Bernard. Special quasirandom structures. *Physical Review Letters*, 65(3):353, 1990.

[65] Lance Jacob Nelson. Cluster expansion models via bayesian compressive sensing. 2013.

[66] ZW Lu, S-H Wei, Alex Zunger, S Frota-Pessoa, and LG Ferreira. First-principles statistical mechanics of structural stability of intermetallic compounds. *Physical Review B*, 44(2):512, 1991.

[67] Alexei A Maradudin, Elliott Waters Montroll, George Herbert Weiss, and IP Ipatova. *Theory of lattice dynamics in the harmonic approximation*, volume 3. Academic press New York, 1963.

[68] GJ Ackland, MC Warren, and SJ Clark. Practical methods in ab initio lattice dynamics. *Journal of Physics: Condensed Matter*, 9(37):7861, 1997.

[69] Eleftherios N Economou. *Green's functions in quantum physics*, volume 3. Springer, 1980.

[70] R Haydock, Volker Heine, and MJ Kelly. Electronic structure based on the local atomic environment for tight-binding bands. *Journal of Physics C: Solid State Physics*, 5(20):2845, 1972.

[71] PE Meek. Vibrational spectra and topological structure of tetrahedrally bonded amorphous semiconductors. *Philosophical Magazine*, 33(6):897–908, 1976.

[72] MJ Kelly. Applications of the recursion method to the electronic structure from an atomic point of view. In *Solid State Physics*, volume 35, pages 295–383. Elsevier, 1980.

[73] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282, 1950.

[74] Rolf Landauer. Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM Journal of Research and Development*, 1(3):223–231, 1957.

[75] M Büttiker, Y Imry, R Landauer, and S Pinhas. Generalized many-channel conductance formula with application to small rings. *Physical Review B*, 31(10):6207, 1985.

[76] Supriyo Datta. Steady-state quantum kinetic equation. *Physical Review B*, 40(8):5830, 1989.

[77] Supriyo Datta. A simple kinetic equation for steady-state quantum transport. *Journal of Physics: Condensed Matter*, 2(40):8023, 1990.

[78] Yigal Meir and Ned S Wingreen. Landauer formula for the current through an interacting electron region. *Physical review letters*, 68(16):2512, 1992.

[79] Timothy Hughbanks and Roald Hoffmann. Chains of trans-edge-sharing molybdenum octahedra: metal-metal bonding in extended systems. *Journal of the American Chemical Society*, 105(11):3528–3537, 1983.

[80] Richard Dronskowski and Peter E Blöchl. Crystal orbital hamilton populations (cohp): energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *The Journal of Physical Chemistry*, 97(33):8617–8624, 1993.

[81] Volker L Deringer, Andrei L Tchougréeff, and Richard Dronskowski. Crystal orbital hamilton population (cohp) analysis as projected from plane-wave basis sets. *The journal of physical chemistry A*, 115(21):5461–5466, 2011.

CHAPTER 4

AB INITIO MODELING OF VACANCIES, ANTISITES, AND SI DOPANTS IN ORDERED INGAAS

The work presented in this chapter is published in: **J. Wang**, B. Lukose, M. O. Thompson, and P. Clancy, Journal of Applied Physics **121**, 045106 (2017).

4.1 Introduction

Since the earliest days of semiconductor technologies, III-V materials have been recognized and extensively studied for their superior electrical and optical properties. Not surprisingly then, III-V devices have found broad commercial application, ranging from high electron-mobility transistors (HEMT) to solid state lasers [1]. As advances in CMOS technology shrink transistor nodes below 10 nm, silicon-based MOSFETs are beginning to see performance degradation due to pronounced quantum tunneling effects. In comparison, $In_{0.53}Ga_{0.47}As$, a ternary III-V compound semiconductor, possesses superior electron transport properties, such as high electron mobility (8450 cm^2/V-sec at 300 K) [2] and high injection velocity ($\sim 3 \times 10^7$ cm/sec for gate length < 30 nm), making it particularly appealing for next-generation MOSFET channel materials [3].

Currently, several major challenges remain for InGaAs to be considered suitable for practical use, including the low density of trap states at the channel-dielectric interface, low Ohmic contact resistivity, and hetero-integration on a silicon platform [4]. In particular, in order for sub-10 nm n-MOSFETs to work properly, the contact resistance must be less than 5×10^{-19} Ω-cm^2, corresponding to a lower limit of carrier concentration above $\sim 1 \times 10^{20}$ cm^{-3} [5]. However,

experimental studies have shown that thermodynamically stable carrier concentrations in heavily Si-doped n-type $In_{0.53}Ga_{0.47}As$ have, so far, not exceeded $\sim 1.4 \times 10^{19}$ cm^{-3}, largely regardless of annealing temperature [6]. Possible limiting factors of donor activation include the low solid solubility of dopants, self compensation, and compensation by intrinsic defects. To date, the origin of low dopant activation in n-type InGaAs is still not known from a theoretical point of view. The urgency and technological importance of addressing this fundamental challenge is reflected by its explicit inclusion in the 2013 edition of the *International Technology Roadmap for Semiconductors* [4].

In recent years, some *ab initio* theoretical studies have also been carried out on defects in InGaAs. [7, 8, 9, 10, 11, 12] Each of these studies considers a selected portion of all possible defects in InGaAs. This approach is suitable for investigating properties of individual defects, but does not provide a holistic picture of defect-dopant interactions. In particular, the dominant species of compensating defect – the main interest of experimental investigation – cannot be properly identified without taking into account all possible species of intrinsic defects. Furthermore, qualitative predictions of maximum carrier and defect concentrations cannot be properly determined.

In this paper, we provide a detailed and quantitative description of dopant activation and deactivation mechanisms in InGaAs on atomic and electronic scales, taking into account both dopant-induced defects and native point defects in InGaAs. Using Density Functional Theory (DFT) in conjunction with a band gap correction employing many-body perturbation theory (*GW* approximation), we calculate formation energies and thermodynamic transition levels of Si on substitutional and interstitial sites, as well as of vacancies and antisites

in InGaAs. From these *ab initio* results, we estimate the equilibrium concentration of various defects and free electrons after annealing at higher temperature, and thereby identify the predominant compensation mechanism in InGaAs under various growth conditions. These results help us understand the atomistic origin of low dopant activation, and could serve as a guide in determining the optimal conditions for n-type doping in InGaAs.

4.2 Methods

4.2.1 Defect formation energy

In the supercell formalism, the formation energy of a defect in a crystalline solid is given by

$$E^f(D^q) = E(D^q) - E_{\text{bulk}} - \sum_i \Delta n_i \mu_i + q(E_V + E_F) + E_{\text{corr}}, \tag{4.1}$$

where $E(D^q)$ is the total energy of the supercell containing a defect D in charge state q, and E_{bulk} is the total energy of the pristine bulk supercell. Δn_i is the number of atoms of species i added to ($\Delta n_i > 0$) or removed from ($\Delta n_i < 0$) the supercell as a result of forming the defect, and $\mu_i = \mu_i^{\text{bulk}} + \Delta\mu_i$ is the chemical potential of element species i. For charged defects, the formation energy also contains a contribution from the chemical potential of the electrons, also known as the Fermi level, E_F. The Fermi level of a semiconductor is treated as an independent variable that can assume any value within the band gap, and it is referenced to E_V, the valence band maximum (VBM) of the bulk material. The correction term, E_{corr}, takes into account the errors introduced by finite size effects and the periodic boundary conditions, such as spurious overlaps of

neighboring defect wave functions and, in the case of charged defects, Coulomb interactions between image charges. We use a combined correction scheme by Lany and Zunger [13] and Freysoldt *et al.* [14, 15], as this approach does not require additional higher-level DFT or *GW* calculations. (For details, see Sec. 4.2.4)

4.2.2 Details of DFT and *GW* calculations

The composition of InGaAs alloys that is most commonly considered for n-channel devices, $In_{0.53}Ga_{0.47}As$, is lattice-matched to an InP substrate, and falls close to an equimolar InAs/GaAs ratio. However, the stoichiometry reveals nothing of the way that the cations (In and Ga) are arranged on an atomic scale; the cations could, in principle, be either randomly arranged or ordered. Experimental observations have repeatedly confirmed the presence of long-range order in ternary III-V alloys, including InGaAs [16, 17, 18, 19]; theoretically, it is also understood that ordered structures reduce total bulk strain energy compared with that of a random solid solution, as bond distortion is minimized. [20] For ternary III-V alloys, the two most common *ordered* structures fall into the categories of CuAu-I and CuPt-B [21], which correspond to periodically alternating layers of In and Ga atoms in the $\langle 001 \rangle$ and $\langle 111 \rangle$ directions, respectively, as shown in Fig. 4.1. In this study, we focus only on CuAu-I ordered $In_{0.5}Ga_{0.5}As$ alloys.

We use Density Functional Theory (DFT) to obtain energy-minimized structures and the energetics of defects in InGaAs. All the DFT calculations in this study are performed within the local density approximation (LDA) in the

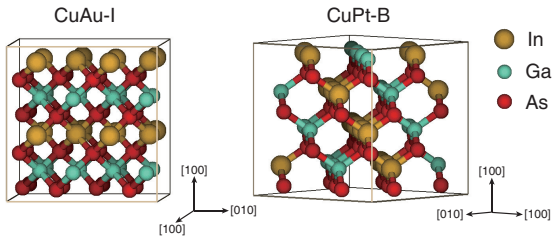

Figure 4.1: 64-atom supercells of CuAu-I and CuPt-B ordered $In_{0.5}Ga_{0.5}As$ superlattices, showing alternating layers of In and Ga in the [001] and [111] direction.

Perdew-Zunger parameterization [22], with the norm-conserving pseudopotential of a Goedecker-Hartwigsen-Hutter-Teter (GHHT) type [23, 24], as implemented in the plane-wave DFT code QUANTUM ESPRESSO [25]. Compared to our previous study (Lee *et al.* [10]) which used the Projector Augmented Wave (PAW) method within the Generalized Gradient Approximation (GGA), we find that using a norm-conserving pseudopotential in LDA produces a better description of parameters such as the bulk modulus and band gap for InGaAs. The 4*d* electron states of In and the 3*d* states of Ga are treated as core states, since inclusion of those electrons as valence electrons would require an infeasibly large (> 250 Ry) cut-off energy in order to achieve adequate convergence. It has been shown that such simplifications yield a systematic error of around 0.1 eV in the final value of the defect formation energy, independent of supercell size, but changes monotonically with the charge state of the defect [26].

Thermodynamic arguments suggest that a certain amount of native defects will always be present within a crystal under normal conditions. Hence, in this study, we investigated a large number of native defects (vacancies, cation-anion and anion-cation antisites) in InGaAs, together with Si-induced defects (substitutional and interstitial). Those defects are chosen for their relatively low defect formation energies (compared with, say, self-interstitials) as well as the consid-

eration of different charge states (in contrast to cation-cation antisites In_{Ga} and Ga_{In}), which make them possible donors and acceptors in InGaAs. The complexity of the ternary compound itself and the sheer number of defect species and charge states that need to be modeled makes it imperative to use the most efficient method of calculation as well as maintaining sufficient accuracy. In that regard, our first test was to check for convergence of results with respect to supercell size. We considered supercells of InGaAs containing 64 and 128 atoms, corresponding respectively to a $2 \times 2 \times 2$ simple cubic unit cell and a $4 \times 4 \times 4$ fcc unit cell of zincblende structure. As we will show later for Si substitutional defects, the formation energies converge to within ~0.05 eV for both a 128-atom supercell and a 64-atom cell, indicating that considering a system size larger than 64 atoms is unnecessary. In contrast, the formation energies of charged interstitials, vacancies and antisites do not converge even for a 128-atom supercell, due to pronounced finite-size effects; hence, we apply an adequate correction scheme for charged defects. We adopt the correction method proposed by Freysoldt, Neugebauer, and Van de Walle (FNV) [14, 27], as this scheme is purely *ab initio* and requires no additional DFT calculations.

Self-consistent calculations determine the cutoff energy and **k**-point sampling necessary to satisfy the convergence criteria we chose for the total energy ($\leq$ 5 meV/atom) and force ($\leq$ 10 meV/(A·atom)) on each atom. Based on these criteria, we use an energy cutoff E_{cut} = 50 Ry for all our calculations. The **k**-points are a Γ-centered $3 \times 3 \times 3$ Monkhorst-Pack mesh for both 64-atom and 128-atom cells. Using these input parameters, a cell relaxation optimizes the structure of a bulk InGaAs supercell. This is followed by an ionic relaxation after introducing the defect to the cell. All atoms are slightly displaced in random directions for the defective supercell to remove any spurious symmetry.

Our convergence criteria for relaxation are 10^{-5} Ry for the total energy and 10^{-4} Ry/Bohr for the force for 64-atom supercells, and 10^{-4} Ry for the total energy and 10^{-4} Ry/Bohr for the force for larger cells.

To reproduce the correct doping behavior and accurately predict the maximum free carrier concentration, requires that the defect formation energies are calculated accurately. Unfortunately, it is well known that, as a ground state theory, LDA DFT not only systematically underestimates the band gap of semiconductors and insulators, but invariably predicts the wrong absolute band energies, which, in turn, critically affects the value of the defect formation energy. On the other hand, many-body perturbation theory – a higher-level *ab initio* method – takes into account the exchange and correlation effects of electrons, and thus provides a much more accurate approximation to both the band energies and the band gaps than DFT results in general [28]. Since, to our knowledge, the experimental energy of valence band maximum is unknown for $In_{0.5}Ga_{0.5}As$, we resort to many-body perturbation theory to obtain its VBM energy.

We use the code YAMBO [29] as the algorithm for many-body perturbation theory calculations. These calculations use non-self-consistent one-shot *GW* (G_0W_0) under a plasmon-pole approximation (PPA) for dynamically screened Coulomb interactions (W). Due to the extremely high computational cost of *GW* methods, we use the four-atom primitive cell of CuAu-I type $In_{0.5}Ga_{0.5}As$ for the calculation. The preceding LDA DFT calculations use a 60 Ry energy cutoff and a Γ-centered 21×21×15 **k**-mesh (corresponding to 528 **k**-points in the irreducible Brillouin zone). We use 96 bands to calculate the Green's function, G, and the irreducible polarizability, χ. We use up to 6 Ry of **G**-vectors in the dielectric

function and 12 Ry of **G**-vectors in the exchange part of the self energy. To test the accuracy of the G_0W_0 calculation, its predicted band gap is compared to the experimental value. Our choices for the parameters described above yield an uncertainty of less than 0.05 eV for both the energy of VBM and the band gap.

4.2.3 Constraints on equilibrium chemical potentials

From Eq. (1), the defect formation energy is dependent on the elemental chemical potentials, μ_i. As the relative abundance of each atomic species during growth is unknown, the exact values of each μ_i cannot be determined. Nevertheless, in thermal equilibrium, the value of μ_i in a compound is subjected to multiple constraints. First, each μ_i must be no greater than the chemical potential of the bulk elemental solid ($\Delta\mu_i \equiv \mu_i - \mu_i^{\text{bulk}} \leq 0$). Next, the sum of the chemical potentials of any two elements A and B must be no greater than the chemical potential of the bulk binary compound A_mB_n ($m\Delta\mu_A + n\Delta\mu_B \leq \Delta H_{A_mB_n}$, where ΔH denotes heat of formation). Lastly, the sum of the chemical potentials (weighted by stoichiometry) of all three species must be equal to the chemical potential of the bulk ternary compound ($\Delta\mu_{In} + \Delta\mu_{Ga} + 2\Delta\mu_{As} = \Delta H_{InGaAs_2}$). Essentially, these constraints ensure that no elemental and binary bulk phase is thermodynamically preferred than the ternary alloy itself. The boundary of the allowed region of elemental chemical potentials as defined by these constraints corresponds to the limits of growth conditions of $In_{0.5}Ga_{0.5}As$ in experiments.

Table 4.1 lists heats of formation per formula unit for all possible binary compounds that can be formed with In, Ga, As and Si, and the unit cell of CuAu-I ordered $In_{0.5}Ga_{0.5}As$. Apart from $SiAs_2$, whose very large positive

heat of formation is a consequence of its instability below 45 kbar of pressure [30], all other compounds are stable under standard conditions. However, we found that the requirements $\Delta\mu_{\mathrm{In}} + \Delta\mu_{\mathrm{As}} \leq \Delta H_{\mathrm{InAs}}$, $\Delta\mu_{\mathrm{Ga}} + \Delta\mu_{\mathrm{As}} \leq \Delta H_{\mathrm{GaAs}}$ and $\Delta\mu_{\mathrm{In}} + \Delta\mu_{\mathrm{Ga}} + 2\Delta\mu_{\mathrm{As}} = \Delta H_{\mathrm{InGaAs_2}}$ cannot mutually hold. Indeed, The heat of formation of $InGaAs_2$ is greater than the sum of the heats of formation of GaAs and InAs by a small amount, $\sim$ 0.08 eV. Fig. 4.2 shows that the phase space for GaAs and InAs overlap with each other, leaving no space for CuAu-I ordered $In_{0.5}Ga_{0.5}As$ to be thermodynamically favorable. This indicates that, from a purely enthalpic point of view, bulk CuAu-I ordered $In_{0.5}Ga_{0.5}As$ cannot be stably formed by mixing GaAs and InAs compounds. The endothermic formation of ordered InGaAs from binaries is also supported by the calculations of Chakrabarti *et al.* [31] for chalcopyrite-type $In_{0.5}Ga_{0.5}As$. According to Chakrabarti *et al.*, surface reconstruction is the driving force behind cation ordering in the lattice, thus such ordered alloys are favored by growth kinetics rather than thermodynamics. For our purposes, this simply means that we should discard the constraints imposed by binaries GaAs and InAs (constraint 2), and only stipulate that the elemental chemical potentials be limited by their bulk values and by the heat of formation of InGaAs (constraints 1 and 3). Thus, the limiting growth conditions now correspond to the vertices of the triangular region in Fig. 4.2, namely In-poor ($\Delta\mu_{\mathrm{In}} = \Delta H_{\mathrm{InGaAs_2}}$), Ga-poor ($\Delta\mu_{\mathrm{Ga}} = \Delta H_{\mathrm{InGaAs_2}}$), and In/Ga-rich ($\Delta\mu_{\mathrm{In}} = \Delta\mu_{\mathrm{Ga}} = 0$) growth conditions.

4.2.4 Corrections to the defect formation energy

Accurate determination of defect formation energies is a challenging task, since it involves the difference between two total energies that are orders of mag-

Substance	ΔH (eV)	Ref. ΔH (eV)
InAs	−0.65	-0.62 ± 0.02^a
GaAs	−0.76	-0.85 ± 0.01^a
SiAs	−0.13	-0.12^b
$SiAs_2$	151.65	N/A
$InGaAs_2$	−1.33	N/A

Table 4.1: Heats of formation per formula unit of all possible binary compounds formed by In, Ga, As and Si, and CuAu-I ordered $InGaAs_2$. [a]Experimental values [32]; [b]theoretical value [33].

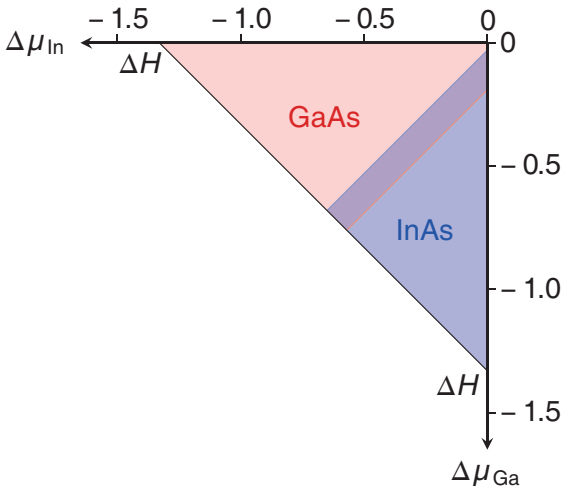

Figure 4.2: Parameter space spanned by chemical potentials of In and Ga, constrained by $\Delta H \leq \Delta\mu_{In} \leq 0$ and $\Delta H \leq \Delta\mu_{Ga} \leq 0$ (see Sec. 4.2.3 for detail). $\Delta H = -1.33$eV is the heat of formation of $InGaAs_2$, a unit cell of CuAu-I ordered $In_{0.5}Ga_{0.5}As$. Chemical potential domains of GaAs (red) and of InAs (blue) overlap, indicating that CuAu-I ordered $In_{0.5}Ga_{0.5}As$ is a metastable phase.

nitude greater than the formation energy itself, as is evident from Eq. (1). Freysoldt *et al.* [34] argues that the accuracy of calculated defect formation energies in a semiconductor can be no better than about 0.1 eV using DFT. Errors usually arise from two sources. The first kind of error is due to insufficient precision of the calculation, such as finite-size effects, including band-filling effects, defect wave function overlap, and image charge interactions. Such errors can

usually be kept to a minimum if the supercell size is large enough, or some careful correction scheme is applied *a posteriori*. The second kind of error is inherent in the approximate nature of the exchange-correlation (xc) functional in DFT. This error, in contrast, cannot be systematically reduced, without using more accurate, and far more computationally costly, methods such as *GW* or DFT with hybrid functionals. These errors can, in principle, be removed by a general correction term as follows: [13]

$$E_{\text{corr}} = (\Delta E_{\text{BF}} + \Delta E_{\text{el}}) + (q\Delta E_V + z_e \Delta E_C - z_h \Delta E_V). \tag{4.2}$$

The first two terms, the band-filling correction ΔE_{BF} and the electrostatic correction ΔE_{el}, attempt to remove the first kind of error described above. The band-filling correction is specific for shallow dopants; it corrects the Moss-Burstein type [35, 36] band-filling effect of donor electron/acceptor hole, which arises from impurity interactions in very highly doped semiconductors. Since the purpose of this study is to investigate the limit of doping for a channel material, the dilute-limit case scenario is of limited interest, and hence we do not adopt the band-filling correction here. On the other hand, the electrostatic correction remedies the slow convergence of Coulomb interaction between image charges of the defect. To date, several schemes for electrostatic correction have been proposed [37, 14, 15, 13]. The most widely used scheme consists of an alignment of the average electrostatic potential between the defective and the bulk supercell, together with a Makov-Payne correction term:

$$\Delta E_{\text{el}}^{\text{MP}} = \frac{q^2 \alpha}{2\varepsilon L} - \frac{2\pi q Q}{3\varepsilon L^3}, \tag{4.3}$$

where ε is the dielectric constant of the bulk solid, α the Madelung constant of the supercell (calculated from a periodic array of point charges q), and Q the second radial moment of the localized defect charge density, ρ_D, contained

in the supercell. The Makov-Payne scheme assumes that the defect charge can be approximated as a point charge, hence it gives an upper bound on the amount of Coulomb interactions between image charges. For real physical defects, the amount of the correction is usually smaller than $\Delta E_{\text{el}}^{\text{MP}}$ [15], implying that Makov-Payne correction generally overestimates the finite-size error. Some studies [38] have shown that a Makov-Payne-like term, $a_1 L^{-1} + a_n L^{-n}$, agrees with the trend for extrapolation of defect formation energies to the infinite-cell limit, but the coefficients a_1, a_n and n are simply fitting parameters, and thus cannot be determined from first principles. These issues limit the accuracy of the Makov-Payne scheme. A recent alternative option is the Freysoldt-Neugebauer-Van de Walle (FNV) scheme [14, 15], which subtracts from the formation energy the interaction energies between periodic images of defect charges (E_{inter}) and between isolated defect charge and compensating background charge (E_{intra}). This correction already takes into account the alignment of plane-averaged electrostatic potential between defective and bulk supercells, so no separate alignment needs to be performed. Komsa *et al.* [39] has shown that, compared to the Makov-Payne scheme, the FNV scheme generally produces more accurate correction energies. Since the FNV scheme does not need any information other than the DFT electrostatic potentials, we use this scheme in our study for all charged defects.

Unlike the finite-size effects which can largely be eliminated using simple *a posteriori* correction methods, the error in band edge positions (CBM and VBM) is due to the inherent deficiency of Kohn-Sham DFT in describing quasiparticle energy levels, and thus must be determined from other methods. For defect formation energy calculations, we need to obtain the correct VBM, E_V, and band gap value, E_g. The deviation of the exact value of E_V from the DFT value is

denoted by ΔE_V. This adjustment will also change the formation energy of shallow donors and acceptors, as the localized defect levels are hybridized with the conduction or valence band. The amount of change is $z_e \Delta E_C$ and $-z_h \Delta E_V$ for donors and acceptors, respectively, where z_e (z_h) is the number of electrons (holes) occupying the perturbed CBM (VBM) caused by the shallow donor (acceptor). (Details see Lany and Zunger [13] Sec. III F) For deep defects, no such shift is necessary, since the defect level is localized within the band gap.

4.2.5 Maximum dopant and carrier concentration

Under equilibrium conditions, the formation energy of a defect D in charge state q, $E^f(D^q)$, is directly related to the defect concentration in a crystal, which is given by:

$$c(D^q) = N_{\text{site}}\, g(D^q)\, \exp(-E^f(D^q)/k_B T), \tag{4.4}$$

where N_{site} is the concentration of sites available for the defect, $g(D^q)$ is the degeneracy of the defect, and k_B the Boltzmann constant. Since the equilibrium Fermi level, E_F, depends on the doping level, it must be calculated self-consistently using the charge neutrality condition for semiconductors:

$$\sum_{D,q} q \cdot c(D^q) - n + p = 0, \tag{4.5}$$

where the electron concentration, n, and the hole concentration, p, are calculated from the density of states $g(E)$ of bulk $In_{0.5}Ga_{0.5}As$ and the Fermi-Dirac distribution $f(E)$ as:

$$n = \int_{E_C}^{\infty} f(E, E_F, T) g(E) dE; \tag{4.6}$$

$$p = \int_{-\infty}^{E_V} [1 - f(E, E_F, T)] g(E) dE. \tag{4.7}$$

Note that, to attain maximum accuracy under degenerate doping, we use the general expressions with the density of states, rather than the approximation using effective masses of the bands, to calculate these concentrations. By choosing the optimal growth condition for n-type dopants such as Si in InGaAs, the formation energies of charged donors would decrease, while those of charged acceptors would increase, together with a decrease in donor compensation effect. These would lead to an increase in the equilibrium value of E_F towards the conduction band edge (E_C), and hence achieve the maximum electron concentration.

4.3 Results

4.3.1 Bulk Properties of CuAu-I type $In_{0.5}Ga_{0.5}As$

As a ternary alloy of binary compounds GaAs and InAs, $In_xGa_{1-x}As$ adopts a structure derived from the zincblende structure of the binary compounds; yet, the atomic positions of In, Ga and As are distorted from the original Wyckoff positions 4*a* and 4*c*. For CuAu-I type $In_{0.5}Ga_{0.5}As$, the additional ordering in [001]-direction reduces the degrees of freedom of atomic positions in a random alloy. The primitive cell of CuAu-I type $In_{0.5}Ga_{0.5}As$ assumes simple tetragonal structure with space group $P\bar{4}m2$ [40], which is different from the space group of a zincblende structure ($F\bar{4}3m$) due to lattice distortion. The primitive cell vectors are (1/2, 1/2, 0), (-1/2, 1/2, 0), and (0, 0, η); the two cation species (In, Ga) occupy lattice sites $(0, 0, 0)$ and $(0, 1/2, \eta/2)$, and the anions (As) occupy lattice sites $(1/4, 1/4, \eta u)$ and $(1/4, 3/4, \eta(1-u))$ (coordinates are in units of conventional

Defect	q	Volume (Å^3)	Vol. relax. (%)
Si_{In}	+1	6.95	−16.9
	0	6.95	−16.9
Si_{Ga}	+1	6.92	−4.5
	0	6.95	−3.9
Si_{As}	0	7.15	−8.3
	−1	7.07	−9.4
Si^{a}_{T1}	+2	7.55	−9.7
	+1	7.53	−10.0
	0	7.53	−10.0
Si^{b}_{T1}	+2	8.63	+3.2
	+1	8.80	+5.2
	0	9.57	+14.4
Si_{T2}	+2	9.98	+28.0
	+1	9.65	+23.7
	0	8.39	+7.5
V_{In}	0	5.26	−37.0
	−1	5.19	−38.0
	−2	5.08	−39.3
	−3	4.93	−41.0
V_{Ga}	0	5.03	−30.5
	−1	5.00	−30.9
	−2	4.95	−31.5
	−3	4.89	−32.4
V_{As}	+1	5.03	−7.8
	0	5.03	−25.0
	−1	5.00	−48.0
	−2	4.95	−57.6
	−3	4.89	−64.8
As_{In}	+2	7.82	−6.6
	+1	8.24	−1.5
	0	8.62	−3.0
As_{Ga}	+2	7.81	+7.9
	+1	8.23	+13.7
	0	8.64	+19.4
In_{As}	+2	9.49	+21.7
	+1	8.91	+14.2
	0	8.40	+7.6
	−1	8.06	+3.3
	−2	7.87	+0.8
Ga_{As}	+2	7.92	+1.5
	+1	7.63	−2.1
	0	7.35	−5.9
	−1	7.04	−9.6
	−2	6.75	−13.5

Table 4.2: Relaxation effects on defect volume for the defects studies in this work. Column 3 gives the relaxed defect volume (defined as the tetrahedron volume contained by the four nearest-neighbor atoms of the defect); column 4 gives the percentage of change in the defect volume (+ means expansion, − means contraction).

lattice constant a in the xy plane). The resulting structural parameters, a, η, u of CuAu-I type $In_{0.5}Ga_{0.5}As$, are determined by energy minimization using LDA DFT, as shown in Table 4.3. The calculated lattice constants and bond lengths agree with reference values within 2-3%.

As expected, the LDA band structure of the primitive cell of $In_{0.5}Ga_{0.5}As$ shows a direct band gap at Γ of only 0.41 eV, which underestimates the experimental value of 0.813±0.001 eV (at 4.2 K) [41] by about 50%. This large discrepancy reflects the well-known inability of LDA DFT to reproduce band gaps accurately. In contrast, our G_0W_0-derived band gap value, 0.84 eV, agrees well with the experimental result. Table 4.4 lists the energy of the valence-band maximum (VBM), conduction-band minimum (CBM), referenced to the DFT VBM, as well as the band gap for $In_{0.5}Ga_{0.5}As$ as calculated by LDA DFT and G_0W_0 method.

4.3.2 Point defects in $In_{0.5}Ga_{0.5}As$

Si substitutional defects

Effective incorporation of silicon as a dopant in InGaAs requires a change of bonding in the local crystal structure by replacing a native atom by a Si atom. Such substitutional Si defects are the most relevant ones to our study. Considering that the covalent radius of Si atom is smaller than that of In, Ga, and As (r_{Si} = 1.11Å, r_{In} = 1.42Å, r_{Ga} = 1.22Å, r_{As} = 1.19Å [45]), we would expect that Si could readily replace all three kinds of atoms in the crystal if size was the only defining metric. If a Si atom replaces a cation (In or Ga), it can donate an electron to the system, influencing the crystal in the direction of n-type doping. A

	this work	reference
a (Å)	5.71	5.85 [42]
η	1.01	1.005 [43]
u	0.268	N/A
In-As bond length (Å)	2.54	2.61 [44]
Ga-As bond length (Å)	2.42	2.47 [44]

Table 4.3: Structural parameters of bulk *ordered* CuAu-I type $In_{0.5}Ga_{0.5}As$ as calculated by LDA DFT. In the reference column, the lattice constant a and the bond lengths are experimental results of a *random* $In_{0.5}Ga_{0.5}As$ alloy, whereas η is a calculated value for CuAu-I-type *ordered* $In_{0.5}Ga_{0.5}As$.

Energy level (eV)	LDA	G_0W_0
VBM	0.0	−0.04
CBM	0.41	0.80
Band gap	0.41	0.84

Table 4.4: Energy levels of band edges and the band gap of $In_{0.5}Ga_{0.5}As$ as calculated from LDA DFT and a G_0W_0 approximation. The reference energy is set to be the value corresponding to the LDA VBM.

Si atom replacing an anion (As) accepts an electron (or donates a hole), thereby making the system more *p*-type. This amphoteric behavior of Si is not ideal for *n*-type doping since, for a sufficiently high concentration of dopants in the crystal, self-compensation of Si would prevent further incorporation of Si onto cation sites.

The atomic structure of Si substitutional defects does not vary from the corresponding local structure of the bulk InGaAs lattice. For pristine InGaAs, the slight lattice distortion of the zincblende structure results in a reduction of sym-

metry for the tetrahedral bonds surrounding As atoms, from perfect tetrahedral symmetry T_d to C_{2v}. On the other hand, the bonds surrounding In or Ga atoms form a higher symmetry D_{2d} since in this case, all four nearest-neighbor atoms are arsenic. We found that all relaxed structures containing Si substitional defects preserve the original symmetries at the defect site. The conservation of symmetry indicates that the bonding electron wave functions at the defect site share the same spatial distribution as in the bulk crystal. Upon relaxation, we observed that the bond lengths for all charge states decreased compared with the unrelaxed values. (Volumes for all relaxed and unrelaxed defects are given in Table 4.2.)

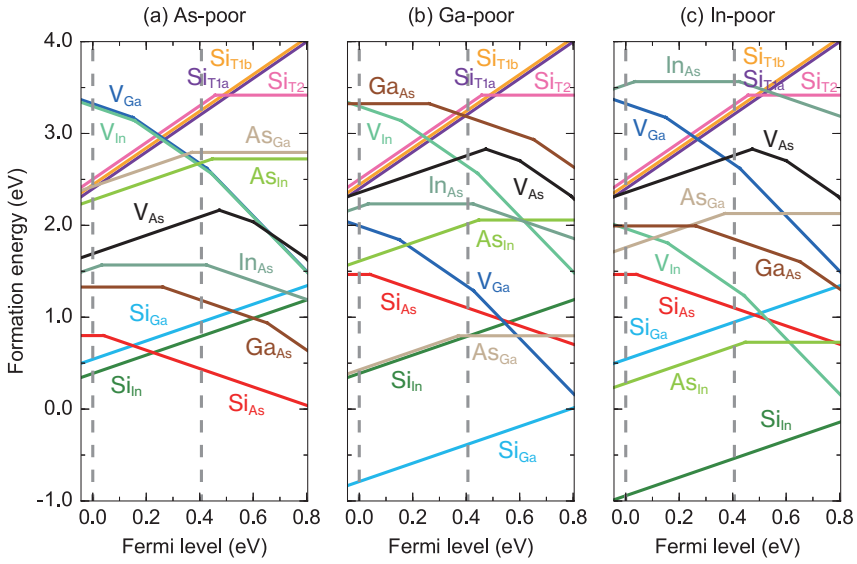

Figure 4.3: Formation energy of Si-induced and intrinsic defects in CuAu-I-type $In_{0.5}Ga_{0.5}As$ under different limiting growth conditions: (a) As-poor; (b) Ga-poor; (c) In-poor. Gray dashed lines indicate the VBM and CBM of the DFT band gap, while the solid vertical boundaries of the figure indicate the VBM and CBM of the *GW*-corrected band gap. The slope of each line segment equals the charge state of the defect.

The formation energies of Si substitutional defects are shown in Fig. 4.3,

under three limiting growth conditions (As-poor, Ga-poor, In-poor) outlined in Sec. 4.2.3. The slope of each line segment represents the charge state of the corresponding defect; and for any given Fermi level, only the most stable charge state for each defect is shown. The energies are extrapolated to within the *GW* band gap, using the scheme detailed by Lany and Zunger [13] (cf. Sec. II D). Note that the thermodynamic transition levels remain unchanged under different growth conditions due to the fact that the elemental chemical potentials, μ_i, do not enter the definition of the transition level $\varepsilon(q/q')$ (Eq.(2)). It is clear from Fig. 4.3 that both Si_{In} and Si_{Ga} are stable within the band gap only under charge state +1, indicating that both defects are shallow monovalent donors. Similarly, Si_{As} is stable within the gap mostly in the -1 state, with $\varepsilon(0/-1)$ being very close to the VBM, suggesting that it is a shallow monovalent acceptor. These characteristics agree with our intuition for the behavior of Si substitutionals in GaAs [46]. Quantitatively, the *GW-corrected* thermodynamic transition levels for Si_{In}, Si_{Ga} and Si_{As} between the stable charged state and the neutral state are $\varepsilon(+1/0) = E_C + 0.023$ eV, $\varepsilon(+1/0) = E_C + 0.026$ eV and $\varepsilon(0/-1) = E_V + 0.084$ eV, respectively.

Regarding the influence of growth conditions, we find that the self-compensation effect is strong under As-poor growth conditions; see Fig. 4.3(a). This is expected, as more Si atoms tend to occupy anion sites under As-poor conditions, and *vice versa*. Specifically, the Fermi level at which the most stable defect transfers from Si_{In} to Si_{As} occurs at $E_V + 0.28$ eV, which lies in the *p*-type doping regime. This suggests that under As-poor growth conditions, Si would behave more as an acceptor than as a donor, which would lead to *p*-type doping rather than the desired *n*-type. In contrast, Si exhibits much weaker self-compensation behavior under Ga-poor and In-poor growth conditions. (Fig.

4.3(b,c)) This is due to the fact that, while the formation energy of Si_{As} is higher compared to that under As-poor conditions, the formation energy of Si_{Ga} (and Si_{In}) is lower. Hence Si_{As} would compensate only the Si-cation substitutional defect with a higher energy, but not that with a lower energy, resulting in a strong tendency for Si to occupy cation sites and become a donor. It is especially encouraging that the formation energy of Si_{Ga} (and Si_{In}) remains below zero within the band gap under Ga (and In)-poor conditions, meaning that Si donors will form spontaneously across all doping regimes.

Si interstitial defects

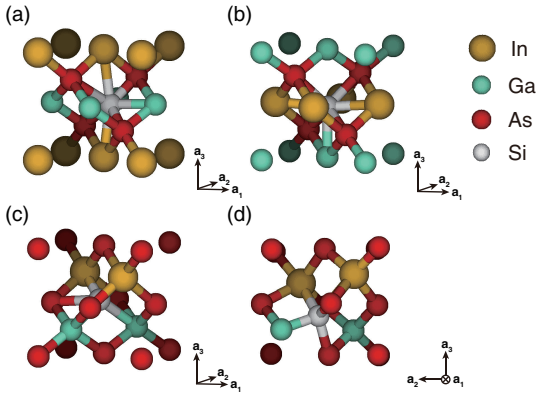

Figure 4.4: Atomic configuration of (a-c) unrelaxed Si tetrahedral interstitials Si_{T1a}, Si_{T1b}, and Si_{T2}, showing regular tetrahedral symmetry; (d) relaxed Si^{0}_{T2}, showing split interstitial-like distortion.

In addition to substitutional sites, Si atoms may also occupy interstitial sites in InGaAs. Although multiple possible interstitial sites exist, we focus on the three possible tetrahedral defects: Si_{T1a}, Si_{T1b} and Si_{T2}. The T1*a* and T1*b* interstitials are tetrahedrally coordinated with four As atoms, whereas the T2 interstitial is tetrahedrally coordinated with two In and two Ga atoms; see Fig. 4.4(a-c). The difference between Si_{T1a} and Si_{T1b} lies in the second-nearest neighbor atoms;

for Si_{T1a} these are two In and four Ga atoms, while for Si_{T1b} these are two Ga and four In atoms. The four tetrahedral bonds in the initial unrelaxed configuration satisfy D_{2d} symmetry for Si_{T1a} and Si_{T1b}, and C_{2v} symmetry for Si_{T2}. For relaxed defects, we observe a clear trend of expansion of these bonds (with a notable exception stated below). The respective symmetries are largely preserved for almost all *stable* charge states for each defect, with the only exception being the neutral state of Si_{T2}, in which case the Si atom is significantly deviated from its tetrahedral position; the fully relaxed configuration resembles a split-interstitial consisting of a Si and a Ga atom; see Fig. 4.4(d). This split configuration has a formation energy $\sim$ 0.7 eV lower than that of the tetrahedral configuration. A similar proclivity has been observed for dopants, especially boron, to form stable split interstitials in a silicon matrix under neutral state. [47] The reason behind this behavior may be that, without the symmetric bonding due to extra electrons, the repulsion from In atoms pushes the Si atom toward the Ga atoms, creating a highly asymmetric configuration. It is interesting to observe that this behavior is also observed for the more unlikely negative charge states (-1, -2) of Si_{T2}. On the other hand, the lack of such asymmetrical distortion for positive charge states of Si_{T2} probably indicates the presence of an energy barrier between the tetrahedral and the split-interstitial configuration, suggesting that interstitial diffusion of charged Si either from or to the T2 interstitial sites is unlikely to occur.

Regarding the electronic properties, type T1 and T2 Si-interstitials also exhibit qualitatively different behaviors. The T1 interstitials are both shallow donors, with +2 being the most stable charge state throughout the band gap; see Fig. 4.3. In terms of formation energy, the three +2-charged interstitials come very close, with the greatest difference between them being less than $\sim$ 0.1

eV. However, due to the above mentioned large relaxation effect of the T2 interstitial, the formation energy of the neutral Si_{T2} is considerably lower than those of the tetrahedral neutral T1 interstitials (by $\sim 0.6 - 0.8$ eV), resulting in a deep-donor-like and negative-U behavior ($\varepsilon(+2/0) = E_C - 0.35$ eV) for the T2 interstitial. In any case, the formation energies of all Si-interstitials are independent of growth conditions, as no addition / removal of In, Ga, and As atom is involved in forming these defects. Although Si-interstitials are all donor-like defects, their relatively high formation energies (> 2.0 eV) prevent them from being nearly as prevalent as Si-substitutionals, and hence their contribution to free electrons will be much smaller.

Vacancies

In CuAu-I type $In_{0.5}Ga_{0.5}As$, the nearest neighbors of a cation vacancy (V_{In}, V_{Ga}) are all As atoms, and the arrangement of second nearest neighbors (In and Ga) on the four sides is symmetric. In GaAs, it is observed from *ab initio* calculations that the single vacancy, V_{Ga}, adopts approximate T_d symmetry under all allowed charge states, ranging from -3 to $+1$ [48]. In our calculations, however, we observe different degrees of distortion from an ideal T_d symmetry for both V_{In} and V_{Ga} under charge states from -3 to 0. We find that this peculiar phenomenon is caused by the fact that neither an In atom nor a Ga atom lies at the center of a 64-atom or an 128-atom supercell. Hence, upon full atomic relaxation but no cell relaxation, the local structure of the vacancy assumes the asymmetry of *global* atomic arrangements surrounding the vacancy. This is in contrast to defects with an on-site atom (substitutionals, interstitials, and antisites), since the bonding interactions between the central atom and the nearest neighbor atoms

essentially balance out the forces exerted by atoms further away. But such bonding forces are absent for a vacancy. The observed deviation from T_d symmetry should therefore be seen as an artifact of the supercell scheme, not a physical feature of the vacancy itself. Indeed, in 128-atom supercells, the difference in the vacancy "bond lengths" decreases compared to those in 64-atom supercells. As a general trend, the nearest neighbor As atoms relax toward the vacancy's original site under all charge states, and the "bond lengths" decrease with lower charge states.

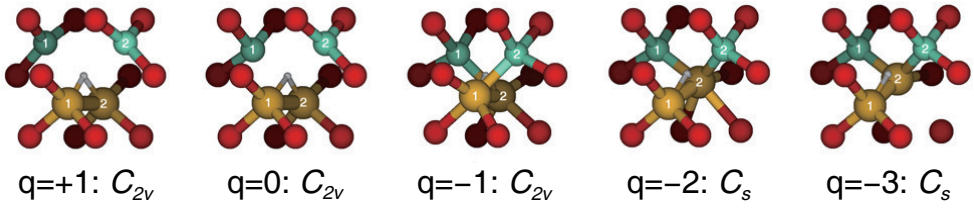

Figure 4.5: Atomic structures of a relaxed arsenic vacancy (V_{As}) in InGaAs in charge states $+1, 0, -1, -2, -3$, and their respective symmetry. The vacancy site is denoted by a small grey dot.

The atomic structure of an anion vacancy (V_{As}) shows a qualitatively different picture. Instead of the T_d-symmetric configuration of the cation vacancies, V_{As} exhibits large structure distortions. In GaAs, V_{As} assumes a C_{2v} symmetry in $+1$ charge state, a D_{2d} symmetry in 0 and -1 states, and becomes a split vacancy with $\sim C_{3v}$ symmetry for -2 and -3 states [48]. Each configuration stated above is found to be the lowest energy structure under the specified charge state. The multiple-symmetry nature of V_{As} suggests that it should be a deep trap state. In InGaAs, we find a very similar pattern of structural change for the same defect (Fig. 4.5). The nearest-neighbor atoms around V_{As} relax inward and assume a C_{2v} symmetry in $+1$, 0 and -1 charge states. For -2 and -3 states of V_{As}, the vacancy goes through a large structural distortion *via* moving one of the nearest neighbor atoms significantly towards the original vacancy site, effectively

creating a "split-vacancy". In GaAs, the other three As atoms relax about the same distance toward the vacancy, thereby creating the C_{3v} symmetry for the four As atoms. In InGaAs, however, the other three atoms do not belong to the same species. In this case, an In atom relaxes significantly towards the vacancy, making the In-In bond length much shorter than the Ga-Ga bond length. This asymmetric configuration eliminates the rotational symmetries altogether, keeping only the reflection symmetry. Therefore, the nearest-neighbor atoms of V_{As} assume a lower C_s symmetry.

The formation energies of vacancies are plotted in Fig. 4.3. V_{In} and V_{Ga} are both shallow acceptors, since only negative charge states are thermodynamically favorable within the gap. The transition levels for V_{In} are $\varepsilon(0/-1) = E_V - 0.056$ eV, $\varepsilon(-1/-2) = E_V + 0.16$ eV, and $\varepsilon(-2/-3) = E_V + 0.44$ eV; those for V_{Ga} are $\varepsilon(0/-1) = E_V - 0.057$ eV, $\varepsilon(-1/-2) = E_V + 0.15$ eV, and $\varepsilon(-2/-3) = E_V + 0.43$ eV. The closeness of the transition levels suggest the chemical similarity between the two defects. In contrast, the deep-trap character of V_{As} is evident from its transition levels. Charge states ranging from (+1) to (-3) can be accommodated within the band gap, making V_{As} one of the most complex defects surveyed in this work. The transition levels are $\varepsilon(+1/-1) = E_V + 0.52$ eV, $\varepsilon(-1/-2) = E_V + 0.64$ eV, and $\varepsilon(-2/-3) = E_V + 0.84$ eV ($E_C - 0.01$ eV). Note that according to our results, the defect exhibits one negative-U behavior at the $+1/-1$ transition, which agrees with predictions for V_{As} in GaAs, but lacks the other negative-$-U$ behavior predicted at the $-1/-3$ transition in GaAs [49, 48]. This disagreement could be explained by the bistability of the V_{As} [50]; namely, the structural distortion stabilizes the defect in the corresponding charge state (-2 or -3), thereby making the "metastable ground state" with the undistorted structure slightly less energetically favorable.

As with the substitutional defects, the formation energies of vacancies are heavily influenced by the growth condition. Under In (Ga)-poor conditions, In (Ga) vacancies dominate the defect population under n-type doping regimes, compensating the substitutional Si_{In} (Si_{Ga}) at E_C + 0.073 eV (E_C + 0.036 eV); on the other hand, V_{As} has a relatively high formation energy, rendering its compensation effect negligible. On the other hand, under As-poor conditions, cation Si-substitutionals are compensated by all three types of vacancies, but none of them are significant compared with the self-compensation from Si_{As}.

Antisites

Antisites are a class of well-studied defects in GaAs. Experimentally, the presence of the so-called *EL2* defect, a deep defect responsible for the semi-insulating character of GaAs under certain growth conditions, is known to be closely related to the anion antisite As_{Ga}. Opinions in the literature are divided as to the exact composition of the *EL2* defect; while some claim that *EL2* is simply the single antisite As_{Ga} [48, 51, 52], others have proposed more complex models, such as As_{Ga} + V_{Ga} [53] or As_{Ga} + As_i [54]. Despite the inconclusiveness regarding the atomic structure of *EL2*, evidence suggests that As_{Ga} should be a deep donor in GaAs [55]. The other antisite defect in GaAs, Ga_{As}, is less experimentally investigated, but is predicted to be a deep amphoteric defect by theoretical calculations [48, 56, 57].

In GaAs, both As_{Ga} and Ga_{As} are predicted to preserve the T_d symmetry upon relaxation to the lowest energy state [48]. This behavior is also observed in our calculation for all antisites under all possible charge states in InGaAs, as none of the antisites display any significant Jahn-Teller distortion. Specifi-

cally, cation antisites (In_{As}, Ga_{As}) preserve C_{2v} symmetry, while anion antisites (As_{In}, As_{Ga}) preserve D_{2d} symmetry, as do the Si substitutionals on the corresponding lattice sites.

Formation energies of antisite defects are shown in Fig. 4.3. We find that, just as has been experimentally verified for As_{Ga} in GaAs, both anion antisites in InGaAs exhibit deep-donor-like behavior in the *GW* band gap. Both defects share two possible stable charge states $(+1, 0)$ within the gap, and the transition levels of the two defects are very close: ($\varepsilon(+1/0) = E_C - 0.36$ eV for As_{In} and $E_C - 0.43$ eV for As_{Ga}). As for cation antisites, besides the positive and the neutral charge states, these two defects can also be stable within the gap under negative charge states ($-1, -2$ for In_{As}, -1 for Ga_{As}), making them behave as deep traps. The transition levels of In_{As} occur at $\varepsilon(+1/0) = E_V + 0.077$ eV and $\varepsilon(0/-1) = E_V + 0.47$ eV, and those of Ga_{As} occur at $\varepsilon(0/-1) = E_V + 0.31$ eV and $\varepsilon(-1/-2) = E_V + 0.70$ eV. Note that except for the -2 state of Ga_{As}, none of the formal charge states of the other antisites become stable within the band gap.

The formation energies of antisites sensitively depend on the growth condition. Under As-poor conditions (Fig. 4.5(a)), anion antisites are unlikely to form, therefore their formation energies are significantly greater, especially under n-type doping. On the other hand, cation antisites are much more likely to form, as evidenced by their low formation energy. In particular, under As-poor conditions, Ga_{As} is the most stable intrinsic defect, compensating Si_{In} at $E_V + 0.65$ eV and Si_{Ga} at $E_V + 0.57$ eV. This is another reason why an As-poor growth condition is extremely unfavorable for n-type doping. In contrast, the formation energy of Ga_{As} (In_{As}) is raised to among the highest of all defects under Ga(In)-poor conditions. Hence for n-type doping, the only charge-compensating single

point defect is the corresponding vacancy.

Dependence of free carrier concentration on growth conditions

Under thermal equilibrium, doped semiconductors satisfy the charge neutrality condition (Eq. (4.2)). Using the formalism in Sec. 4.2.5, we obtain the maximum net carrier concentration (*e.g.*, net carrier concentration at the Si solubility limit) under thermal equilibrium as a function of annealing temperature (Fig. 4.6). The results agree well with the knowledge we obtained from formation energy calculations (Fig. 4.3). For As-poor growth conditions, the charge compensation of donor (Si_{In}) by acceptor (Si_{As}) occurs in the p-type regime at $E_F = E_V + 0.28$ eV (Fig. 4.3(a)). Indeed, Fig. 4.6 shows that the equilibrium Fermi level at temperatures from 500 K to 1200 K is pinned at 0.2-0.3 eV above the VBM. The same correspondence can also be seen for Ga-poor and In-poor growth conditions, where the Fermi level pinning occurs near the conduction band under both circumstances.

In contrast to the slightly varying equilibrium Fermi level at various temperatures, the equilibrium net carrier concentration varies significantly with temperature, due to its exponential sensitivity to the Fermi level. Under As-poor conditions, as the equilibrium Fermi level lies in p-type regime, holes become the majority carrier and the electrons the minority carrier, whereas under Ga and In-poor conditions, electrons become the majority carriers. We find that the maximum net carrier concentration increases with annealing temperature under all growth conditions. The net carrier concentration under As-poor conditions ($p - n$) can reach 1.7×10^{19} cm^{-3} at 1200 K, while the net carrier concentration under Ga- and In-poor conditions ($n - p$) can only reach 3.7×10^{18}

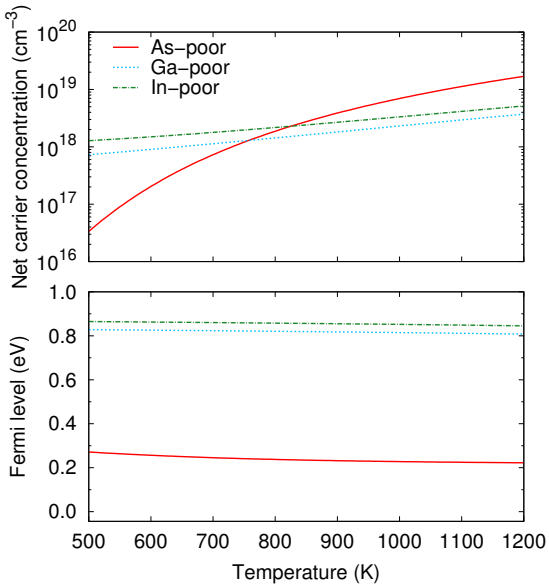

Figure 4.6: Top: Equilibrium net carrier concentration of Si-doped $In_{0.5}Ga_{0.5}As$ containing all defects studied in this work as a function of annealing temperature, under various limits of growth conditions. Bottom: Equilibrium Fermi level of Si-doped $In_{0.5}Ga_{0.5}As$ as a function of annealing temperature.

cm^{-3} and 5.2×10^{18} cm^{-3}, respectively, at the same temperature. This behavior can be explained by the fact that the effective density of states of the valence band of $In_{0.5}Ga_{0.5}As$ is much greater than the effective density of states of the conduction band, making p-type doping easier than n-type doping for InGaAs under their respective optimal growth conditions. Our predicted maximum net electron concentration under In-poor conditions agrees with the experimentally observed "thermodynamic limit," 1.4×10^{19} cm^{-3} [58], within a factor of about three, suggesting that our equilibrium approach is quite accurate. On the other hand, using non-equilibrium growth and annealing techniques may be able to achieve majority carrier concentration and dopant activation beyond this limit (for example, cf. [59]).

Fig. 4.7 shows the trend of dopant activation as a function of dopant concen-

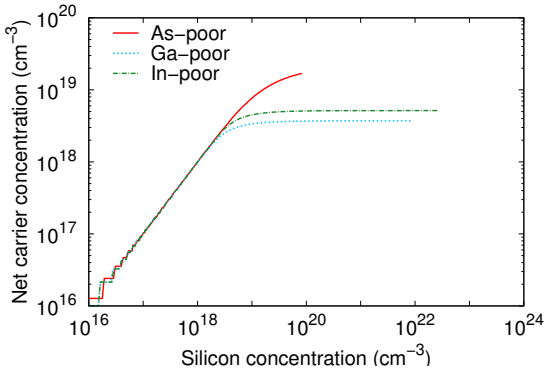

Figure 4.7: Net free carrier concentration in Si-doped InGaAs as a function of Si concentration under thermal equilibrium at T = 1200K.

tration at 1200 K. When the total Si concentration in InGaAs is below a certain threshold value ($\sim 5 \times 10^{18}$ cm^{-3} for In- and Ga-poor conditions, $\sim 1 \times 10^{19}$ cm^{-3} for As-poor condition), the net carrier concentration is equal to the Si concentration; this means that all incorporated dopants are ionized and contributing to the conductivity. After the Si concentration accumulates beyond this threshold, however, the dopant activation rate quickly starts to decrease, causing the net free carrier concentration to saturate. This suggests that the concentration of charge-compensating defects only becomes significant after the Si concentration reaches this threshold. This is confirmed in Fig. 4.8, which shows that the concentration of the dominant charge-compensating acceptor (V_{Ga} under Ga-poor conditions and V_{In} under Ga-poor conditions) rises to comparable levels to that of the primary Si donor after the total Si concentration reaches the threshold. Fig. 4.8 also shows the most abundant species of single point defects in InGaAs and their respective concentrations under cation-poor conditions; the trends agree very well with the formation energies plotted in Fig. 4.3. In addition, Fig. 4.7 suggests that additional Si concentration beyond the threshold level does not contribute more free carriers.

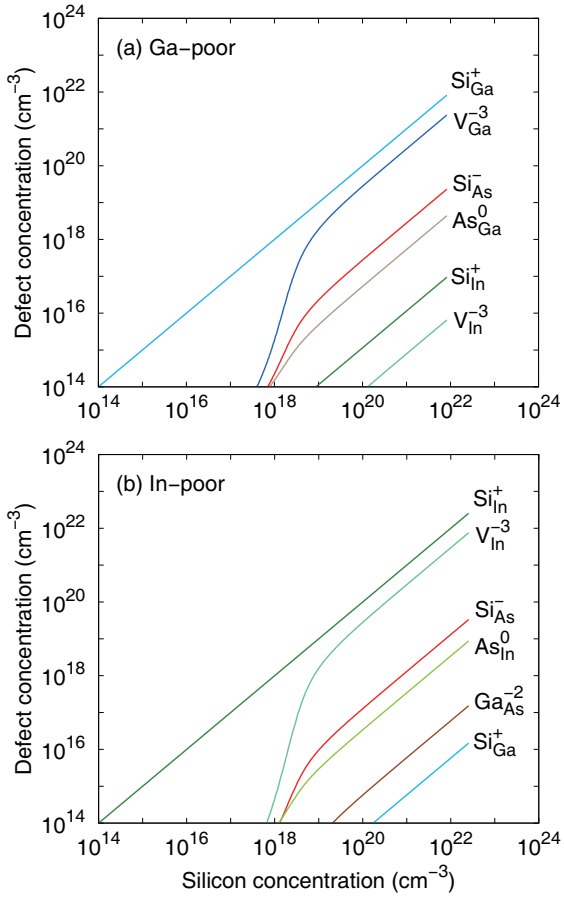

Figure 4.8: Concentration of dominant species of single point defects in Si-doped InGaAs as a function of Si concentration under thermal equilibrium at T = 1200K.

Based on the defect concentrations in Fig. 4.8, it is clear that most of the Si atoms still form donors on cation sites, instead of becoming acceptors on anion sites or interstitials. Hence these additional Si dopant atoms should get deactivated primarily by forming doubly negatively charged donor-vacancy pairs $(Si_{III} - V_{III})^{-2}$ (III = In/Ga, depending on growth conditions) with the dominant triply-charged cation vacancies. This is confirmed in Fig. 4.9, which shows that, under In-poor growth conditions, $(Si_{In} - V_{In})^{-2}$ does indeed have a lower formation energy than Si^{+}_{In} for a Fermi level $E_F > 0.68$eV. Note that the formation energy of $(Si_{In} - V_{In})^{-2}$ is much smaller than that of the V_{In} alone. In addition, the

formation energy of $(Si_{In} - Si_{As})^0$ is also much smaller than that of Si_{As} within the band gap (-0.03 eV). These suggest that both defect pairs have a small binding energy, making them easy to form from single point defects. Our conclusion strongly corroborates the statement by Lind *et al.* [60] that Si activation in heavily-doped n-type InGaAs is not limited by chemical solubility or auto-compensation, but by defect-complex formation with charged cation vacancies.

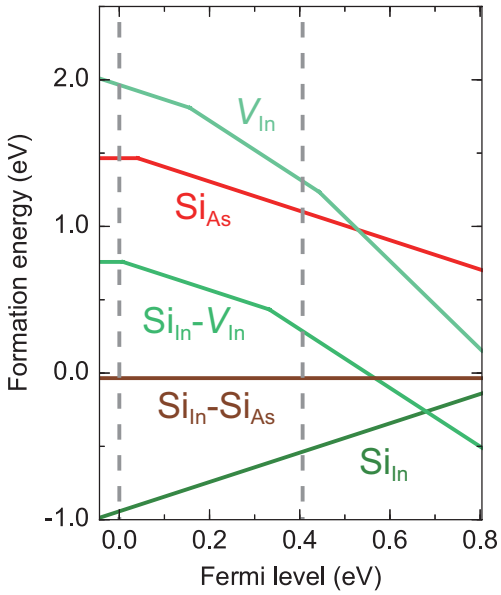

Figure 4.9: Formation energies of dominant species of single point defects and defect pairs in CuAu-I-type $In_{0.5}Ga_{0.5}As$, under In-poor growth conditions.

4.4 Conclusions

In this paper, we have performed a comprehensive set of *ab initio* calculations of formation energies of Si-induced and intrinsic single point defects in an ordered (CuAu-I-type) bulk $In_{0.5}Ga_{0.5}As$. We have used the purely *ab initio* G_0W_0

method to correct the band gap underestimated by LDA DFT, thereby obtaining excellent agreement with experimental band gap, and reducing the error in defect transition levels due to an incorrect band gap. We find that Si substitutional defects, Si_{T1} interstitials, and cation vacancies are shallow defects, while Si_{T2} interstitial, arsenic vacancy (V_{As}), and all antisite defects are deep. We observe that under all growth conditions, Si-interstitials remain difficult to form under n-type doping, while the formation energies of other defects vary significantly under different growth conditions. As-poor conditions are extremely unfavorable for n-type doping, as both Si_{As} and Ga_{As} participate in the charge compensation process, and the Fermi level is pinned near valence band maximum. In-poor and Ga-poor conditions are much more beneficial for n-type doping, since the formation energy of the most stable donor (Si_{In} and Si_{Ga}, respectively) is considerably lowered compared with under As-poor conditions, and the charge compensation effect is significantly reduced. However, in this case the corresponding cation vacancy (V_{In}^{-3} and V_{Ga}^{-3}) has a formation energy comparable to that of the Si donor. In addition, cation vacancies tend to form pairs with the Si donor due to their small binding energy, which compensates the donor and inhibits the n-type doping, thereby limiting the maximum net electron concentration in InGaAs. We have predicted the maximum net carrier concentration and Fermi level under thermal equilibrium as a function of annealing temperature, and found good agreement with experimental results. We suggest that non-equilibrium annealing process would be an effective way to increase dopant activation.

Bibliography

[1] Roger J Malik. *III-V Semiconductor Materials and Devices*. Elsevier, 2012.

[2] Yoshikazu Takeda, Akio Sasaki, Yujiro Imamura, and Toshinori Takagi. Electron mobility and energy gap of in0.53ga0.47as on inp substrate. *J. Appl. Phys.*, 47(12):5405–5408, 1976.

[3] Jesús A Del Alamo. Nanometre-scale electronics with iii-v compound semiconductors. *Nature*, 479(7373):317–323, 2011.

[4] ITRS 2013 Modeling and Simulation Difficult Challenges. `http://www.itrs.net/ITRS%201999-2014%20Mtgs,%20Presentations%20&%20Links/2013ITRS/2013TableSummaries/2013Modeling_SummaryTable.pdf`, 2013.

[5] Ashish Baraskar, AC Gossard, and Mark JW Rodwell. Lower limits to metal-semiconductor contact resistance: Theoretical models and experimental data. *J. Appl. Phys.*, 114(15):154516, 2013.

[6] Aaron G Lind, Henry L Aldridge Jr, Cory C Bomberger, Christopher Hatem, Joshua MO Zide, and Kevin S Jones. Comparison of thermal annealing effects on electrical activation of mbe grown and ion implant si-doped in0.53ga0.47as. *J. Vac. Sci. Technol., B*, 33(2):021206, 2015.

[7] Hannu-Pekka Komsa and Alfredo Pasquarello. Intrinsic defects in gaas and ingaas through hybrid functional calculations. *Physica B*, 407(15):2833–2837, 2012.

[8] Hannu-Pekka Komsa and Alfredo Pasquarello. Comparison of vacancy and antisite defects in gaas and ingaas through hybrid functionals. *J. Phys.: Condens. Matter*, 24(4):045801, 2012.

[9] SR Lee, AF Wright, NA Modine, CC Battaile, SM Foiles, JC Thomas, and A Van der Ven. First-principles survey of the structure, formation energies, and transition levels of as-interstitial defects in ingaas. *Phys. Rev. B*, 92(4):045205, 2015.

[10] Cheng-Wei Lee, Binit Lukose, Michael O Thompson, and Paulette Clancy. Energetics of neutral si dopants in ingaas: An ab initio and semiempirical tersoff model study. *Phys. Rev. B*, 91(9):094108, 2015.

[11] Henry Aldridge Jr, Aaron G Lind, Cory C Bomberger, Yevgeniy Puzyrev, Christopher Hatem, Russell M Gwilliam, Joshua MO Zide, Sokrates T Pantelides, Mark E Law, and Kevin S Jones. Implantation and diffusion of silicon marker layers in in0. 53ga0. 47as. *J. Electron. Mater.*, 45:4282, 2016.

[12] Henry Aldridge, Aaron G Lind, Cory C Bomberger, Yevgeniy Puzyrev, Joshua MO Zide, Sokrates T Pantelides, Mark E Law, and Kevin S Jones. N-type doping strategies for ingaas. *Mater. Sci. Semicond. Process.*, 57:39–47, 2017.

[13] 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. *Phys. Rev. B*, 78(23):235104, 2008.

[14] Christoph Freysoldt, Jörg Neugebauer, and Chris G Van de Walle. Fully ab initio finite-size corrections for charged-defect supercell calculations. *Phys. Rev. Lett.*, 102(1):016402, 2009.

[15] Chris G Van de Walle and Jörg Neugebauer. First-principles calculations for defects and impurities: Applications to iii-nitrides. *J. Appl. Phys.*, 95(8):3851–3879, 2004.

[16] TS Kuan, WI Wang, and EL Wilkie. Long-range order in inxga1-xas. *Appl. Phys. Lett.*, 51(1):51–53, 1987.

[17] MA Shahid, S Mahajan, DE Laughlin, and HM Cox. Atomic ordering in ga 0.47 in 0.53 as and ga x in 1- x as y p 1- y alloy semiconductors. *Phys. Rev. Lett.*, 58(24):2567, 1987.

[18] Akiko Gomyo, Tohru Suzuki, and Sumio Iijima. Observation of strong ordering in ga x in 1- x p alloy semiconductors. *Phys. Rev. Lett.*, 60(25):2645, 1988.

[19] Osamu Ueda, Masahiko Takikawa, Junji Komeno, and Itsuo Umebu. Atomic structure of ordered ingap crystals grown on (001) gaas substrates by metalorganic chemical vapor deposition. *Jpn. J. Appl. Phys.*, 26(11A):L1824, 1987.

[20] Masaya Ichimura and Akio Sasaki. Short-range order in iii-v ternary alloy semiconductors. *J. Appl. Phys.*, 60(11):3850–3855, 1986.

[21] Angelo Mascarenhas. *Spontaneous Ordering in Semiconductor Alloys*. Springer Science & Business Media, 2012.

[22] John P Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B*, 33(12):8822, 1986.

[23] C Hartwigsen, Sephen Gœdecker, and Jürg Hutter. Relativistic separable dual-space gaussian pseudopotentials from h to rn. *Phys. Rev. B*, 58(7):3641, 1998.

[24] S Goedecker, M Teter, and Jürg Hutter. Separable dual-space gaussian pseudopotentials. *Phys. Rev. B*, 54(3):1703, 1996.

[25] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source

software project for quantum simulations of materials. *J. Phys.: Condens. Matter*, 21(39):395502, 2009.

[26] CWM Castleton, A Höglund, and Susanne Mirbt. Density functional theory calculations of defect energies using supercells. *Modell. Simul. Mater. Sci. Eng.*, 17(8):084003, 2009.

[27] Christoph Freysoldt, Jörg Neugebauer, and Chris G Van de Walle. Electrostatic interactions between charged defects in supercells. *Phys. Status Solidi B*, 248(5):1067–1076, 2011.

[28] Ferdi Aryasetiawan and Olle Gunnarsson. The gw method. *Rep. Prog. Phys.*, 61(3):237, 1998.

[29] Andrea Marini, Conor Hogan, Myrta Grüning, and Daniele Varsano. Yambo: an ab initio tool for excited state calculations. *Comput. Phys. Commun.*, 180(8):1392–1403, 2009.

[30] PC Donohue, WJ Siemons, and JL Gillson. Preparation and properties of pyrite-type sip 2 and sias 2. *J. Phys. Chem. Solids*, 29(5):807–813, 1968.

[31] Aparna Chakrabarti, Peter Kratzer, and Matthias Scheffler. Surface reconstructions and atomic ordering in in x ga 1- x as (001) films: A density-functional theory study. *Phys. Rev. B*, 74(24):245328, 2006.

[32] Mohamed Tmar, Armand Gabriel, Christian Chatillon, and Ibrahim Ansara. Critical analysis and optimization of the thermodynamic properties and phase diagrams of the iii–v compounds ii. the ga-as and in-as systems. *J. Cryst. Growth*, 69(2):421–441, 1984.

[33] DK Biegelsen, RD Bringans, JE Northrup, MC Schabel, and L-E Swartz. Arsenic termination of the si (110) surface. *Phys. Rev. B*, 47(15):9589, 1993.

[34] Christoph Freysoldt, Blazej Grabowski, Tilmann Hickel, Jörg Neugebauer, Georg Kresse, Anderson Janotti, and Chris G Van de Walle. First-principles calculations for point defects in solids. *Rev. Mod. Phys.*, 86(1):253, 2014.

[35] TS Moss. The interpretation of the properties of indium antimonide. *Proc. Phys. Soc. B*, 67(10):775, 1954.

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

[37] G Makov and MC Payne. Periodic boundary conditions in ab initio calculations. *Phys. Rev. B*, 51(7):4014, 1995.

[38] Christopher WM Castleton, Andreas Höglund, and Susanne Mirbt. Managing the supercell approximation for charged defects in semiconductors: Finite-size scaling, charge correction factors, the band-gap problem, and the ab initio dielectric constant. *Phys. Rev. B*, 73(3):035215, 2006.

[39] Hannu-Pekka Komsa, Tapio T Rantala, and Alfredo Pasquarello. Finite-size supercell correction schemes for charged defect calculations. *Phys. Rev. B*, 86(4):045112, 2012.

[40] AA Mbaye, DM Wood, and Alex Zunger. Stability of bulk and pseudomorphic epitaxial semiconductors and their alloys. *Phys. Rev. B*, 37(6):3008, 1988.

[41] K Alavi, RL Aggarwal, and SH Groves. Interband magnetoabsorption of $in_{0.53}ga_{0.47}as$. *Phys. Rev. B*, 21(3):1311, 1980.

[42] JC Woolley and BA Smith. Solid solution in the gaas-inas system. *Proc. Phys. Soc. B*, 70(1):153, 1957.

[43] Dan Teng, Jun Shen, Kathie E Newman, and Bing-Lin Gu. Effects of ordering on the band structure of iii–v semiconductors. *J. Phys. Chem. Solids*, 52(9):1109–1128, 1991.

[44] JC Mikkelsen Jr and JB Boyce. Atomic-scale structure of random solid solutions: Extended x-ray-absorption fine-structure study of ga 1- x in x as. *Phys. Rev. Lett.*, 49(19):1412, 1982.

[45] Beatriz Cordero, Verónica Gómez, Ana E Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. *Dalton Trans.*, (21):2832–2838, 2008.

[46] John E Northrup and SB Zhang. Dopant and defect energetics: Si in gaas. *Phys. Rev. B*, 47(11):6791, 1993.

[47] GD Watkins. Defects in irradiated silicon: Epr and electron-nuclear double resonance of interstitial boron. *Phys. Rev. B*, 12(12):5824, 1975.

[48] Peter A Schultz and O Anatole Von Lilienfeld. Simple intrinsic defects in gallium arsenide. *Modell. Simul. Mater. Sci. Eng.*, 17(8):084007, 2009.

[49] Fedwa El-Mellouhi and Normand Mousseau. Self-vacancies in gallium arsenide: An ab initio calculation. *Phys. Rev. B*, 71(12):125207, 2005.

[50] GA Baraff and M Schluter. Bistability and metastability of the gallium vacancy in gaas: the actuator of el 2? *Phys. Rev. Lett.*, 55(21):2340, 1985.

[51] DJ Chadi and KJ Chang. Metastability of the isolated arsenic-antisite defect in gaas. *Phys. Rev. Lett.*, 60(21):2187, 1988.

[52] Jaroslaw Dabrowski and Matthias Scheffler. Theoretical evidence for an optically inducible structural transition of the isolated as antisite in gaas: Identification and explanation of el 2? *Phys. Rev. Lett.*, 60(21):2183, 1988.

[53] Zhongji Zou and Yuanxi Zou. Identification of two bands in the pl spectra of si lec gaas on the basis of a strain model. *Mater. Lett.*, 4(5):286–289, 1986.

[54] HJ Von Bardeleben, D Stievenard, D Deresmes, A Huber, and JC Bourgoin. Identification of a defect in a semiconductor: El2 in gaas. *Phys. Rev. B*, 34(10):7192, 1986.

[55] V Ortiz, J Nagle, J-F Lampin, E Péronne, and Antigoni Alexandrou. Low-temperature-grown gaas: Modeling of transient reflectivity experiments. *J. Appl. Phys.*, 102(4):043515, 2007.

[56] SB Zhang and DJ Chadi. Cation antisite defects and antisite-interstitial complexes in gallium arsenide. *Phys. Rev. Lett.*, 64(15):1789, 1990.

[57] Hyangsuk Seong and Laurent J Lewis. Tight-binding molecular-dynamics study of point defects in gaas. *Phys. Rev. B*, 52(8):5675, 1995.

[58] Aaron Gregg Lind, Henry Lee Aldridge, Cory Carl Bomberger, Chris Hatem, Joshua MO Zide, and Kevin Scott Jones. Annealing effects on the electrical activation of si dopants in ingaas. *ECS Trans.*, 66(7):23–27, 2015.

[59] T Fujii, T Inata, K Ishii, and S Hiyamizu. Heavily si-doped ingaas lattice-matched to inp grown by mbe. *Electron. Lett.*, 22:191, 1986.

[60] AG Lind, HL Aldridge, C Hatem, ME Law, and KS Jones. Review—dopant selection considerations and equilibrium thermal processing limits for n+-in0. 53ga0. 47as. *ECS J. Solid State Sci. Technol.*, 5(5):Q125–Q131, 2016.

CHAPTER 5

FINGERPRINTING THE VIBRATIONAL SIGNATURES OF DOPANTS AND DEFECTS IN A FULLY RANDOM ALLOY: AN *AB INITIO* CASE STUDY OF SI, SE, AND VACANCIES IN $IN_{0.5}GA_{0.5}AS$

The work presented in this chapter is published in: H. Jia, **J. Wang**, and P. Clancy, submitted to Physical Review B (2019). Part of the work is reproduced with permission from Haili Jia.

5.1 Introduction

Dopants and defects play a major role in altering electronic and optical properties of semiconductors. In many scenarios, dopant activation of semiconductors is limited either by the presence of charge-compensating defects, or the occupation of dopants on lattice sites that generate minority charge carriers ("amphoteric doping"). Insufficient dopant activation is often a bottleneck for achieving higher levels of device performance; in such situations, obtaining a detailed picture of dopant and defect distribution in the crystal becomes critically important. Experimentally, spectroscopic methods such as infrared (IR) spectroscopy and Raman spectroscopy are conventionally deployed for this purpose [1, 2]. By directly probing the vibrational modes of atoms in the crystal, these methods are capable of identifying distinct local bonding configurations of dopant atoms and defects from the frequencies and characteristics of corresponding vibrational modes (known as "vibrational fingerprints"). Due to the high accuracy of these methods, many experimental analyses utilizing vibrational spectroscopy have achieved great success in identifying certain dopant and defect lattice positions in binary III-V semiconductors. [3, 4, 5, 6, 7, 8, 9]

In a typical semiconductor alloy, the long-range ordering of constituent species on their common sublattice is usually not preserved, replaced by a short-ranged or near-random fluctuation of local atomic environments. [10, 11, 12] Such variations in local atomic configuration induce a variety of vibrational modes at different frequencies for a substitutional dopant or defect in the alloy, making it difficult to accurately assign dopant and defect lattice positions to a given mode.

Despite the multitude of experimental and computational studies of dopant- and defect-induced vibrational modes in pure semiconductor compounds, to date there have been limited number of studies on the dopant- or defect-induced vibrational modes in *near-random* semiconductor alloys. One prominent example is the resolution of preferred local bonding for nitrogen impurities in high In/Ga content InGaAs alloys. Experimental studies using Fourier transform infrared (FTIR) spectroscopy [13, 14, 15] and Raman spectroscopy [16, 17, 18] have measured the local vibrational modes of nitrogen in InGaAs; however, due to random variations in the local environment, these experimental studies provide conflicting and ambiguous results. On the computational side, numerical simulations have attempted to address the same challenge [19] using approximate schemes such as empirical lattice dynamical models. These models require *a priori* calibration with experimental data and thus are system-dependent and lack generality and transferability. Besides, the Abell-Tersoff semi-empirical model predictions of the extrinsic "substitution energy" of a Si dopant on a cationic lattice site in InGaAs were found to be strongly dependent on the strain induced by a local arrangement of In and Ga cationic atoms [22]. Moreover, current state-of-the-art *ab initio* density functional theory (DFT) codes often make extensive use of point group and space group symmetries for

treating phonons in crystals,[20] which are not applicable to highly disordered random alloys. Several *ab initio* computational studies of defects in alloys have been performed. [21, 23] However, these studies only considered variations among chemical species of nearest-neighbor atoms, while ignoring the randomness of atomic arrangements in the larger host crystal. They also ignored any influence of dopants and defects on lattice structure that extends over several shells of neighbors. A more rigorous theoretical treatment must take these crucial aspects into account, while maintaining a feasible computational cost.

In this work, we present a systematic approach to modeling the vibrational modes of impurity atoms in a random alloy. For our case study, we choose two prototypical *n*-type dopants, silicon and selenium, and one prevalent compensating defect, a cation vacancy, in a random InGaAs alloy. Specifically, we use special quasirandom structures (SQS) to model a random alloy within the framework of DFT, and real space Green's function to calculate local phonon density of states of a dopant or defect within an extended supercell that is beyond the capabilities of DFT calculation. This approach allows us to accurately determine the local vibrational modes of defects in a random alloy directly from DFT calculations, without any empirical parameters or extensive computational cost. Specifically, our method reveals the key qualitative difference between dopant and defect modes in *random* semiconductor alloys and in pure semiconductor compounds, namely the formation of satellite peaks across a continuous range of frequencies.[19] We will discuss the methodology in detail in Sec. 5.2 and provide results in Sec. 5.3.

5.2 Methodology

5.2.1 Structural model of random $In_{0.5}Ga_{0.5}As$

For density functional theory (DFT) calculations, we use 216-atom cubic supercells of $In_{0.5}Ga_{0.5}As$ containing one impurity (substitutional silicon on cation site (Si_{III}), substitutional selenium on cation site (Se_{III}), or cation vacancy (V_{III})). This particular size is chosen, as we verified that if the impurity is placed at the center of the supercell, the impurity strain field is almost fully contained within the cell boundaries (above a small force threshold of 0.03 eV/Å) for all defect/dopant species in this work. (details see Sec. 5.3.2) This is crucial since the finite-size effect of elastic interaction between neighboring images of impurities need to be reduced as much as possible.

The random arrangement of the cation species is modeled within the formalism of special quasirandom structures. [37] We used an in-house code to generate candidate structures with minimizing the sum of cluster functions for six clusters: pairs with first, second, third, and fourth nearest neighbor separations on the cation sublattice, as well as triplet and quadruplets with first nearest-neighbor separation on the cation sublattice. We use simulated annealing [38] to find the optimal 216-atom SQS model such that the sum of cluster functions lies below a chosen threshold of 10^{-2}. Fig. 5.1 shows the SQS model of random $In_{0.5}Ga_{0.5}As$ we use in all subsequent calculations and analysis.

We use the plane-wave density functional code QUANTUM ESPRESSO for all calculations in this work. We apply PBEsol functional with ultrasoft pseudopotentials, with an energy cutoff of 50 Rydberg and a density cutoff of 400

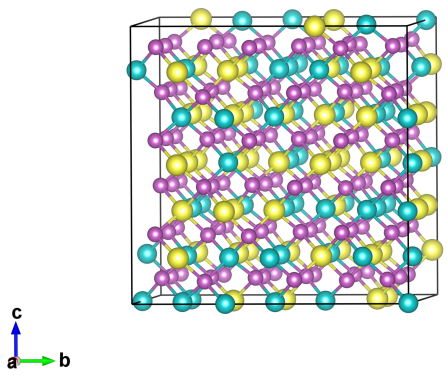

Figure 5.1: The 216-atom cubic special quasirandom structure (SQS) model of random $In_{0.5}Ga_{0.5}As$ used in this work, where yellow, blue, and magenta spheres represent indium (In), gallium (Ga), and arsenic (As) atoms respectively.

Rydberg. Due to the large number of expensive calculations, We use a single **k**-point $\frac{2\pi}{a}(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ in order to minimize computational cost while retaining computational accuracy, as it is shown to be the most accurate single **k**-point sampling for simple cubic systems [39]. We have verified that upon atomic relaxation, this setting yields convergence of atomic forces to within 5 meV/Å. We use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [40] for atomic and cell relaxation while keeping the cubic symmetry of the cell, with a force criterion of 10^{-3} Rydberg/Bohr and a stress criterion of 0.5 kbar. The relaxed supercell of our $In_{0.5}Ga_{0.5}As$ SQS assumes an equal side length of 16.96Å.

5.2.2 Dynamic matrix

To calculate interatomic force constants required for analyzing phonon density of states, we apply frozen phonon method [41] by displacing each atom within the impurity strain field by ±0.05Å relative to its equilibrium position in each of the Cartesian directions, with the impurity placed at the center of the supercell. The thus calculated force constants are then used to construct the corresponding entries of the much larger 3000×3000 force constant matrix for the 1000-atom effective supercell of defective random $In_{0.5}Ga_{0.5}As$, which is in turn used in calculation of atom-projected local phonon density of states using the real-space Green's function formalism. (details see Sec. 5.2.3)

A simplified picture that illustrates our formulation is shown in Fig. 5.2. We devide all the atoms in the 216-atom defective supercell into region 1 and region 2. Specifically, region 1 is designated to consist of atoms that are close enough to the impurity, such that their interatomic interactions are strongly influenced by the impurity; region 2 consists of the rest of atoms whose interatomic interactions remain largely unaffected by the impurity. In our calculations, region 1 consists of all atoms within the strain field of the impurity as determined by DFT. (details see Sec. 5.3.1) We then embed this 216-atom supercell at the center of a larger, 1000-atom cubic *effective* supercell of random $In_{0.5}Ga_{0.5}As$ to construct the total dynamic matrix. We denote all the atoms outside of the 216-atom supercell as region 3. The reason to use a larger effective supercell is that, the size of the 216-atom supercell imposes a cutoff on the range of the interatomic force constants, which is insufficient for real-space Green's function to yield a converged local phonon density of states. The total dynamic matrix, $\mathbf{D}$, of the 1000-atom effective supercell can then be formally expressed in the following

form:

$$\mathbf{D} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{H}_2 & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{H}_3 \end{bmatrix}. \tag{5.1}$$

In this representation, $\mathbf{H}_i$ are *intra*-region dynamic submatrices, describing interactions between atom pairs both belonging to the same region; $\mathbf{V}_{ij} = \mathbf{V}_{ji}^T$ are the *inter*-region dynamic submatrices, describing interactions between atom pairs belonging separately to two regions i and j. Since from our previous 216-atom DFT calculations, we have already calculated the force constants for all pairs of atoms with at least one atom lying within the strain field of the impurity (region 1, designated as "inner region"), we have determined all elements of $\mathbf{H}_1$, $\mathbf{V}_{12}$ and $\mathbf{V}_{21}$. To determine the rest part of the total dynamic matrix, we resort to virtual crystal approximation (VCA), namely treating the interatomic interactions in regions less impacted by the impurity (regions 2 + 3, designated as "outer region") as the corresponding interatomic interactions within a homogeneous medium of random $In_{0.5}Ga_{0.5}As$. The VCA force constants are calculated by averaging the force constants of bulk GaAs and of bulk InAs, and then mapped onto the corresponding lattice positions of the 1000-atom effective supercell using space group symmetry operations. Note that since the impurity is sufficiently far from any atom in region 3, we can effectively approximate the interatomic force constant between the impurity and any atom in region 3 as zero. This treatment has minimal impact on the local vibrational modes of the impurity, but is necessary in order to construct the larger dynamic matrix needed for proper convergence in real-space Green's function method. Finally, we symmetrize **D** by iteratively applying the force constant sum rules together with the

diagonal symmetrization procedure. [41] Our formalism effectively reduces the heavy computational cost of phonon calculations, while retaining the required accuracy for the local vibrational spectrum of the impurities. We note that our approach is similar to the "combined dynamic matrix" method proposed by Shi and Wang. [42] In all of our calculations, we use the charge-neutral dopants and defects, since the Coulomb interactions between the supercell images are long-ranged and therefore requires special treatment in phonon calculations. [43] Experimentally, it is found that the vibrational modes of charged shallow dopants (such as Si and Se) are only slightly different from those of the neutral ones, as the dopant charge is delocalized throughout space. [44] The vibrational modes of deep-level defects (such as cation vacancies) can shift by as much as several tens of cm^{-1} under different charge states [44], but this effect does not modify our qualitative conclusions (see Sec. 5.3.3 and Sec. 5.4).

5.2.3 Local phonon density of states from real-space Green's function

To calculate the vibrational density of states in a crystalline solid, the conventional approach is to uniformly sample a sufficient number of **k**-points in the Brillouin zone. The vibrational density of states can then be constructed either from the sum of states over eigenvalues and **k**-points, or from the imaginary part of the trace of the static lattice Green's function. However, for defective crystals, this approach is not practical, since the real-space "unit cell" must be large enough to minimize interactions between periodic images of the defect, which renders such method prohibitively expensive. On the other hand, ran-

dom crystals are characterized by their local environments, instead of periodicity. This suggests that the optimal method should take advantage of the locality of the atomic interactions in the real space, rather than the periodicity of the supercell.

Based on these considerations, we choose to use the real-space Green's method [45] that is particularly suitable for treating large disordered systems. With the total dynamic matrix **D** calculated following the procedure in Sec. 5.2.2, we can construct the real-space Green's function, **G**, defined as [46]

$$\mathbf{G} = (\omega^2\mathbf{I} - \mathbf{D})^{-1}. \tag{5.2}$$

where ω is the phonon frequency, and **I** is the identity matrix with same dimensions as **D**. This expression is of little practical use, as inversion operation on a large matrix such as **D** is infeasible to compute. On the other hand, since **D** is symmetric by construction, we can use Haydock's recursion method [47] to transform **D** to a tridiagonal matrix $\tilde{\mathbf{D}}$. From there, we can easily construct the diagonal elements of the Green's function by a continued fraction expansion using entries of the tridiagonal $\tilde{\mathbf{D}}$, [48] and obtain the local phonon density of states (LPDOS) of the defect using the spectral expression: [49]

$$g(\omega, \mathbf{r}_\alpha) = 2\omega\left(-\frac{1}{\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}\sum_{j=1}^{3} G_{\alpha j,\alpha j}(\omega^2 + i\epsilon)\right) \tag{5.3}$$

where $G_{\alpha j,\alpha j}(\omega^2 + i\epsilon)$ is the real space Green's function for atom α in direction j.

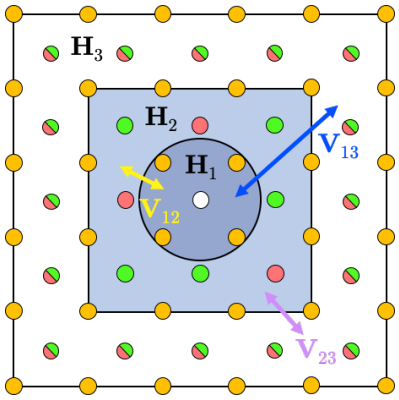

Figure 5.2: Schematic illustration of the total dynamic matrix for phonon calculations in this work. The 1000-atom effective supercell of random $In_{0.5}Ga_{0.5}As$ is divided into three regions, whose intra-region and inter-region dynamic submatrices are defined as $\mathbf{H}_i$ and $\mathbf{V}_{ij}$ respectively. (described in Sec. 5.2.3)

5.3 Results

5.3.1 Local atomic environments

The vibrational signature of a substitutional dopant or defect (henceforth called an "impurity") in a random crystal is strongly affected by its local atomic environment, as its interatomic interaction with different species of atoms generally have different strengths. An impurity can sit on either group III lattice sites or group V lattice sites in InGaAs. We searched and recorded the local environments throughout the 216-atom quasirandom supercell of random $In_{0.5}Ga_{0.5}As$, described by the number of In and Ga atoms in the nearest group III sites of the unrelaxed lattice. The results of the local environments for group III sites and group V sites are shown in Fig. 5.3.

Based on these results, we made the assumption that we can ignore any local environment with a ¡2% probability of being substituted by a defect. Based on

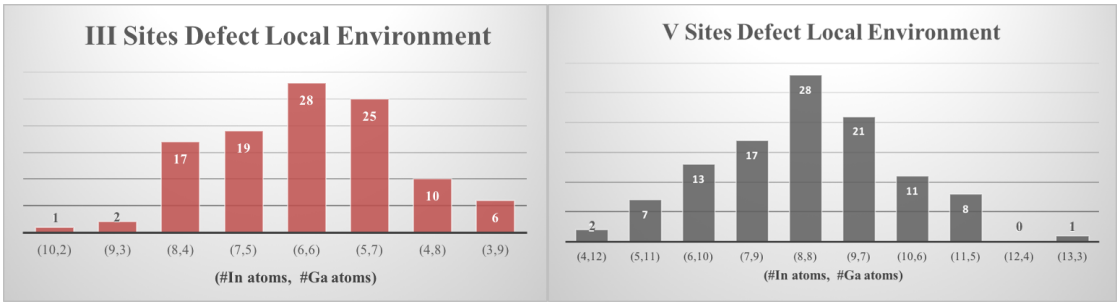

Figure 5.3: Local environments for group III and V sites in a 216-atom quasirandom supercell of $In_{0.5}Ga_{0.5}As$. For the histogram on the left, (x, y) on the bottom represents the local environment in which x In atoms and y Ga atoms reside in this kind of group III/V site's second nearest neighbors, and the number on each bar represents the number of occurrences (108 in total).

this assumption, we were able to mark six local environments for a group III site defect (Si and vacancy) and seven local environments for a group V site defect (Se and Te), which are considered subsequently for vibrational calculations.

5.3.2 Strain field induced by dopants and defects

Introducing an impurity into a pristine crystal reshapes the electrostatic interactions and atomic force constants in the neighborhood of the impurity. As a consequence, a strain field is created, defined as the force response of each atom on the *relaxed pristine* lattice due to the presence of the defect. In DFT simulations of periodic crystals, if the strain field of the impurity is not fully contained within the supercell, elastic interactions between image impurities could occur, which introduces a finite-size error to the calculated phonon spectrum. Furthermore, the interatomic force constant between any two atoms within the strain field of the defect may be different from the bulk value due to the change in

local lattice geometry. As a result, in order to accurately study the vibrational properties of an impurity, we must ensure that the supercell size is sufficiently large to fully contain the strain field of the impurity. The impact of the strain field surrounding the impurity on the interatomic force constants must also be fully taken into account in the DFT calculations.

In this work, we study three kinds of substitutional defects: Si_{III}, V_{III} (cation vacancy) and Se_{As} (subscript III denotes group III atom (In/Ga)). We use a force criterion of 0.03 eV/A to determine if a particular atom lies within the strain field of the impurity. We verified that for all three impurities and all local atomic environment considered in this work, all atoms within the strain field are fully contained within the 216-atom supercell boundary.

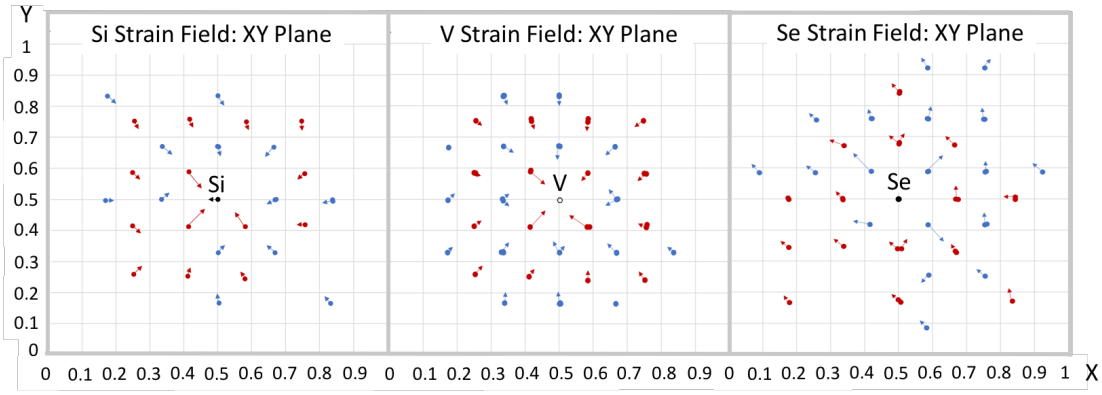

Figure 5.4: XY-plane projected strain fields for three types of defects (Si_{III}, V_{III}, Se_{As}) for selected local atomic environments in $In_{0.5}Ga_{0.5}As$, with blue dots representing In/Ga atoms and red dots representing As atoms. Each arrow represents the direction and relative magnitude of force on each atom. A few atoms are shown with more than one arrow due to overlapping, and a few atoms have no arrow since the defect does not stress them in the XY-plane.

Fig. 4 shows the strain fields of selected local atomic environments for all

three impurities considered, projected onto the XY-plane. Specifically, Si_{III} exhibits an ellipsoidally symmetric compressive field. This field induces strain on 30 surrounding atoms (reaching out to its fifth nearest neighbors), as shown in the leftmost picture of Fig. 5.4. Three of the four first nearest neighbors of the defect moved *towards* the Si dopant atom, while the remaining atom stays in its original position due to geometrical reasons. The presence of Si creates an off-center ion in which its equilibrium position is shifted away from the original lattice site. This is because Si is much smaller than the regular group III ions that it replaces (the atomic radius of Si is $r_{Si} = 1.11$Å, compared to $r_{In} = 1.42$Å and $r_{Ga} = 1.22$Å [50]). Consequently, this significantly weakens the repulsive forces between the impurity ion and its nearest neighbors that stabilize the ion at a regular site.

The strain field around a V_{III} defect also exhibits ellipsoidal symmetry. On the other hand, V_{III} exerts a stronger and more isotropic compressive field with less symmetry reduction than that created by Si_{III}. The vacancy compressed around 55 atoms, extending its influence out to its sixth nearest neighbors. Similar to the Si-induced case, three of the vacancy's four first nearest neighbors moved significantly towards the unoccupied site, while the remaining atoms remain in their original positions.

Unlike the case for the two defects mentioned above, the strain field around a Se_{As} defect is asymmetric and more dispersive. Se creates a long-range tensile strain field which distorts the zincblende structure and reduces the symmetry of tetrahedral bonds to a greater extent than either Si or vacancy. Atoms in a Se-doped crystal are more likely to shift towards one particular direction (+y). This can be understood by the fact that, in binary compounds such as InSe, selenium

atoms tends to form a wurtzite structure, which is non-centrosymmetric, unlike the tetrahedral bonds in a zincblende lattice. [51]

5.3.3 Local Phonon Density of States

Method Validation

As mentioned in section 2.2, we determine the *intra*-region force constants by displacing every atom within the defect's strain field. Empirically, all the atoms within the defect's first three nearest neighbor shells may have non-negligible effects on the local density of states of the defect. But we noticed that insertion of the impurity does not stress all the atoms inside these three shells. This is caused by the fact that the strains imposed on the host atoms may balance some forces bring by the presence of the defect. Thus, some atoms that are not stressed by the defect may still have non-negligible interatomic interactions with the defect. To ensure we have considered all non-negligible interatomic interactions in our calculations, we compute a few more displacement jobs for those "non-overlapping" atoms (that is, atoms within the defect's third-nearest-neighbor distance but not inside the defect's strain fields). Since Si has the most non-overlapping atoms, compared to the other cases we studied, we conduct a test calculation to examine the contribution of those non-overlapping atoms to the LPDOS of the defect.

As shown in Fig. 5.5, the results confirm our assumption that it is unnecessary to calculate additional displacement jobs for the non-overlapping atoms since it will not change the local vibrational modes of the defect. The two curves shown have very similar shape with local modes (peaks) occuring at nearly the

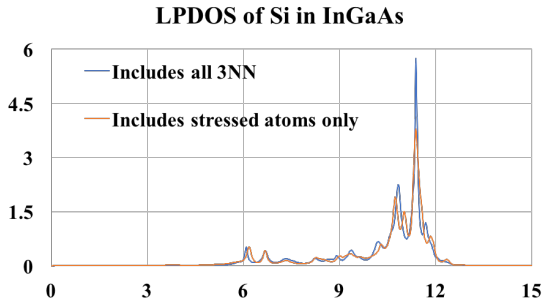

Figure 5.5: The local phonon density of states (LPDOS) of Si in quasirandom InGaAs under a specific local environment in which 4 In + 8 Ga atoms are located on group III sites of the first three nearest neighbor shells of Si. The blue line shows the LPDOS obtained by displacing the atoms in the defect's strain field plus all other atoms within the defect's third-nearest-neighbor distance. The orange curve represents the LPDOS obtained by displacing atoms in the defect's strain field only.

same frequencies. This validates our method that displacing atoms within the defect's strain field only brings accurate results can save significant computational cost without sacrificing accuracy.

Sensitivity Analysis

To study how uncertainty in the force constants affects the vibrational density of states, we conduct several sensitivity tests based on a Gaussian process in which random variables have a multivariate distribution. Specifically, this distribution is a joint distribution of all those random variables and, as such, it depicts a distribution over functions with space. Thus, in our case, this process is a homogeneous (stationary and isotropic) process with zero mean, and the covariance function depends only on the Euclidean distance between each atom and the defect. We implement the Gaussian process module from scikit-learn (v0.20.1) [52] to add noise to the force constants of each pair of atoms.

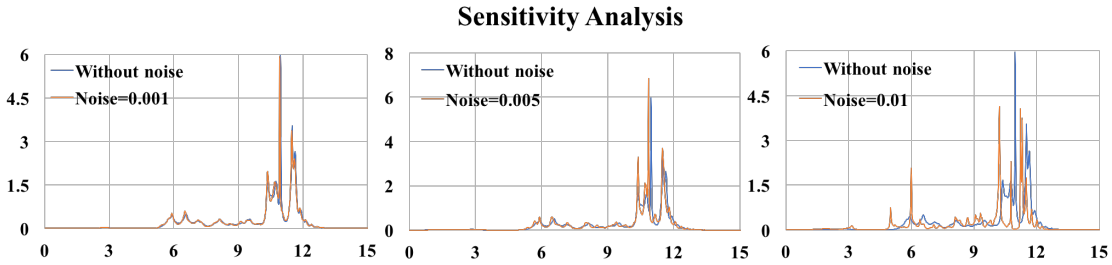

Figure 5.6: Local phonon density of states of Si in quasirandom InGaAs under a specific local environment in which 6 In + 6 Ga atoms are located on group III sites of the first three nearest neighbor shells of the Si atom. The orange lines show the result obtained from force constants with the application of 0.001/0.005/0.01 (in Rydberg atomic units) of noise, while the blue lines show the results without noise.

The results, shown in Fig. 5.6, provide evidence that our model has robust confidence and that local vibrations remain unchanged even when the introduced noise on the force constants is as high as 0.005 in Rydberg atomic units. In particular, if the uncertainty of force constants is 0.001, the vibrational density of states perfectly matches the results without noise. Local vibrations start to exhibit minor inconsistencies with the original output when the introduced noise is 0.005, shown in the middle plot of Fig. 5.6, while the peaks still occur at the same frequencies. These results confirm the robustness of our approach.

Binary cases: Local phonon density of states of defects in GaAs and InAs

To validate our method, we first apply our method to calculate the local phonon density of states for defects in binary alloys: GaAs and InAs. Our results, as shown in Fig. 5.7 and Fig. 5.8, are in good agreement with the existing experimental data. Specifically, the calculated vibrational modes of Si occur at 12.124 and 10.762 THz in GaAs and InAs, while the experimentally observed absorp-

tion peaks occur at 11.512 [6] and 10.763 THz, [7, 8] respectively. For Se_V, the peaks of local phonon density of states in GaAs and InAs occur at 9.181 and 8.010 THz. The vibrational modes of the vacancy's first-NNs are all resonant states where the peaks occur at frequencies ranging from 3.503 to 5.627 THz and 3.073 to 3.912 THz, respectively, in GaAs and InAs alloys.

We showed that our methodology gives a quantitatively good description of phonon modes and these results will be our reference in the following subsections to interpret the vibrational properties of defects in a random host crystal.

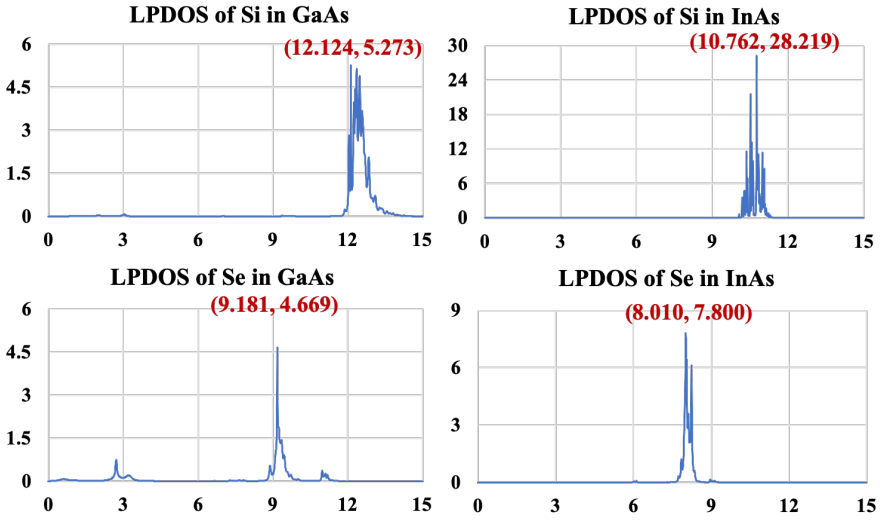

Figure 5.7: Local phonon density of states of Si_{III}, Se_V and V_{III} in ordered GaAs and InAs. The coordinates of the highest peak (x, y) under each scenario are shown in red, where x represents the frequency in THz and y represents the intensity (dimensionless).

Local phonon density of states of Si_{III}

The local phonon density of states of Si_{III} in the quasirandom $In_{0.5}Ga_{0.5}As$ alloy are shown in Fig. 5.9 for all the local environments considered. Since the first nearest neighbors of Si, which have the most significant interations with the

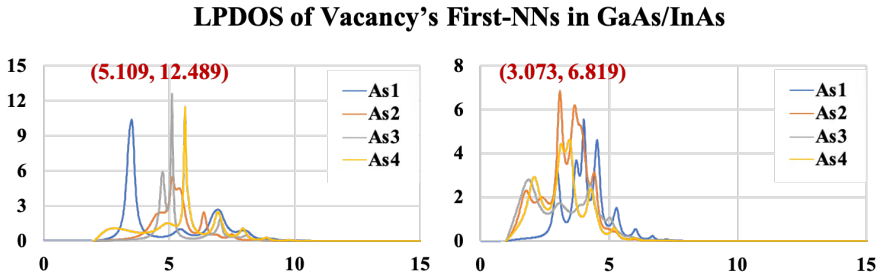

Figure 5.8: Local phonon density of states of a cation vacancy's first four nearest neighbors in ordered GaAs and InAs. The coordinates of the highest peak (x, y) under each scenario are shown in red, where x represents the frequency in THz and y represents the intensity (dimensionless).

defect, are all occupied by As atoms, the peaks in different local environments occur at similar frequencies. Specifically, the most intense vibrations occur at 11.492, 12.075, 10.964, 10.860, 11.389 and 11.544 THz, respectively, in each local environment. As expected, all the obtained vibrational modes of Si_{III} in a quasirandom $In_{0.5}Ga_{0.5}As$ alloy occur at frequencies between 10.762 THz and 12.123 THz, corresponding respectively to the frequency of Si_{In} in InAs and that of Si_{Ga} in GaAs based on the results of our control group. We observe that when the number of Ga atoms equals, or is close to, the number of In atoms in the Si atom's neighborhood (within 3rd-NN distance), the peak occurs at a lower frequency. Besides, we do not see a specific pattern of vibrational modes with an increasing number of In/Ga atoms in Si's neighborhood. These indicate that the phonon modes of Si in random III-V alloy does not depend very sensitively on the nearest-neighbor atom species.

The local vibrational modes of Si's first nearest neighbors and a "bulk" atom (an atom which sits in the virtual crystal and is far from the defect) are shown in Fig. 5.10. The phonon modes of the bulk atom exhibit the global vibrations of the quasirandom $In_{0.5}Ga_{0.5}As$ crystal since this atom has negligible interactions

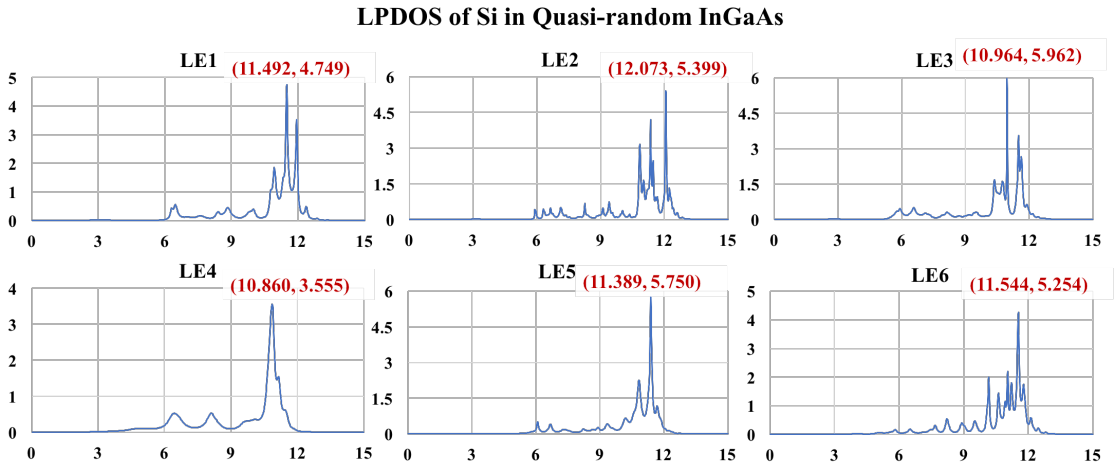

Figure 5.9: Local phonon density of states of Si_{III} in quasirandom InGaAs in all possible local environments (LE). Specifically, LE1, LE2, LE3, LE4, LE5, and LE6 represents there are 8In + 4Ga, 7In + 5Ga, 6In + 6Ga, 5In + 7Ga, 4In + 8Ga, and 3In + 9Ga, respectively, occupying group III sites in the first three nearest neighbor shells of Si. The coordinates of the highest peak (x, y) under each local environment are shown in red, where x represents the frequency in THz and y represents the intensity (dimensionless).

with the defect. We can clearly see that the insertion of Si breaks the symmetry and disrupts the global vibrations. All four nearest neighbors react strongly to this interruption and show similar phonon modes with those of Si_{III}. The local modes in all scenarios occur at frequencies ranging from 10.3 to 11.8 Thz. One of the four first nearest neighbors in each scenario, such as the one shown in the upper right plot in Fig. 5.10, displays weak vibrations (reflected as low intensity). It does not respond as strongly as the other three first nearest neighbors to the introduction of Si, which is consistent with our results for the strain fields (one atom is not strained by the defect), as discussed in Sec. 5.3.3.

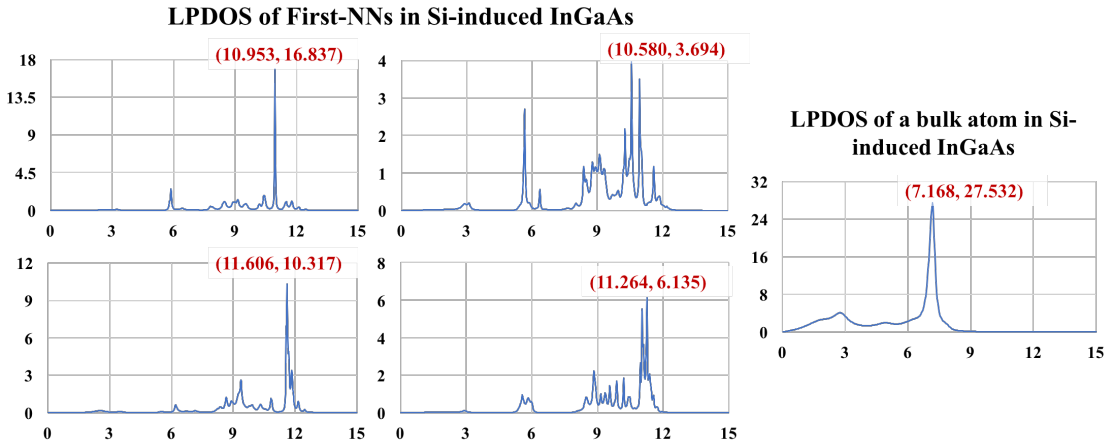

Figure 5.10: Local phonon density of states of Si_{III}'s first four nearest neighbors and a bulk atom in quasirandom InGaAs under a specific local environment in which 6 In and 6 Ga atoms are located on the III sites of the first three nearest neighbors of Si.

Local phonon density of states of Se_{As}

The local phonon density of states of Se_{As} in the quasirandom $In_{0.5}Ga_{0.5}As$ alloy are shown in Fig. 5.11 for all the local environments considered. Compared to the vibrational modes of Si_{III}, those of Se_{As} depend more sensitively on the local environment in both frequency and intensity due to variations of its first nearest neighbors in different scenarios. We notice that when the first-NN shell sites are all occupied by In atoms and most of the third-NN shell sites are occupied by Ga atoms (such as LE1 and LE2), the most intense peaks occur inside the range of the vibrational frequencies of the host lattice, making them resonant states. This may be because the difference in covalent radius of Se and As is much smaller when compared to In than when compared to Ga (r_{Ga} = 1.26Å, r_{In} = 1.44Å [50]). Hence the lattice structure around the Se dopant is similar to a central As atom with In atoms as all first-NNs, As atoms as all second-NNs and Ga atoms as all third-NNs, which makes the local structure ordered and Se_{As} showing stronger

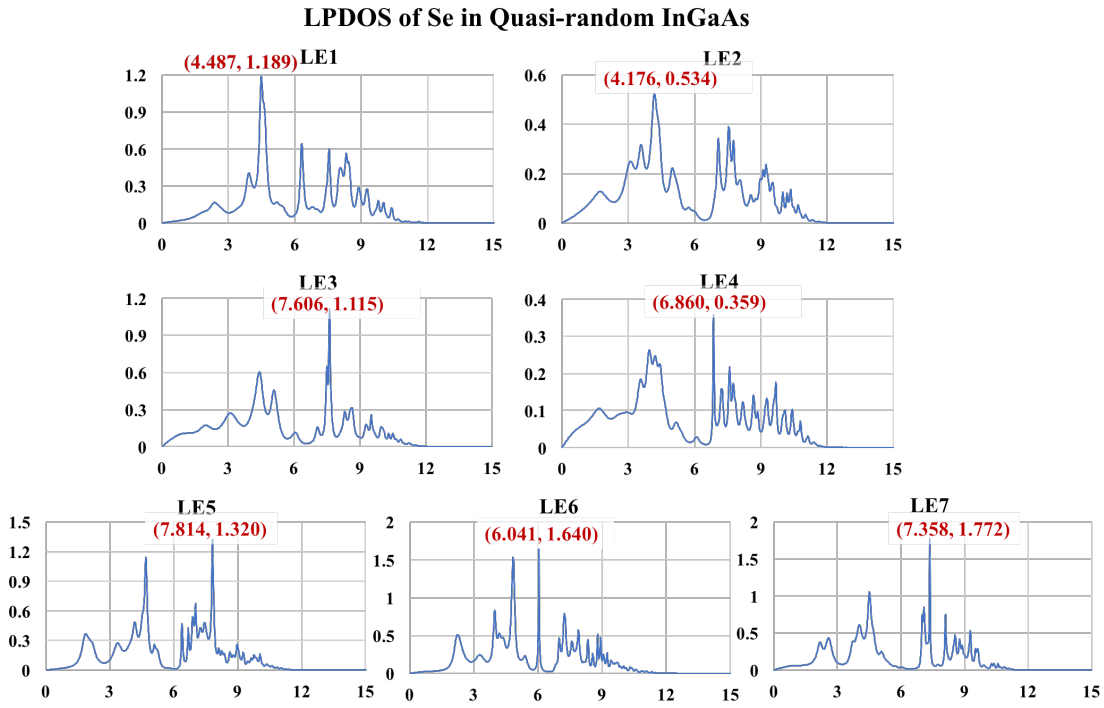

Figure 5.11: Local phonon density of states of Se_{As} in quasirandom InGaAs under all possible local environments, LE1, LE2, LE3, LE4, LE5, LE6, and LE7 correspond to 5In + 11Ga, 6In + 10Ga, 7In + 9Ga, 8In + 8Ga, 9In + 7Ga, 10In + 6Ga, and 11In + 5Ga, respectively, occupying group III sites in the first three nearest neighbor shells of Se.

global vibration. Under this circumstance, with the increasing number of In in the third-NNs, the local structure becomes more disordered and Se shows relatively stronger local vibrations, which is reflected from the much weaker global vibration ad more evident peaks in local frequency zone in LE2 compared to LE1. Particularly, the peaks occur at 4.487 and 4.176 THz (stronger global vibrations) under LE1 and LE2; and 7.606, 6.680, 7.814, 6.041, and 7.358 THz (stronger local vibrations) under LE3-LE7, respectively. Comparing with our binary group results, these peaks occur at slightly lower frequencies than that in an ordered host lattice. Since more energy (a higher frequency) is required to

vibrate a stronger bond, our results suggest that the bond strengths between Se atom and host atoms in the random lattice are weaker than those in an ordered lattice.

Since local atomic environments involve random atomic arrangements, multiple vibrational modes are observed even for the same type of neighboring atoms of the defect. The intensity of Se_{As}'s vibrations is much lower than that of Si_{III} due to its larger atomic mass (m_{Si} = 28.08 amu, m_{Se} = 79.92 amu [53]) and its similar covalent radius to As (r_{Se} = 1.16Å, r_{As} = 1.19Å [50]). These peaks occur at lower frequencies than those in the Si case, which is consistent with empirical research in which the vibrations of heavier defects result in lower frequencies.

The local vibrational modes of the first nearest neighbors of Se and a bulk atom are shown in Fig. 5.12. The most intense peaks in different local environments occur at frequencies ranging from 9.2 to 10.7 THz. The peaks of these first nearest neighbors shift to higher frequencies compared to that of the defect. In contrast to the results of Si case, three of four nearest neighbors of Se still have some global oscillations reflected as smaller peaks in the low-frequency zone. As expected, the bulk-like atom shows similar phonon density of states curves to those for the Si case.

Local phonon density of states of V_{III}

The local phonon density of states of a cation vacancy's first nearest neighbors in the quasirandom $In_{0.5}Ga_{0.5}As$ alloy are shown in Fig. 5.13 for all the local environments considered.

All the first nearest neighbors show a complex distribution of local vibra-

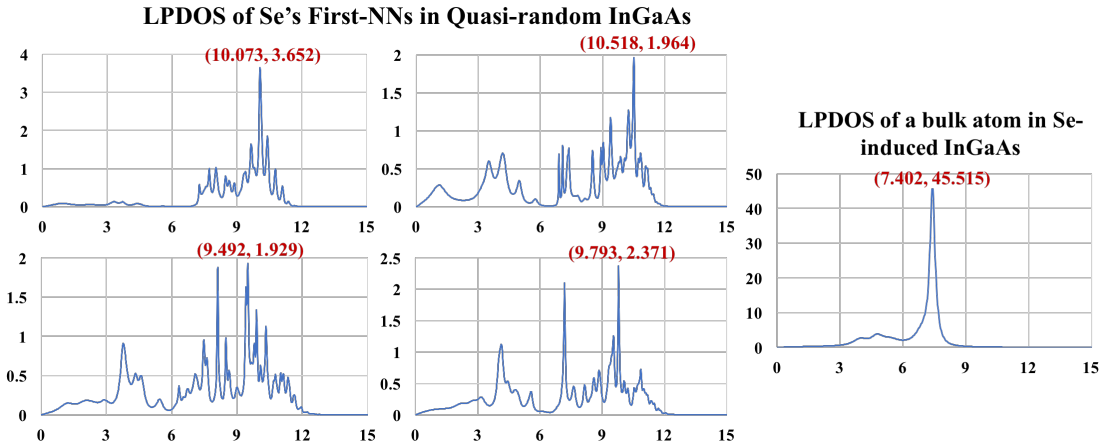

Figure 5.12: Local phonon density of states of Se_{As}'s first four nearest neighbors and a bulk atom in quasirandom InGaAs under a specific local environment where there are 8 In and 8 Ga atoms within Se's first three nearest neighbor shells.

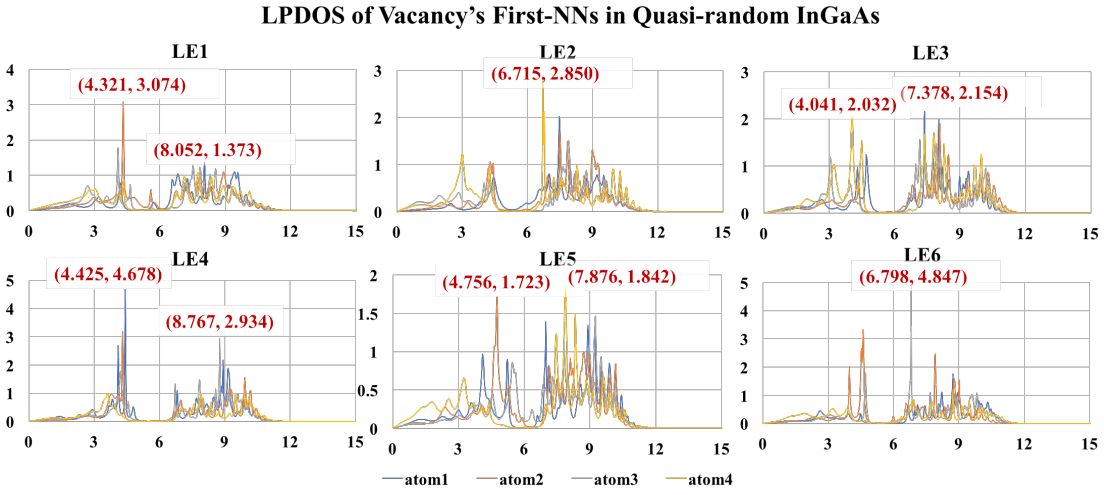

Figure 5.13: Local phonon density of states of a cation vacancy's first four nearest neighbors and a 'bulk' atom in quasirandom InGaAs under two specific local environments. LE1, LE2, LE3, LE4, LE5 and LE6 refer to cases considering 8In + 4Ga, 7In + 5Ga and 6In + 6Ga, 5In + 7Ga, 4In + 8Ga and 3In + 9Ga, respectively, occupying group III sites in the vacancy's first three nearest neighbor shells.

tions, to different extents, in addition to global vibrations, as shown by the varying intensities of the bulk-like peaks in the LPDOS. Comparing with all first-NNs of vacancy show resonant modes in both GaAs and InAs cases, we find that local vibrations arise due to the disordered atomic arrangements in quasi-random InGaAs. Unlike the cases for dopants where localized modes often show higher intensities compared to bulk-like modes, global vibrations are more pronounced relative to local vibrational modes in the first nearest-neighbor projected modes. This is mainly because in the case of vacancy, there is no impurity *atoms* to which the nearest-neighbor atoms are bonded; hence the local vibrational modes related to the impurity atom are not present. A vacancy removes less local lattice symmetry compared to Si and Se and this is also reflected in a more symmetrical strain field, shown in Section 3.1. The vibrational signatures of the cation vacancies suggest that the presence of such defects in a random InGaAs would introduce a noisy "background" of signals on top of the more distinguishable peaks of the dopants such as Si and Se.

5.4 Conclusions

The novel computational approach presented in this paper allowed us to determine the vibrational signatures of single-atom dopants/defects in a random alloy with fully *ab initio* accuracy. The impact of this approach is that it provides a way to predict the distinctive vibrational signatures of dopant/defect species in the random alloys commonly observed in experiments. These vibrational signatures may then serve as a guide for experimentalists to assess the dopant activation efficiency of novel doping schemes for alloys such as InGaAs. Our calculations take into account all possible local environments for a defect site in

a random alloy, which is of great importance for highly disordered system but which have not been performed in previous studies due to the computational difficulty. We demonstrate that our methodology provides a quantitative good description of phonon modes by the excellent agreement between our binary alloy results and the experimental data.

For our case study, we consider two prototypical *n*-type dopants, silicon and selenium, and one prevalent compensating defect, the cation vacancy, in a random $In_{0.5}Ga_{0.5}As$ alloy. We demonstrated (and visualized) the strain fields and vibrational modes of these defects. We find that a random alloy introduces strong variations in the vibrational mode of the dopant/defect and the degree of dependence of dopant/defect vibrational modes on the local atomic arrangement varies across impurity type. Besides, comparing with the results from binary alloys, we find that the phonon mode of dopant/defect in a ternary alloy like InGaAs tends to occur at an intermediate frequency of the binary cases (GaAs and InAs), and the local phonon density of states typically have more satellite peaks. In addition, we find that the *local* vibrational modes of Si under all possible local environments occur at frequencies higher than that of host atoms, ranging from 10.9 to 12.1 THz. On contrary, Se can exhibit resonant vibrational modes for some local configurations and the most intense peak occur at very low frequencies ranging from 4.2 to 4.5 THz, while for other local configurations, the vibrational modes occur at higher frequencies (6.0 to 7.8 THz) than that of host atoms, assuming localized characters. Compared to that of Si and Se, the nearest neighbors of a vacancy respond much more weakly to the defect center. More than half of the vacancy's first nearest neighbors show stronger *global* oscillations than local vibrations, while all the nearest neighbors of Si and Se react strongly to the insertion of the defect and show dominant local vibra-

tions. Our study reveals the significant influence of random local environments on the vibrational signature of impurities in solids, and our method opens the door to investigating such phenomenon in crystalline materials.

Bibliography

[1] J. Newman, in Semiconductors and Semimetals, Vol. 38, edited by E. R. Weber (Academic, New York, 1993), p. 59.

[2] M. D. McCluskey, J. Appl. Phys., **87**, 8 (2000).

[3] J. Maguire, R. Murray, and R. C. Newman, Appl. Phys. Lett. **50**, 516 (1987).

[4] M. Ramsteiner, J. Wagner, H. Ennen, and M. Maier, Phys. Rev. B 38, 10669 (1988)

[5] R. Addinall, R. Murray, R. C. Newman, J. Wagner, S. D. Parker, R. L. Williams, R. Droopad, A. G. DeOliveira, I. Ferguson and R. A. Stradling, Semicond. Sci. & Technol. **6**, 147 (1991).

[6] R. Murray, R. C. Newman, M. J. L. Sangster, R. B. Beall, J. J. Harris, P. J. Wright, J. Wagner and M. Ramsteiner, J. Appl. Phys. **66**, 2589 (1989)

[7] J. Wagner, P. Koidl, and R. C. Newman, Appl. Phys. Lett. **59**, 1729 (1991).

[8] M. Uematsu, J. Appl. Phys., **69**, 1781 (1991).

[9] S. Najmi, X. K. Chen, A. Yang, M. Steger, M. L. W. Thewalt, and S. P. Watkins, Phys. Rev. B **74**, 113202 (2006).

[10] J. C. Mikkelsen and J. B. Boyce, Phys. Rev. Lett. **49**, 1412 (1982).

[11] M. Ichimura and A. Sasaki, J. Appl. Phys. **60**, 3850 (1986).

[12] K. Shin, J. Yoo, S. Joo, T. Mori, D. Shindo, T. Hanada, H. Makino, M. Cho, T. Yao, and Y.-G. Park, Mater. trans. **47**, 4 (2006).

[13] H.C. Alt, J. Phys. Condens. Matter **16**, S3037 (2004); ibid., H.C. Alt, Y.V. Gomeniuk, B. Wiedemann, Phys. Rev. B **69**, 125214 (2004); ibid., Semicond. Sci. Technol. **18**, 303 (2003); ibid., Appl. Phys. Lett. **77**, 3331 (2000); ibid., Mat. Sci. Forum 258–263, 867 (1997).

[14] S. Kurtz, J. Webb, L. Gedvilas, D. Friedman, J. Geisz, J. Olsen, R. King, D. Joslin, N. Karam, Appl. Phys. Lett. **78**, 748 (2001).

[15] H.C. Alt, Y.V. Gomeniuk, Phys. Rev. B **70**, 161314 (2004).

[16] K. Köhler, J. Wagner, P. Gesner, D. Serries, T. Geppert, M. Maier, L. Kirste, IEEE Proc. Optoelectron. **151**, 247 (2004); ibid., J. Appl. Phys. **90**, 2576 (2004); ibid., Solid State Electron. **47**, 461 (2003); ibid., Mater. Res. Symp. Proc. **744**, 627 (2003); ibid., Appl. Phys. Lett. **83**, 2799 (2003); ibid., Appl. Phys. Lett. **80**, 2081 (2002); ibid., J. Appl, Phys. **90**, 5027 (2001); ibid., Appl. Phys. Lett. **77**, 3592 (2000).

[17] J. Wagner, K. Köhler, P. Ganser, M. Maier, Appl. Phys. Lett. 87, 051913 (2005).

[18] T. Kitatani, M. Kondow, M. Kudo, Jpn. J. Appl. Phys. 40, L750 (2001).

[19] D. N. Talwar, J. Appl. Phys. **99**, 123505 (2006).

[20] A. Togo and I. Tanaka, Scr. Mater. **108**, 1 (2015).

[21] A. M. Teweldeberhan, and S. Fahy, Phys. Rev. B **73**, 245215 (2006).

[22] C.-W. Lee, B. Lukose, M. O. Thompson, and P. Clancy, Phys. Rev. B **91**, 094108 (2015).

[23] A. M. Teweldeberhan, G. Stenuit, S. Fahy, E. Gallardo, S. Lazić, J. M. Calleja, J. Miguel-Sánchez, M. Montes, A. Hierro, R. Gargallo-Caballero, A. Guzmán, and E. Muñoz, Phys. Rev. B **77**, 155208 (2008).

[24] G. E. Moore, Electronics **38**, 114 (1965).

[25] "International roadmap for devices and systems, 2017," (`https://irds.ieee.org/images/files/pdf/2017/2017IRDS_MM.pdf`).

[26] J. A. Del Alamo, Nature **479**, 317 (2011).

[27] H. Aldridge, A. G. Lind, C. C. Bomberger, Y. Puzyrev, J. M. Zide, S. T. Pantelides, M. E. Law, and K. S. Jones, Mater. Sci. Semicond. Process. **57**, 39 (2017).

[28] J. O'Connell, E. Napolitani, G. Impellizzeri, C. Glynn, G. P. McGlacken, C. O'Dwyer, R. Duffy, and J. D. Holmes, ACS Omega **2**, 1750 (2017).

[29] J. J. M. Law, A. D. Carter, S. Lee, C. Y. Huang, H. Lub, M. J. W. Rodwell, and A. C. Gossard, J. Cryst. Growth **378**, 92 (2013).

[30] A. Maassdorf, M. Hoffmann, and M. Weyers, J. Cryst. Growth **315**, 57 (2011)

[31] J. Wang, B. Lukose, M. O. Thompson, and P. Clancy, J. Appl. Phys. **121**, 045106 (2017).

[32] M. Reveil, J. Wang, M. O. Thompson, and P. Clancy, Acta. Materialia 140 (2017).

[33] T. S. Kuan, W. I. Wang, and E. L. Wilkie, Appl. Phys. Lett. **51**, 51 (1987).

[34] M. A. Shahid, S. Mahajan, D. E. Laughlin, and H. M. Cox, Phys. Rev. Lett **58**, 2567 (1987).

[35] H. Ch. Alt, A. Yu. Egorov, H. Riechert, B. Wiedemann, J. D. Meyer, R. W. Michelmann, and K. Bethge, Physica B Condens Matter **302** (2001).

[36] S. R. Kurtz, A. A. Allerman, J. F. Klein, R. M. Sieg, C. H. Seager, and E. D. Jones, Mat. Res. Soc. Symp. Proc. **692** (2002).

[37] A. Zunger, S.-H. Wei, L. G. Ferreira, and J. E. Bernard, Phys. Rev. Lett. **65**, 353 (1990).

[38] W. M. Spears, in Cliques, Coloring and Satisfiability: 2nd DIMACS Implementation Challenge. (1993) p. 533-558.

[39] A. Baldereschi, Phys. Rev. B **7**, 5212 (1973).

[40] B. G. Pfrommer, M. Côté, S. G. Louie, and M. L. Cohen, J. Comput. Phys. **131**, 233 (1997).

[41] G. J. Ackland, M. C. Warren, and S. J. Clark, J. Phys.: Condens. Matter **9**, 7861 (1997).

[42] L. Shi, and L.-W. Wang, Phys. Rev. Lett. **109**, 245501 (2012).

[43] Y. Wang, S.-L. Shang, H. Fang, Z.-K. Liu, and L.-Q. Chen, Npj Comput. Mater. **2**, 16006 (1996).

[44] M. Stavola, in Semiconductors and Semimetals, Vol. 51B, edited by M. Stavola (Academic, New York, 1999) p. 153.

[45] J. J. Sinai, Phys. Rev. B **54**, 7937 (1996).

[46] P. H. Dederichs, R. Zeller, K. Schroeder. Point defects in metals II: Dynamical Properties and Diffusion Controlled Reactions. Vol. 87. Springer, 1980.

[47] R. Haydock, V. Heine, and M. J. Kelly, J. Phys. C **5**, 2845 (1972).

[48] H. S. Wall, Analytical Theory of Continued Fractions (Van Nostrand-Reinhold, Princeton, NJ, 1948).

[49] Z. Tang and N. R. Aluru, Phys. Rev. B **74**, 235441 (2006).

[50] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barrag, and S. Alvarez, Dalton Trans. 2832 (2008).

[51] Tao, X and Gu, Y. Nano Letters, **13(8)**, 3501-3505 (2013).

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[53] G. Audi, A. H. Wapstra, and C. Thibault, "The AME2003 Atomic Mass Evaluation:(II). Tables, Graphs and References," Nuclear Physics A, **729**(1): 337-676 (2003).

CHAPTER 6

AB INITIO STUDIES OF CONTACT RESISTIVITY OF ALLOYED AND NON-ALLOYED NI/IN$_{1-X}$GA$_X$AS CONTACT

The work presented in this chapter originates from the author's work during his internships at IBM T. J. Watson Research Center.

NOTE: The results presented in this chapter is affected by a known bug in the simulation software ATK QuantumWise (version 2017.0), and hence the exact numbers are not reliable. Only the trends are meaningful.

6.1 Introduction

As the dimensions of transistors shrinks below 10nm according to Moore's law, [1] the device performance becomes limited by multiple materials-dependent factors. One of the major performance-limiting factors is the (specific) contact resistivity, defined as the partial derivative of the voltage with respect to the current density evaluated at zero current limit. The contact resistivity is a quantitative, area-independent measure of the ability to conduct electron flow across an interface. Contact resistivity is a major source of parasitic resistance in an electronic device, which limits the extent of ON current for a given voltage, and in turn the ratio of Joule heating over total power consumption. Such a problem must be overcome for nanometer-scale devices in order to make them practical in very-large-scale integrated circuits.

III-V materials such as InGaAs have been heralded as a promising candidate for use in future transistors, mainly due to its high electron mobility compared to silicon. Nevertheless, the metal-InGaAs contact has not yet achieved

the $10^{-9}\Omega$-cm^2 requirement for contact resistivity as of today, which severely limits the potential of wide-scale industrial deployment of InGaAs in transistor applications. [2] Although there exists several experimental realizations of low-resistivity metal-InGaAs contacts with different metals [3, 4, 5, 6], as well as computational studies on the ideal lower (Landauer) limit of metal-III-V contact resistivity based on generic quantum mechanical arguments [7], a systematic approach to improve the contact resistivity taking into account multiple parameters has, to our best knowledge, yet to be conducted.

In this work, we aim to gain a comprehensive understanding of the impact of materials-dependent factors on the contact resistivity of metal/III-V interface. Specifically, we consider five different knobs: composition of $In_{1-x}Ga_xAs$, surface termination of In(Ga)As, doping in In(Ga)As, presence of compositional grading in $In_{1-x}Ga_xAs$, and presence of metal-semiconductor alloying. Using ab initio atomistic modeling based on non-equilibrium Green's function (Sec. 3.5.3) method, we separate the effect of each factor in a controlled manner. Such an approach is critical in gaining a detailed roadmap towards optimization of contact resistivity of metal/III-V systems.

6.2 Methods

To model the metal-semiconductor contact, a standard two-terminal device configuration is used. This type of device configuration consists of a "central" region sandwiched in between two "electrode" regions. (Fig. 3.20) Specifically, the left electrode regions contain the metal, the right electrode region contains the semiconductor, and the central region contains the metal-semiconductor inter-

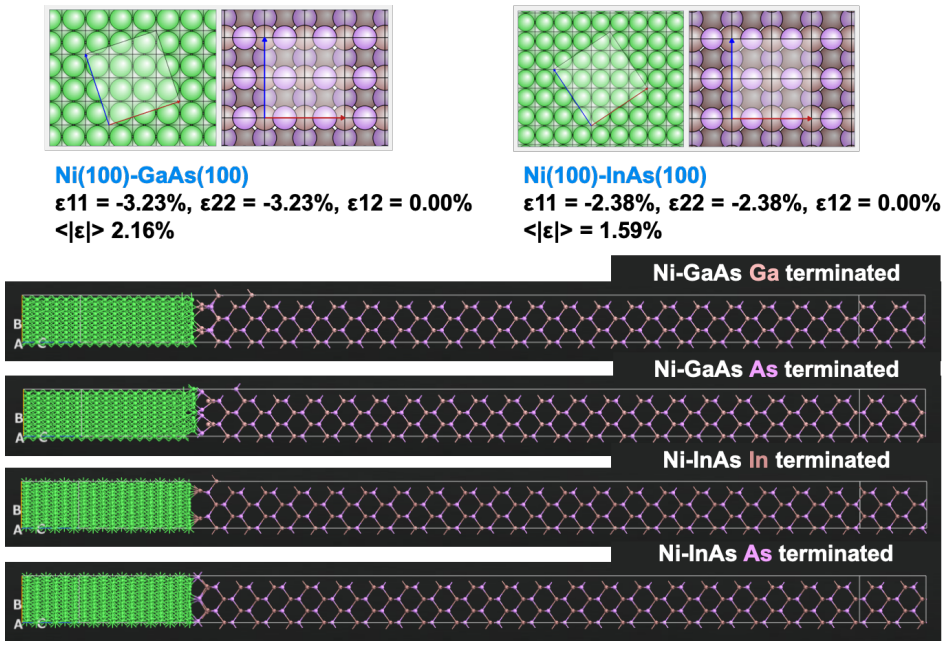

Figure 6.1: Surface unit cells and two-probe device configurations of Ni-GaAs and Ni-InAs used in the simulation, with As and In/Ga termination for both configurations.

face. This setup assumes periodic boundary conditions for the electrodes, hence the electron density of each electrode is replicated *ad infinitum* in the open end direction. The electrodes are chosen to have lengths of $\sim 10 - 20$Å (minimum-length unit cell allowing periodicity), and the central region is chosen to have length of ~ 150Å. (Fig. 6.1) The relatively long length of the central region is crucial to ensure that the interfacial band bending on the semiconductor side is completely screened out within the right electrode.

All calculations in this work are performed with density functional theory (DFT) using the software ATOMISTIX TOOLKIT version 2017.0 [8]. The Kohn-Sham wavefunctions are linearly expanded in the SG15 double-ζ basis set with medium accuracy. The electron density mesh cutoff is set to be 100 Hartree, or 2.7×10^3 eV for convergence of numerical results. The semiconductor right elec-

trode is constructed by extracting a bulk portion of a surface slab in a particular orientation (in this case, (001)). The metal left electrode is constructed by the same procedure, plus an additional constrained geometry optimization procedure where the transverse lattice vectors are strained and fixed to be equal to those of the semiconductor electrode. For each specific orientation of the semiconductor, we choose the cross-sections of the metal and the semiconductor so that the absolute value of lattice mismatch of the bulk metal relative to the bulk semiconductor is less than 3% in both transverse directions, corresponding to the experimental criteria of stable interface with minimal occurence of misfit dislocations. Once the metal and the semiconductor electrodes have been fully relaxed, we relax the central region by imposing a rigid condition (i.e. fixing the internal degrees of freedom) on the outermost three (six) atomic layers of metal (semiconductor) electrode, while allowing the other atoms to optimize their positions through the BFGS algorithm, until the residual force on each atom becomes less than 0.05 eV/Å. For geometry optimization, the PBEsol functional [9] is used to obtain an accurate geometry for the metal-semiconductor interface; a Monkhorst-Pack **k**-point mesh with density $\sim$ 4Å is used for the left electrode, and a **k**-mesh of $3 \times 3 \times 1$ is used for the central region (one **k**-point in the transport direction).

The device simulation is performed with the density functional theory (DFT)-based non-equilibrium Green's function (NEGF) method, which is a widely used method for simulating ballistic transport based on Landauer-Büttiker formalism. (Sec. 3.5.3) In order to simulate the realistic device, a doping concentration corresponding to experimentally relevant values (1×10^{19}-1×10^{20}cm^{-3}) in InGaAs is introduced inside the semiconductor region. The doping is implemented by an effective scheme where atomically localized elec-

tron densities corresponding to the doping concentration are uniformly added in the semiconductor region.

For transport calculations, the meta-GGA functional TB09-MBJ [10, 11] is used to obtain the correct band gap; a Monkhorst-Pack grid of $6\times6\times150$ is used for all configurations, as a very dense **k**-point grid in the transport direction for the electrodes is required for the transmission coefficient $T(E)$ to converge. The definition of contact resistance at the metal-semiconductor interface is

$$R_c = R_{\text{dev}} - R_{\text{metal}} - R_{\text{semi}}, \tag{6.1}$$

where R_{dev}, R_{metal}, and R_{semi} denote the resistance of the full two-terminal device, the intrinsic resistance of bulk metal, and the intrinsic resistance of bulk semiconductor, respectively. As the resistivity of a typical metal is several orders of magnitude lower compared to that of the semiconductor, we may approximate R_c by

$$R_c = R_{\text{dev}} - R_{\text{semi}}, \tag{6.2}$$

R_{dev} is readily calculated from the device transmission coefficient $T(E)$ obtained from the central region's Green's function $\mathbf{G}_C$ (Eq. 3.72) at zero-bias limit, as:

$$R_{\text{dev}} = \left[\frac{2e^2}{h}\int \mathrm{d}E\, T_{\text{dev}}(E)\left(-\frac{\partial f}{\partial E}\right)\right]^{-1}, \tag{6.3}$$

where $f(E)$ is the Fermi-Dirac distribution. R_{semi}, on the other hand, can be obtained by a direct bulk transmission simulation of the semiconductor electrode.

6.3 Results

As a case study of the Nickel/InGaAs interface, five independent factors are taken into account: (1) semiconductor alloy composition; (2) semiconductor sur-

face termination; (3) doping concentration in semiconductor; (4) semiconductor compositional grading; (5) metal-semiconductor alloying.

6.3.1 Semiconductor alloy composition

As a ternary compound, $In_{1-x}Ga_xAs$ crystals may assume any composition ranging from GaAs to InAs. Even though $In_{0.53}Ga_{0.47}As$ is the most common variant of InGaAs for its perfect lattice match with indium phosphide (InP) substrate, other compositions such as $In_{0.3}Ga_{0.7}As$ are also used in electronic devices. In our case study, we choose four different compositions along the GaAs-InAs tie line: GaAs, InAs, $In_{0.5}Ga_{0.5}As$, and $In_{0.3}Ga_{0.7}As$. An orientation of (100) is used for all four InGaAs compositions. Furthermore, a doping concentration of $1 \times 10^{19} cm^{-3}$ is used for all four configurations.

Fig. 6.1 shows the cross-sections of the semiconductors and their respective metal electrodes, as well as the device configurations used in the transport simulation for GaAs and InAs semiconductors. As the lattice constant of InAs (6.06Å) is 7% larger than the lattice constant of GaAs (5.65Å), it is crucial to choose different surface unit cells for the metal in order to maintain minimal strain with the semiconductor surface. As shown in Fig. 6.1, for the (2×2) unit cell of GaAs (InAs), the minimal-strain surface unit cell for Nickel is the $(\sqrt{10} \times \sqrt{10})R18.4^\circ$ ($(\sqrt{13} \times \sqrt{13})R33.7^\circ$) in Wood's notation, respectively. This leads to a compressive strain of 3.23% and 2.38%, respectively, both below the 4% threshold.

The local densities of states (LDOS) of the Ni/In(Ga)As(100) contact at equilibrium (zero bias) condition are plotted in Fig. 6.2. The most noticeable features

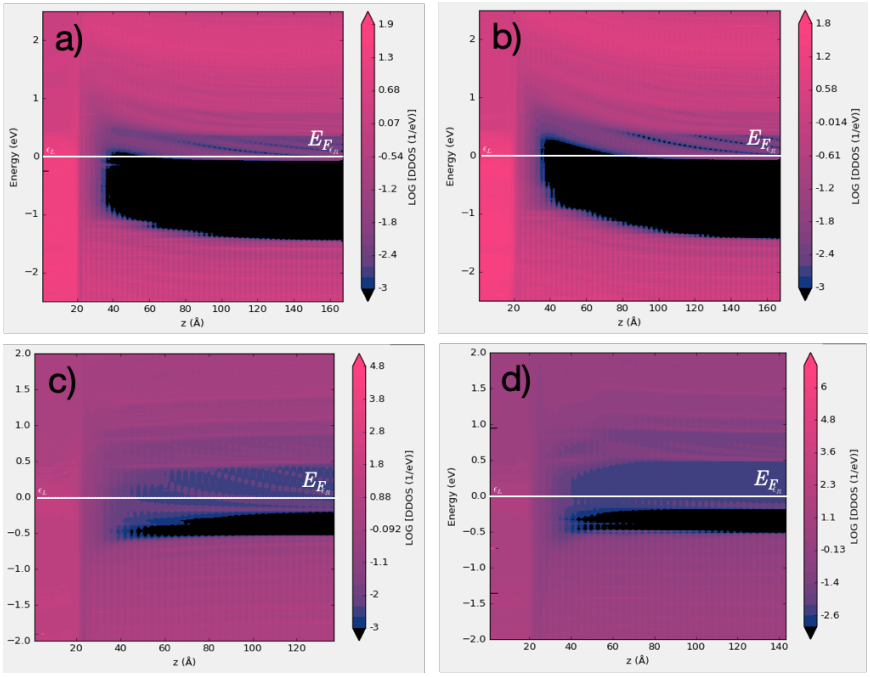

Figure 6.2: The local density of states (LDOS) for: (a) Ni/GaAs(100) (As-terminated); (b) Ni/GaAs(100) (Ga-terminated); (c) Ni/InAs(100) (As-terminated); and (d) Ni/InAs(100) (In-terminated).

of these plot are the strong band bending at the Ni/GaAs(100) interface, and the lack thereof at the Ni/InAs(100) interface. According to the Schottky-Mott rule, [12, 13] without taking into account the interfacial Fermi level pinning, the amount of semiconductor band bending, or the Schottky barrier height (SBH), is given by the difference between the work function of the metal ϕ_{M} and the electron affinity of the semiconductor χ_{SC}. As $\chi_{\mathrm{GaAs}} = 4.07\mathrm{eV}$ and $\chi_{\mathrm{InAs}} = 4.90\mathrm{eV}$, [14] it is evident that the SBH of GaAs should be higher than that of InAs with any metal, to which our simulation results agree. This difference in the magnitude of SBH results in the considerable difference in the specific contact resistivity (Table 6.1): while the contact resistivity of Ni/InAs system lies on the order $10^{-9}\Omega\text{-cm}^2$, the contact resistivity of Ni/GaAs system lies at least two orders of magnitude higher.

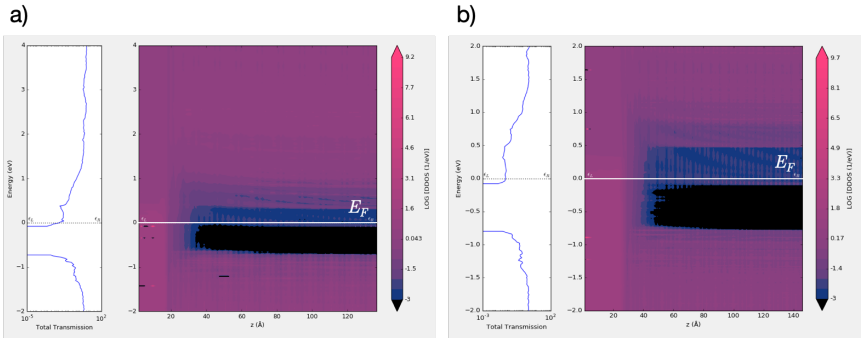

Figure 6.3: The transmission spectrum and local density of states (LDOS) for: (a) Ni/$In_{0.5}Ga_{0.5}As$(100) (As-terminated); (b) Ni/$In_{0.3}Ga_{0.7}As$(100) (As-terminated).

Last but not least, we focus on two more technologically relevant intermediate compositions: $In_{0.5}Ga_{0.5}As$ and $In_{0.3}Ga_{0.7}As$. In order to investigate the effect of the composition only, we fix the surface termination to arsenic (As) layer for both configurations. The calculated LDOS are shown in Fig. 6.3. The SBH due to band bending of $In_{0.3}Ga_{0.7}As$ is significantly smaller than that of GaAs, but still greater than that of $In_{0.5}Ga_{0.5}As$. The contact resistivity, as a consequence, follows the same trend (Table 6.2): as the indium concentration increases, the contact resistivity decreases.

6.3.2 Semiconductor surface termination

The atomic termination of a surface directly determines the chemical properties of the surface. In particular, different surface terminations for the semiconductor result in different bonding charge distribution when interfaced with metal, which leads to distinct transmission characteristics and specific contact resistivities.

This general rule is best illustrated by the difference in specific contact resistivity of Ni/(As-terminated GaAs(100)) and Ni/(Ga-terminated GaAs(100)) interfaces. (Table 6.1) From Fig. 6.2, it is clear that this difference in ρ_c is a direct consequence of the degree of band bending at the interface; for Ga-terminated GaAs(100) surface, the SBH is higher than for As-terminated GaAs(100) surface when interfaced with Ni (by about 0.2 eV). This difference in SBH cannot be explained by the Schottky-Mott rule alone, as the electron affinity of both the metal and the semiconductor are fixed. Rather, it is attributed to the presence of metal-induced gap states (MIGS) [15, 16] resulted from the exponentially-decaying metallic wavefunctions into the band gap of the semiconductor near the interface boundary. For the Ga-terminated GaAs(010) surface, the MIGS comes mainly from the metallic Ni-Ga bonds; compared with the Ni-As bonds for the As-terminated GaAs(100) surface, the Ni-Ga bonds contribute more wavefunctions to the MIGS states, resulting in a greater extent of Fermi level pinning and hence a lower Fermi level E_F relative to the conduction band edge at the interface.

On the other hand, for InAs, the contact resistivities are almost equal for the As-terminated and the In-terminated InAs(100) surface. This could be attributed to the fact that the Fermi level lies just 10 meV above the conduction band for GaAs, but 300 meV above the conduction band for InAs, due to the significantly lower energy level of the conduction band edge of InAs compared to that of GaAs. [17] This implies that the Fermi level would not be pinned by the **mid-gap** MIGS states for metal-InAs interface, in contrast to the metal-GaAs interface, as indicated by experimental measurements. [18]

ρ_c (Ω-cm^2)	Ni/GaAs (100)	Ni/InAs (100)
As-terminated	4.24×10^{-7}	3.28×10^{-9}
Ga/In-terminated	4.24×10^{-5}	3.42×10^{-9}

Table 6.1: Specific contact resistivity of commensurate Ni/GaAs(100) and Ni/InAs(100) interface, with different surface termination on the semiconductor side at 1×10^{19}cm^{-3} active doping concentration.

ρ_c (Ω-cm^2)	Ni/InAs (100)	Ni/$In_{0.5}Ga_{0.5}As$ (100)	Ni/$In_{0.3}Ga_{0.7}As$ (100)
As-terminated	3.28×10^{-9}	5.20×10^{-9}	1.45×10^{-8}

Table 6.2: Specific contact resistivity of commensurate Ni/$In_{1-x}Ga_xAs$(100) at three different compositions: InAs, $In_{0.5}Ga_{0.5}As$, and $In_{0.3}Ga_{0.7}As$, with As termination at 1×10^{19}cm^{-3} active doping concentration.

6.3.3 Doping concentration in semiconductor

Doping is one of the most effective factors that determines the contact resistance. For Ohmic contacts, the contact resistivity ρ_c is negatively correlated with the doping concentration N_d via the relation $\rho_c \propto \exp(N_d^{-1/2})$ (Eq. 2.11). For this study, we choose InAs as the semiconductor, as we have shown that ρ_c of Ni/InAs interface is almost independent of the surface termination. We investigate selected experimentally relevant doping concentrations, namely 1×10^{19}cm^{-3}, 2×10^{19}cm^{-3}, 4×10^{19}cm^{-3}, 6×10^{19}cm^{-3}, 8×10^{19}cm^{-3}, and 1×10^{20}cm^{-3}.

The results are shown in Fig. 6.4. It is clear that ρ_c of Ni-InAs interface follows the correct decreasing trend as the doping concentration in InAs increases. As the doping concentration approaches the 1×10^{20}cm^{-3} limit, the decrease of ρ_c slows down exponentially, and ρ_c reaches the ITRS threshold of $1 \times 10^{-9}\Omega$-cm^2 at $N_d \geq 6 \times 10^{19}$cm^{-3}. The calculated trend and values agree well with those

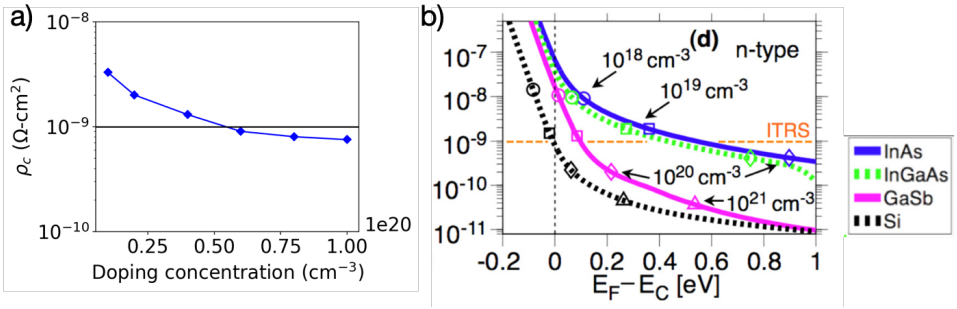

Figure 6.4: Contact resistivity (ρ_c) of Ni/InAs interface as a function of active doping concentration N_d in InAs. (b) is reproduced from Fig. 3(c), [19].

of previous simulations of the Landauer limit of ρ_c for InAs. [19, 7] According to these simulations, the Landauer-limit contact resistivity of $In_{0.5}Ga_{0.5}As$ is approximately equal to that of InAs. Hence, it is reasonable to imply that the contact resistivity of metal-$In_{0.5}Ga_{0.5}As$ interface can, at least in principle, reach the ITRS requirement of $1 \times 10^{-9}\Omega$-cm^2 for the doping level above 6×10^{19}cm^{-3}. Since experiments have shown that achieving active post-annealing doping concentration of more than 1.5×10^{19}cm^{-3} is a challenging task, we have to resort to other strategies such as optimizing the interfacial structure to further lower the contact resistivity. [20]

6.3.4 Semiconductor compositional grading

In Sec. 6.3.1, we found that the contact resistivity of an Ni/InGaAs interface is negatively correlated with the concentration of indium in InGaAs. Nonetheless, it may not be completely advantageous to adopt the pure InAs as the source/drain material as intuition suggests, as intermediate compositions such as $In_{0.5}Ga_{0.5}As$ possess additional benefits such as lattice matching with InP

substrate. Ideally, we would like to combine the low SBH of the InAs with the lattice matching of $In_{0.5}Ga_{0.5}As$ in a single device. This could in principle be achieved through compositional grading, where the indium concentration gradually decreases from 100% to 50% along the transport direction moving away from the metal-semiconductor interface. Introducing compositional grading has been shown to effectively reduce the contact resistivity of metal/III-V contacts. [21, 22]

Fig. 6.5 shows the LDOS for the Ni/InGaAs interface with both uniform $In_{0.5}Ga_{0.5}As$ and graded $In_{1-x}Ga_xAs$, with doping concentration $N_d = 1 \times 10^{19}cm^{-3}$ in the semiconductor region.. To simulate the linearly graded $In_{1-x}Ga_xAs$, we increase the number of indium atoms in each transverse atomic layer in a gradual manner as the location moves further away from the interface. As is expected, the Fermi level near the interface of the Ni/(graded InGaAs) corresponds to that of the Ni/InAs interface, whereas the Fermi level near the bulk InGaAs electrode corresponds to that of the pure doped InGaAs. The Ni/InAs interface ensures that the Fermi level is not affected by the midgap MIGS states that could be present near the Ni/InGaAs interface, while the smooth transition from InAs to $In_{0.5}Ga_{0.5}As$ results in a gradual decrease of strain in the transverse direction, which induces less inelastic scattering compared with the abrupt InAs/$In_{0.5}Ga_{0.5}As$ junction. Compared with uniform $In_{0.5}Ga_{0.5}As$, graded $In_{1-x}Ga_xAs$ lowers the contact resistivity with Ni by 19% at $N_d = 1 \times 10^{19}cm^{-3}$. (Table 6.3)

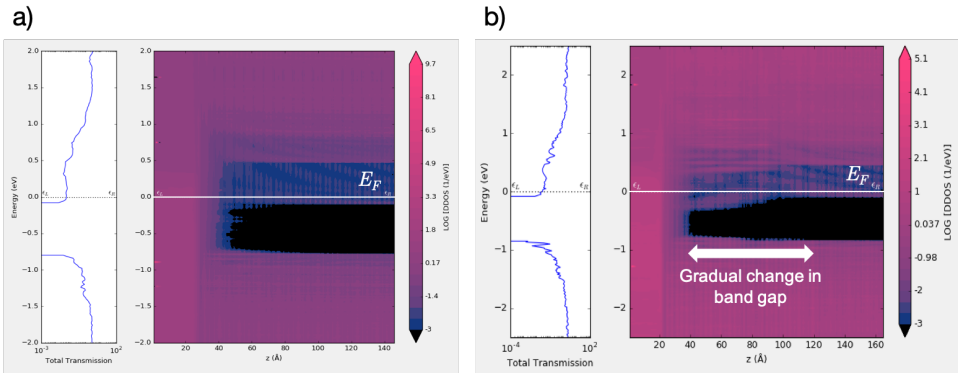

Figure 6.5: The transmission spectrum and local density of states (LDOS) for: (a) Ni/InAs(100) (As-terminated); (b) Ni/$In_xGa_{1-x}As$(100) (x changes linearly from 1 at the interface to 0.5 at the electrode) (As-terminated).

ρ_c (Ω-cm^2)	Ni/$In_{0.5}Ga_{0.5}As$ (100)	Ni/($In_{1-x}Ga_xAs$) (100)
As-terminated	5.20×10^{-9}	4.18×10^{-9}

Table 6.3: Specific contact resistivity of commensurate Ni/$In_{0.5}Ga_{0.5}As$ (100) and Ni/(linearly graded $In_{1-x}Ga_xAs$) (100), with As termination at 1×10^{19}cm^{-3} active doping concentration.

6.3.5 Metal-semiconductor alloying

The final factor of consideration in this work is the metal-semiconductor alloying. Common contact metals such as tungsten, titanium, nickel and cobalt are well known to react with silicon at moderate temperatures to form intermetallic silicides. [23, 24] Such metallization reaction is a standard procedure in silicon processing to reduce the source/drain contact resistance, due to its ability to produce a more uniform interfacial crystal structure compared to the nonalloyed interface. [23, 24] For III-V substrates, metallization is much less studied compared with silicon; notably, such procedure has been demonstrated with Nickel. [25, 26] Like with the self-aligned silicide, the self-aligned Ni-InGaAs

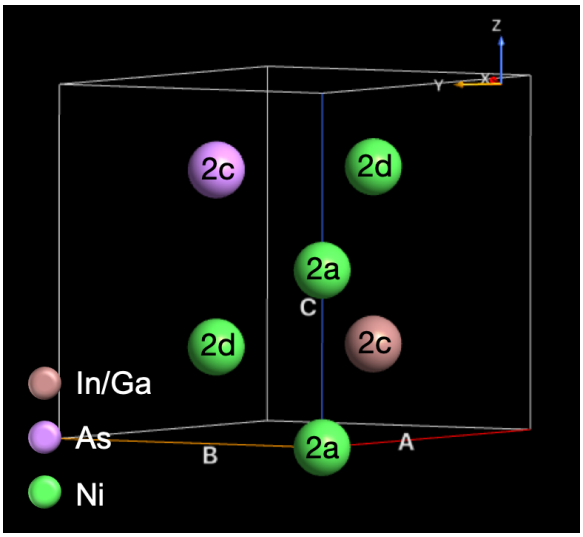

Figure 6.6: The unit cell of $Ni_xInGaAs$ ($x = 2, 3, 4$). Ni atoms occupy $2a$ sites and possibly $2d$ sites; In/Ga and As atoms occupy $2c$ sites.

has been shown to reduce the source/drain contact resistivity. [27] The Ni-InGaAs alloy has been experimentally observed to adopt a NiAs (B8) crystal structure, with the Bravais lattice assuming hexagonal symmetry. In the unit cell of Ni-InGaAs, Ni atoms can occupy up to four unique lattice positions: (0, 0, 0), (0, 0, 1/2) ($2a$ sites), (2/3, 1/3, 3/4), (1/3, 2/3, 1/4) ($2d$ sites), whereas the In/Ga and As atom respectively occupy the lattice sites (1/3, 2/3, 1/4) and (2/3, 1/3, 3/4) ($2c$ sites). (Fig. 6.6) [28, 29] The composition x of $Ni_xInGaAs$ can range from 2 to 4, depending on the thermal processing condition on the as-deposited metal/InGaAs contact. The $2a$ sites must be fully occupies, whereas the $2d$ sites may or may not be fully occupied depending on the Ni composition x.

In our study, we choose Ni_2InAs as the contact metal in the device model. The dimensions of the optimized unit cell are $a = b = 3.85$Å, $c = 5.29$Å, which agree with the experimentally measured values $a = b = 3.81$Å, $c = 5.13$Å for $Ni_3InGaAs$. [30] For the interface, we adopt the experimentally observed orientation $Ni_2InAs(10\bar{1}0)$-InAs(100), $Ni_2InAs[0001]$-InAs$[\bar{1}10]$. [29] The atomic

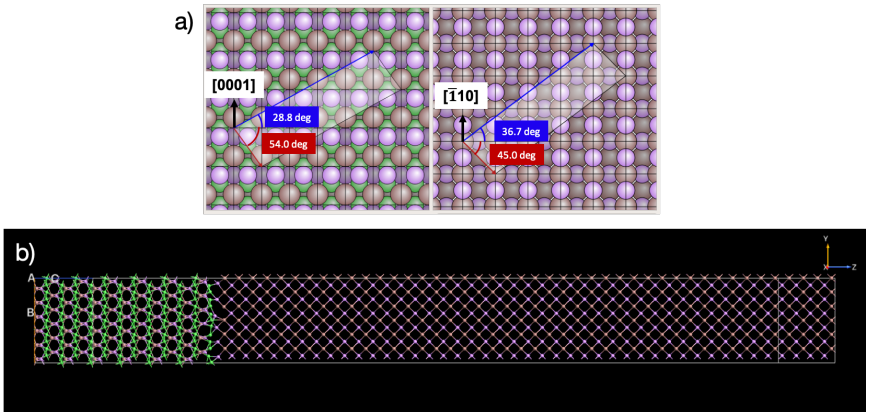

Figure 6.7: (a) The commensurate surface unit cells of Ni_2InAs and InAs satisfying the orientation relation $Ni_2InAs(10\bar{1}0)$-InAs(100), $Ni_2InAs[0001]$-InAs$[\bar{1}10]$; (b) the two-probe configuration of an Ni_2InAs/InAs interface with the orientation above.

structure of the Ni_2InAs/InAs device is shown in Fig. 6.7. Note that the alignment $Ni_2InAs[0001]$-InAs$[\bar{1}10]$ is not exact, as there exists a 1.8° difference in the transverse cell angles, leading to a shear strain ϵ_{12} of 1.16% in Ni_2InAs on a fixed InAs surface. The normal strain ϵ_{11} and ϵ_{22} induced in Ni_2InAs are −3.61% and 1.33%, respectively, both beneath the 4% threshold.

Fig. 6.8 shows the LDOS for the Ni_2InAs/InAs and Ni/InAs interface, with doping concentration $N_d = 1\times10^{19}\text{cm}^{-3}$ in the semiconductor region. Compared with the Ni/InAs interface, we find more MIGS states penetrating the interface from the metal to the semiconductor. These MIGS states, however, does not affect the transmission characteristics near the Fermi level, as the position of the Fermi level lies well above the conduction band of InAs. Moreover, the intensity of the LDOS changes gradually from high to low across the Ni_2InAs/InAs interface, whereas at the Ni/InAs interface the change is abrupt. Despite the advantage, the contact resistivity of Ni_2InAs/InAs is slightly higher than that of Ni/InAs (Table 6.4). This may be due to the fact that, as an intermetallic com-

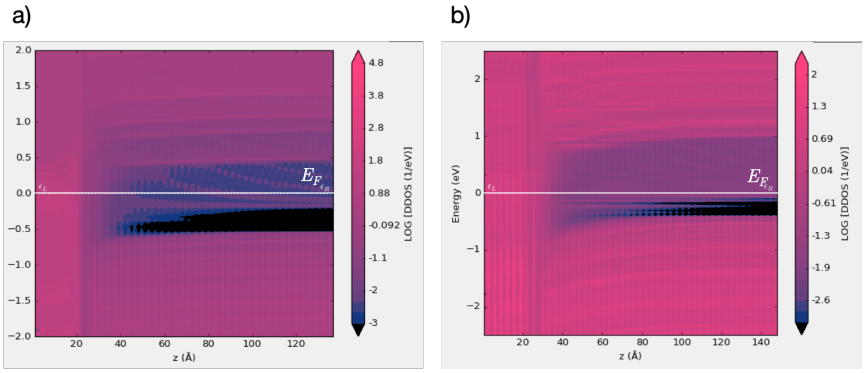

Figure 6.8: The local density of states (LDOS) for: (a) Ni/InAs(100) (As-terminated); (b) Ni_2InAs/InAs(100) (As-terminated).

pound, Ni_2InAs has lower electrical conductivity compared with pure Ni (as indicated by the lower intensity in the LDOS). In practice, it is often observed that unlike many M_xSi/Si interfaces (M = metal), the Ni-InGaAs reaction does not usually produce a smooth interface; this rough interface causes the contact resistivity to be much higher (in the $10^{-6}\Omega$-cm^2 range) [31, 32] than the predicted value in our work. In recent years, researchers have found that ammonium sulfide surface treatment [33] or insertion of a conducting interlayer between NiInGaAs and InGaAs [5, 34] could significantly reduce ρ_c down to $\sim 10^{-8}\Omega$-cm^2.

ρ_c (Ω-cm^2)	Ni/InAs (100)	Ni_2InAs/InAs (100)
As-terminated	3.28×10^{-9}	4.26×10^{-9}

Table 6.4: Specific contact resistivity of commensurate Ni/InAs (100) and Ni_2InAs/InAs (100) interface, with As termination at 1×10^{19}cm^{-3} active doping concentration.

6.4 Conclusions

We have conducted a comprehensive computational study of the impact of five factors (semiconductor alloy composition, semiconductor surface termination, doping concentration in semiconductor, semiconductor compositional grading, metal-semiconductor alloying.) on the contact resistivity of the Ni/In(Ga)As interface. Our key findings are: (1) increasing indium concentration helps reduce contact resistivity; (2) As-terminated In(Ga)As has lower contact resistivity than cation-terminated In(Ga)As; (3) higher active doping concentration leads to lower contact resistivity; (4) compositional grading with increasing indium concentration towards the Ni/In(Ga)As interface lowers the contact resistivity; and (5) alloying of In(Ga)As with Ni alone does not have noticeable impact on supressing contact resistivity. To better assist experimental efforts on improving metal/InGaAs contact resistivity, it is vital to investigate other material- and structure-dependent factors as the next step, such as interfacial roughness and inclusion of conducting interlayers.

Bibliography

[1] Gordon E Moore et al. Cramming more components onto integrated circuits.

[2] Jesús A Del Alamo, Dimitri A Antoniadis, Jianqiang Lin, Wenjie Lu, Alon Vardi, and Xin Zhao. Nanometer-scale iii-v mosfets. *IEEE Journal of the Electron Devices Society*, 4(5):205–214, 2016.

[3] Uttam Singisetti, Mark A Wistey, Jeramy D Zimmerman, Brian J Thibeault, Mark JW Rodwell, Arthur C Gossard, and Seth R Bank. Ultralow resistance in situ ohmic contacts to ingaas/inp. *Applied Physics Letters*, 93(18):183502, 2008.

[4] Ashish K Baraskar, Mark A Wistey, Vibhor Jain, Uttam Singisetti, Greg Burek, Brian J Thibeault, Yong Ju Lee, Arthur C Gossard, and Mark JW Rodwell. Ultralow resistance, nonalloyed ohmic contacts to n-in ga as. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 27(4):2036–2039, 2009.

[5] Meng Li, Jeyoung Kim, Jungwoo Oh, and Hi-Deok Lee. Reduction of contact resistance between ni–ingaas and n-ingaas by ge2sb2te5 interlayer. *Applied Physics Express*, 10(4):041201, 2017.

[6] Saeid Masudy-Panah, Ying Wu, Dian Lei, Annie Kumar, Yee-Chia Yeo, and Xiao Gong. Nanoscale metal-ingaas contacts with ultra-low specific contact resistivity: Improved interfacial quality and extraction methodology. *Journal of Applied Physics*, 123(2):024508, 2018.

[7] Ashish Baraskar, AC Gossard, and Mark JW Rodwell. Lower limits to metal-semiconductor contact resistance: Theoretical models and experimental data. *Journal of Applied Physics*, 114(15):154516, 2013.

[8] Quantumwise A/S. Atomistix toolkit version 2017. `www.quantumwise.com`, cited November 2019.

[9] John P Perdew, Adrienn Ruzsinszky, Gábor I Csonka, Oleg A Vydrov, Gustavo E Scuseria, Lucian A Constantin, Xiaolan Zhou, and Kieron Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical review letters*, 100(13):136406, 2008.

[10] Axel D Becke and Erin R Johnson. A simple effective potential for exchange. 2006.

[11] Fabien Tran and Peter Blaha. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Physical review letters*, 102(22):226401, 2009.

[12] W Schottky. Deviations from ohm's law in semiconductors. *Phys. Z*, 41:570, 1940.

[13] NF Mott. Note on the contact between a metal and an insulator or semiconductor. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 34, pages 568–572. Cambridge University Press, 1938.

[14] Semiconductors on nsm. http://www.ioffe.ru/sva/nsm/semicond/, cited November 2019.

[15] Volker Heine. Theory of surface states. *Physical Review*, 138(6A):A1689, 1965.

[16] Steven G Louie, James R Chelikowsky, and Marvin L Cohen. Ionicity and the theory of schottky barriers. *Physical Review B*, 15(4):2154, 1977.

[17] Rita Magri, Alex Zunger, and H Kroemer. Evolution of the band-gap and band-edge energies of the lattice-matched ga in as sb/ ga sb and ga in as sb/ in as alloys as a function of composition. *Journal of applied physics*, 98(4):043701, 2005.

[18] LÖ Olsson, CBM Andersson, MC Hkansson, J Kanski, L Ilver, and Ulf O Karlsson. Charge accumulation at inas surfaces. *Physical review letters*, 76(19):3626, 1996.

[19] Jesse Maassen, C Jeong, A Baraskar, M Rodwell, and M Lundstrom. Full band calculations of the intrinsic lower limit of contact resistivity. *Applied Physics Letters*, 102(11):111605, 2013.

[20] Jian Zhang, Lin-Lin Wang, Hao Yu, Clement Merckling, Yves Mols, Abhitosh Vais, Siva Ramesh, Tsvetan Ivanov, Marc Schaekers, Naoto Horiguchi, et al. Effective contact resistivity reduction for mo/pd/n-in 0.53 ga 0.47 as contact. *IEEE Electron Device Letters*, 40(11):1800–1803, 2019.

[21] Takumi Nittono, Hiroshi Ito, Osaake Nakajima, and Tadao Ishibashi. Extremely low resistance non-alloyed ohmic contacts to n-gaas using compositionally graded inxga1-xas layers. *Japanese Journal of Applied Physics*, 25(10A):L865, 1986.

[22] Takumi Nittono, Hiroshi Ito, Osaake Nakajima, and Tadao Ishibashi. Non-alloyed ohmic contacts to n-gaas using compositionally graded inxga1-xas layers. *Japanese journal of applied physics*, 27(9R):1718, 1988.

[23] LJ Chen. Metal silicides: An integral part of microelectronics. *Jom*, 57(9):24–30, 2005.

[24] Raymond T Tung. Silicides for s/d contacts. In KHJ Buschow, editor, *Encyclopedia of Materials: Science and Technology*, pages 8479–8486. Elsevier, Amsterdam; New York, 2001.

[25] SH Kim, M Yokoyama, N Taoka, R Iida, S Lee, R Nakane, Y Urabe, N Miyata, T Yasuda, H Yamada, et al. Self-aligned metal source/drain in x ga 1-x as n-mosfets using ni-ingaas alloy. In *2010 International Electron Devices Meeting*, pages 26–6. IEEE, 2010.

[26] SangHyeon Kim, Masafumi Yokoyama, Noriyuki Taoka, Ryo Iida, Sunghoon Lee, Ryosho Nakane, Yuji Urabe, Noriyuki Miyata, Tetsuji Yasuda, Hisashi Yamada, et al. Self-aligned metal source/drain inxga1-xas

n-metal–oxide–semiconductor field-effect transistors using ni–ingaas alloy. *Applied Physics Express*, 4(2):024201, 2011.

[27] SangHyeon Kim, Masafumi Yokoyama, Ryosho Nakane, Osamu Ichikawa, Takenori Osada, Masahiko Hata, Mitsuru Takenaka, and Shinichi Takagi. High-performance inas-on-insulator n-mosfets with ni-ingaas s/d realized by contact resistance reduction technology. *IEEE Transactions on Electron Devices*, 60(10):3342–3350, 2013.

[28] Ivana, Yong Lim Foo, Xingui Zhang, Qian Zhou, Jisheng Pan, Eugene Kong, Man Hon Samuel Owen, and Yee-Chia Yeo. Crystal structure and epitaxial relationship of ni4ingaas2 films formed on ingaas by annealing. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, 31(1):012202, 2013.

[29] Seifeddine Zhiou, Tra Nguyen-Thanh, Philippe Rodriguez, Fabrice Nemouchi, Laetitia Rapenne, Nils Blanc, Nathalie Boudet, and Patrice Gergaud. Reaction of ni film with in0. 53ga0. 47as: Phase formation and texture. *Journal of Applied Physics*, 120(13):135304, 2016.

[30] C Perrin, E Ghegin, S Zhiou, F Nemouchi, Ph Rodriguez, P Gergaud, P Maugis, D Mangelinck, and K Hoummada. Formation of ni3ingaas phase in ni/ingaas contact at low temperature. *Applied Physics Letters*, 109(13):131902, 2016.

[31] Shlomo Mehari, Arkady Gavrilov, Shimon Cohen, Pini Shekhter, Moshe Eizenberg, and Dan Ritter. Measurement of the schottky barrier height between ni-ingaas alloy and in0. 53ga0. 47as. *Applied Physics Letters*, 101(7):072103, 2012.

[32] Xingui Zhang, Hua Xin Guo, Xiao Gong, and Yee-Chia Yeo. Multiple-gate in0. 53ga0. 47as channel n-mosfets with self-aligned ni-ingaas contacts. *ECS Journal of Solid State Science and Technology*, 1(2):P82–P85, 2012.

[33] Michael Abraham, Shih-Ying Yu, Won Hyuck Choi, Rinus TP Lee, and Suzanne E Mohney. Very low resistance alloyed ni-based ohmic contacts to inp-capped and uncapped n+-in0. 53 ga0. 47as. *Journal of Applied Physics*, 116(16):164506, 2014.

[34] Sunil Babu Eadi, Jeong Chan Lee, Hyeong-Sub Song, Jungwoo Oh, and Hi-Deok Lee. Critical role of thulium metal interlayer in ultra-low contact resistance reduction in ni-ingaas/n-ingaas for n-mosfets. *Vacuum*, 166:151–154, 2019.

CHAPTER 7

AB INITIO STUDIES OF SEGREGATION OF *N*-TYPE DOPANTS AND VACANCIES NEAR A β-GA_2O_3 (010) SURFACE

The work presented in this chapter is published in: **J. Wang**, and P. Clancy, submitted to Applied Surface Science (2019).

7.1 Introduction

The β-phase of gallium oxide (β-Ga_2O_3) is a wide band-gap semiconductor with a variety of potential applications, especially in power electronic devices and optoelectronic devices. [1, 2, 3] Such applications benefit from its wide band gap (4.8 eV) and relatively high breakdown field (8 MV/cm).[1] Successful performance of β-Ga_2O_3 in these applications is critically determined by the efficiency of *n*-type doping. For example, the electrical performance of a Schottky barrier diode or high-voltage field effect transistors incorporating Ga_2O_3 would increase by decreasing the Schottky barrier height at the metal-Ga_2O_3 interface; this is achievable by increasing the active doping concentration. Experiments [4, 5, 6] and first principles calculations [7, 8] have demonstrated that Si, Ge, and Sn are shallow *n*-type donors in *bulk* β-Ga_2O_3. A number of intrinsic defects also play important roles: positively charged oxygen vacancies (V_O) act as deep donors, [7, 9] while negatively charged gallium vacancies (V_{Ga}) [7, 10, 11] compensate the donors at degenerate *n*-type doping levels, hence limiting the maximum attainable free electron concentration.

In electronic devices, it is important to be able to have effective control of electrical conductivity by the ability to dope across a wide range of concentra-

tions, not only inside the bulk region but also near the surface. The specific contact resistivity at the interface with the metal layer must be sufficiently low to achieve satisfactory performance. For a doped semiconductor, two major factors can negatively affect the electronic properties of its surface. First, when the surface is rich in electrically compensating intrinsic defects such as vacancies, these defects can behave as charge traps that contribute to the excess surface charge density, causing band-bending and carrier depletion near the surface. Second, it may be thermodynamically favorable for the dopant atoms to segregate towards the surface, and eventually become deactivated by clustering together and forming a separate, secondary phase; this leads to reduction in the active doping concentration in the bulk semiconductor. Both phenomena have been observed in experiments on β-Ga_2O_3 surfaces during growth or after thermal annealing. [12, 13, 14, 15, 16] Clearly, such segregation behavior of defects and dopants needs to be suppressed in order to achieve the maximum performance potential for β-Ga_2O_3.

Previous computational work regarding *n*-type dopants and intrinsic defects in β-Ga_2O_3 has focused on the bulk region. [7, 8, 9, 11, 17] In particular, work by Lany [8] offered an explanation of the formation of secondary phases of dopant oxide inside Ga_2O_3 from the perspective of defect-dopant thermodynamics in bulk Ga_2O_3. However, this cannot be the full picture as formation energies of defects and dopants on the surface could be drastically different from those in the bulk. In this work, we aim to redress this deficiency using a comprehensive computational analysis of the following: (1) the energetics of charged shallow donors (Si, Ge, Sn) and intrinsic defects (interstitials, vacancies) in the (010) surface of β-Ga_2O_3 in comparison to the bulk; (2) the driving force for segregation for these dopants towards a β-Ga_2O_3 (010) surface; and (3) the strategy of re-

ducing segregation tendency of dopants in Ga_2O_3. Such knowledge will be important in understanding the limiting factors of device performance in β-Ga_2O_3 materials.

7.2 Computational methods

We use the plane-wave Density Functional Theory (DFT) code QUANTUM ESPRESSO with the PBEsol functional and ultrasoft pseudopotentials [18] for our calculations. An energy cutoff of 816 eV was chosen to achieve good convergence of both energy and forces. A $2 \times 8 \times 4$ **k**-point grid is used to obtain optimized structural parameters for the Ga_2O_3 primitive cell. This produces values for the primitive cell lengths and angles: a = 12.30Å, b = 3.05Å, c = 5.82Å, $\alpha = 90°$, $\beta = 103.7°$, and $\gamma = 90°$. These lattice parameter values are in excellent agreement with experimental values: a = 12.214(3)Å, b = 3.0371(9)Å, c = 5.7981(9)Å, $\beta = 103.83(2)°$. [19] For defect calculations in bulk Ga_2O_3, we extend the cell into a 160-atom supercell (Fig. 7.2(a)) containing $1 \times 2 \times 4$ of the primitive cell (Fig. 7.1), and a $2 \times 2 \times 2$ Monkhorst-Pack **k**-point grid is used for the bulk supercell calculations. This results in approximately equal lengths in each crystallographic direction, which ensures that the defect image charge interaction will not become too large in any direction.

The thermodynamic preference for any "impurity" atom (whether dopant or defect) to occupy a given position in bulk β-Ga_2O_3 is quantified by its bulk formation energy, E^f_{bulk}, given by Eq. (4.1). The charged defect correction is performed *via* the Freysoldt-Neugebauer-van de Walle (FNV) scheme, [20, 21] as implemented in CoFFEE (Corrections For Formation Energy and Eigenvalues

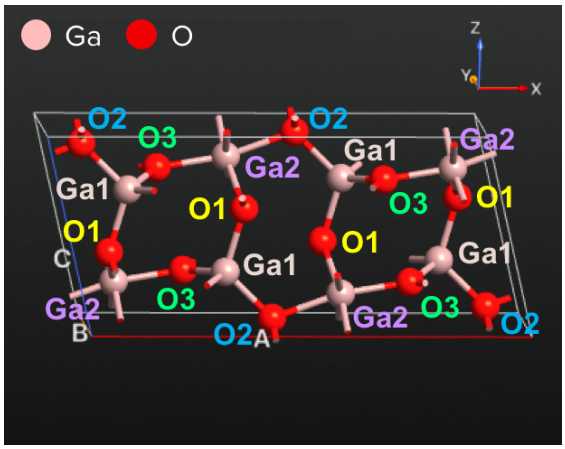

Figure 7.1: 20-atom primitive unit cell of β-Ga_2O_3 with a $C2/m$ space group. Each atom belongs to one of the five types of non-equivalent atoms (Ga_1. Ga_2, O_1, O_2, O_3).

for charged defect simulations) [22]. In order to correct the error in the band gap commonly experienced in DFT, which will affect the formation energies of charged dopant/defect species, we use the HSE06 functional. [23] This functional is known to yield accurate band gaps for semiconductors and insulators. However, instead of performing large-scale HSE06 calculations for the supercells, we adopt a much simpler and more efficient scheme by applying a "scissor operator," which simply shifts the position of the conduction and valence band edges to the band edge positions of the bulk Ga_2O_3 unit cell. In our case, the HSE06 band gap happens to reproduce the experimental value of the band gap of Ga_2O_3, 4.8 eV. We will show in Sec. 2 that our results calculated using this scheme show excellent agreement with previous results by Varley *et al.* [7]

Varley *et al.* [7] used *ab initio* calculations to demonstrate that many group IV species (Si, Ge, Sn) act as shallow donors in bulk Ga_2O_3, while V_{Ga} and V_O stabilize in charge state -3 and 0, respectively, under *n*-type doping conditions. In order to test the validity of our approach, we reproduced the bulk dopant and vacancy formation energies observed by Varley *et al.* [7] (see Sec. 7.3.1)

However, experimental results have shown that Sn tends to segregate toward the (010) surface of β-Ga_2O_3 upon thermal annealing at high temperatures. [16] The extent of segregation under prolonged high-temperature annealing conditions can be so pronounced that it forms a separate SnO_2 phase with oxygen. [16] One key indicator of the impurity's tendency for surface segregation is the impurity segregation energy.

Our calculations used a 280-atom "slab" model of the Ga_2O_3, consisting of a periodic supercell configuration in which 14 atomic layers of Ga_2O_3 are sandwiched between two vacuum regions. (Fig. 7.2(b)) The thickness is chosen such that the electron density of the slab center converges to that of bulk Ga_2O_3. (see Sec. 7.3.2, Supplementary Information) The segregation energy of an impurity is then defined as the difference between the total energy of the supercell with an impurity near the slab surface and that with the same impurity in the slab center (a bulk-like region):

$$E_{\text{segr}}(D^q) = E_{\text{tot}}(D^q, \text{surf}) - E_{\text{tot}}(D^q, \text{bulk}). \tag{7.1}$$

The segregation energy, defined in Eq. (7.1), represents the degree of *thermodynamic* preference for a dopant species, D^q, with a charge state q to be located near the β-Ga_2O_3(010) surface, rather than in the bulk. Dopant segregation is determined by both thermodynamic and kinetic considerations. Due to the highly complex lattice structures of β-Ga_2O_3(010) surface, a complete description of migration kinetics from the bulk to the (010)-surface of Ga_2O_3 lies beyond the scope of this paper. Hence, in this work, we focus only on the thermodynamic aspects of dopant segregation.

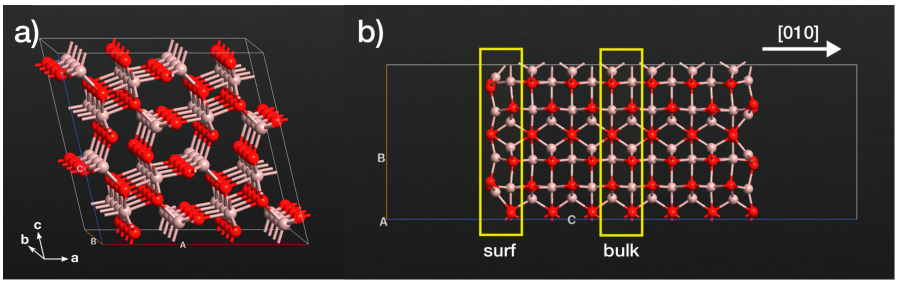

Figure 7.2: Images depicting (a) a 160-atom bulk β-Ga_2O_3 supercell; (b) 280-atom slab model of the β-Ga_2O_3(010) surface, with surface and bulk-like regions outlined using yellow rectangles.

To obtain the segregation energy, we use a 320-atom β-Ga_2O_3 (010) surface slab model, as shown in Fig. 7.2(b). The model consists of seven layers of the Ga_2O_3 unit cell stacked together in the [010]-direction, with each layer containing 4×2 of the β-Ga_2O_3 unit cell in the (010) plane. The dimensions of the slab are 11.65Å and 12.30Å in the transverse directions, and 36.39Å in the longitudinal ([010]) direction with a vacuum space of 15Å. Note that each *atomic* layer of the slab is equivalent under periodic boundary conditions, implying that the slab is nonpolar. The same density functional and cutoff energy as those used for bulk Ga_2O_3 are used. A single **k**-point at the Γ-point is used to speed up the calculations; we have verified that using a larger $2\times2\times1$ Monkhorst-Pack **k**-point grid makes no noticeable difference ($<$ 10 meV) compared to the result for a single **k**-point. We found it very difficult to converge the electron density of the setup, especially for the first few SCF cycles, due to the well-known "charge sloshing" problem resulted from the surface-atom charge spilling into the vacuum. [24] To facilitate convergence, we used a modified version of the Thomas-Fermi charge mixing scheme [25], which adaptively applies a random perturbation to the residual output electron density whenever the rate of convergence drops below a certain threshold. We have verified that, for bulk systems, our method leads to the same final total energy as that calculated using the original charge

mixing method.

We consider three different scenarios for dopant segregation towards the β-Ga_2O_3(010) surface. In the first, simplest scenario, the slab contains only one dopant atom, free to segregate from the bulk-like region to the surface. In the second scenario, an intrinsic defect in its preferred bulk charge state (under heavy *n*-type doping) is fixed inside the bulk-like region, while the dopant is free to segregate. In the final scenario, two positively charged dopant atoms are present in the slab, with one dopant atom fixed inside the bulk-like region and the other one free to segregate. These scenarios cover some of the most representative elementary mechanisms through which the *local environment* affects dopant segregation, and help distinguish the roles played by each factor (lattice strain, defect, dopant) in dopant segregation.

As first proposed by Lee *et al.* [26], the driving mechanism of dopant segregation can be decomposed into an elastic contribution and an electrostatic contribution. The elastic contribution comes from the lattice strain induced by the dopant atom, whereas the electrostatic contribution originates from the Coulombic interaction between the dopant charge and other impurity or background charges. While this decoupling scheme correctly captures the different physical origins of the segregation mechanisms, the quantitative expressions as in Lee *et al.* (Eq. (1) and (2) in [26]) have several inherent limitations. These include: (1) they only apply to certain ideal situations, such as dilute dopant concentration for the Friedel elastic energy expression, [27] and purely ionic host material for the Coulombic electrostatic energy expression; and (2) the elastic energy term describes the interaction of a *single* dopant atom with its host lattice, but fails to account for the change in lattice strain when other defects are present.

These limitations could be the reasons that the sum of the elastic and the electrostatic energies does not, in general, equal the segregation energy in Lee *et al.*'s work. In this work, we propose an alternative decomposition scheme, which can be obtained straightforwardly from DFT calculations. Our expressions for elastic and electrostatic energies of a segregating (charged) dopant atom are:

$$E_{\text{elastic}} = E_{\text{segr}}(D_*^0), \tag{7.2}$$

$$E_{\text{electrostatic}} = E_{\text{segr}}(D^q) - E_{\text{segr}}(D_*^0), \tag{7.3}$$

where $E_{\text{segr}}(D_*^0)$ is the DFT-calculated dopant segregation energy for *charge-neutral* slabs, with the relaxed atomic positions of the corresponding charged slabs. This definition recognizes the fact that defect-induced local lattice strain is, in general, dependent on the charge state of the defect. Our definition ensures that the effect of the lattice strain *as induced by the charged dopant* is correctly captured by the elastic term, from which the purely electrostatic interactions is completely separated. It is clear from equations (2) and (3) that our definition does not assume any requirement on the physical and chemical nature of the system, and that the sum of these two contributions is exactly the DFT segregation energy of the charged dopant.

Finally, in order to further quantify the underlying mechanism of elastic interactions, we adopt the "crystal orbital Hamilton population" method (COHP) for the chemical bonding analysis. [28] The projected COHP (pCOHP) [29] is an analytical scheme that partitions the electron wavefunction into regions of bonding and anti-bonding characteristics by weighting the density of states (DOS) with the Hamiltonian matrix elements projected onto atomic orbitals from the plane waves. The integrated COHP (ICOHP), defined as the integral of COHP up to the Fermi level, provides a measure of bond interaction strength

between neighboring atoms. In particular, large negative values of ICOHP indicate strong bonding (stabilizing interactions), whereas large positive ICOHP values indicate strong anti-bonding (destablizing) interactions. In this work, we use the LOBSTER package [30] to calculate the ICOHP values for all neighboring Ga-O and D-O (D = Si, Ge, Sn) bonds in the Ga_2O_3(010) slab.

7.3 Results and Discussion

7.3.1 Formation energy of shallow donors and intrinsic defects in bulk β-Ga_2O_3

We begin by validating our formation energy results for the well-studied bulk β-Ga_2O_3 against previous work. [7] The formation energies of substitutional dopants (Si, Sn, Ge) in bulk β-Ga_2O_3 are shown as a function of Fermi level in Fig. 7.3. Each colored line segment represents the most stable charge state for the particular dopant/defect species at the corresponding Fermi level. Fig. 7.3 shows that all the dopants considered here (Si, Sn, Ge) stabilize in the +1 charge state throughout the band gap, confirming that they are shallow donors in bulk β-Ga_2O_3. However, the preference of lattice sites for each donor species differ. Specifically, Si prefers a Ga_1 (tetrahedral) site over a Ga_2 (octahedral) site, while Sn prefers the Ga_2 (octahedral) site over Ga_1 (tetrahedral) site. Ge exhibits an almost equal preference for occupying either the Ga_1 (tetrahedral) or Ga_2 (octahedral) sites. These observations are consistent with the results reported by Varley *et al*. [7] These results are consistent with the fact that Si forms native oxide SiO_2 with silica structure in tetrahedral coordination, Sn forms SnO_2 in rutile

crystal structure with octahedral coordination, whereas Ge forms GeO_2 in either silica structure with tetrahedral coordination or rutile structure with octahedral coordination, depending on the pressure. [31] Echoing Varley *et al.*, [7] we confirm that the relative order of formation energies depends on growth condition: Under Ga-rich conditions, Si is the most efficient electron donor, whereas under O-rich conditions, Ge and Sn are slightly more efficient as electron donors compared to Si. Furthermore, the absolute values of the formation energies increase for all dopants from Ga-rich to O-rich growth conditions.

In Fig. 7.4, the bulk formation energies of the vacancies are plotted as a function of Fermi level. We find out that as *n*-type doping level increases, the most stable charge state for all oxygen vacancies (V_{O1}, V_{O2}, V_{O3}) transitions from a +2 state to a neutral state, with the charge transition level relative to the conduction band edge $\varepsilon(+2/0)$ = 3.67 eV, 3.09 eV, and 3.81 eV for V_{O1}, V_{O2}, and V_{O3}, respectively. These results are in excellent agreement with those in Varley *et al* [7] and Zacherle *et al* [32]. This indicates that under heavy *n*-type doping, like that considered in this work, none of the oxygen vacancies are electrically active. On the other hand, the gallium vacancies (V_{Ga1}, V_{Ga2}) become stabilized in a -3 charge state under the same doping condition, making them potentially detrimental to dopant activation at very high concentrations.

Fortunately, under Ga-rich growth conditions, which we showed to be the more favorable condition for *n*-type dopant activation, the bulk formation energies of gallium vacancies are much higher than those under O-rich conditions, implying that the inhibition effect of dopant activation is much less severe under Ga-rich growth conditions, at least in bulk Ga_2O_3. Finally, we note that under Ga-rich growth conditions, among all six donor types (Si, Ge, Sn on two

non-equivalent gallium sites) only Si_{Ga1} is more stable than the triply negatively charged gallium vacancies at the conduction band edge. From a thermodynamic point of view, this implies that, at extremely high dopant concentrations, Si is the most favorable among all three dopant species for *n*-type doping.

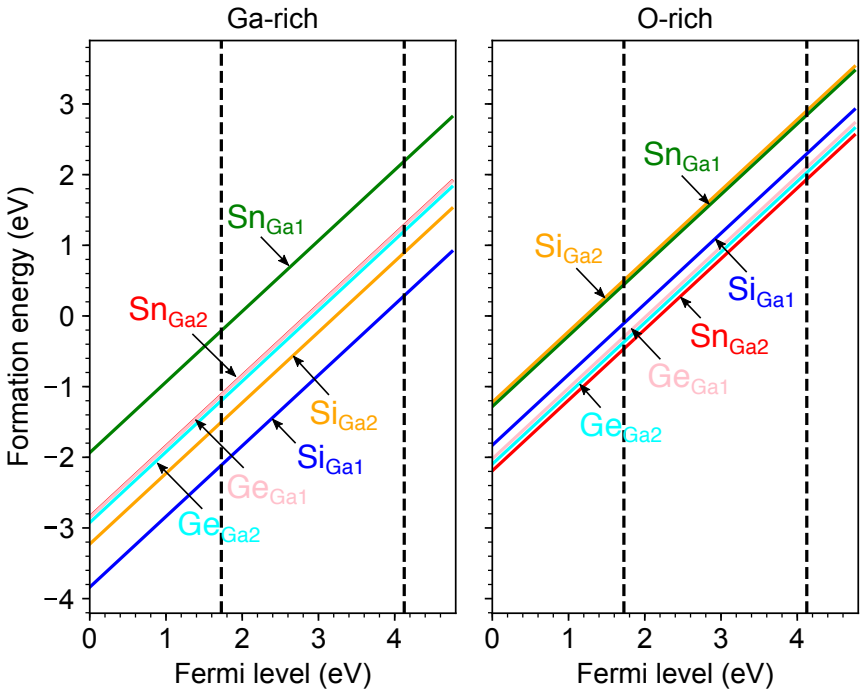

Figure 7.3: Formation energy of Si_{Ga}, Ge_{Ga}, and Sn_{Ga} in bulk β-Ga_2O_3, under gallium-rich (left) and oxygen-rich (right) growth conditions. The slope of each line segment represents the charge state of the dopant ion. The dashed lines indicate the positions of the conduction band edges calculated using DFT with the PBEsol functional, which underestimates the experimental band gap.

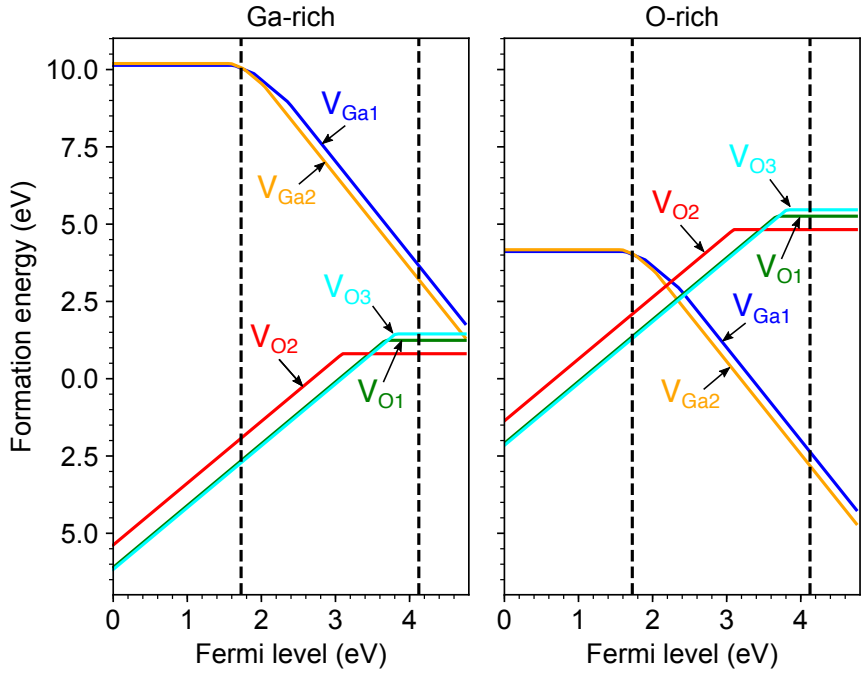

Figure 7.4: Formation energy of vacancies V_{Ga} and V_O in bulk β-Ga_2O_3, under gallium-rich (left) and oxygen-rich (right) growth conditions. The slope of each line segment represents the charge state of the dopant ion. The dashed lines indicate positions of conduction band edges calculated using DFT with PBEsol functional, which underestimates the experimental band gap.

7.3.2 Dopant segregation towards β-Ga_2O_3(010) surface

Pristine β-Ga_2O_3(010) surface

The pristine β-Ga_2O_3 (010) surface has only one unique termination; as shown in Fig. 7.1, each unit cell of β-Ga_2O_3 contains two atomic layers in the [010] direction, each layer containing two of Ga_1, Ga_2, O_1, O_2, and O_3 atoms respectively. The two atomic layers are entirely equivalent, as they differ only by a distance of $a/2$ in the [100] direction and $c/2$ in the [001] direction. Therefore, a pristine β-Ga_2O_3 (010) slab containing any number of atomic layers in the longitudinal direction is non-polar. Bermudez [33] used DFT to study the structures of sev-

eral low-index β-Ga_2O_3 surfaces, including the (010) surface. He found that the β-Ga_2O_3 (010) surface shows limited degree of reconstruction upon structural relaxation. Furthermore, he found that the atomic coordinates at the center of the β-Ga_2O_3 (010) slab converge to those of bulk β-Ga_2O_3 when the number of atomic layers $N \geq 7$. Specifically, for $N = 15$, the displacement from bulk equilibrium positions for all atomic species is less than 0.01Å at the center of the slab.

To verify that our model faithfully represents both the surface and bulk local environment, the displacement relative to bulk equilibrium (unrelaxed) atomic positions is plotted as a function of the layer index in Fig. 7.5. Our result agrees excellently with the result of Bermudez [33]. Specifically, the transverse displacement $\delta(x, y)$ of all atomic species converges to less than 0.01Å at four atomic layers beneath the surface, while the longitudinal displacement of all atomic species converges to the same level at six atomic layers beneath the surface. The macroscopic average electrostatic potential profile along the longitudinal direction (using a convolution window of 3Å) also shows a convergence towards a stable bulk-like value within five layers of distance. We have verified that increasing transverse **k**-point grid density to 2×2 shifts the positions of all atoms by no more than 0.01Å. These results confirm the validity of our relaxed slab structure as an accurate representation of both the bulk and the surface local environments. The stability of the surface can be quantified by the surface energy γ, defined as

$$\gamma = \frac{E_{\text{tot}}(\text{surf}) - \tilde{E}_{\text{tot}}(\text{bulk})}{2A} \tag{7.4}$$

where $E_{\text{tot}}(\text{surf})$ is the total energy of the slab, $\tilde{E}_{\text{tot}}(\text{bulk})$ is the scaled total energy of the corresponding bulk material with the same number of formula units as in the slab, and A is the cross-sectional surface area of the slab. Using DFT, the

formation energy of the relaxed $Ga_2O_3(010)$ surface is calculated to be 0.10 eV $Å^{-2}$, suggesting its relative stability under appropriate growth conditions.

In a crystal slab, due to presence of surface termination, atomic bonding characteristics and coordination numbers are typically changed at the surface compared to in bulk. Hence it is important to investigate the effect of local coordination on dopant/defect segregation. The locations and characters of bonding in a periodic crystal can be determined through QTAIM (quantum theory of atoms in molecules) analysis, by locating the bond critical points of electron density and calculate the Laplacian of the electron density at those points. We use the code Critic2 [34, 35] to perform such analysis. We find that the $Ga_2O_3(010)$ surface introduces a host of lattice sites with varying coordination, as listed in Table 7.1. Note that while the notations for bulk lattice sites are used for surface sites, they do not mean that these surface sites are symmetrically equivalent to those in bulk, due to reduction of coordination and geometry upon surface termination. Indeed, the coordination numbers of the surface sites are lower than those of the corresponding bulk sites.

Single-site dopants and vacancies

We considered segregation of three *n*-type shallow dopants (Si, Ge, Sn) and two vacancy species (V_{Ga}, V_O) from the bulk to the (010) surface of Ga_2O_3. We focus first on the segregation energy of each dopant or defect from a bulk lattice site to the *corresponding* surface site. Specifically, we consider only the most stable *bulk* impurity charge states under *n*-type doping: +1 for dopants, −3 for V_{Ga} and 0 for V_O.

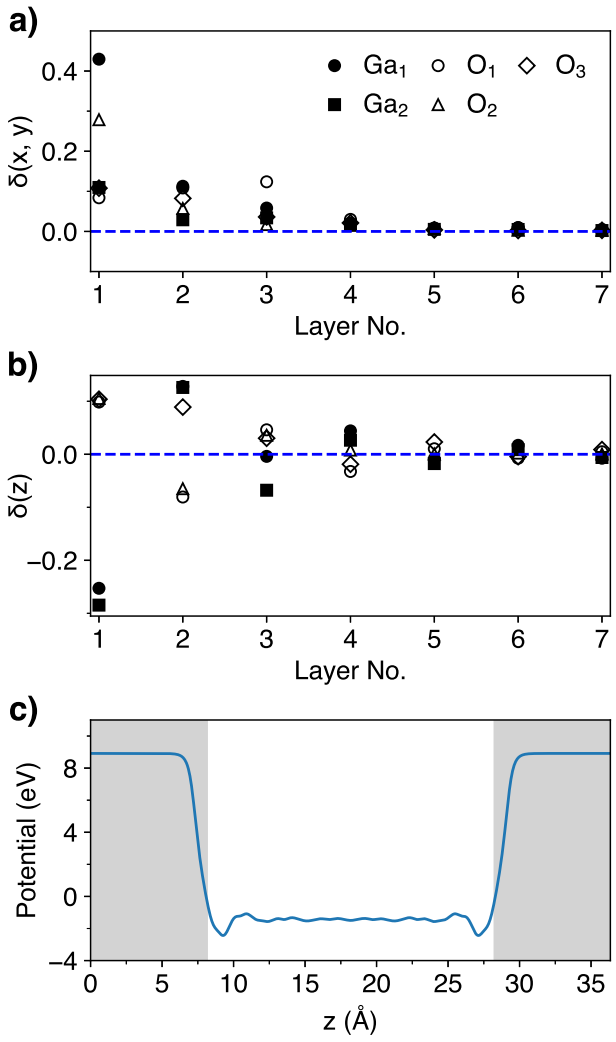

Figure 7.5: The displacement of atoms in each layer of the $Ga_2O_3(010)$ slab, with respect to bulk equilibrium positions, as a function of layer number, in (a) transverse and (b) longitudinal directions; (c) macroscopically-averaged electrostatic potential of the $Ga_2O_3(010)$ slab, as a function of longitudinal coordinate z.

In a crystal slab, surface termination introduces changes in coordination number of lattice sites compared to the bulk. Using a QTAIM analysis (quantum theory of atoms in molecules) with the code Critic2 [34, 35], we determine the coordination number of surface lattice sites of $Ga_2O_3(010)$ (listed in Table

7.1). While the notations for bulk lattice sites are used for surface sites, these surface sites are symmetrically inequivalent to their counterparts in bulk, due to reduction of coordination and geometry upon surface termination.

Lattice site	Ga_1	Ga_2	O_1	O_2	O_3
Coordination (bulk)	4	6	3	4	3
Coordination (surface)	3	4	2	2	2

Table 7.1: Lattice sites and their respective coordination numbers in the bulk and on the top layer of the (010) surface of β-Ga_2O_3.

For dopants located in the top layer of the Ga_2O_3(010) surface, the local atomic geometry of dopants are shown in Fig. 7.6. It is evident that all dopant species induce large distortions of the bonding structure of neighboring oxygen atoms during geometry optimization, due to redistribution of the electron density. In particular, for dopants located in the top surface layer, the coordination number of the dopant lattice site changes from 3-fold in the unrelaxed surface to 4-fold in the relaxed surface. The extra bond occurs between the dopant atom and its second nearest neighbor O_1 atom. This happens as a result of the significant movement of the dopant atom and its nearest-neighbor O_2 atom away from their equilibrium positions and towards the O_1 atom. This movement has the effect of stabilizing the dopant on top of the surface, since Si, Ge, and Sn all have coordination numbers greater than three in their native oxides. On the other hand, according to Table 7.2, the average length of the dopant-oxygen (D-O) bonds on the surface are shorter for all dopants compared to those in the bulk, indicating a tendency of neighboring O atoms to contract towards the dopant on the surface.

A clearer trend across the dopant species can be found when considering the average Ga-O bond length on the surface. The ratio of the average D-O bond

length to the average Ga-O bond length in the topmost surface layer increases as the row number in the periodic table increases (Si → Ge → Sn). These ratios indicate that Si experiences strong compressive strain, Ge experiences light compressive strain, and Sn experiences strong tensile strain in the Ga_2O_3(010) top surface layer. This trend is in reasonable agreement with the Shannon-Prewitt effective ionic radii of the corresponding elements (see Table 7.3) For comparison, the local atomic geometry of vacancies are shown in Fig. 7.7. Compared to dopants, the spatial extents of the distortion of neighboring atoms are much greater: certain neighboring atoms can even relax into near-interstitial positions. It can be inferred that the coordination number of different surface lattice sites, as well as the charge distribution associated with the vacancy, determines the degree of local atomic distortion.

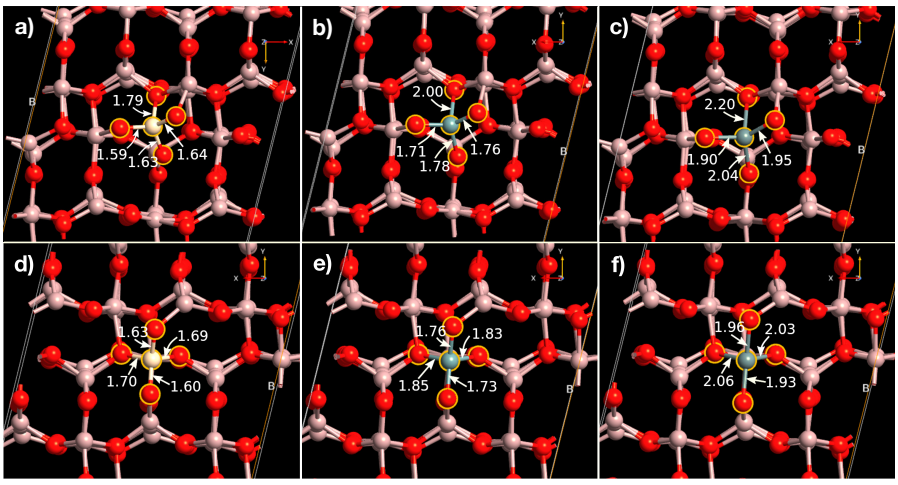

Figure 7.6: Local atomic configuration of dopants in the top surface layer of Ga_2O_3(010): (a) Si_{Ga1}, (b) Ge_{Ga1}, (c) Sn_{Ga1}, (d) Si_{Ga2}, (e) Ge_{Ga2}, and (f) Sn_{Ga2}. The dopant and its first nearest neighbors are highlighted with yellow circles. Numbers indicate the bond length between the dopant atom and the corresponding nearest-neighbor oxygen atom (in Å).

The segregation energies for each of the three dopant species from bulk to the (010) surface of Ga_2O_3 are plotted in Fig. 7.8. As the strain transitions from com-

Dopant species D^q	$\bar{r}_{\text{D-O}}$(top)/$\bar{r}_{\text{D-O}}$(bulk)	$\bar{r}_{\text{D-O}}$(top)/$\bar{r}_{\text{Ga-O}}$(top)
Si^{+}_{Ga1}	97.7%	90.1%
Si^{+}_{Ga2}	90.3%	88.2%
Ge^{+}_{Ga1}	98.4%	97.3%
Ge^{+}_{Ga2}	92.7%	95.2%
Sn^{+}_{Ga1}	97.1%	109.0%
Sn^{+}_{Ga2}	96.3%	106.1%

Table 7.2: Ratio of the average D-O bond length (D = Si, Ge, Sn) in the top surface layer to that in the bulk (second column), and the ratio of average D-O bond length vs. average Ga-O bond length in the top surface layer (third column), as calculated by DFT.

Ion species D^q	Si^{+4}	Ge^{+4}	Sn^{+4}	Ga^{+3}
$r_{\text{4-fold}}$ (pm)	26	39	55	47
$r_{\text{6-fold}}$ (pm)	40	53	69	62

Table 7.3: Ion species and their respective Shannon-Prewitt radii at lattice sites with 4-fold and 6-fold coordination. [36]

pressive to tensile as we change the dopant, the segregation energy decreases. This trend is consistent with previous computational studies of dopant segregation in crystals. [26, 37] In the bulk lattice, as the effective size of dopant atom increases, so does the elastic strain induced on the neighboring atoms, making it thermodynamically less favorable for the dopant to form. However, near the surface, such dopant-induced strain can largely be alleviated by local reconstruction, hence the elastic energy associated with the dopant is reduced. The overall effect is a decrease in the segregation energy, corresponding to a stronger tendency for the dopant to segregate towards the surface.

It is apparent in Fig. 7.8 that the segregation energy varies across distinct surface lattice sites. The relative magnitude of the segregation energy of single dopants is determined mainly by the strength of the elastic interaction between

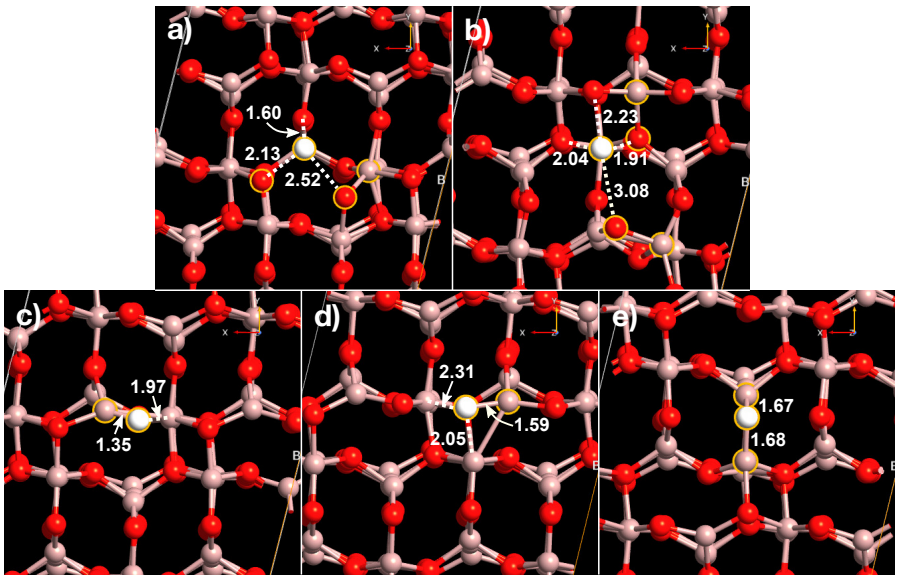

Figure 7.7: Local atomic configuration of two vacancy types in the topmost surface layer of Ga_2O_3(010): (a) V_{Ga1}, (b) V_{Ga2}, (c) V_{O1}, (d) V_{O2}, (e) V_{O3}. The vacancy (represented by a white sphere) and the atoms experiencing the most distortion are highlighted with yellow circles. Numbers indicate the bond length between the vacancy and the corresponding nearest-neighbor Ga/O atom (in Å).

the dopant and lattice. In the literature, the strength of the elastic interaction is reflected by the dopant size relative to that of the matrix atoms, as measured by the ionic/covalent radius of the atomic species. [26, 37] This approach cannot distinguish between different lattice sites. But this limitation can be resolved by a COHP analysis, which quantifies the bond strength at any lattice site. The total D-O ICOHP, defined as the sum of ICOHP values for all D-O bonds, are listed in Table 7.4. The most notable implication of Fig. 7.8 and Table 7.4 is that, for a particular dopant species in charge state q, D^q, the absolute value of E_{segr} to a particular surface site is, in general, correlated with Δ(ICOHP), the difference between the total D-O ICOHP at the surface site and that at the corresponding bulk site. This trend is followed closely by Si and Sn, but less so by Ge. The segregation energy of Ge to all non-equivalent surface Ga sites are similar in value,

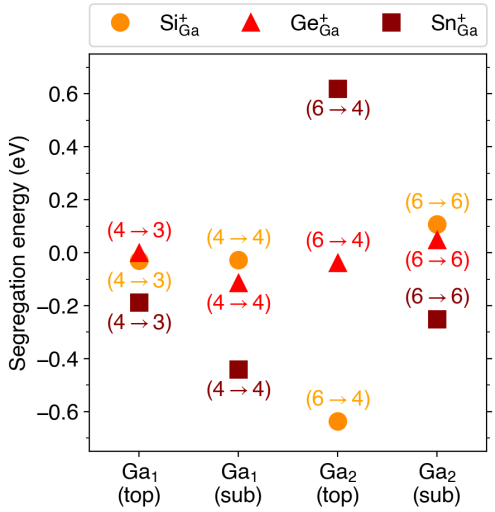

Figure 7.8: Segregation energies of Si, Ge, and Sn from their most stable bulk lattice site to surface lattice sites, located on the top layer (top) and one layer beneath the surface (sub). The labels ($m \rightarrow n$) denote segregation from an m-coordinated site in bulk to an n-coordinated site at surface.

regardless of the relative magnitude of the COHP. This seemingly paradoxical scenario may be explained by taking into account the Coulombic interaction between the dopant charge and the compensating uniform background. Using eqns. (3-4), we found that for Si and Sn, the Coulombic contribution is much smaller than the elastic contribution across all Ga surface sites. In contrast, for Ge, the Coulombic contribution is comparable to the elastic contribution, and varies across different surface Ga sites (Table 7.6). This is a clear sign that both elastic effects associated with dopant size, and electrostatic effects associated with local dopant charge distribution, determine the segregation energy of Ge.

Fig. 7.8 shows the site-specific segregation energy for dopants, which depends strongly on the local bonding strength. Hence it is difficult to compare the overall trend of segregation across the dopant species based on these energies alone. In order to compare the segregational tendency of dopants on an

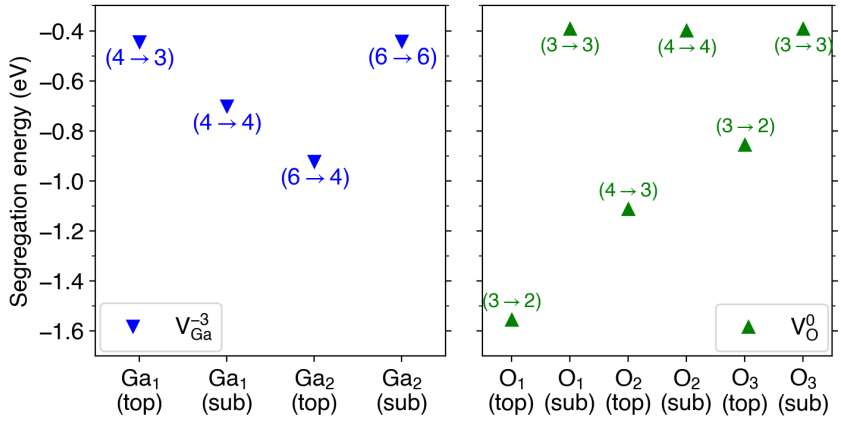

Figure 7.9: Segregation energies of V_{Ga} and V_O from the most stable bulk lattice site to surface lattice sites, located on top layer (top) and one layer beneath the surface (sub). The labels $(m \rightarrow n)$ assume the same meaning as in Fig. 7.8.

Dopant species D^q	surface (top)	surface (sub)	bulk
Si^+_{Ga1}	-30.14	-31.05	-30.97
Si^+_{Ga2}	-30.31	-32.84	-32.75
Ge^+_{Ga1}	-26.59	-27.81	-27.93
Ge^+_{Ga2}	-27.06	-29.11	-29.71
Sn^+_{Ga1}	-20.34	-19.70	-21.93
Sn^+_{Ga2}	-20.67	-24.19	-24.95

Table 7.4: Total ICOHP values (in eV) for different dopant configurations on the top surface layer, the sub-surface layer, and in the bulk.

absolute scale, we consider an "effective" segregation energy, defined as

$$E^{\text{eff}}_{\text{segr}}(D^q) = E^{\text{lowest}}_{\text{tot}}(D^q, \text{surf}) - E^{\text{lowest}}_{\text{tot}}(D^q, \text{bulk}). \tag{7.5}$$

where $E^{\text{lowest}}_{\text{tot}}$ is the lowest total energy of all impurity-containing slab configurations for an impurity located in either the surface or the bulk.

We calculated these effective segregational energies for the three dopants we

studied located on a Ga lattice site:

$$\begin{aligned} E^{\text{eff}}_{\text{segr}}(\text{Si}_{\text{Ga}}) &= E_{\text{tot}}(\text{Si}_{\text{Ga1}}, \text{sub}) - E_{\text{tot}}(\text{Si}_{\text{Ga1}}, \text{bulk}) = -0.14\text{eV}; \\ E^{\text{eff}}_{\text{segr}}(\text{Ge}_{\text{Ga}}) &= E_{\text{tot}}(\text{Ge}_{\text{Ga1}}, \text{sub}) - E_{\text{tot}}(\text{Ge}_{\text{Ga1}}, \text{bulk}) = -0.11\text{eV}; \\ E^{\text{eff}}_{\text{segr}}(\text{Sn}_{\text{Ga}}) &= E_{\text{tot}}(\text{Sn}_{\text{Ga2}}, \text{sub}) - E_{\text{tot}}(\text{Sn}_{\text{Ga2}}, \text{bulk}) = -0.25\text{eV}. \end{aligned} \tag{7.6}$$

This shows that Sn is the most likely dopant species to segregate from the bulk to the (010) surface of Ga_2O_3. This result agrees well with experimental observations. [16]

We now consider native vacancies (V_{Ga} and V_{O}) in Ga_2O_3. The segregation energies of vacancies (V_{Ga} and V_{O}) in all non-equivalent surface sites are plotted in Fig. 7.9. Again, it is clear that the segregation energy is tied to a specific lattice site, which can be understood by a COHP analysis of the local chemical bonding. From Table 7.5 and Fig. 7.9, we see that, for V_{O}, the greater the total D-O ICOHP value for a particular oxygen surface site (*i.e.*, having a smaller magnitude), the more likely it is that a vacancy will segregate to this site from the bulk. In contrast, V_{Ga} does not follow a similar trend very well. Like the explanation we offered for Ge, these results can be explained by the Coulombic interaction. Specifically, V_{O} does not introduce localized defect charge in the band gap, regardless of the location of V_{O} in the lattice. Thus, the Coulombic interaction does not change significantly when the defect (and its associated charge) is placed at the surface compared to the bulk. In contrast, a relatively localized charge density is introduced when V^{-3}_{Ga} is formed, with the specific charge distribution strongly dependent on the lattice location of V^{-3}_{Ga}. (Fig. 7.10) As a result, the Coulombic interaction between the localized charge and the compensating charge background changes greatly depending on the location of the defect.

Defect species D^q	surface (top)	surface (sub)	bulk
V^{-3}_{Ga1}	-19.03	-22.02	-22.29
V^{-3}_{Ga2}	-20.26	-23.33	-23.99
V^{-3}_{O1}	-13.21	-14.12	-14.95
V^{-3}_{O2}	-14.37	-15.57	-15.59
V^{-3}_{O3}	-13.20	-14.99	-14.69

Table 7.5: Total ICOHP values (in eV) for different vacancy configurations on the top surface layer, sub-surface layer, and in the bulk.

Impurity species D^q	surface (top)		surface (sub)	
	elastic	Coulombic	elastic	Coulombic
Si^{+}_{Ga1}	-0.09	0.06	-0.15	-0.01
Si^{+}_{Ga2}	-0.68	0.04	0.09	0.02
Ge^{+}_{Ga1}	-0.04	0.04	-0.13	0.01
Ge^{+}_{Ga2}	-0.08	0.04	0.02	0.03
Sn^{+}_{Ga1}	-0.25	0.06	-0.47	0.03
Sn^{+}_{Ga2}	0.50	0.12	-0.29	0.04
V^{-3}_{Ga1}	-1.64	1.20	-1.34	0.64
V^{-3}_{Ga2}	-1.82	0.90	-1.07	0.63

Table 7.6: Elastic vs. Coulombic contributions to the segregation energy (in eV) for Si_{Ga}, Ge_{Ga}, Sn_{Ga}, and V_{Ga} on Ga lattice sites, calculated from eqns. (3-4).

Effect of surface vacancies on dopant segregation

The results presented in the previous section showed the strong preference of vacancies to occupy surface lattice sites compared to bulk sites. These surface vacancies may then affect dopant segregation in Ga_2O_3. To look at this, we fixed a single vacancy (V_{Ga} or V_O) in the top or sub-surface layers of the Ga_2O_3(010) surface, and placed a dopant either in the surface (within a second nearest-neighbor distance of the vacancy) or in the bulk region of the slab. (Fig. 7.11) In particular, due to the combinatorial complexity of possible dopant-defect pair

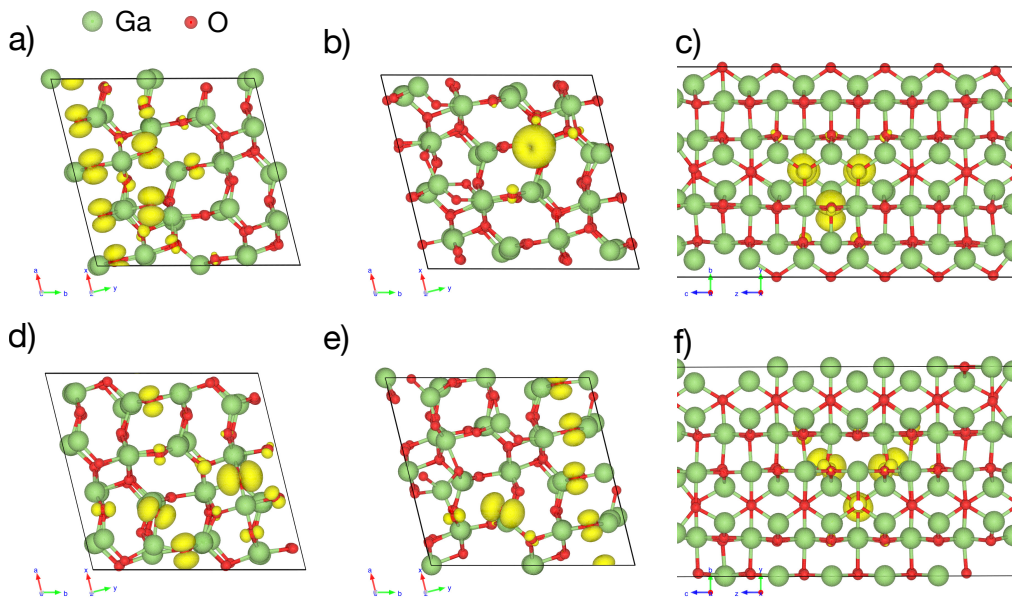

Figure 7.10: Charge density of the localized defect states associated with V_{Ga}^{-3}, located at: (a) Ga_1 (top) site; (b) Ga_1 (sub) site; (c) Ga_1 (bulk) site; (d) Ga_2 (top) site; (e) Ga_2 (sub) site; (f) Ga_2 (bulk) site; isosurfaces at constant value 3×10^{-3} Bohr^{-3} are shown in yellow.

configurations on the $Ga_2O_3(010)$ surface, we only consider the most stable surface site for each dopant species, *i.e.*, the one with the lowest total energy for a given dopant species. This is Ga_1 (sub) for both Si and Ge, and Ga_2 (sub) for Sn. The net charge state of the system is fixed to be the sum of the dopant's charge state (+1) and the vacancy's charge state (−3 for V_{Ga}, and 0 for V_O).

Fig. 7.12 shows the corresponding dopant segregation energies. Two important trends can be drawn from this figure: Firstly, as the dopant size increases (Si → Ge → Sn), the segregation energy towards an V_O-filled $Ga_2O_3(010)$ surface decreases (consistent with the trend for single-dopant segregation). In contrast, the dopant's corresponding segregation energy towards an V_{Ga}-filled $Ga_2O_3(010)$ surface does not exhibit a similarly strong overall decrease. Secondly, the segregation energy for a given dopant species is, in general, much

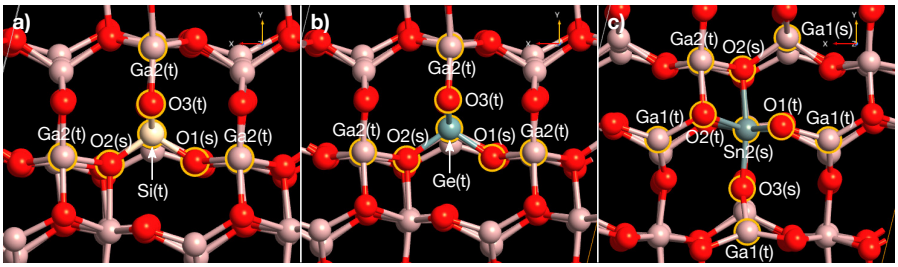

Figure 7.11: Surface dopant-vacancy pairs considered in this work for three dopant species: (a) Si_{Ga1} (sub), (b) Ge_{Ga1} (sub), and (c) Sn_{Ga2} (sub). The lattice positions of the dopant and their first- and second-nearest-neighbor vacancies are annotated with labels ("t" and "s" denote "top" and "sub", as in Fig. 7.8, respectively) and highlighted with yellow circles.

lower (by about 1-2 eV) for a V_{Ga}-filled $Ga_2O_3(010)$ surface compared to that for a V_O-filled $Ga_2O_3(010)$ surface.

To understand these trends, we performed the same analysis of decoupling elastic *vs.* electrostatic contributions, with the results shown in Table 7.6. We find that, while the elastic contribution dominates the segregation energy for the V_O^0 cases, the Coulombic contribution becomes significant (equalling the elastic contribution) for V_{Ga}^{-3} cases. This result confirms the intuition that Coulombic interactions between oppositely charged dopant-defect pair is a major driving force for dopant segregation, together with the elastic energy arising from lattice mismatch. Among the V_O^0 cases, the elastic contribution to the dopant segregation energy decreases as the dopant size increases (from Si to Ge to Sn), while the much smaller Coulombic contributions show little change. This explains the negative correlation of dopant segregation energy to dopant size for a V_O-filled $Ga_2O_3(010)$ surface.

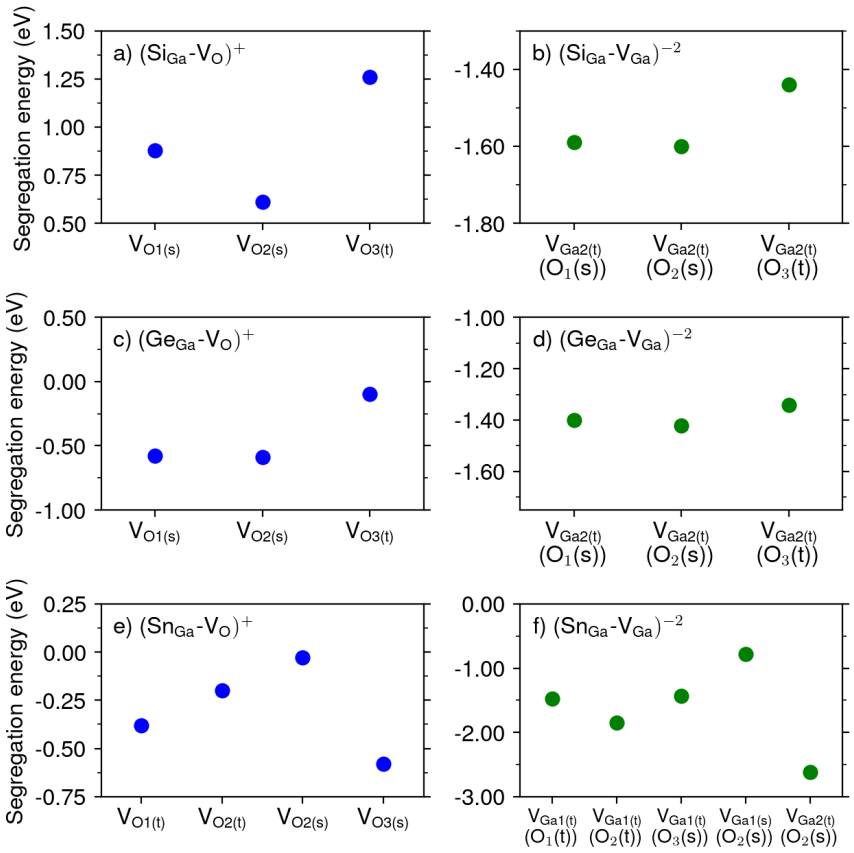

Figure 7.12: Segregation energies of Si^+_{Ga}, Ge^+_{Ga}, and Sn^+_{Ga} from the most stable bulk lattice site to surface lattice sites with a vacancy (V^{-3}_{Ga} or V^0_O) within the second nearest-neighbor distance, located on the top layer (top) and one layer beneath the surface (sub).

Strategies to suppress Sn segregation

We have seen that Sn dopants are particularly prone to segregate toward the β-Ga_2O_3(010) surface and that the presence of V^{-3}_{Ga} greatly increases this proclivity. In order to suppress this undesirable phenomenon, we investigated the effect of co-doping with Si or Ge on the segregation energy of Sn to the surface. To do so, we calculate the segregation energy of Sn toward the surface in a slab with one Si (or Ge) atom in the bulk region, $E_{segr}(Sn_{Ga}|D_{Ga})$, by:

Dopant-vacancy pair (D-V)^q	Elastic	Coulombic
$(Si_{Ga1}\text{-}V_{O3(t)})^{+}$	1.26	0.00
$(Si_{Ga1}\text{-}V_{Ga2(t)}(O_2(s)))^{-2}$	-0.82	-0.78
$(Ge_{Ga1}\text{-}V_{O3(t)})^{+}$	-0.08	-0.02
$(Ge_{Ga1}\text{-}V_{Ga2(t)}(O_2(s)))^{-2}$	-0.68	-0.74
$(Sn_{Ga2}\text{-}V_{O1(t)})^{+}$	-0.33	-0.05
$(Sn_{Ga2}\text{-}V_{Ga2(t)}(O_2(s)))^{-2}$	-1.28	-1.34

Table 7.7: Elastic vs. Coulombic contributions to the segregation energy (in eV) for Si_{Ga} (from Ga_1(bulk) to Ga_1(sub)), Ge_{Ga} (from Ga_1(bulk) to Ga_1(sub)), Sn_{Ga} (from Ga_2(bulk) to Ga_2(sub)) for selected surface vacancy configurations.

$$E_{segr}(Sn_{Ga}|D_{Ga}) = E_{segr}(Sn_{Ga}) - E_{bind}(Sn_{Ga}\text{-}D_{Ga}), \tag{7.7}$$

where $E_{segr}(Sn_{Ga})$ is the segregation energy of Sn_{Ga} calculated in the defect-free slab. The final term in equation 6, the binding energy of Sn_{Ga} and D_{Ga} (where D = Si or Ge) in *bulk* Ga_2O_3, as calculated in a 160-atom bulk supercell, is given by:

$$E_{bind}(Sn_{Ga}\text{-}D_{Ga}) = E^f_{bulk}(Sn_{Ga}\text{-}D_{Ga}) - E^f_{bulk}(Sn_{Ga}) - E^f_{bulk}(D_{Ga}) \tag{7.8}$$

where $E^f_{bulk}(Sn_{Ga}\text{-}D_{Ga})$, $E^f_{bulk}(Sn_{Ga})$ and $E^f_{bulk}(D_{Ga})$ refer to the formation energy of the Sn_{Ga}-D_{Ga} pair, Sn_{Ga}, and D_{Ga} (D = Si, Ge) in bulk Ga_2O_3 respectively.

To evaluate the likelihood of Sn segregation on an absolute scale, we consider the segregation of Sn from a bulk Ga_2 site to a sub-surface Ga_2 site for all the dopant-pair configurations we considered. We find $E_{segr}(Sn_{Ga}) = E^{eff}_{segr}(Sn_{Ga}) = -0.25$ eV. Furthermore, we only consider configurations where a Si (or Ge) atom is located at a second-nearest-neighbor position of the Sn_{Ga2} atom; see Fig. 7.11. The corresponding binding and segregation energies are presented in Table 7.8. For all such configurations, the binding energy, E_{bind},

is less than zero, meaning that both Si_{Ga} and Ge_{Ga} are likely to bind to Sn_{Ga} in a second-nearest-neighbor configuration. Encouragingly, these negative binding energies increase the Sn segregation energies (see Eq. (7.5)). The Sn segregation energies, $E_{segr}(Sn_{Ga}|D_{Ga})$, are comparable to $E^{eff}_{segr}(Si_{Ga}) = -0.14$ eV and $E^{eff}_{segr}(Ge_{Ga}) = -0.11$ eV. Experimentally, Si and Ge exhibit a very limited degree of segregation toward Ga_2O_3 surfaces, in contrast to Sn. [38] This observation suggests that co-doping with Si and Ge may be able to reduce Sn segregation significantly and increase net dopant activation in bulk Ga_2O_3.

$(Sn_{Ga}\text{-}D_{Ga})^q$	D = Si		D = Ge	
	E_{bind}(eV)	E_{segr}(eV)	E_{bind}(eV)	E_{segr}(eV)
$(Sn_{Ga2}\text{-}O_1\text{-}D_{Ga1})^{+2}$	-0.08	-0.17	-0.06	-0.19
$(Sn_{Ga2}\text{-}O_2\text{-}D_{Ga1})^{+2}_{I}$	-0.16	-0.09	-0.12	-0.13
$(Sn_{Ga2}\text{-}O_2\text{-}D_{Ga1})^{+2}_{II}$	-0.09	-0.16	-0.09	-0.16
$(Sn_{Ga2}\text{-}O_3\text{-}D_{Ga1})^{+2}$	-0.05	-0.20	-0.10	-0.15
$(Sn_{Ga2}\text{-}O_2\text{-}D_{Ga2})^{+2}$	-0.06	-0.19	-0.09	-0.16
$(Sn_{Ga2}\text{-}(O_1, O_2)\text{-}D_{Ga2})^{+2}$	-0.09	-0.16	-0.09	-0.16

Table 7.8: Binding energies of Sn_{Ga} and D_{Ga} in various inequivalent second-nearest-neighbor pair configurations, and the corresponding segregation energy of Sn_{Ga} from a bound position in bulk Ga_2O_3 compared to that at a Ga_2O_3(010) surface.

7.4 Conclusions

We have completed a detailed *ab initio* study of the thermodynamics of segregation of shallow dopants (Si, Ge, Sn) towards a β-Ga_2O_3(010) surface. Si and Ge exhibit a weaker tendency to segregate to surface sites than Sn, which overwhelmingly prefers octahedral sites in the sub-surface layer of a Ga_2O_3(010) surface. Gallium and oxygen vacancies stabilize dopants on all surface sites relative

to their counterparts in the bulk. A COHP analysis showed that the preference of particular surface sites over others can be explained mainly by a difference in local coordination number and bonding strength. The presence of surface vacancies enhances the overall trends of dopants to segregate towards the surface, which is disadvantageous to increasing the doping efficiency of bulk Ga_2O_3. Finally, we note that co-doping with Si and Ge could effectively increase the segregation energy (make it less negative), thereby reducing the tendency of Sn to segregate from the bulk to the (010) surface of Ga_2O_3.

Bibliography

[1] M. Higashiwaki, K. Sasaki, A. Kuramata, T. Masui, and S. Yamakoshi, Appl. Phys. Lett. **100**(1), 013504 (2012).

[2] M. A. Mastro, A. Kuramata, J. Calkins, J. Kim, F. Ren, and S. J. Pearton, ECS Jour. Solid State Sci. Technol., **6**, P356 (2017).

[3] S. J. Pearton, J. Yang, P. H. Cary IV, F. Ren, J. Kim, M. J. Tadjer, and M. A. Mastro, Appl. Phys. Rev., **5**, 011301 (2018).

[4] E. G. Villora, K. Shimamura, Y. Yoshikawa, T. Ujiie, and K. Aoki, Appl. Phys. Lett. **92**(20), 202120 (2008).

[5] K. Sasaki, A. Kuramata, T. Masui, E. G. Vıllora, K. Shimamura, and S. Yamakoshi, Appl. Phys. Express **5**(3), 035502 (2012).

[6] E. Ahmadi, O. S. Koksaldi, S. W. Kaun, Y. Oshima, D. B. Short, U. K. Mishra, and J. S. Speck, Appl. Phys. Express **10**(4), 041102 (2017).

[7] J. B. Varley, J. R. Weber, A. Janotti, and C. G. Van de Walle, Appl. Phys. Lett. **97**, 142106 (2010).

[8] S. Lany, APL Mater. **6**, 046103 (2018).

[9] L. Dong, R. Jia, B. Xin, B. Peng, and Y. Zhang, Sci. Rep. **7**, 40160 (2017).

[10] B. E. Kananen, L. E. Halliburton, K. T. Stevens, G. K. Foundos, and N. C. Giles, Appl. Phys. Lett. **110**, 202104 (2017).

[11] A. Kyrtsos, M. Matsubara, and E. Bellotti, Phys. Rev. B **95**, 245202 (2017).

[12] T. C. Lovejoy, R. Chen, X. Zheng, E. G. Villora, K. Shimamura, H. Yoshikawa, Y. Yamashita, S. Ueda, K. Kobayashi, S. T. Dunham, F. S. Ohuchi, and M. A. Olmstead, Appl. Phys. Lett. **100**, 181602 (2012).

[13] S. Ohira, N. Suzuki, N. Arai, M. Tanaka, T. Sugawara, K. Nakajima, and T. Shishido, Thin Solid Films, **516**, 17 (2008).

[14] K. Sasaki, M. Higashiwaki, A. Kuramata, T. Masui and S. Yamakoshi, J. Cryst. Growth **392**, 30 (2014).

[15] A. Y. Polyakov, I.-H. Lee, N. B. Smirnov, E. B. Yakimov, I. V. Shchemerov, A. V. Chernykh, A. I. Kochkova, A. A. Vasilev, P. H. Carey, F. Ren, D. J. Smith, and S. J. Pearton, APL Materials **7**, 061102 (2019).

[16] C. Chang, and D. A. Muller, private communication (2019).

[17] P. Deák, Q. D. Ho, F. Seemann, B. Aradi, M. Lorke, and T. Frauenheim, Phys. Rev. B **95**, 075208 (2017).

[18] A. Dal Corso, Comput. Mater. Sci. **95**, 337 (2014).

[19] J. Åhman, G. Svensson, and J. Albertsson, Acta Crystallogr., Sect. C: Cryst. Struct. Commun. **52**, 6 (1996).

[20] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).

[21] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Status Solidi B **248**, 5 (2011).

[22] M. H. Naik, and M. Jain, Comput. Phys. Commun. **226** (2018).

[23] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003); **124**, 219906 (2006).

[24] G. P. Kerker, Phys. Rev. B, **23**, 6 (1981).

[25] D. Raczkowski, A. Canning, and L. W. Wang, Phys. Rev. B **64**, 121101(R) (2001).

[26] W. Lee, J. W. Han, Y. Chen, Z. Cai, and B. Yildiz, J. Am. Chem. Soc., **135**, 21 (2013).

[27] J. Friedel, Adv. Phys. **3**, 446 (1954).

[28] R. Dronskowski, and P. E. Blöchl, J. Phys. Chem. **97**, 33 (1993).

[29] V. L. Deringer, A. L. Tchougréeff, and R. Dronskowski, J. Phys. Chem. A **115**, 21 (2011).

[30] S. Maintz, V. L. Deringer, A. L. Tchougreeff, R. Dronskowski, J. Comput. Chem. **37**, 1030 (2016).

[31] A. R. West, Basic Solid State Chemistry, Chapter 2, p 78. John Wiley, Chichester, UK, 1999.

[32] T. Zacherle, P. C. Schmidt, and M. Martin, Phys. Rev. B **87**, 235206 (2013).

[33] V. M. Bermudez, Chem. Phys. **323**, 193 (2006).

[34] A. Otero-de-la-Roza, E. R. Johnson and V. Luaña, Comput. Phys. Commun. **185**, 1007-1018 (2014).

[35] A. Otero-de-la-Roza, M. A. Blanco, A. Martín Pendás and V. Luaña, Comput. Phys. Commun. **180**, 157-166 (2009).

[36] R. D. Shannon and C. T. Prewitt, Acta Crystallogr. B **25**, 925-946 (1969).

[37] H. Kwon, W. Lee, and J. W. Han, RSC Adv. **6**, 69782 (2016).

[38] H. G. Xing, and M. O. Thompson, private communication (2019).

CHAPTER 8

SUMMARY AND FUTURE WORK

8.1 Summary

This thesis is a computational survey of various aspects of important electronic materials (InGaAs and Ga_2O_3) with direct applications and implications in device performance. The unifying focus of the projects presented in this thesis is the consideration of chemical doping of materials. This is one of the key determining factors of charge carrier concentration in semiconductors and, hence, the performance of electronic devices. Each of the projects in this thesis utilizes different techniques and methodologies, and shines light on a different aspect of doped materials: the impact of intrinsic point defects on doping efficiency (Chap. 4), the vibrational signatures of dopants and defects in random alloys (Chap. 5), the effect of doping and other materials factors on contact resistivity (Chap. 6), and the thermodynamics of surface segregation of dopants and defects (Chap. 7). A more detailed summary of each part of the thesis are as follows.

In Chapter 4, we are concerned with the concentration saturation of silicon in InGaAs. The commercialization of InGaAs as a source/drain material in InGaAs has been hindered in part by the insufficiently high active silicon concentration achievable as an *n*-type dopant. However, the underlying atomistic mechanism of this phenomenon remains unclear. Using formation energy calculations, our *ab initio* calculations for a model CuAu-I ordered $In_{0.5}Ga_{0.5}As$ suggests that charge compensation from the triply negatively charged cation vacancy ($V^{-3}_{In/Ga}$) is responsible for this undesirable phenomenon. Specifically, the

conclusions reached in this chapter include: (1) In-poor and Ga-poor growth conditions are best suited for *n*-type doping as, in these cases, $Si^{+}_{In/Ga}$ has the lowest formation energy among all Si configurations on the InGaAs lattice; (2) under heavy *n*-type doping, $V^{-3}_{In/Ga}$ has the lowest formation energy among all defect species; (3) $V^{-3}_{In/Ga}$ can be further stabilized on the InGaAs lattice by binding with second-nearest-neighbor $Si^{+}_{In/Ga}$ atoms to form dopant-defect pairs; (4) due to charge compensation effects of $V^{-3}_{In/Ga}$, the net free electron concentration in Si-doped InGaAs saturates at the predicted value of $5 \times 10^{18}\text{cm}^{-3}$, which is roughly on the same order of magnitude as the experimentally measured value $1.5 \times 10^{19}\text{cm}^{-3}$.

In Chapter 5, we propose a novel method of modeling the vibrational modes of impurities in randomly alloyed materials. Local vibrational mode (LVM) detection using spectroscopic techniques is a routine procedure to discern different impurity types and lattice locations. But such procedures meet difficulties in dealing with random alloys, as the strong variations in mode coupling between an impurity and the local atomic environment produce complex vibrational signatures that are difficult to interpret. Utilizing a combination of techniques such as special quasi-random structure (SQS), virtual crystal approximation (VCA), and lattice-static Green's function method, we have modeled, for the first time, the vibrational patterns of two common *n*-type dopant species (Si, Se) and the cation vacancy (the most important compensating defect according to Chap. 4) in a random $In_{0.5}Ga_{0.5}As$. We found that: (1) instead of showing a single one-frequency peak characteristic of LVMs in pure non-alloyed material (such as GaAs and InAs), the vibrational modes of impurities in a random environment exhibit a continuous band of peaks; (2) the frequency range of the band of peaks for dopants is similar to the single frequency (Si: 10.9 to 12.1 THz, Se: 4.2 to 7.8

THz); (3) for each specific impurity, the frequency of the most intense peak is dependent on the local atomic environment.

In Chapter 6, we turn our focus onto contact resistivity – an important limiting factor of device performance at sub-10nm scale. InGaAs-based devices are known for insufficiently low ($> 10^{-9}$ cm^{-3}) specific contact resistivity due to a lack of a coherent defect-free interface. But there is a quantitative influence on contact resistivity by a multitude of materials parameters such as composition, doping, termination, grading, and metal-semiconductor alloying. Using a non-equilibrium Green's function (NEGF) approach, we take a holistic look at how these important parameters change the contact resistivity in a better or worse direction. We find that some parameters affect the contact resistivity in a direct and intuitive way; for example, increasing the doping level in the III-V semiconductor decreases the contact resistivity. Other methods are also effective in reducing the contact resistivity, such as increasing the indium concentration in III-V semiconductors, as well as using the As-terminated III-V semiconductor rather than In-/Ga- terminated. In particular, compositional grading from InAs to $In_{0.5}Ga_{0.5}As$ introduces a gradual change in the relative Fermi level to the CBM of the semiconductor, thereby increasing the current flow and reducing the resistivity at the contact interface. Alloying with Ni, on the other hand, seems to have limited effect in reducing contact resistivity, despite its alleged importance in forming a self-aligned contact.

Finally, in Chapter 7, we elucidate surface segregation phenomena of *n*-type dopants upon thermal annealing in the wide-gap semiconductor Ga_2O_3. Dopant activation in bulk Ga_2O_3 cannot reach its optimal theoretical limit if the tendency of surface dopant segregation is strong. By first principles calcula-

tions, we pinpoint the thermodynamic preference of common group-IV *n*-type dopants (Si, Ge, Sn) as well as vacancies (V_{Ga}, V_O) near the Ga_2O_3(010) surface compared to their behavior in bulk Ga_2O_3. We note that, although the effective segregation energy for the group-IV dopants are similar (Si: -0.14 eV, Ge: -0.11 eV, Sn: -0.25 eV), their preference for different lattice sites can vary drastically. This is due to differences in the effective radius of the dopant species as well as the local bonding strength. One aspect of particular note is the strong preference of vacancies for surface sites: this is similar to the case in bulk InGaAs, in which the triply negatively charged V_{Ga} is the most stable compensating defect on the Ga_2O_3(010) surface. The presence of this surface defect has the unwanted consequence of greatly lowering the segregation energy (and increasing the likelihood of segregation) of group-IV dopants. Fortunately, the strategy of co-doping Sn with Si/Ge can effectively increase the segregation energy of both Sn and Si/Ge by a small amount ($\sim$ 0.1eV). Co-doping could, we believe, potentially be a remedy for the segregation problem.

8.2 Future work

Although this thesis has touched on some of the important aspects of semiconductor modeling, it is far from a complete treatise even of the topics covered. The scope of this thesis is inevitably limited by many factors, including but not limited to: (1) computational resources available at the time of research; (2) knowledge of the author in the relevant subjects and techniques; and (3) access to direct experimental validation of the computational results. Based on these considerations, future work that can be extended from the body of work that composes this thesis include:

(1) A detailed investigation of the kinetics of impurities in semiconductors.

Thermodynamics and kinetics both play crucial roles in determining the spatial distribution of impurities. (Sec. 3.4) A detailed kinetics study would provide an important complement to the thermodynamic perspective in our work, as the kinetic barrier is often the rate-limiting factor in many important chemical reactions. However, we have found early on that simulating the kinetics of impurities in complex materials systems such as β–Ga_2O_3 requires an extensive amount of computational resources beyond our current capacity. Algorithms that can efficiently perform constrained optimization on the potential energy surface (PES) could offer a way forward. For example, an automated variation [1] and a Gaussian process regression (GPR) variation [2] of the nudged elastic band (NEB) calculations [3, 4, 5] would be very valuable in providing insights on the preferred low-energy pathways of dopant migration. Combined with mesoscale simulation methods such as kinetic Monte Carlo, this approach could provide a clear picture of the evolution of dopants in semiconductors in realistic time scales.

(2) A more accurate treatment of charged impurities in semiconductors.

Charged impurities (dopants, point defects) are one of the central topics in our work, due to their key role in controlling the doping efficiency in semiconductors. Currently, there are several notable difficulties in achieving an accurate *ab initio* description of charged impurities in semiconductors. First, the widely used FNV correction scheme for charged impurities works only for single-atom impurities, while it fails for the important case of dopant/defect pairs and complexes. A more general treatment of charged impurities must be able to correct defects of any morphology. Second, the current density functionals often yield

inconsistent thermodynamic transition level for charged impurities, which produce contradictory and inconclusive information regarding the electronic properties of the impurities. [6] Many of these problems hinge on the so-called delocalization error or self-interaction error, [7, 8] namely conventional density functionals tend to over-delocalize the electron density due to an inadequate treatment of electron correlation. Modern methods such as hybrid functional theory could ameliorate this issue, however a more efficient algorithm to compute hybrid functionals is in urgent need in order to treat defective large supercells.

(3) A direct validation by experimental results.

This thesis consists of works that are purely computational in nature. The experimental results with which we compare in our studies come from existing literature. Although such comparisons are necessary for the purpose of validating our results, they are not quite sufficient, as the earlier experiments are not tailored towards the scope of the computations. As a result, they often do not constitute a direct validation to our numerical results. For example, since the experimentally measured vibrational modes of Si, Se, and a vacancy in *random* InGaAs alloys are not yet available, our computational result serves as a useful prediction, but they cannot be directly validated at the present time. In the future, a collaboration with experimentalists would be very beneficial in obtaining a more direct interpretation of the computational results to realistic situations.

Bibliography

[1] Esben L Kolsbjerg, Michael N Groves, and Bjørk Hammer. An automated nudged elastic band method. *The Journal of chemical physics*, 145(9):094107, 2016.

[2] Olli-Pekka Koistinen, Freyja B Dagbjartsdóttir, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Nudged elastic band calculations accelerated with gaussian process regression. *The Journal of chemical physics*, 147(15):152720, 2017.

[3] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions.

[4] Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics*, 113(22):9978–9985, 2000.

[5] Graeme Henkelman, Blas P Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of chemical physics*, 113(22):9901–9904, 2000.

[6] 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*, 78(23):235104, 2008.

[7] Paula Mori-Sánchez, Aron J Cohen, and Weitao Yang. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Physical review letters*, 100(14):146401, 2008.

[8] Stephan Lany and Alex Zunger. Polaronic hole localization and multiple hole binding of acceptors in oxide wide-gap semiconductors. *Physical Review B*, 80(8):085202, 2009.